Distributionally Robust Optimization with Decision-Dependent Ambiguity Set

Nilay Noyan
Industrial Engineering Program, Sabancı University, Istanbul, Turkey, nnoyan@sabanciuniv.edu

Gábor Rudolf
Department of Industrial Engineering, Koc University, Istanbul, Turkey, grudolf@ku.edu.tr

Miguel Lejeune
Department of Decision Sciences, George Washington University, USA, mlejeune@gwu.edu

September 10, 2018

Abstract: We introduce a new class of distributionally robust optimization problems under decision-dependent ambiguity sets. In particular, as our ambiguity sets we consider balls centered on a decision-dependent probability distribution. The balls are based on a class of *earth mover's distances* that includes both the total variation distance and the Wasserstein metrics. We discuss the main computational challenges in solving the problems of interest, and provide an overview of various settings leading to tractable formulations. Some of the arising side results are also of independent interest, including mathematical programming expressions for robustified risk measures in a discrete space. Finally, we rely on state-of-the-art modeling techniques from machine scheduling and humanitarian logistics to arrive at potentially practical applications.

Keywords: stochastic programming; distributionally robust optimization; decision-dependent ambiguity; earth mover's distances; Wasserstein metric; endogenous uncertainty; decision-dependent probabilities; risk-averse; robustified risk; stochastic scheduling; robust scheduling; robust pre-disaster; random link failures; network interdiction

1. Introduction The classical stochastic programming literature relies on the assumption that the probability distribution of uncertain model parameters is given as a model input, often as set of scenarios along with their probabilities. However, in many decision-making applications the true parameter distribution is unknown. *Distributionally robust optimization* (DRO) is a recent and appreciated approach to hedge against such distributional uncertainty. Instead of assuming that there is a known underlying probability distribution, in DRO one considers an *ambiguity set* that consists of probability distributions, and solves a minimax-type problem to determine decisions that provide hedging against the worst-case parameter distribution in the ambiguity set (see, e.g., Goh and Sim, 2010; Wiesemann et al., 2014).

Another common fundamental assumption in the stochastic programming literature is that the underlying probability space is independent of the decisions. In other words, it is usually assumed that the probability distributions of random model parameters are exogenously given. In the DRO setting this attitude translates to the assumption that the specified ambiguity set of distributions is decision-independent. However, in certain situations decisions can directly affect the distribution of the parameters, either by changing the parameter realizations or by changing the probabilities of underlying random events that occur after the decisions are taken. This phenomenon is known as *endogeneous uncertainty*. For example, in the context of pre-disaster planning, if the links of a transportation network are subject to random failure in case of a disaster, then the investment decisions on strengthening such links (seismic retrofitting of bridges/viaducts on links) can reduce the failure probabilities and improve network survivability (Peeta et al., 2010).

In our study we aim to address both distributional and endogeneous uncertainty. We next provide a brief overview of the relevant literature on these two concepts.

Distributionally robust optimization. The two most widely used types of ambiguity sets in the DRO literature are moment-based and statistical distance-based ones (for a review, see Postek et al., 2016). Moment-based ambiguity sets contain all probability distributions that satisfy certain general moment conditions (see, e.g., Delage and Ye, 2010; Zymler et al., 2013; Wiesemann et al., 2014). A common example is the ambiguity set consisting of distributions that exactly match the empirical first and second moments; however, such exact moment-based ambiguity sets typically do not contain the true distribution. In addition, as very different distributions can have the same (or similar) lower moments, and the use of higher moments can be impractical, moment-based approaches often lead to overly conservative solutions.

In the present work we therefore limit our attention to statistical distance-based ambiguity sets. These sets consist of probability distributions that are in the vicinity of a nominal distribution—often the empirical one—thought to approximate the true distribution. The vicinity is defined here as a ball centered on the nominal distribution. A wide variety of statistical distances, which provide a measure of dissimilarity between two probability distributions, have been employed to construct such balls. These range from Wasserstein metrics (see, e.g., Pflug and Wozabal, 2007; Esfahani and Kuhn, 2018), the Prohorov metric (see, e.g., Erdoğan and Iyengar, 2006), and ζ-structures (Zhao and Guan, 2018) such as the Kolmogorov–Smirnov statistic and the bounded Lipschitz metric, to the class of ϕ-divergences (see, e.g., Jiang and Guan, 2015). Distances in the latter class are frequently employed in a data-driven context (for an overview see Bayraksan and Love, 2015), and include the Kullback–Leibler divergence (see, e.g., Calafiore, 2007; Hu and Hong, 2012), the Burg entropy (Wang et al., 2016), the total variation distance, the Hellinger distance, the χ^2 distance, and the modified-χ^2 distance. Several studies have highlighted the fact that utilizing a distance-based approach in DRO leads to desirable statistical properties, including consistency and good out-of-sample performance (see, e.g., Lam, 2016; Esfahani and Kuhn, 2018; Van Parys et al., 2017). On a related note, many of the popular regularization methods that are utilized in the machine learning literature to improve out-of-sample performance have recently been shown to be equivalent to statistical distance-based DRO models (see, e.g., Blanchet et al., 2017; Gao et al., 2017). A significant number of DRO studies focus on ϕ-divergences, which are often shown to work well if the uncertain parameters are known to be supported on a discrete set. However, when the possible realizations of parameters form a continuous spectrum, the use of ϕ-divergences can be problematic (as pointed out by, e.g., Blanchet et al., 2017; Gao and Kleywegt, 2016) due to ignoring the metric structure of the realization space, and limiting the support of the measures in the ambiguity set.

An attractive feature of any distance-based approach is that one can control the degree of conservatism simply by adjusting the radius. When using certain distances, such the Wasserstein-1 metric, an appropriate choice of radius can also guarantee that, with a prescribed level of confidence, the true probability distribution belongs to the ambiguity set (see Esfahani and Kuhn, 2018; Zhao and Guan, 2015). This is in contrast to, for example, the Kullback-Leibler divergence, which does not permit the construction of a confidence set that includes the true probability distribution (Esfahani and Kuhn, 2018). We refer to Gao and Kleywegt (2016) for a more elaborate discussion of the pros and cons associated with various ambiguity sets, and in particular the advantages of Wasserstein metrics over ϕ-divergences. Due to these advantages the use of Wasserstein distances in DRO has seen a recent sharp increase, including studies by Pflug and Wozabal (2007), Wozabal (2014), Zhao and Guan (2018), Gao and Kleywegt (2016), Esfahani and Kuhn (2018), Ji and Lejeune (2017), Gao and Kleywegt (2017), and Luo and Mehrotra (2017). In line with these developments our focus will be on a general class of *earth mover's distances*, introduced in a discrete context by Rubner et al. (1998). Our chosen class includes both the total variation distance and the Wasserstein-1 metric (also known simply as the Wasserstein metric, or the Kantorovich-Rubinstein metric, Kantorovich and Rubinshtein, 1958), allows the construction of ambiguity sets based on higher-order Wasserstein distances, and also has favorable tractability properties.

Endogenous uncertainty. As highlighted in Haus et al. (2017), while decision-dependent uncertainty—endogenous uncertainty—is straightforward to express in the framework of Markov decision processes, its use in stochastic programming remains a tough endeavor, and is far from being a well-resolved issue. Hellemo et al. (2014) and Hellemo (2016) discuss the modeling and applications of decision-dependent uncertainty in mathematical programming, and present a taxonomy of stochastic programming approaches with decision-dependent uncertainty. The relevant literature primarily focuses on two types of optimization problems (Goel and Grossmann, 2006): problems with decision-dependent information revelation, and problems with decision-dependent probabilities. In problems of the first type, decisions can partially resolve the uncertainty, affect the timing of uncertainty resolution, and alter the set of possible future random outcomes. In problems of the second-type, decisions alter the probability measures. The first problem type has been addressed more widely (see, e.g., Jonsbråten et al., 1998; Goel and Grossmann, 2004; 2006; Khaligh and MirHassani, 2016) in the literature. Accordingly, in our study, we aim to contribute to the literature by focusing on problems of the second type, where decisions can affect the likelihood of underlying random future events and/or can affect the possible realizations of the random parameters.

Stochastic problems with decision-dependent probability measures are notoriously difficult to model and solve, and, not surprisingly, the relevant literature is quite sparse. Dupacova (2006) briefly discusses optimization under endogenous uncertainty, without providing specific formulations or solution methods. Studies that feature algorithmic developments are relatively recent, and typically rely on additional structural properties that are specific to their problems of interest. A significant part of the literature focuses on one particular stochastic pre-disaster investment problem, where the links of a transportation network are subject to probabilistic failures. This problem—originally introduced by Peeta et al. (2010)—aims to use a limited budget to increase the survival probabilities of selected links in such a way that the total expected shortest-path distance between a number of origin-destination pairs is minimized. Modeling the problem in a straightforward fashion involves expressing probabilities as non-linear functions of decision variables, which gives rise to highly non-linear models that are often intractable. Several relevant studies (Flach and Poggi, 2010; Laumanns et al., 2014; Schichl and Sellmann, 2015; Haus et al., 2017) have instead focused on developing efficient alternative solution methods for this particular pre-disaster investment problem. Among these studies, the working papers by Laumanns et al. (2014) and Haus et al. (2017) consider a class of problems where the decisions are binary, and the inherent uncertainty is characterized by a set of binary vectors whose components are independent random variables. They develop effective and exact mixed-integer programming formulations for this class by introducing novel distribution shaping and scenario bundling techniques. These techniques enable an efficient characterization of the decision-dependent scenario probabilities via a set of linear constraints.

DRO with endogenous uncertainty. In this study, we incorporate endogenous uncertainty into distributionally robust stochastic programming problems via decision-dependent ambiguity sets. Until recently, DRO with decision-dependent ambiguity sets has been an almost untouched research area. Zhang et al. (2016) consider decision-dependent ambiguity sets defined via parametric moment conditions with generic cone constraints. Adopting the total variation metric, the authors establish quantitative stability results for the ambiguity set, the optimal values and solutions. Royset and Wets (2017) utilize recent developments from the variational theory of bivariate functions to establish convergence results for approximations of a class of DRO problems with decision-dependent ambiguity sets. Their discussion covers a variety of ambiguity sets, including moment-based and stochastic dominance-based ones. A major part of their toolset relies on the so-called hypo-distance between CDFs, which is shown to be a metrization of weak convergence. Finally, we highlight the recent interest in the tangentially related area of traditional robust optimization models with decision-dependent uncertainty sets (Bertsimas and Vayanos, 2014; Lappas and Gounaris, 2018; Nohadani and Sharma, 2016).

Our contributions. We present a unified modeling framework for a class of DRO problems with decision-dependent EMD-based ambiguity sets. Our models typically give rise to non-convex non-linear programs, which are in general very hard to solve. However, we provide an overview of several settings where it is possible to obtain tractable formulations. Some of the side results that make these formulations possible are also of independent interest, including novel mathematical programming expressions for robustified risk measures in a discrete space. We also discuss potential practical applications, utilizing state-of-the-art modeling techniques from the fields of machine scheduling and humanitarian logistics.

Outline. The rest of the paper is organized as follows. In Section 2 we establish necessary notation and recall some basic definitions. Section 3 describes the class of DRO problems of interest. Sections 4–6 are dedicated to developing the corresponding mathematical programming formulations, with Section 5 in particular dedicated to the aforementioned side results about robustified risk measures in a discrete space. Section 7 presents potential applications, and Section 8 contains our concluding remarks regarding future research directions.

2. Preliminaries The set of the first n positive integers is denoted by $[n]=\{1,\dots,n\}$, while the positive part of a number $\eta\in\mathbb{R}$ is denoted by $[\eta]_+=\max\{\eta,0\}$. The extended real numbers are denoted by $\bar{\mathbb{R}}=\mathbb{R}\bigcup\{-\infty,+\infty\}$.

The family of all probability measures on a measurable space $(\Omega,\mathcal{A})$ is denoted by $\mathcal{P}(\Omega,\mathcal{A})$. Let us denote by $\mathcal{L}^m(\Omega,\mathcal{A})$ the family of all measurable mappings from $(\Omega,\mathcal{A})$ to $(\mathbb{R}^m,\mathcal{A}_B^m)$, where $\mathcal{A}_B^m$ is the σ-algebra of m-dimensional Borel sets, and denote the set of m-dimensional random vectors by $\mathcal{V}^m(\Omega,\mathcal{A})=\mathcal{P}(\Omega,\mathcal{A})\times\mathcal{L}^m(\Omega,\mathcal{A})$. In a pair $[\mathbb{P},\boldsymbol{\xi}]\in\mathcal{V}^m(\Omega,\mathcal{A})$ we view the mapping $\boldsymbol{\xi}:\Omega\to\mathbb{R}^m$ as a random variable on the probability space $(\Omega,\mathcal{A},\mathbb{P})$, with corresponding CDF $F_{[\mathbb{P},\boldsymbol{\xi}]}$. If we denote the family of all m-variate CDFs by $\mathcal{F}^m$, then we trivially have $\{F_{[\mathbb{P},\boldsymbol{\xi}]}\mid[\mathbb{P},\boldsymbol{\xi}]\in\mathcal{V}^m(\Omega,\mathcal{A})\}\subset\mathcal{F}^m$, and equality holds if and only if $(\Omega,\mathcal{A})$ is a continuous space, i.e., if it admits a standard continuous uniform random variable. One such continuous space is the standard Borel space $((0,1),\mathcal{A}_B)$ on the unit interval; we will denote the Borel probability measure on this standard space (i.e., the restriction of the Lebesgue measure to $\mathcal{A}_B$) by $\mathbb{B}$. For $1\le p\le\infty$ we introduce the standard L_p-space $L_p^S=\left\{X\in\mathcal{L}^1((0,1),\mathcal{A}_B,\mathbb{B})\ :\ \|X\|_{L_p}<\infty\right\}$, where $\|\cdot\|_{L_p}$ is the L_p-norm for random variables on the standard probability space. In the remainder of the paper we use the common convention whereby p and q refer to a dual pair of values that satisfy $1\le p<\infty$, $1<q\le\infty$, and $\frac{1}{p}+\frac{1}{q}=1$.

We define the *law* of a random vector $[\mathbb{P},\boldsymbol{\xi}]\in\mathcal{V}^m(\Omega,\mathcal{A})$ as the push-forward probability measure $\mathrm{law}[\mathbb{P},\boldsymbol{\xi}]\in\mathcal{P}(\mathbb{R}^m,\mathcal{A}_B^m)$ given by $[\mathrm{law}[\mathbb{P},\boldsymbol{\xi}]](A)=\mathbb{P}(\boldsymbol{\xi}\in A)$ for $A\in\mathcal{A}_B^m$. Two random vectors have the same law if and only if they have the same CDF. By definition, for any measurable space $(\Omega,\mathcal{A})$ we have $\{\mathrm{law}[\mathbb{P},\boldsymbol{\xi}]\ :\ [\mathbb{P},\boldsymbol{\xi}]\in\mathcal{V}^m(\Omega,\mathcal{A})\}\subset\mathcal{P}(\mathbb{R}^m,\mathcal{A}_B^m)$, and equality again holds for continuous spaces such as $((0,1),\mathcal{A}_B)$.

Finally, we establish some notational conventions for working with a finite sample space $\Omega=\{\omega^1,\dots,\omega^n\}$. Probability measures on $(\Omega,2^\Omega)$ are denoted by blackboard bold characters, and the probabilities of elementary events by corresponding lowercase letters, e.g., given $\mathbb{P}\in\mathcal{P}(\Omega,2^\Omega)$ we write $p^i=\mathbb{P}\left(\{\omega^i\}\right)$. Similarly, we use uppercase letters for scalar-valued random variables, and use the corresponding lowercase letters for their realizations, e.g., given $Z:\Omega\to\mathbb{R}$ we write $z^i=Z(\omega^i)$. Finally, random vectors are typically denoted by bold Greek letters, and upper indices are used to refer to their realizations, e.g., given $\boldsymbol{\xi}:\Omega\to\mathbb{R}^m$ we write $\boldsymbol{\xi}^i=\boldsymbol{\xi}(\omega^i)$.

2.1 Earth mover's distances We now introduce a general class of earth mover's distances (EMDs). Consider a function $\delta:\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}_+$, typically chosen to be a symmetric measure of dissimilarity (or distance) between m-dimensional real vectors. We will always assume that δ is *reflexive*, i.e., that

$\delta(\mathbf{x},\mathbf{x}) = 0$ holds for all $\mathbf{x} \in \mathbb{R}^m$. If the stronger condition $\delta(\mathbf{x}_1,\mathbf{x}_2) = 0 \Leftrightarrow \mathbf{x}_1 = \mathbf{x}_2$ holds for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$, we say that ρ is *definite*. We remark that the choice of $\mathbb{R}^m$ as the native space of realizations is somewhat arbitrary, and for the purposes of the following definitions the space $\mathbb{R}^m$ could be replaced with a general ground set; in the literature the ground set is commonly assumed to be a Polish space with distinguished metric δ. However, we restrict ourselves to working with real vectors, as all of the examples and applications that we discuss later will naturally fit this framework.

The function δ, which measures dissimilarities among vectors in the ground space $\mathbb{R}^m$, induces an EMD $\Delta : \mathcal{P}(\mathbb{R}^m, \mathcal{A}_B^m) \times \mathcal{P}(\mathbb{R}^m, \mathcal{A}_B^m) \to \bar{\mathbb{R}}_+$ that measures dissimilarities among probability distributions on $\mathbb{R}^m$. The EMD between distributions $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathbb{R}^m, \mathcal{A}_B^m)$ is given by

$$\Delta(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\mathbb{P}^* \in \Pi(\mathbb{P}_1,\mathbb{P}_2)} \int\limits_{\mathbb{R}^m \times \mathbb{R}^m} \delta(\mathbf{x}_1,\mathbf{x}_2)\, \mathbb{P}^*\left(\mathrm{d}(\mathbf{x}_1,\mathbf{x}_2)\right), \tag{1}$$

where the infimum is taken over the family of distributions with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$,

$$\Pi(\mathbb{P}_1,\mathbb{P}_2) = \left\{ \mathbb{P}^* \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m, \mathcal{A}_B^m \times \mathcal{A}_B^m) \ : \ \begin{array}{l} \mathbb{P}^*(S \times \mathbb{R}^m) = \mathbb{P}_1(S), \\ \mathbb{P}^*(\mathbb{R}^m \times S) = \mathbb{P}_2(S) \end{array} \ \text{for all } S \in \mathcal{A}_B^m \right\}.$$

The above definition can naturally be extended to quantify dissimilarities between any two m-dimensional random vectors. With a slight abuse of notation, for any two measurable spaces $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ the EMD $\Delta : \mathcal{V}^m(\Omega_1, \mathcal{A}_1) \times \mathcal{V}^m(\Omega_1, \mathcal{A}_2) \to \bar{\mathbb{R}}_+$ will be given by

$$\begin{aligned} \Delta\left([\mathbb{P}_1, \boldsymbol{\xi}_1], [\mathbb{P}_2, \boldsymbol{\xi}_2]\right) &= \Delta\left(\mathrm{law}[\mathbb{P}_1, \boldsymbol{\xi}_1], \mathrm{law}[\mathbb{P}_2, \boldsymbol{\xi}_2]\right) \\ &= \inf \left\{ \int\limits_{\mathbb{R}^m \times \mathbb{R}^m} \delta(\mathbf{x}_1,\mathbf{x}_2)\, \mathbb{P}^*\left(\mathrm{d}(\mathbf{x}_1,\mathbf{x}_2)\right) \ : \ \mathbb{P}^* \in \Pi\left(\mathrm{law}[\mathbb{P}_1, \boldsymbol{\xi}_1], \mathrm{law}[\mathbb{P}_2, \boldsymbol{\xi}_2]\right) \right\}. \end{aligned} \tag{2}$$

We aim to incorporate distributional uncertainty into decision problems via EMD balls centered on a nominal random vector $[\mathbb{P}, \boldsymbol{\xi}] \in \mathcal{V}^m(\Omega, \mathcal{A})$. To model cases where there is ambiguity both in the probability measure and in the realizations, we construct the EMD ball on the standard probability space, and refer to it as a *continuous EMD ball*. This ball will represent all possible m-dimensional distributions within κ distance from the nominal one:

$$\mathcal{B}_{\delta,\kappa}^{\mathbb{P}}(\boldsymbol{\xi}) = \left\{ \boldsymbol{\zeta} \in \mathcal{L}^m((0,1), \mathcal{A}_B) \ : \ \Delta\left([\mathbb{P}, \boldsymbol{\xi}], [\mathbb{B}, \boldsymbol{\zeta}]\right) \le \kappa \right\}. \tag{BALL-C}$$

On the other hand, if the realizations of random vectors always belong to some discrete set (e.g., if they are binary), it is not meaningful to consider small variations in realizations. For such cases a natural approach is to construct the EMD ball on the native measurable space of the nominal random vector by allowing the probability measure to change while keeping the realization mapping $\boldsymbol{\xi}$ fixed. We will refer these balls given below as *discrete EMD balls.*

$$\mathcal{B}_{\delta,\kappa}^{\boldsymbol{\xi}}(\mathbb{P}) = \left\{ \mathbb{Q} \in \mathcal{P}(\Omega, \mathcal{A}) \ : \ \Delta\left([\mathbb{P}, \boldsymbol{\xi}], [\mathbb{Q}, \boldsymbol{\xi}]\right) \le \kappa \right\}. \tag{BALL-D}$$

A similar approach is seen, for example, in Pflug and Pichler (2011), where the Wasserstein distance to a reference distribution is minimized among probability distributions with a fixed finite support.

Remark 2.1 *We introduced the definition* (BALL-C) *instead of the perhaps more natural*

$$\mathcal{B}_{\delta,\kappa}\left([\mathbb{P}, \boldsymbol{\xi}]\right) = \left\{ \mathbb{Q} \in \mathcal{P}(\mathbb{R}^m, \mathcal{A}_B^m) \ : \ \Delta\left(\mathrm{law}[\mathbb{P}, \boldsymbol{\xi}], \mathbb{Q}\right) \le \kappa \right\}.$$

The two definitions are essentially equivalent, as it is easy to see that $\zeta \in \mathcal{B}_{\delta,\kappa}^{\mathbb{P}}(\boldsymbol{\xi})$ *holds if and only if we have* $\mathrm{law}[\mathbb{B}, \boldsymbol{\zeta}] \in \mathcal{B}_{\delta,\kappa}\left([\mathbb{P}, \boldsymbol{\xi}]\right)$. *The definition* (BALL-C) *was chosen both for notational convenience, and to emphasize that distributions in continuous spaces can be specified via varying outcome mappings (as opposed to varying probability measures). This approach is taken by* Pflug et al. (2012) *to constructively prove the crucially important Proposition* 4.2*, which underlies our development in Section* 4*.*

The EMD balls defined in (BALL-C) and (BALL-D) are non-empty for any $\kappa \geq 0$, since due to the reflexivity of δ they always contain the nominal distribution. We also note that the domain of the EMD Δ implicitly depends on the construction used: In (BALL-C) we have $\Delta : \mathcal{V}^m(\Omega, \mathcal{A}) \times \mathcal{V}^m((0,1), \mathcal{A}_B) \to \bar{\mathbb{R}}_+$, while in (BALL-D) we have $\Delta : \mathcal{V}^m(\Omega, \mathcal{A}) \times \mathcal{V}^m(\Omega, \mathcal{A}) \to \bar{\mathbb{R}}_+$. Unless specified otherwise, outside of this preliminary section we will always assume that the sample space Ω is finite, with $\mathcal{A} = 2^{\Omega}$.

The family of EMDs includes widely used metrics such as the *total variation distance*, which (see, e.g., Lindvall, 1992, Theorem 5.2) is the EMD induced by the discrete metric

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} 0 & \text{if } \mathbf{x}_1 = \mathbf{x}_2 \\ 1 & \text{if } \mathbf{x}_1 \neq \mathbf{x}_2. \end{cases} \tag{3}$$

Wasserstein metrics are also closely related to EMDs. For $p \in [1, \infty)$ the Wasserstein-p metric $W_p : \mathcal{V}^m(\Omega_1, \mathcal{A}_1) \times \mathcal{V}^m(\Omega_2, \mathcal{A}_2) \to \bar{\mathbb{R}}_+$ is defined as

$$W_p\left([\mathbb{P}_1, \boldsymbol{\xi}_1], [\mathbb{P}_2, \boldsymbol{\xi}_2]\right) = \inf \left\{ \left(\int\limits_{\Omega_1 \times \Omega_2} \|\boldsymbol{\xi}_1(\omega_1) - \boldsymbol{\xi}_2(\omega_2)\|_p^p \, \mathbb{P}^*(\mathrm{d}\omega_1, \mathrm{d}\omega_2) \right)^{1/p} \; : \; \mathbb{P}^* \in \Pi\left(\operatorname{law}[\mathbb{P}_1, \boldsymbol{\xi}_1], \operatorname{law}[\mathbb{P}_2, \boldsymbol{\xi}_2]\right) \right\}.$$

It is easy to see that the Wasserstein-1 metric, is the EMD induced by the 1-norm distance $\delta(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_1$. More generally, for any $p \in [1, \infty)$ we have $W_p\left([\mathbb{P}_1, \boldsymbol{\xi}_1], [\mathbb{P}_2, \boldsymbol{\xi}_2]\right) = \Delta^p\left([\mathbb{P}_1, \boldsymbol{\xi}_1], [\mathbb{P}_2, \boldsymbol{\xi}_2]\right)^{\frac{1}{p}}$, where Δ^p is the EMD induced by $\delta^p(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_p^p$. It follows that a Wasserstein-p ball of radius κ is identical to the corresponding EMD ball with the same center, and a radius of κ^p.

2.2 Risk measures Unless specified otherwise, the definitions and results in this section are presented for risk measures that are natively defined on a standard L_p space. Any such risk measure $\rho : L_p^S \to \mathbb{R}$ can be naturally extended to p-integrable random variables defined on an arbitrary probability space $(\Omega, \mathcal{A}, \mathbb{P})$ via *inverse transform sampling* as follows: It is well-known that if $X : \Omega \to \mathbb{R}$ is a random variable, then its *generalized inverse CDF* $F_X^{(-1)} : (0,1) \to \mathbb{R}$, given by $F_X^{(-1)}(\alpha) = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}$ and viewed as a random variable on the standard space $((0,1), \mathcal{A}_B, \mathbb{B})$, has the same CDF as X itself. Consequently, X is p-integrable if and only if we have $F_X^{(-1)} \in L_p^S$, in which case with a slight abuse of notation we will write $\rho\left([\mathbb{P}, X]\right) = \rho(X) = \rho\left(F_X^{(-1)}\right)$.

Risk measures are functionals that represent the risk associated with a random variable by a scalar value, and their desirable properties, such as law invariance and coherence, are axiomatized in Artzner et al. (1999). Throughout this paper we limit our attention to law invariant coherent risk measures. We say that a mapping $\rho : L_p^S \to \mathbb{R}$ is a *coherent risk measure* if ρ has the following properties (for all $V, V_1, V_2 \in L_p^S$):

- *Monotone*: $V_1 \leq V_2 \;\Rightarrow\; \rho(V_1) \leq \rho(V_2)$.
- *Convexity*: $\rho(\lambda V_1 + (1-\lambda)V_2) \leq \lambda\rho(V_1) + (1-\lambda)\rho(V_2)$ for all $\lambda \in [0,1]$.
- *Translation equivariant*: $\rho(V + \lambda) = \rho(V) + \lambda$ for all $\lambda \in \mathbb{R}$.
- *Positive homogeneous*: $\rho(\lambda V) = \lambda\rho(V)$ for all $\lambda \geq 0$.

The more general class of convex risk measures is obtained by dropping positive homogeneity (Föllmer and Schied, 2002). For a more general discussion on quantifying risk we refer to Müller and Stoyan (2002), Pflug and Römisch (2007), and Shapiro et al. (2009). We now introduce an important family of coherent risk measures. The *conditional value-at-risk* at confidence level $\alpha \in [0,1)$ for a random variable Z is defined (Rockafellar and Uryasev, 2000) as

$$\mathrm{CVaR}_\alpha(Z) = \min\left\{ \eta + \frac{1}{1-\alpha}\mathbb{E}\left([Z - \eta]_+\right) \;:\; \eta \in \mathbb{R} \right\}. \tag{4}$$

The minimum in (4) is attained at the α-quantile, which is known as the *value-at-risk* (VaR) at confidence level α: $\mathrm{VaR}_\alpha(Z) = \min\{\eta \in \mathbb{R} \ : \ P(Z \leq \eta) \geq \alpha\}$. For risk-averse decision makers typical choices for the confidence level are large values such as $\alpha = 0.9$.

Suppose that Z is a discrete random variable with realizations $z^1, \dots, z^n$, and corresponding probabilities $p^1, \dots, p^n$. Then $\mathrm{VaR}_\alpha(Z) = z^j$ holds for at least one $j \in [n]$, which implies

$$\mathrm{CVaR}_\alpha(Z) = \min_{j\in[n]} z^j + \frac{1}{1-\alpha}\sum_{i\in[n]} p^i[z^i - z^j]_+. \tag{5}$$

It is also well known that the optimization problem in (4) can equivalently be formulated as the following linear program:

$$\min\left\{\eta + \frac{1}{1-\alpha}\sum_{i\in[n]} p^i v^i \ : \ v^i \geq z^i - \eta \quad \forall\, i \in [n], \quad \mathbf{v} \in \mathbb{R}^n_+, \ \eta \in \mathbb{R}\right\}. \tag{6}$$

CVaR has been widely used in decision-making problems under uncertainty due to a number of useful properties. It captures a wide range of risk preferences, including risk-neutral (for $\alpha = 0$) and pessimistic worst-case (for sufficiently large values of α, $\alpha \to 1$) preferences. It is also a spectral risk measure (Acerbi, 2002) and thus can be viewed as a weighted sum of the least favorable outcomes as illustrated by the following dual representations of CVaR_α:

$$\max\left\{\frac{1}{1-\alpha}\sum_{i\in[n]}\beta^i z^i \ : \ \sum_{i\in[n]}\beta^i = 1-\alpha, \quad 0 \leq \beta^i \leq p^i \ \ \forall\, i \in [n]\right\} = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_a(Z)\,\mathrm{d}a. \tag{7}$$

The knapsack-type maximization problem in (7) is equivalent to the linear programming dual of (6), and to the well-known *risk envelope-based dual representation* of CVaR (see, e.g., Rockafellar, 2007). Due to the last equality, CVaR is also known in the literature as *average value-at-risk* and *tail value-at-risk*.

CVaR is of particular importance as it serves as a fundamental building block for other coherent risk measures (Kusuoka, 2001). It was shown in Noyan and Rudolf (2015) that the class of risk measures that can be obtained by extending a law invariant coherent risk measure from L_p^S via inverse transform sampling coincide with the class of operators with so-called *Kusuoka representations* of the form

$$\rho(X) = \sup_{\mu\in\mathcal{M}}\int_0^1 \mathrm{CVaR}_\alpha(X)\,\mu(\mathrm{d}\alpha) \qquad \text{for all } X \in L_p(\Omega, \mathcal{A}, \mathbb{P}), \tag{8}$$

where $\mathcal{M}$ is a family of probability measures on $(0,1)$. When this family consists of finitely many finitely supported measures, we say that ρ is *finitely representable* (we note that such risk measures are dense among coherent ones, see Noyan and Rudolf, 2013). If the family $\mathcal{M}$ consist only of a *single* such measure, i.e., if ρ is a convex combination of finitely many CVaRs, then ρ is called a *mixed* CVaR *measure*. Finally, we note that for finite probability spaces the class of mixed CVaR measures coincides with the class of spectral risk measures (Noyan and Rudolf, 2015).

3. Distributionally Robust Optimization Models We are now ready to introduce the main focus of the present work, a class of distributionally robust stochastic optimization problems with decision-dependent ambiguity sets. To begin, let us consider a simple stochastic optimization problem: The decision maker aims to minimize the expected value of an outcome $G(\mathbf{x}, \boldsymbol{\xi})$, where $\mathbf{x}$ is a decision belonging to some feasible set $\mathcal{X}$, and the outcome, given by the mapping $G : \mathcal{X} \times \mathbb{R}^m \to \mathbb{R}$, depends on an m-dimensional random vector $\boldsymbol{\xi}$. In particular, we are interested in problems with *endogenous uncertainty*, where the distribution of the parameter vector $\boldsymbol{\xi}$ can depend on the decision $\mathbf{x}$. More precisely, given mappings $\mathbb{P} : \mathcal{X} \to \mathcal{P}(\Omega, \mathcal{A})$ and $\boldsymbol{\xi} : \mathcal{X} \to \mathcal{L}^m(\Omega, \mathcal{A})$ the problem takes the form

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}(\mathbf{x})}\left(G(\mathbf{x}, \boldsymbol{\xi}(\mathbf{x}))\right). \tag{9}$$

The next step is to account for uncertainty about the distribution of the parameters. To this end, we introduce as our ambiguity set an EMD ball, either of type (BALL-C) or of type (BALL-D), centered on the *nominal* random parameter vector $[\mathbb{P}(\mathbf{x}), \boldsymbol{\xi}(\mathbf{x})] \in \mathcal{V}^m(\Omega, \mathcal{A})$. This leads to the following DRO variants of the (risk-neutral) *underlying problem* (9):

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\boldsymbol{\zeta}\in\mathcal{B}^{\mathbb{P}(\mathbf{x})}_{\delta,\kappa}(\boldsymbol{\xi}(\mathbf{x}))} \mathbb{E}_{\mathbb{B}}\left(G(\mathbf{x},\boldsymbol{\zeta})\right) \tag{DRO-RNC}$$

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{B}^{\boldsymbol{\xi}(\mathbf{x})}_{\delta,\kappa}(\mathbb{P}(\mathbf{x}))} \mathbb{E}_{\mathbb{Q}}\left(G(\mathbf{x},\boldsymbol{\xi}(\mathbf{x}))\right). \tag{DRO-RND}$$

Recalling our notation from Section 2.1, here κ is the radius of the ball, and δ is the underlying distance or dissimilarity measure on $\mathbb{R}^m$. In an applied context the appropriate choices of κ and δ, as well as the choice between the models (DRO-RNC) and (DRO-RND) will be driven both by the specifics of the base problem and by tractability concerns.

Aiming to minimize the expected value of an outcome represents a risk-neutral attitude. To incorporate risk-aversion into our decision problems we can replace the expected value operator in (DRO-RNC) and (DRO-RNC) with an appropriately chosen risk measure ρ, leading to the problems

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\boldsymbol{\zeta}\in\mathcal{B}^{\mathbb{P}(\mathbf{x})}_{\delta,\kappa}(\boldsymbol{\xi}(\mathbf{x}))} \rho\left(G(\mathbf{x},\boldsymbol{\zeta})\right), \tag{DRO-RAC}$$

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{B}^{\boldsymbol{\xi}(\mathbf{x})}_{\delta,\kappa}(\mathbb{P}(\mathbf{x}))} \rho\left([\mathbb{Q}, G(\mathbf{x},\boldsymbol{\xi}(\mathbf{x}))]\right). \tag{DRO-RAD}$$

Remark 3.1 *While our focus in this paper is on the decision-dependent nominal distribution of the parameter vector, our framework could allow for the radius κ of the ambiguity set to also be decision dependent. One possible approach is to make κ itself a decision variable and add to the objective function a term that penalizes low values of κ, effectively introducing a cost of robustness (analogous to the cost associated with the reliability level in chance-constrained optimization, see, e.g., Lejeune and Shen, 2016).*

3.1 Specifying the nominal distribution One of the main distinguishing features of our approach is that the nominal distribution at the center of the ambiguity set is decision-dependent; in this section we briefly discuss possible ways to describe this dependence.

The case when parameter realizations are decision-dependent, but the probabilities of underlying events are not, is fairly straightforward, as it is sufficient to specify the mappings $\mathbf{x} \mapsto \boldsymbol{\xi}^i(\mathbf{x})$ for each scenario $i \in [n]$. In Section 7.1 we present two representative examples of such mappings in the context of machine scheduling problems, where the uncertain parameters are the processing times of jobs. The first example introduces *linearly compressible processing times* with continuous control decisions, while the second example—*control with discrete resources*—features binary control decisions.

We next turn our attention to the opposite case, when parameter realizations are fixed, but probabilities are decision-dependent. While this setting formally appears quite similar to the one discussed above, it is typically very challenging to construct scenario probability mappings $\mathbf{x} \mapsto p^i(\mathbf{x})$ that can properly model problems of practical interest while maintaining a reasonable level of tractability. In Section 7.2 we discuss the state-of-the-art technique of *distribution shaping*, which allows one to express multiplicative probabilities via linear constraints for certain problem classes with binary decisions. Another interesting special case is when the random parameter vector is drawn from a population that consists of subpopulations whose proportions are decision-dependent (see, e.g., Dupacova, 2006; Hellemo, 2016). For example, in a revenue management context the subpopulations would correspond to various customer types or market segments whose proportions are influenced by marketing or pricing decisions. More precisely, given a fixed outcome mapping $\boldsymbol{\xi} : \Omega \to \mathbb{R}$ let $\mathbb{P}_1, \ldots, \mathbb{P}_S \in \mathcal{P}(\Omega)$ denote the probability measures associated with the S subpopulations, and let $\pi_1(\mathbf{x}), \ldots, \pi_S(\mathbf{x})$ denote the corresponding proportions of each subpopulation in the population. Then the nominal parameter vector follows a *mixture distribution* $[\mathbb{P}, \boldsymbol{\xi}]$ with $\mathbb{P} = \sum_{s=1}^{S} \pi_s(\mathbf{x})\mathbb{P}_s$. If we have $\mathcal{X} \subset \mathbb{R}^r$ for some $r \in \mathbb{N}$, and the mappings π_s are affine, with $\pi_s(\mathbf{x}) = \pi_s^0 + \boldsymbol{\pi}_s^\top \mathbf{x}$ for some $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_S \in \mathbb{R}^r$, then scenario probabilities can be expressed via the linear constraints $p^i(\mathbf{x}) = \sum_{s=1}^{S} \boldsymbol{\pi}_s^0 p_s^i + \boldsymbol{\pi}_s^\top \mathbf{x} p_s^i$ for $i \in [n]$.

4. Formulations for continuous Wasserstein balls We now turn our attention to a class of problems where outcome mapping G has a bilinear structure, and the ambiguity set is a continuous Wasserstein-p ball. Our principal tool to obtain potentially tractable formulations for problems in this class will be Proposition 4.2, due to Pflug et al. (2012), which generalizes the following well-known consequence of Hölder's inequality to a stochastic context.

Proposition 4.1 *For any two vectors $\mathbf{v}, \mathbf{y}_0 \in \mathbb{R}^m$ and $\kappa \geq 0$ we have*

$$\sup_{\mathbf{y} \in B_\kappa^p(\mathbf{y}_0)} \mathbf{y}^\top \mathbf{v} = \mathbf{y}_0^\top \mathbf{v} + \kappa \|\mathbf{v}\|_q,$$

where $B_\kappa^p(\mathbf{y}_0) = \{\mathbf{y} \in \mathbb{R}^m \ : \ \|\mathbf{y} - \mathbf{y}_0\|_p \leq \kappa\}$ is the p-norm ball of radius κ centered on $\mathbf{y}_0$.

The above proposition concerns the robustification of a scalar product with respect to one of its factors, using a p-norm ball as the ambiguity set. We next consider a stochastic variant of this problem where we replace the central vector $\mathbf{y}_0$ with a nominal random vector $[\mathbb{B}, \boldsymbol{\xi}]$, and replace the p-norm ball with a Wasserstein-p ball as the ambiguity set. When working in a risk-averse framework, our focus will be on an appropriate risk measure of the arising random scalar products. Following along the lines of Pflug et al. (2012) we introduce an important class of risk measures.

Definition 4.1 *Let $\rho : L_p^S \to \mathbb{R}$ be a law-invariant convex risk measure that admits a representation of the form $\rho(V) = \max\left\{\mathbb{E}_{\mathbb{B}}(VZ) - R(Z) \ : \ Z \in L_q^S\right\}$ where $R : L_q^S \to \bar{\mathbb{R}}$ is a convex functional. When $p > 1$, we say that ρ is* well-behaved with factor $C \in \mathbb{R}_+$ *if*

$$\|Z\|_{L_q} = C \textit{ holds for all } Z \in \bigcup_{V \in L_p^S} \arg\max\left\{\mathbb{E}_{\mathbb{B}}(VZ) - R(Z) \ : \ Z \in L_q^S\right\}.$$

When $p = 1$, we say that ρ is well-behaved with factor C if, for the random variables Z specified in the above condition, in addition to $\|Z\|_{L_\infty} = C$ we also have $Z \in \{0, C\}$ almost everywhere.

Before we state the following key result from Pflug et al. (2012), we recall from Section 2.1 that the Wasserstein-p ball of radius κ centered on a random vector $[\mathbb{B}, \boldsymbol{\xi}] \in \mathcal{V}^m((0,1), \mathcal{A}_B)$ is identical to the EMD ball $\mathcal{B}_{\delta^p, \kappa^p}^{\mathbb{B}}(\boldsymbol{\xi})$ with radius κ^p, where δ^p is the measure of dissimilarity induced by the p-th power of the p-norm.

Proposition 4.2 *Consider a random vector $[\mathbb{B}, \boldsymbol{\xi}] \in \mathcal{V}^m((0,1), \mathcal{A}_B)$, and assume that the law invariant convex risk measure $\rho : L_p \to \mathbb{R}$ is well-behaved with factor C. Then for any $\mathbf{v} \in \mathbb{R}^m$ such that $\boldsymbol{\xi}^\top \mathbf{v} \in L_p^S$ we have*

$$\sup_{\boldsymbol{\zeta} \in \mathcal{B}_{\delta^p, \kappa^p}^{\mathbb{B}}(\boldsymbol{\xi})} \rho(\boldsymbol{\zeta}^\top \mathbf{v}) = \rho(\boldsymbol{\xi}^\top \mathbf{v}) + C\kappa \|\mathbf{v}\|_q. \tag{10}$$

Wozabal (2014) applies this result to provide robustified versions of many popular risk measures; here we only mention the following important corollary:

$$\sup_{\boldsymbol{\zeta} \in \mathcal{B}_{\delta^1, \kappa}^{\mathbb{B}}(\boldsymbol{\xi})} \mathrm{CVaR}_\alpha(\boldsymbol{\zeta}^\top \mathbf{v}) = \mathrm{CVaR}_\alpha(\boldsymbol{\xi}^\top \mathbf{v}) + \frac{1}{1-\alpha} \kappa \|\mathbf{v}\|_\infty. \tag{11}$$

We next examine the implications of this result on the optimization problems introduced in Section 3, focusing on the case when the outcome mapping has a bilinear structure. More precisely, we assume that the outcome mapping is of the form $G(\mathbf{x},\boldsymbol{\zeta}) = \boldsymbol{\zeta}^\top \mathbf{v}(\mathbf{x})$ for some vector-valued mapping $\mathbf{v} : \mathcal{X} \to \mathbb{R}^m$. We first observe that in this case, due to the linearity of expectation, the risk-neutral underlying problem (9) is equivalent to the deterministic problem $\min_{\mathbf{x}\in\mathcal{X}} \bar{\boldsymbol{\xi}}^\top(\mathbf{x})\mathbf{v}(\mathbf{x})$, where the mapping $\bar{\boldsymbol{\xi}} : \mathcal{X} \to \mathbb{R}^m$ is given by $\bar{\boldsymbol{\xi}}(\mathbf{x}) = \mathbb{E}(\boldsymbol{\xi}(\mathbf{x}))$. Noting that the expected value operator is trivially well-behaved with factor 1, it is easy to verify that for $\rho = \mathbb{E}$ the formula (10) becomes equivalent to the conclusion of Proposition 4.1 with $\mathbf{y}_0 = \mathbb{E}(\boldsymbol{\xi})$. Therefore the risk-neutral DRO problem (DRO-RNC) with decision-dependent ambiguity set $\mathcal{B}^{\mathbb{B}}_{\delta^p,\kappa^p}(\boldsymbol{\xi})$ can be equivalently reformulated as the following deterministic optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \bar{\boldsymbol{\xi}}^\top(\mathbf{x})\mathbf{v}(\mathbf{x}) + \kappa\|\mathbf{v}(\mathbf{x})\|_q. \tag{12}$$

The risk-averse variant of the problem, where ρ is an arbitrary law invariant convex risk measure that is well-behaved with some factor C, can be similarly reformulated using Proposition 4.2, leading to

$$\min_{\mathbf{x}\in\mathcal{X}} \rho\left(\boldsymbol{\xi}^\top(\mathbf{x})\mathbf{v}(\mathbf{x})\right) + C\kappa\|\mathbf{v}(\mathbf{x})\|_q. \tag{13}$$

In contrast to the risk-neutral case, this reformulated problem typically remains inherently stochastic.

5. Robustified risk measures in finite spaces

In Section 4 we managed to convert the minimax DRO problem (DRO-RAC), which features a continuous EMD ball of type (BALL-C) as its ambiguity set, to a straightforward minimization. Our eventual goal is to similarly convert the problem (DRO-RAD), which arises when the ambiguity set is a discrete EMD ball of type (BALL-D). The primary difficulty lies in the fact that Proposition 4.2, which provided an elegant way to robustify risk measures in a continuous context by replacing the supremum over the ambiguity set with the closed-form formula (10), is no longer valid in a discrete setting, as the following example shows.

Example 5.1 *Let $\boldsymbol{\xi}$ be a 2-dimensional random vector with possible realizations $(1,0)^\top$ and $(0,1)^\top$, and let $\mathbf{x} = (1,1)^\top$. Then $\mathbb{E}_{\mathbb{Q}}(\mathbf{x}^\top\boldsymbol{\xi}) = 1 < 1 + \kappa\|\mathbf{x}\|_q$ for any probability distribution $\mathbb{Q}$.*

We mention that a one-sided version of Proposition 4.2, analogous to Lemma 1 of Pflug et al. (2012), remains true for discrete EMD balls.

Proposition 5.1 *Consider an arbitrary measurable space $(\Omega,\mathcal{A})$ and a random vector $[\mathbb{P},\boldsymbol{\xi}] \in \mathcal{V}^m(\Omega,\mathcal{A})$. If the law invariant convex risk measure $\rho : L_p \to \mathbb{R}$ is well-behaved with factor C, then for any $\mathbf{v} \in \mathbb{R}^m$ such that $\boldsymbol{\xi}^\top\mathbf{v} \in L^S_p$ we have*

$$\sup_{\mathbb{Q}\in\mathcal{B}^{\boldsymbol{\xi}}_{\delta^p,\kappa^p}(\mathbb{P})} \rho\left([\mathbb{Q},\boldsymbol{\zeta}^\top\mathbf{v}]\right) \le \rho\left([\mathbb{P},\boldsymbol{\xi}^\top\mathbf{v}]\right) + C\kappa\|\mathbf{v}\|_q. \tag{14}$$

Proof. The discrete EMD ball $\mathcal{B}^{\boldsymbol{\xi}}_{\delta^p,\kappa^p}(\mathbb{P})$ can be trivially embedded into the continuous ball $\mathcal{B}^{\mathbb{P}}_{\delta^p,\kappa^p}(\boldsymbol{\xi})$ as follows. Consider a probability measure $\mathbb{Q} \in \mathcal{B}^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$. Using the well-known fact that every finite-dimensional distribution can be realized on a probability space that admits a continuous uniform distribution, there exists a mapping $\boldsymbol{\zeta} \in \mathcal{L}^m([0,1])$ such that $\text{law}[\mathbb{B},\boldsymbol{\zeta}] = \text{law}[\mathbb{Q},\boldsymbol{\xi}]$. As EMDs are defined in a law-invariant fashion, $\boldsymbol{\zeta} \in \mathcal{B}^{\mathbb{P}}_{\delta,\kappa}(\boldsymbol{\xi})$ immediately follows. Furthermore, since ρ is also law-invariant, we have $\rho\left([\mathbb{Q},\boldsymbol{\xi}^\top\mathbf{v}]\right) = \rho\left([\mathbb{B},\boldsymbol{\zeta}^\top\mathbf{v}]\right)$. Therefore the supremum in (14) is taken over a smaller set than the one in (10), which implies our proposition. □

While we do not have closed-form analogue to formula (10) for discrete spaces, in this section we develop some mathematical tools to replace the supremum involved in the robustification of certain risk measures with an equivalent minimization. These tools will then be utilized to recast (DRO-RAD) as a conventional optimization problem; in Section 6.1 we examine certain important cases where this approach leads to potentially tractable formulations. Throughout the remainder of this section $\boldsymbol{\xi} : \Omega \to \mathbb{R}^m$ will denote a fixed mapping from a finite sample space of size n, and we will use the notation $\delta^{ij} = \delta\left(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j\right)$ for distances among the realizations of $\boldsymbol{\xi}$, where $i, j \in [n]$.

5.1 A parametric relation between random variables For two scalar-valued random variables $X, Y \in \mathcal{L}^1(\Omega, 2^\Omega)$ the usual ordering relation $X \geq Y$ holds if and only if we have $x^i \geq y^j$ for all $i, j \in [n]$. A key idea behind the developments of this section is that one can robustify certain risk expressions by replacing the usual ordering with a parametric family of relations, and introducing a corresponding "penalty term".

Definition 5.1 *Given a threshold $\tau \geq 0$ we define the relation $\succeq_\tau$ as follows. For $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A})$*

$$X \succeq_\tau Y \quad \textit{holds if and only if we have} \quad x^i \geq y^j - \delta^{ij}\tau \quad \textit{for all } i, j \in [n]. \tag{15}$$

While $\succeq_\tau$ is typically not a preorder among random variables, it is closely related to the usual ordering. The properties below are easily verified:

(i) The relation $X \succeq_\tau Y$ implies $X \geq Y$, due to the reflexivity of δ.

(ii) If δ is definite, then for sufficiently high values of τ the relations $X \succeq_\tau Y$ and $X \geq Y$ are equivalent. In particular, the equivalence holds when $\tau \geq \max\limits_{i,j \in [n]} \frac{y^j - x^i}{\delta^{ij}}$.

(iii) The relation $X \succeq_0 Y$ is equivalent to $X \geq \sup(Y)$.

(iv) When δ is the discrete metric, the relation $X \succeq_\tau Y$ is equivalent to the conventional inequality $X \geq \max\left(Y,\ \sup(Y) - \tau\right)$.

We next present and discuss the main results of Section 5, which will then be proved in Section 5.3.

5.2 Robustified risk formulas Let $\mathbb{P} \in \mathcal{P}(\Omega, 2^\Omega)$ be a fixed nominal probability measure. Given a risk measure $\rho : L_p^S \to \mathbb{R}$ and a radius $\kappa \geq 0$ we define the *robustified risk measure* $\rho^\kappa : \mathcal{L}^1(\Omega, 2^\Omega) \to \mathbb{R}$ on our finite probability space by

$$\rho^\kappa(Z) = \sup\left\{\rho\left([\mathbb{Q}, Z]\right) \ : \ \mathbb{Q} \in \mathcal{B}^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})\right\} \quad \text{for } Z \in \mathcal{L}^1(\Omega, 2^\Omega). \tag{16}$$

We now present the robustified versions of several important risk measures; the corresponding proofs can be found in the next section.

5.2.1 Robustified expectation The following expression closely parallels the trivial formula $\mathbb{E}_\mathbb{P}(Z) = \inf\left\{\mathbb{E}_\mathbb{P}(V) \ : \ V \geq Z\right\}$ for the nominal expectation, with the relation $\succeq_\tau$ playing a similar role to that of the usual ordering $\geq$:

$$\mathbb{E}^\kappa(Z) = \inf\left\{\mathbb{E}_\mathbb{P}(V) + \kappa\tau \ : \ \tau \geq 0,\ V \succeq_\tau Z\right\}. \tag{17}$$

The additional "robustification term" $\kappa\tau$, which also appears in the results below, is analogous to the term seen when robustifying the expected value operator in a continuous space (see Section 4).

Example 5.2 (Total variation distance) *When the ambiguity set is based on the total variation distance, it is easy to identify the worst-case distribution, as it can be obtained by greedily "transferring probability" from lower outcomes (starting with the lowest one) to the worst-case outcome, until either the boundary of the ambiguity set is reached, or all probability is transferred to the worst case. As it has been observed in the literature (Jiang and Guan, 2018, Theorem 1; see also Rahimian et al., 2018, Proposition 3), this implies that the robustified expectation is a convex combination of the worst-case outcome and the nominal* CVaR *at an appropriate level, and thus a coherent risk measure of the outcome.*

More precisely, if δ is the discrete metric, then, introducing the notation $z^+ = \sup(Z)$, for $\kappa \in [0,1]$ we have $\mathbb{E}^\kappa(Z) = \kappa z^+ + (1-\kappa)\,\mathrm{CVaR}_\kappa(Z)$. Using the representation (6) for CVaR*, we can then express $\mathbb{E}^\kappa(Z)$ as the optimum of the following LP:*

$$\begin{aligned}
\min \quad & \kappa z^+ + (1-\kappa)\left(\eta + \frac{1}{1-\kappa}\sum_{i\in[n]} p^i \hat{v}^i\right) && && (18a)\\
s.t. \quad & \hat{v}^i \geq z^i - \eta, && \forall i \in [n] && (18b)\\
& \hat{v}^i \geq 0, && \forall i \in [n] && (18c)\\
& \eta \leq z^+. && && (18d)
\end{aligned}$$

Here the redundant constraint (18d) reflects the trivial inequality $\mathrm{VaR}_\kappa(Z) \leq \sup(Z)$. The above formulation turns out to be essentially the same as the LP formulation of (17) given in (29). To see the correspondence between these two LPs, we first note that in accordance with Property (iv) we can rewrite constraints (29b) as

$$\begin{aligned}
& v^i \geq z^i, && \forall i \in [n]\\
& v^i \geq z^+ - \tau, && \forall i \in [n].
\end{aligned}$$

Let us introduce the change of variables $\eta = z^+ - \tau$, $\hat{v}^i = v^i + \tau - z^+$ for $i \in [n]$. It is now easy to verify that the formulations (18) and (29) are equivalent. We note that the preceding argument constitutes an alternative proof for Theorem 1 of Jiang and Guan (2018) in our discrete setting. Additionally, it follows that the optimum in (17) can be attained when we have $\tau = \sup(Z) - \mathrm{VaR}_\kappa(Z)$.

5.2.2 Robustified CVaR. Recalling the definition of CVaR from (4), for a probability level $\alpha \in [0,1)$ we have

$$\mathrm{CVaR}_\alpha^\kappa(Z) = \inf\left\{\eta + \mathbb{E}_{\mathbb{P}}(S) + \kappa\tau \ : \eta \in \mathbb{R},\ \tau \geq 0,\ S \succeq_\tau \frac{1}{1-\alpha}[Z-\eta]_+\right\}. \tag{19}$$

This robustified expression exhibits a similar structure to (4), again with an additional robustification term. By applying a scaling factor of $(1-\alpha)$ to S and τ, we can also rewrite (19) as

$$\mathrm{CVaR}_\alpha^\kappa(Z) = \inf\left\{\eta + \mathbb{E}_{\mathbb{P}}\left(\frac{1}{1-\alpha}S\right) + \frac{1}{1-\alpha}\kappa\tau \ : \ \eta \in \mathbb{R},\ \tau \geq 0,\ S \succeq_\tau [Z-\eta]_+\right\}. \tag{20}$$

This version better highlights the parallels with the corresponding continuous result in (11), where the robustification term for $\mathrm{CVaR}_\alpha(\boldsymbol{\xi}^\top \mathbf{z})$ took the form $\frac{1}{1-\alpha}\kappa\|\mathbf{z}\|_\infty$. However, in contrast to (19), the formula (20) does not generalize in a straightforward fashion to mixed CVaR measures.

Example 5.3 (Total variation distance) *Similarly to the case of robustified expectation, when the ambiguity set is based on the total variation distance, we can express $\mathrm{CVaR}_\alpha^\kappa(Z)$ as a convex combination of the worst-case outcome, and a nominal* CVaR *of the outcome at an appropriate level. Recalling our notation from Example 5.2, we first observe that if $\kappa \geq 1-\alpha$ holds, then the ambiguity set contains a distribution where Z takes value z^+ with a probability of at least $1-\alpha$, which immediately implies $\mathrm{CVaR}_\alpha^\kappa(Z) = z^+$. On the other hand, in the non-trivial case when $\kappa \leq 1-\alpha$ holds, we have*

$$\mathrm{CVaR}_\alpha^\kappa(Z) = \frac{\kappa}{1-\alpha}z^+ + \frac{1-\alpha-\kappa}{1-\alpha}\,\mathrm{CVaR}_{\alpha+\kappa}(Z). \tag{21}$$

While we are not aware of the above formula appearing elsewhere in the literature, it can be proved analogously to Theorem 1 in Jiang and Guan (2018), because the worst-case distribution is obviously the same as for the case of robustified expectation. To obtain an alternative proof, we can also start from the LP representation (31) of the formula (19), and apply the same change of variables as in Example 5.2 to obtain an LP representation of (21). Like before, this approach also shows that the optimum in (19) can be obtained when we have $\tau = z^+ - \mathrm{VaR}_{\alpha+\kappa}(Z)$.

5.2.3 Robustified mixed CVaR. Making explicit the definition from Section 2.2, given a finitely supported probability measure μ on the interval $[0,1)$, the *mixed* CVaR risk measure $\rho_{\{\mu\}} : L_p^S \to \mathbb{R}$ is given by

$$\rho_{\{\mu\}}(Z) = \int_0^1 \mathrm{CVaR}_\alpha(Z)\,\mu(\mathrm{d}\alpha) = \sum_{\alpha\in\mathrm{supp}(\mu)} \mu\left(\{\alpha\}\right)\mathrm{CVaR}_\alpha(Z). \tag{22}$$

We note that, according to the above expression, the risk measure $\rho_{\{\mu\}}$ can be interpreted as the expected value of CVaR_α when the level α is randomly selected from the interval $[0,1)$ according to the probability measure μ. More precisely, if we denote the identity function of the interval by $A : [0,1) \to [0,1)$, then we have $\rho_{\{\mu\}}\left([\mathbb{P}, Z]\right) = \mathbb{E}_\mu\left(\mathrm{CVaR}_A\left([\mathbb{P}, Z]\right)\right)$. The robustification of $\rho^\kappa_{\{\mu\}}$ is now given by the following generalization of (19):

$$\rho^\kappa_{\{\mu\}}(Z) = \inf\left\{\mathbb{E}_\mu(H) + \mathbb{E}_\mathbb{P}(S) + \kappa\tau \ : \ H \in \mathbb{R}^{[0,1)},\ \tau \geq 0,\ S \succeq_\tau \mathbb{E}_\mu\left(\frac{1}{1-A}[Z-H]_+\right)\right\}. \tag{23}$$

Here A ("capital alpha") is viewed as the probability level of CVaR, selected randomly according to μ. Similarly, the random variable H ("capital eta") plays the role of the VaR value at level A.

5.2.4 Robustified finitely representable risk measures. As discussed in Section 2.2, a finite family $\mathcal{M}$ of finitely supported probability measures on $[0,1)$ defines a finitely representable risk measure $\rho_\mathcal{M} : L_p^S \to \mathbb{R}$ given by

$$\rho_\mathcal{M}(Z) = \sup_{\mu\in\mathcal{M}} \rho_{\{\mu\}}(Z). \tag{24}$$

While the motivation behind the next formula is to robustify this important class of risk measures, it remains valid even when the cardinality of the family $\mathcal{M}$ is infinite.

$$\rho^\kappa_\mathcal{M}(Z) = \inf\left\{R\in\mathbb{R} \ : \ H\in\mathbb{R}^{[0,1)},\, \boldsymbol{\tau}\in\mathbb{R}^\mathcal{M}_+,\ \begin{array}{l} S_\mu \succeq_{\tau_\mu} \mathbb{E}_\mu\left(\frac{1}{1-A}[Z-H]_+\right), \\ R \geq \mathbb{E}_\mu(H) + \mathbb{E}_\mathbb{P}(S_\mu) + \kappa\tau_\mu \end{array} \quad \forall\mu\in\mathcal{M}\right\}. \tag{25}$$

We remark that the domain of the mapping $H : [0,1) \to \mathbb{R}$ in the above formulas can be restricted from $[0,1)$ to the support set $\bigcup_{\mu\in\mathcal{M}} \mathrm{supp}(\mu)$. Similarly to the role of the threshold η in the expected excess-based representation (4) of CVaR, we can view H as representing the VaR functional under the worst-case distribution in the ambiguity set. More precisely, if for $\rho = \rho_\mathcal{M}$ the supremum in (16) is attained at $\mathbb{P}^* \in \mathcal{B}^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$, then the choice $H^*(\alpha) = \mathrm{VaR}_\alpha\left([\mathbb{P}^*, Z]\right)$ is optimal in (23) and (25).

5.2.5 Robustification in discrete and continuous cases. We would like to highlight that the above robustification formulas exhibit fundamentally different qualitative properties than their counterparts in continuous spaces, despite the similar formal structures. In more detail, Pflug et al. (2012) show that, when taking the supremum in a Wasserstein ball of type (BALL-C), the worst-case distribution can be obtained by starting from the nominal random realization vector, and moving in a fixed direction until we reach the boundary of the ball. This leads to the robustified risk growing linearly in terms of the ball radius, as seen in (10). By contrast, when considering balls of type (BALL-D), the supremum is bounded by the risk achieved at the degenerate distribution where all probability is concentrated on the worst-case outcome. Therefore, if the ambiguity ball is large enough to contain this degenerate distribution, further increasing the radius has no impact on the robustified risk. These behaviors are illustrated in Figure 1, which compares the Wasserstein-1 robustifications of $\mathrm{CVaR}_{0.5}$ for an equal-weight three-asset portfolio, where the nominal asset loss realizations have been randomly generated, and are equally likely.

5.3 Proof of robustified risk formulas We will use linear programming duality to derive the formulas of the previous section. To this end, let us begin by establishing a characterization of EMD balls in finite probability spaces via a system of linear inequalities.

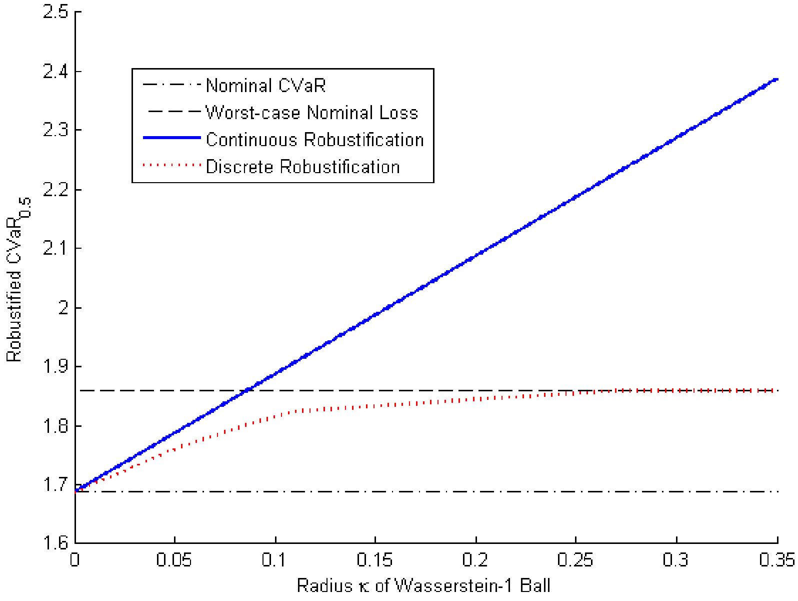

Figure 1: Continuous vs. discrete robustification

Lemma 5.1 *For two probability measures $\mathbb{P},\mathbb{Q}\in\mathcal{P}(\Omega,2^\Omega)$ and a radius $\kappa\geq 0$ we have $\mathbb{Q}\in B^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$ if and only if the following system of inequalities is feasible.*

$$\sum_{j\in[n]}\gamma^{ij}=p^i,\qquad \forall\ i\in[n] \tag{26a}$$

$$\sum_{i\in[n]}\gamma^{ij}=q^j,\qquad \forall\ j\in[n] \tag{26b}$$

$$\sum_{i\in[n]}\sum_{j\in[n]}\delta^{ij}\gamma^{ij}\leq\kappa, \tag{26c}$$

$$\boldsymbol{\gamma}\in\mathbb{R}^{n\times n}_{+}. \tag{26d}$$

Proof. Introducing the notation $\hat{\mathbb{P}}=\text{law}[\boldsymbol{\xi},\mathbb{P}]$ and $\hat{\mathbb{Q}}=\text{law}[\boldsymbol{\xi},\mathbb{Q}]$ the condition $\mathbb{Q}\in B^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$ is by definition equivalent to the inequality $\Delta(\hat{\mathbb{P}},\hat{\mathbb{Q}})\leq\kappa$. This inequality is in turn is equivalent to the feasibility of the following system of inequalities:

$$\sum_{\mathbf{y}\in\text{supp}(\hat{\mathbb{Q}})}\hat{\gamma}(\mathbf{x},\mathbf{y})=\hat{\mathbb{P}}\left(\{\mathbf{x}\}\right),\qquad\forall\ \mathbf{x}\in\text{supp}(\hat{\mathbb{P}}) \tag{27a}$$

$$\sum_{\mathbf{x}\in\text{supp}(\hat{\mathbb{P}})}\hat{\gamma}(\mathbf{x},\mathbf{y})=\hat{\mathbb{Q}}\left(\{\mathbf{y}\}\right),\qquad\forall\ \mathbf{y}\in\text{supp}(\hat{\mathbb{Q}}) \tag{27b}$$

$$\sum_{\mathbf{x}\in\text{supp}(\hat{\mathbb{P}})}\sum_{\mathbf{y}\in\text{supp}(\hat{\mathbb{Q}})}\delta(\mathbf{x},\mathbf{y})\hat{\gamma}(\mathbf{x},\mathbf{y})\leq\kappa, \tag{27c}$$

$$\hat{\gamma}:\text{supp}(\hat{\mathbb{P}})\times\text{supp}(\hat{\mathbb{Q}})\to\mathbb{R}_{+}. \tag{27d}$$

We can obtain this second equivalence by directly applying the EMD definition (1) to finitely supported measures, with the joint probability measure $\mathbb{P}^*$ supported on $\text{supp}(\hat{\mathbb{P}})\times\text{supp}(\hat{\mathbb{Q}})$ and given there by $\mathbb{P}^*\left(\{(\mathbf{x},\mathbf{y})\}\right)=\hat{\gamma}(\mathbf{x},\mathbf{y})$. The lemma then follows immediately from the two observations below:

- Assume that the system (26) has a feasible solution $\boldsymbol{\gamma}$. Keeping in mind the trivial equalities $\hat{\mathbb{P}}\left(\{\mathbf{x}\}\right)=\sum\limits_{i:\boldsymbol{\xi}^i=\mathbf{x}}p^i$ and $\hat{\mathbb{Q}}\left(\{\mathbf{y}\}\right)=\sum\limits_{j:\boldsymbol{\xi}^j=\mathbf{y}}q^j$, it is easy to verify that the aggregated values

$\hat{\gamma}(\mathbf{x},\mathbf{y}) = \sum\limits_{i\,:\,\boldsymbol{\xi}^i=\mathbf{x}} \sum\limits_{j\,:\,\boldsymbol{\xi}^j=\mathbf{y}} \gamma^{ij}$ solve the system (27), which implies $\mathbb{Q} \in B^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$.

- If $\mathbb{Q} \in B^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$ holds, then the system (27) has a feasible solution $\hat{\gamma}$. It is again easy to verify that the disaggregated values $\gamma^{ij} = \hat{\gamma}\left(\boldsymbol{\xi}(\omega^i), \boldsymbol{\xi}(\omega^j)\right) \frac{p^i}{\hat{\mathbb{P}}(\boldsymbol{\xi}^i)} \frac{q^j}{\hat{\mathbb{Q}}(\boldsymbol{\xi}^j)}$, where $\frac{0}{0}$ is understood as zero, solve the system (26).

□

5.3.1 Robustified expectation We first point out that the formula (17) follows directly from applying the CVaR formula (19) with $\alpha = 0$. Here we also present a short stand-alone proof, which will serve as a template for our later more complex arguments. By Lemma 5.1 we can express the robustified expectation $\mathbb{E}^\kappa(Z)$ as the optimum value of the following LP:

$$\max \quad \left\{ \sum_{j\in[n]} z^j q^j \ : \ (26a)\text{–}(26d) \right\}. \tag{28}$$

We can somewhat simplify this LP by replacing each variable q^j with the sum $\sum_{i\in[n]} \gamma^{ij}$, and removing the now redundant defining constraints (26b). By taking the dual of the simplified LP we can express $\mathbb{E}^\kappa(Z)$ via linear minimization as

$$\begin{aligned}
\min \quad & \sum_{i\in[n]} p^i v^i + \kappa\tau && && (29a)\\
\text{s.t.} \quad & v^i \geq z^j - \delta^{ij}\tau, && \forall i,j \in [n] && (29b)\\
& \tau \geq 0. && && (29c)
\end{aligned}$$

Noting that $\sum\limits_{i\in[n]} p^i v^i = \mathbb{E}_{\mathbb{P}}(V)$, and that the constraints (29b) are equivalent to the relation $V \succeq_\tau Z$, the desired formula (17) follows.

5.3.2 Robustified CVaR. Following the same logic as before, we can combine Lemma 5.1 with the dual representation of CVaR given in (7) to obtain the robustified CVaR value $\mathrm{CVaR}^\kappa_\alpha(Z)$ as the optimum value of the LP

$$\begin{aligned}
\max \quad & \frac{1}{1-\alpha} \sum_{j\in[n]} z^j \beta^j && && (30a)\\
\text{s.t.} \quad & (26a)\text{–}(26c), && && (30b)\\
& \beta^j \leq q^j, && \forall\ j \in [n] && (30c)\\
& \sum_{j\in[n]} \beta^j = 1-\alpha, && && (30d)\\
& \boldsymbol{\gamma} \in \mathbb{R}^{n\times n}_+, \quad \boldsymbol{\beta} \in \mathbb{R}^n_+. && && (30e)
\end{aligned}$$

We can again simplify the LP formulation by eliminating the q^j variables, and take the dual afterwards. Applying a scaling factor of $1-\alpha$ to each dual variable, we arrive at the following expression of $\mathrm{CVaR}^\kappa_\alpha(Z)$:

$$\begin{aligned}
\min \quad & \eta + \frac{1}{1-\alpha} \sum_{i\in[n]} p^i v^i + \frac{1}{1-\alpha}\kappa\tau && && (31a)\\
\text{s.t.} \quad & v^i \geq z^j - \eta - \delta^{ij}\tau, && \forall i,j \in [n] && (31b)\\
& \mathbf{v} \in \mathbb{R}^n_+, && && (31c)\\
& \tau \geq 0. && && (31d)
\end{aligned}$$

The constraints (31b) are clearly equivalent to the relation $V \succeq_\tau Z - \eta$, and the non-negativity of V immediately implies $V \succeq_\tau 0$. Combining these two relations we obtain $V \succeq_\tau [Z-\eta]_+$, and the formula

(20), which is trivially equivalent to the desired (19), follows. We mention that, in addition to its role in proving our concise formulas, the LP formulation (31) will also prove valuable as a tool to explicitly incorporate robustified risk into mathematical programming formulations.

5.3.3 Robustified mixed CVaR. Linear formulations involving CVaR can be extended to mixed CVaR measures by introducing duplicate variables and constraints corresponding to each probability level in the (finite) support of the mixing measure (see Noyan and Rudolf 2013 or Noyan and Rudolf 2018 for more detailed discussion and examples). The desired formula (23) follows from these extended linear formulations via LP duality in exactly the same fashion as before, so for the sake of conciseness we omit the lengthy details.

5.3.4 Robustified finitely representable risk measures. Finally, bypassing a direct LP duality argument, the formula (25) follows directly from (23). Combining the observation that we have

$$\rho^{\kappa}_{\mathcal{M}}(Z) = \sup_{\mathcal{Q}} \sup_{\mu} \rho_{\{\mu\}}\left([\mathbb{P}, Z]\right) = \sup_{\mu} \sup_{\mathcal{Q}} \rho_{\{\mu\}}\left([\mathbb{P}, Z]\right) = \sup_{\mu} \rho^{\kappa}_{\{\mu\}}(Z)$$

with the trivial formula $\sup A = \inf\{R \in \mathbb{R} \ : \ R \geq a \quad \forall a \in A\}$ for expressing the supremum of a set $A \subset \mathbb{R}$ we immediately obtain (25), with one slight difference: the formula (25) features a single variable H, while the direct approach we outlined would introduce an indexed family $(H_\mu)_{\mu \in \mathcal{M}}$, similarly to other duplicated variables. However, as discussed at the end of Section 5.2, it can be assumed without loss of generality that these H_μ variables all express the VaR functional under the worst-case distribution, and therefore coincide.

6. Formulations for discrete EMD balls The robustification formula (19) and its LP expression (30) enable us to recast our minimax DRO problem as a conventional minimization problem for the case $\rho = \mathrm{CVaR}_\alpha$. Using the system (30) to represent the supremum in (DRO-RAD) we obtain the formulation

$$\begin{aligned}
\min \quad & \eta + \frac{1}{1-\alpha} \sum_{i \in [n]} p^i(\mathbf{x}) v^i + \frac{1}{1-\alpha} \kappa \tau && && (32a)\\
\text{s.t.} \quad & v^i \geq G(\mathbf{x}, \boldsymbol{\xi}^j(\mathbf{x})) - \eta - \delta^{ij} \tau, && \forall i, j \in [n] && (32b)\\
& \delta^{ij} = \delta\left(\boldsymbol{\xi}^i(\mathbf{x}), \boldsymbol{\xi}^j(\mathbf{x})\right), && \forall i, j \in [n] && (32c)\\
& \mathbf{v} \in \mathbb{R}^n_+, \ \tau \in \mathbb{R}_+, \ \mathbf{x} \in \mathcal{X}. && && (32d)
\end{aligned}$$

The case when we have $\alpha = 0$ and $\rho = \mathrm{CVaR}_0 = \mathbb{E}$ is somewhat simpler, because we can utilize (29) in place of (30) to formulate (DRO-RND) as

$$\begin{aligned}
\min \quad & \sum_{i \in [n]} p^i(\mathbf{x}) v^i + \kappa \tau && && (33a)\\
\text{s.t.} \quad & v^i \geq G(\mathbf{x}, \boldsymbol{\xi}^j(\mathbf{x})) - \delta^{ij} \tau, && \forall i, j \in [n] && (33b)\\
& \delta^{ij} = \delta\left(\boldsymbol{\xi}^i(\mathbf{x}), \boldsymbol{\xi}^j(\mathbf{x})\right), && \forall i, j \in [n] && (33c)\\
& \mathbf{v} \in \mathbb{R}^n, \ \tau \in \mathbb{R}_+, \ \mathbf{x} \in \mathcal{X}. && && (33d)
\end{aligned}$$

We saw in Section 4 that the risk-neutral underlying problem (9) is deterministic. However, this is no longer the case for the above DRO variant. As Example 7.4 shows, it is possible that, given two nominal distributions with the same mean, the arising robustified expectations are different.

Remark 6.1 *The LP expression of the robustified* CVaR *formula* (19) *facilitated a conventional optimization formulation of* (DRO-RAD)*. As discussed in Section 5.3, analogous, although more complex, linear expressions can be obtained for the robustification formulas* (23) *and* (25) *for mixed and finitely* *representable coherent risk measures. Similarly to* (32)*, these linear formulations can then be used to cast* (DRO-RAD) *as a conventional minimization problem when the risk measure ρ belongs to one of these more general classes. As our primary focus in the remainder of this paper is on problems that feature the canonical risk measure $\rho = \mathrm{CVaR}_\alpha$, the arising extended versions of* (32) *are omitted for the sake of brevity.*

6.1 Towards tractable formulations We have seen that, under appropriate assumptions, it is possible to reformulate (DRO-RAD) as a (typically non-linear) optimization problem of the form (32). We now turn our attention to the computational challenges involved in solving such problems, and will examine several important problem classes where these challenges can be mitigated.

6.1.1 Decision-independent nominal realizations If the uncertain vector $\boldsymbol{\xi}(\mathbf{x})$ depends on the decision $\mathbf{x}$ in a non-trivial fashion, then this dependence becomes a significant source of non-linearity in (32). However, if the nominal realizations are decision-independent, then we can drop the argument $\mathbf{x}$ from the terms $\boldsymbol{\xi}^i(\mathbf{x})$, $\boldsymbol{\xi}^j(\mathbf{x})$ for all $i,j \in [n]$, and replace them with a common uncertain vector $\boldsymbol{\xi}$. Consequently, the distance values δ^{ij} can also be viewed as fixed parameters, defined by the equations $\delta^{ij} = \delta\left(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j\right)$. If the set $\mathcal{X}$ of feasible decisions is polyhedral, and the mapping $\mathbf{x} \mapsto G(\mathbf{x}, \boldsymbol{\xi})$ is linear, then (32) becomes a linearly constrained problem (apart from the possible non-linearity implicit in the constraint $\mathbf{x} \in \mathcal{X}$). A more general version of this statement is given precise form in the remark below. Along similar lines, if $\mathbb{P}(\mathbf{x})$ depends on $\mathbf{x}$ in a linear fashion, then the objective function is quadratic.

Remark 6.2 *Consider a feasible set $\mathcal{X} \subset \mathbb{R}^{r_1}$ for some $r_1 \in \mathbb{N}$, and assume that we can express $G(\mathbf{x}, \boldsymbol{\xi}^j)$ as the minimum of an LP. More precisely, we assume that for each $j \in [n]$ there exist matrices $A_j \in \mathbb{R}^{r_3 \times r_2}$, $B_j \in \mathbb{R}^{r_3 \times r_1}$ and vectors $\mathbf{c}_j \in \mathbb{R}^{r_1}$, $\mathbf{d}_j \in \mathbb{R}^{r_2}$, $\mathbf{b}_j \in \mathbb{R}^{r_3}$ for some $r_2, r_3 \in \mathbb{N}$ such that for every decision $\mathbf{x} \in \mathcal{X}$ the outcome $G(\mathbf{x}, \boldsymbol{\xi}^j)$ is the minimum of the LP*

$$\begin{aligned} \min \quad & \mathbf{c}_j^\top \mathbf{x} + \mathbf{d}_j^\top \mathbf{y} \\ s.t. \quad & A_j \mathbf{y} \geq B_j \mathbf{x} + \mathbf{b}_j, \\ & \mathbf{y} \in \mathbb{R}^{r_2}. \end{aligned}$$

Then we can formulate (32) *as the following linearly constrained program:*

$$\begin{aligned} \min \quad & \eta + \frac{1}{1-\alpha} \sum_{i \in [n]} p^i(\mathbf{x}) v^i + \frac{1}{1-\alpha} \kappa\tau \\ s.t. \quad & v^i \geq \mathbf{c}_j^\top \mathbf{x} + \mathbf{d}_j^\top \mathbf{y}_j - \eta - \delta^{ij}\tau, && \forall i,j \in [n] \\ & A_j \mathbf{y}_j \geq B_j \mathbf{x} + \mathbf{b}_j, && \forall j \in [n] \\ & \mathbf{v} \in \mathbb{R}^n_+, \ \tau \in \mathbb{R}_+, \ \mathbf{x} \in \mathcal{X}, \\ & \mathbf{y}_j \in \mathbb{R}^{r_2}, && \forall j \in [n]. \end{aligned}$$

6.1.2 Using the discrete metric Let us assume that δ is the discrete metric given by (3). As discussed in Section 2.1, this choice of δ allows us to use total variation distance-based balls as ambiguity sets. We now present a streamlined formulation of our DRO problem under the additional assumptions that neither the nominal realizations nor the outcomes are decision-dependent. Remarkably, while these assumptions appear to be highly restrictive, the resulting problem class still contains highly non-trivial instances of practical interest, such as our formulations for the pre-disaster planning problems detailed in Section 7.2. Let us again denote the nominal realizations by $\boldsymbol{\xi}^i \in \mathbb{R}^m$, and the corresponding outcome realizations by $G^i \in \mathbb{R}$, for $i \in [n]$. In addition, let $G^+ = \max_{j \in [n]} G^j$. We can then reformulate (32) as follows (matching Property (iv) in Section 5.1):

$$\min \quad \eta + \frac{1}{1-\alpha} \sum_{i \in [n]} p^i(\mathbf{x}) v^i + \frac{1}{1-\alpha} \kappa\tau \tag{34a}$$

$$\begin{aligned}
\text{s.t.} \quad & v^i \geq G^i - \eta, && \forall i \in [n] && \text{(34b)}\\
& v^i \geq G^+ - \eta - \tau, && \forall i \in [n] && \text{(34c)}\\
& \mathbf{v} \in \mathbb{R}^n_+,\ \tau \in \mathbb{R}_+,\ \mathbf{x} \in \mathcal{X}. && && \text{(34d)}
\end{aligned}$$

Analogously to the difference between (32) and (33), when the underlying problem is risk-neutral, i.e., when we have $\alpha = 0$, we can further simplify the above formulation by removing (or setting to zero) the auxiliary variable η, and dropping the non-negativity requirement for the variables $\mathbf{v}$.

6.1.3 Using the Wasserstein-1 metric When the nominal realizations are decision-dependent, the distances between pairs of realizations are represented by the variables δ^{ij} in (32). Whether the corresponding defining constraints (32c) can be represented in a fashion that is amenable to computations depends on the choice of the reflexive mapping $\delta : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$. We now examine the important case when δ is the 1-norm distance, i.e., when the ambiguity set is a Wasserstein-1 ball. Let us assume that the decision-dependent parameters are bounded, i.e., that there exists some $M \in \mathbb{R}_+$ such that we have $\|\boldsymbol{\xi}(\mathbf{x})\|_{L_\infty} < \frac{M}{2}$.

Remark 6.3 *The scaling for the constant $\frac{M}{2}$ in the previous condition was chosen in order to simplify the notation in our optimization formulations. It is easy to see that the condition is satisfied when $\mathcal{X}$ is compact and the mapping $\mathbf{x} \mapsto \boldsymbol{\xi}(\mathbf{x})$ is continuous. In the general case the boundedness condition can be replaced by the following weaker requirement: We assume that the range of each coordinate of the parameter vector is bounded by a decision-independent constant, i.e., that there exists $M \in \mathbb{R}_+$ such that $\left|\xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})\right| < M$ holds for all $i, j \in [n]$ and $k \in [m]$.*

Noting that the equations in (32c) will take the form

$$\delta^{ij} = \delta\left(\boldsymbol{\xi}^i(\mathbf{x}), \boldsymbol{\xi}^j(\mathbf{x})\right) = \left\|\boldsymbol{\xi}^i(\mathbf{x}) - \boldsymbol{\xi}^j(\mathbf{x})\right\|_1 = \sum_{k\in[m]} \left|\xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})\right|, \tag{35}$$

let us introduce the auxiliary variables ν^{ij}_k to represent the values $|\xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})|$ for all $i, j \in [n]$ and $k \in [m]$. We can then equivalently reformulate our problem as

$$\begin{aligned}
\min \quad & \eta + \frac{1}{1-\alpha} \sum_{i\in[n]} p^i(\mathbf{x}) v^i + \frac{1}{1-\alpha}\kappa\tau && && \text{(36a)}\\
s.t. \quad & v^i \geq G(\mathbf{x}, \boldsymbol{\xi}^j(\mathbf{x})) - \eta - \sum_{k\in[m]} \nu^{ij}_k\, \tau, && \forall\ i \in [n],\ j \in [n] && \text{(36b)}\\
& \nu^{ij}_k \leq \xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x}) + M\lambda^{ij}_k, && \forall\ i \in [n],\ j \in [n],\ k \in [m] && \text{(36c)}\\
& \nu^{ij}_k \leq -\xi^i_k(\mathbf{x}) + \xi^j_k(\mathbf{x}) + M(1 - \lambda^{ij}_k), && \forall\ i \in [n],\ j \in [n],\ k \in [m] && \text{(36d)}\\
& \boldsymbol{\lambda} \in \{0,1\}^{n\times n\times m}, \quad \boldsymbol{\nu} \in \mathbb{R}^{n\times n\times m}_+, && && \text{(36e)}\\
& \mathbf{v} \in \mathbb{R}^n_+,\ \tau \in \mathbb{R}_+,\ \mathbf{x} \in \mathcal{X}. && && \text{(36f)}
\end{aligned}$$

We note that the constraints (36c)–(36e) are equivalent to the inequalities $\nu^{ij}_k \leq |\xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})|$ for all $i, j \in [n]$, and $k \in [m]$. It is possible to ensure (without changing the optimum of the problem) that the opposite inequalities $\nu^{ij}_k \geq |\xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})|$ also hold, by adding the corresponding redundant constraints $\nu^{ij}_k \geq \xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})$ and $\nu^{ij}_k \geq -\xi^i_k(\mathbf{x}) + \xi^j_k(\mathbf{x})$ to (36).

6.1.4 Utilizing a comonotone structure The formulation (36) features the auxiliary variables λ^{ij}_k, along with the corresponding constraints (36c)–(36e), which represent the potentially non-convex relations $\nu^{ij}_k \leq |\xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})|$. The introduction of binary variables and big-M constraints often leads to significant computational challenges. However, this issue can be avoided when the mappings $i \mapsto \xi^i_k(\mathbf{x}_1)$ and $i \mapsto \xi^i_k(\mathbf{x}_2)$ are comonotone for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $k \in [m]$. If this condition is satisfied, then for any $i, j \in [n]$ and $k \in [m]$ there are two possibilities: Either $\xi^i_k(\mathbf{x}) \geq \xi^j_k(\mathbf{x})$ holds for all $\mathbf{x} \in \mathcal{X}$, in which case we can set $\nu^{ij}_k = \xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})$, or $\xi^i_k(\mathbf{x}) \leq \xi^j_k(\mathbf{x})$ holds for all $\mathbf{x} \in \mathcal{X}$, in which case we can set $\nu^{ij}_k = -\xi^i_k(\mathbf{x}) + \xi^j_k(\mathbf{x})$. Since these new equality constraints ensure that we have $\nu^{ij}_k = |\xi^i_k(\mathbf{x}) - \xi^j_k(\mathbf{x})|$ for all $i, j \in [n]$ and $k \in [m]$, the auxiliary λ^{ij}_k variables can be dropped from the formulation along with the constraints (36c)–(36e). While the above comonotonicity condition is restrictive, it is naturally satisfied for certain applications, including some of the machine scheduling problems we discuss in Section 7.1.

6.1.5 A parametric programming approach We again consider the general setting where nominal realizations are decision-dependent, and note that the non-convex quadratic terms $\delta^{ij}\tau$ in the constraints (32b) constitute a significant potential obstacle when working toward a tractable approach to solving the problem (32). Fortunately, all of these terms feature the variable τ as a common factor. Therefore, if we fix the value of τ, all of the quadratic terms in question become linear. In certain cases this leads to an optimization problem that belongs to a more tractable class than the original. For example, if the mapping $\mathbf{x} \mapsto G(\mathbf{x}, \boldsymbol{\xi}(\mathbf{x}))$ was linear, then fixing the value of τ would change quadratic constraints into linear ones. We can therefore attempt to solve (32) by performing a single-parameter search over the possible values of τ.

This approach is closely related to the field of of parametric programming. In this context, calculating the optimum of (32b) for a fixed value of τ can be seen as evaluating the *optimum value function* (OVF) of a parametric non-linear program (see, e.g., Kyparisis and Fiacco, 1987, both for a quick introduction to the subject, and for a precise statement of the convexity results discussed below). If the OVF has certain favorable properties, such as convexity or unimodality, then the aforementioned single-parameter search can potentially lead to a viable solution strategy (e.g., by using golden section search) with performance guarantees. While there are a variety of results that prove generalized convexity properties for OVFs, they typically require objective and constraining functions to be jointly convex in all variables. It appears that establishing joint convexity for general problems in the classes that we study is highly non-trivial, except under very restrictive assumptions (such as requiring all probabilities p^i and realization distances δ^{ij} to be decision-independent). However, it still seems plausible that this approach can be leveraged for problems with additional underlying structure. Along similar lines, it can be relatively straightforward to obtain a Lipschitz constant for the OVF in specific problem instances. While the algorithmic consequences are less dramatic than those of, say, unimodularity, efficient derivative-free global optimization methods exist in the literature for minimizing univariate Lipschitz-continuous functions (see, e.g., Hansen et al., 1992).

7. Applications

In this section we provide several examples of how our results can be utilized to provide tractable formulations for specific applied problems.

7.1 Stochastic Single-Machine Scheduling We consider a simple scheduling problem featuring L jobs, with processing times ξ_l and importance weights w_l for $l \in [L]$. Schedules will be evaluated based on the *total weighted completion time* (TWCT) of the jobs, which is a widely used performance measure (see, e.g., Pinedo, 2008). It will be helpful to assume that the TWCT is interpreted on a monetary scale; this can be accomplished by appropriately scaling the weights w_l.

We are primarily interested in the case where the processing times are stochastic, and can be affected by control decisions. Accordingly, let $(\Omega, \mathcal{A}, \mathbb{P})$ be an arbitrary (not necessarily finite) probability space, and let us introduce the mapping $\boldsymbol{\xi} : \mathcal{U} \to \mathcal{L}^L(\Omega, \mathcal{A})$. Here $\mathcal{U}$ is the set of feasible control decisions, and $\xi_l(\mathbf{u}) \in \mathcal{L}^1(\Omega, \mathcal{A})$ is the random processing time of job $l \in [L]$ given decision $\mathbf{u}$. In addition, we denote the cost associated with decision $\mathbf{u}$ by $h(\mathbf{u})$; the cost mapping $h : \mathcal{U} \to \mathbb{R}$ is often chosen to be linear.

In the deterministic scheduling literature a wide variety of schemes have been proposed to control processing times, see, e.g., Shabtay and Steiner (2007). We will now adapt two important models of control to our stochastic setting.

- *Linearly compressible processing times* take the form $\xi_l(\mathbf{u}) = \hat{\xi}_l - \hat{a}_l u_l$, where $\hat{\xi}_l \in \mathcal{L}^1(\Omega, \mathcal{A})$ is the *baseline* random processing time of job $l \in [L]$, and $\hat{a}_l \in \mathcal{L}^1(\Omega, \mathcal{A})$ is the corresponding stochastic compression rate. Feasible control decisions will then constitute a set
$$\mathcal{U} \subset \left\{ \mathbf{u} \in \mathbb{R}^L \,:\, 0 \le u_l \le \operatorname{ess\,inf} \frac{\hat{\xi}_l}{\hat{a}_l} \quad \forall l \in [L] \right\}.$$

  Example 7.1 *In the case $\hat{a}_l = \hat{\xi}_l$ processing times are given by $\xi_l(\mathbf{u}) = (1 - u_l)\hat{\xi}_l$, and the decision $u_l \in [0,1]$ can be interpreted as a proportional decrease in the processing time of job l.*

- *Control with discrete resources*: A finite set of T control options is available for every job, and selecting option $t \in [T]$ for job $l \in [L]$ leads to a random processing time of $\hat{\xi}_{tl}$. Let us introduce the binary decision variables u_{tl} for $t \in [T]$, $l \in [L]$, that take value 1 if and only if control option t is selected for job l. Then the processing time of job l is given by $\xi_l(\mathbf{u}) = \sum_{t\in[T]} u_{tl}\hat{\xi}_{tl}$ for $l \in [L]$, and the feasible control decisions constitute a set
$$\mathcal{U} \subset \left\{ \mathbf{u} \in \{0,1\}^{T\times L} \,:\, \sum_{t\in[T]} u_{tl} = 1 \ \forall l \in [L] \right\}. \tag{37}$$

  Example 7.2 *Assume that for each job the decision maker can choose to apply one of T possible linear compression rates, given by $\hat{a}_{tl} \in [0,1]$ for $t \in [T]$, $l \in [L]$, and let us denote the corresponding speedup factors by $a_{tl} = 1 - \hat{a}_{tl}$. The controllable processing times then take the form $\xi_l^i(\mathbf{u}) = \hat{\xi}_l^i \left(1 - \sum_{t\in[T]} \hat{a}_{tl}u_{tl}\right) = \hat{\xi}_l^i \sum_{t\in[T]} a_{tl}u_{tl}$, where $\hat{\xi}_l$ again denotes the baseline random processing time.*

It is easy to verify that the comonotonicity condition discussed in Section 6.1.4 holds both for Example 7.1 and for Example 7.2.

We next describe the sequencing aspect of our scheduling problems using the well-known *linear ordering formulation*, and remark that the proposed modeling framework can also be naturally adapted to the *assignment and positional date formulation* (see, e.g., Keha et al., 2009). Let us introduce the binary decision variables θ_{kl} for $k, l \in [L]$ that take value 1 if job k precedes job l in the processing sequence, and take value 0 otherwise. Then the set $\mathcal{T}$ of feasible scheduling decisions consists of the binary matrices $\boldsymbol{\theta} \in \{0,1\}^{L\times L}$ that satisfy the system

$$\theta_{ll} = 1, \qquad \forall l \in [L] \tag{38a}$$
$$\theta_{kl} + \theta_{lk} = 1, \qquad \forall k,\ l \in [L] \ :\quad k < l \tag{38b}$$
$$\theta_{kl} + \theta_{lh} + \theta_{hk} \le 2, \qquad \forall k,\ l,\ h \in [L] \ :\ k < l < h. \tag{38c}$$

Here constraints (38a) express the convention that each job is considered to precede itself, constraints (38b) ensure that no job simultaneously precedes and succeeds a different job, while constraints (38c) prevent cyclic subsequences of length three.

If we assume zero release dates for all jobs, then the completion time of job $l \in [L]$ is given by $\sum_{k\in[L]} \xi_k(\mathbf{u})\theta_{kl}$. Introducing the matrix $\Theta = (\theta_{kl})_{k,l\in[L]}$, we can express the TWCT objective as

$$\sum_{l\in[L]} w_l \sum_{k\in[L]} \xi_k(\mathbf{u})\theta_{kl} = \sum_{k\in[L]} \sum_{l\in[L]} \xi_k(\mathbf{u})\theta_{kl}w_l = \boldsymbol{\xi}^\top(\mathbf{u})\Theta\mathbf{w}.$$

The risk-averse version of our stochastic single-machine scheduling problem can now be formulated as

$$\min_{(\boldsymbol{\theta},\mathbf{u})\in\mathcal{T}\times\mathcal{U}} \quad h(\mathbf{u})+\rho\left(\boldsymbol{\xi}^\top(\mathbf{u})\Theta\mathbf{w}\right), \tag{39}$$

where ρ is a law-invariant coherent risk measure. We next proceed to examine DRO variants of this underlying problem.

7.1.1 Continuous Wasserstein balls Let us first consider the case when processing times can take their values from a continuous spectrum and are subject to ambiguity, with a continuous Wasserstein-p ball of radius κ as the ambiguity set. As outlined in Section 4, the DRO variant of the underlying risk-averse problem (39) then takes the form

$$\min_{(\boldsymbol{\theta},\mathbf{u})\in\mathcal{T}\times\mathcal{U}} h(\mathbf{u})+\sup_{\boldsymbol{\zeta}\in\mathcal{B}^{\mathbb{P}}_{\delta^{p},\kappa^{p}}(\boldsymbol{\xi}(\mathbf{u}))} \rho\left(\boldsymbol{\zeta}^\top\Theta\mathbf{w}\right). \tag{40}$$

If the risk measure ρ is well-behaved with some factor C, then it immediately follows from Proposition 4.2 that the problem (40) can be equivalently reformulated as

$$\min_{(\boldsymbol{\theta},\mathbf{u})\in\mathcal{T}\times\mathcal{U}} \quad h(\mathbf{u})+\rho\left(\boldsymbol{\xi}^\top(\mathbf{u})\Theta\mathbf{w}\right)+C\kappa\|\Theta\mathbf{w}\|_q. \tag{41}$$

The only difference between this formulation and the underlying problem (39) is the additional robustification term $C\kappa\|\Theta\mathbf{w}\|_q$, which, due to the convexity of the q-norm, is a convex function of the sequencing variables θ_{kl}. The example below shows that this term can affect the optimal schedule, even when the underlying scheduling problem is deterministic with no compression decisions.

Example 7.3 *Consider the following deterministic instance of the scheduling problem introduced in Section 7.1. There are two jobs (Job* 1 *and Job* 2*) with respective weights* 2 *and* 3*, and respective non-compressible processing times* 21 *and* 32*. Scheduling Job* 1 *before Job* 2 *(*"schedule $1\prec 2$"*) leads to a TWCT of* 201*, which is superior to the TWCT of* 202 *for schedule* $2\prec 1$*. However, in the DRO version of the problem where the ambiguity set for the processing time vector is the* 2*-norm ball* $B^2_4((21,32)^\top)$ *of radius* 4 *around the nominal values, the robustified TWCT for schedule* $1\prec 2$ *becomes (approximately)* 224.32*, which is inferior to the robustified TWCT of* 223.54 *for schedule* $2\prec 1$*. We note that, in accordance with our observations at the end of Section 4, the same results will hold for any risk-neutral stochastic version of the problem with expected nominal processing times* 21 *and* 32*, and a continuous Wasserstein-*2 *ball of radius* 4 *as the ambiguity set.*

However, in certain settings the underlying problem (39) and the robustified problem (41) are guaranteed to have the same solution. In the case $p=1$ it is easy to verify that the robustification term is always equal to the constant $C\kappa\sum_{l\in[L]}w_l$, and thus has no impact on the optimal solution. Along similar lines, if we replace the total weighted completion time in the objective function by the (unweighted) total completion time, i.e., if we set $w_l=1$ for all $l\in[L]$, then the robustification term becomes $C\kappa\left(\sum_{l=1}^{L} l^q\right)^{\frac{1}{q}}$, which again does not depend on the decision variables.

To finish this subsection, we briefly discuss two further cases when (40), our DRO problem with endogenous uncertainty, reduces to a more familiar type of problem.

Observation 7.1 *Assume that the processing times are compressed in a scenario-independent fashion, i.e., that we have* $\boldsymbol{\xi}(\mathbf{u})=\hat{\boldsymbol{\xi}}-\hat{\mathbf{a}}(\mathbf{u})$ *for some baseline random processing time vector* $\hat{\boldsymbol{\xi}}\in\mathcal{L}^m(\Omega,\mathcal{A})$*, and a deterministic compression mapping* $\hat{\mathbf{a}}:\mathcal{U}\to\mathbb{R}^L$*. We can then rewrite* (40) *as a "traditional" DRO problem without endogeneous uncertainty. More precisely, it is easy to verify that a random vector* $[\mathbb{B},\boldsymbol{\zeta}]$ *belongs to the decision-dependent ambiguity set* $\mathcal{B}^{\mathbb{P}}_{\delta,\kappa}(\boldsymbol{\xi}(\mathbf{u}))$ *if and only if it is of the form* $\boldsymbol{\zeta}=\hat{\boldsymbol{\zeta}}-\hat{\mathbf{a}}(\mathbf{u})$*, where* $\hat{\boldsymbol{\zeta}}$ *belongs to the decision-independent ambiguity set* $\mathcal{B}^{\mathbb{P}}_{\delta,\kappa}(\hat{\boldsymbol{\xi}})$*.*

Observation 7.2 *Let us examine the risk-neutral case, where we have $\rho = \mathbb{E}$. In Section 4 we established that, for a general class of problems, the arising DRO instance (DRO-RNC) is equivalent to the deterministic problem (12). In our scheduling context this result leads to the following deterministic reformulation of (40):*

$$\min_{(\boldsymbol{\theta},\mathbf{u})\in\mathcal{T}\times\mathcal{U}} \quad h(\mathbf{u}) + \bar{\boldsymbol{\xi}}^\top(\mathbf{u})\Theta\mathbf{w} + C\kappa\|\Theta\mathbf{w}\|_q,$$

where the operator $\bar{\boldsymbol{\xi}} : \mathcal{U} \to \mathbb{R}^L$ gives the decision-dependent expected processing time vector $\bar{\boldsymbol{\xi}}(\mathbf{u}) = \mathbb{E}\left(\boldsymbol{\xi}(\mathbf{u})\right)$ for $\mathbf{u} \in \mathcal{U}$.

7.1.2 Discrete EMD balls We now consider the case when processing times can take their values from some discrete set, and accordingly the ambiguity set is a discrete EMD ball of type (BALL-D). We begin with a simple example that illustrates the impact of the DRO approach on optimal scheduling decisions, and also shows that equivalent underlying problems can have non-equivalent robustifications.

Example 7.4 *We consider an instance of the stochastic scheduling problem introduced in Section 7.1 with two scenarios, which in the nominal distribution $\mathbb{P}$ both have probability 0.5. There are two jobs, with non-compressible nominal processing times as follows: $\boldsymbol{\xi}_1^1 = 2$, $\boldsymbol{\xi}_1^2 = 4$ for Job 1, and $\boldsymbol{\xi}_2^1 = \boldsymbol{\xi}_2^2 = 6$ for Job 2. We take a risk-neutral approach, and aim to minimize the expected TWCT when the respective weights of Jobs 1 and 2 are 20 and 39. Scheduling Job 1 before Job 2 ("schedule $1 \prec 2$") leads to an expected TWCT of 411, which is superior to the expected TWCT of 414 for schedule $2 \prec 1$. On the other hand, in the DRO variant of the problem where the ambiguity set for the random processing time vector is the discrete Wasserstein-1 ball $\mathcal{B}^{\boldsymbol{\xi}}_{\delta^1,0.1}(\mathbb{P})$ of radius 0.1, the robustified expected TWCT for schedule $1 \prec 2$ is 416.9, which is inferior to the robustified expected TWCT of 416 for schedule $2 \prec 1$.*

To obtain the deterministic counterparts of these problems, we need to replace $[\mathbb{P}, \boldsymbol{\xi}]$ with the trivial distribution where the processing time of Job 1 is changed to its expected value of 3 in both scenarios. As discussed at the end of Section 4, in the risk-neutral case the underlying (i.e., non-robustified) stochastic problem is equivalent to its deterministic counterpart. However, DRO variants of these two equivalent problems are no longer equivalent. More precisely, since any EMD ball of type (BALL-D) around a deterministic nominal vector is trivial (i.e., it contains only its center), all DRO variants of the deterministic problem are equivalent to the underlying non-robustified one. However, as we have just seen, robustifying the original stochastic underlying problem can affect the optimal schedule.

In this section we study the following DRO variant of the risk-averse scheduling problem (39):

$$\min_{(\boldsymbol{\theta},\mathbf{u})\in\mathcal{T}\times\mathcal{U}} h(\mathbf{u}) + \sup_{\mathbb{Q}\in\mathcal{B}^{\boldsymbol{\xi}(\mathbf{u})}_{\delta,\kappa}(\mathbb{P})} \rho\left(\mathbf{w}^\top\Theta^\top\boldsymbol{\xi}(\mathbf{u})\right). \tag{42}$$

For the case $\rho = \mathrm{CVaR}_\alpha$ we can adapt the general formulation (32) to equivalently express our problem (42) as

$$\begin{aligned}
\min \quad & h(\mathbf{u}) + \eta + \frac{1}{1-\alpha}\sum_{i\in[n]} p^i v^i + \frac{1}{1-\alpha}\kappa\tau && && \text{(43a)}\\
\text{s.t.} \quad & v^i \geq \sum_{l\in[L]} w_l \sum_{k\in[L]} \xi_k^j(\mathbf{u})\theta_{kl} - \eta - \delta^{ij}\tau, && \forall i,j\in[n] && \text{(43b)}\\
& \delta^{ij} = \delta\left(\mathbf{w}^\top\Theta^\top\boldsymbol{\xi}^i(\mathbf{u}), \mathbf{w}^\top\Theta^\top\boldsymbol{\xi}^j(\mathbf{u})\right), && \forall i,j\in[n] && \text{(43c)}\\
& (\boldsymbol{\theta},\mathbf{u}) \in \mathcal{T}\times\mathcal{U}, \quad \mathbf{v}\in\mathbb{R}^n_+, \quad \tau \geq 0. && && \text{(43d)}
\end{aligned}$$

Remark 7.1 *In order to keep the presentation simple, we implicitly assumed that the costs associated with our decisions are deterministic. For risk-neutral problems this assumption is without loss of generality,* *because stochastic costs can be equivalently replaced with their expected values. While this is no longer the case in a risk-averse context, we can easily adapt our formulations to a setting with stochastic costs. Denoting the cost of decision $\mathbf{u} \in \mathcal{U}$ under scenario $i \in [n]$ by $h^i(\mathbf{u})$, we can simply remove $h(\mathbf{u})$ from the objective function in (43a), and instead incorporate the costs into the random outcome mapping by adding the term $h^i(\mathbf{u})$ to the right-hand side of constraint (43b).*

The formulation (43) is generally a very challenging non-linear program due in part to the quadratic terms in constraints (43b), and in part to the potential non-linearity in constraints (43c). We next provide potentially tractable forms of this problem for the case of control with discrete resources, when using a Wasserstein-1 ambiguity set.

Let us assume that the processing time of job $l \in [L]$ is $\xi_l(\mathbf{u}) = \sum_{t\in[T]} u_{tl}\hat{\xi}_{tl}$ for $l \in [L]$, where the set $\mathcal{U}$ of feasible control decisions is given as in (37). Then in constraints (43b) we can rewrite $\xi^j_k(\mathbf{u})\theta_{kl}$ as $\sum_{t\in[T]} \hat{\xi}^j_{tl} u_{tk}\theta_{kl}$, and use McCormick envelopes (McCormick, 1976) to linearize the arising quadratic terms $u_{tk}\theta_{kl}$. In addition, when δ is the 1-norm distance, we can express the δ^{ij} values as in (35), and incorporate them into our optimization problem via mixed-integer big-M constraints as in (36). The problem (43) then takes the following form:

$$\min \quad h(\mathbf{u}) + \eta + \frac{1}{1-\alpha}\sum_{i\in[n]} p^i v^i + \frac{1}{1-\alpha}\kappa\tau \tag{44a}$$

$$\text{s.t.} \quad v^i \ge \sum_{l\in[L]}\sum_{k\in[L]}\sum_{t\in[T]} w_l \hat{\xi}^j_{tk} z_{tkl} - \eta - \sum_{l\in[L]} \nu^{ij}_l \tau, \qquad \forall i,j \in [n] \tag{44b}$$

$$z_{tkl} \le u_{tk}, \qquad \forall t \in [T],\ k,l \in [L] \tag{44c}$$

$$z_{tkl} \le \theta_{kl}, \qquad \forall t \in [T],\ k,l \in [L] \tag{44d}$$

$$z_{tkl} \ge u_{tk} + \theta_{kl} - 1, \qquad \forall t \in [T],\ k,l \in [L] \tag{44e}$$

$$\nu^{ij}_l \le \xi^i_l(\mathbf{u}) - \xi^j_l(\mathbf{u}) + M\lambda^{ij}_l, \qquad \forall i,j \in [n],\ l \in [L] \tag{44f}$$

$$\nu^{ij}_l \le -\xi^i_l(\mathbf{u}) + \xi^j_l(\mathbf{u}) + M(1-\lambda^{ij}_l), \qquad \forall i,j \in [n],\ l \in [L] \tag{44g}$$

$$\boldsymbol{\lambda} \in \{0,1\}^{n\times n\times L}, \quad \boldsymbol{\nu} \in \mathbb{R}^{n\times n\times L}_+, \tag{44h}$$

$$(\boldsymbol{\theta},\mathbf{u}) \in \mathcal{T}\times\mathcal{U}, \quad \mathbf{v} \in \mathbb{R}^n_+, \quad \tau \ge 0, \quad \mathbf{z} \in \{0,1\}^{T\times L\times L}. \tag{44i}$$

Keeping in mind that the auxiliary variables $\mathbf{z}$ are binary, the constraints (44c)–(44e) ensure that $z_{tkl} = u_{tk}\theta_{kl}$ holds for all $t \in [T]$, $k,l \in [L]$. Also, since the boundedness condition established at the start of Section 6.1.3 trivially holds in the case of discrete control decisions, for sufficiently high values of the parameter M the constraints (44f)–(44h) are equivalent to the inequalities $\nu^{ij}_k \le |\xi^i_l(\mathbf{u}) - \xi^j_l(\mathbf{u})|$ for all $i,j \in [n]$, $l \in [L]$. We mention that the reformulation-linearization technique (Sherali and Adams, 1994; Sherali et al., 1998) yields the valid inequalities $\sum_{t\in[T]} z_{tkl} = \theta_{kl}$ for all $k,l \in [L]$, which can be added to strengthen the formulation (44). Recalling our discussions from Section 6.1.5 we also point out that, as the remaining quadratic terms in our constraints all involve the common scalar variable τ, if we fix the value of τ, then the above problem becomes a linearly constrained mixed integer program.

Under the comonotonicity assumption of Section 6.1.4 it is possible to significantly simplify (44). In this case the index set of the variable $\boldsymbol{\nu}$ has a partition $[n]\times[n]\times[L] = I_+ \cup^* I_-$ with the following properties: If $(i,j,l) \in I_+$ holds, then we have $\boldsymbol{\xi}^i_l(\mathbf{u}) \ge \boldsymbol{\xi}^j_l(\mathbf{u})$ for all $\mathbf{u} \in \mathcal{U}$, while if $(i,j,l) \in I_-$ holds, then we have $\boldsymbol{\xi}^i_l(\mathbf{u}) \le \boldsymbol{\xi}^j_l(\mathbf{u})$ for all $\mathbf{u} \in \mathcal{U}$. Accordingly, for all $(i,j,l) \in I_+$ we can replace ν^{ij}_l in (44b) with $\Big(\boldsymbol{\xi}^i_l(\mathbf{u}) - \boldsymbol{\xi}^j_l(\mathbf{u})\Big)$, and similarly for all $(i,j,l) \in I_-$ we can replace ν^{ij}_l in (44b) with $\Big(-\boldsymbol{\xi}^i_l(\mathbf{u}) + \boldsymbol{\xi}^j_l(\mathbf{u})\Big)$. The now redundant constraints (44f)–(44h), along with the variables $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$, can then be dropped from the problem formulation.

Additional structure in the underlying scheduling problems can often be exploited to further improve tractability. To demonstrate this, we conclude the section by showing that, for the specific compression scheme introduced in Example 7.2, it is possible to formulate our DRO problem (43) as a *mixed integer linear program* (MIP) when costs are linear. We recall that under this scheme the processing time of job $l \in [L]$ under compression decisions $\mathbf{u} \in \mathcal{U}$ has realizations $\xi_l^i(\mathbf{u}) = \hat{\xi}_l^i \sum_{t\in[T]} u_{tl} a_{tl}$ for $i \in [n]$, where the baseline times $\hat{\xi}_l^i \in \mathbb{R}_+$ and the speedup factors $a_{tl} \in [0,1]$ are given parameters. As before, we will use McCormick envelopes to establish auxiliary variables $z_{tkl} = u_{tk}\theta_{kl}$ for $t \in [T]$, $k, l \in [L]$. Noting that we have $\delta^{ij} = \sum_{l\in[L]} \sum_{t\in[T]} |\hat{\xi}_l^i - \hat{\xi}_l^j| a_{tl} u_{tl}$, we will similarly use McCormick envelopes to help linearize the terms $\delta^{ij}\tau$ in (43b) by introducing the auxiliary variables $y_{tl} = u_{tl}\tau$ for $t \in [T]$, $l \in [L]$. Assuming that the cost of selecting compression option $t \in [T]$ for job $l \in [L]$ is given by $h_{tl} \in \mathbb{R}$, we can now equivalently formulate (43) as the following MIP:

$$
\begin{aligned}
\min \quad & \sum_{l\in[L]} \sum_{t\in[T]} h_{tl} u_{tl} + \eta + \frac{1}{1-\alpha} \sum_{i\in[n]} p^i v^i + \frac{1}{1-\alpha} \kappa\tau \\
\text{s.t.} \quad & v^i \geq \sum_{l\in[L]} \sum_{k\in[L]} \sum_{t\in[T]} w_l \hat{\xi}_k^j a_{tk} z_{tkl} - \eta - \sum_{l\in[L]} \sum_{t\in[T]} |\hat{\xi}_l^i - \hat{\xi}_l^j| a_{tl} y_{tl}, && \forall i, j \in [n] \\
& z_{tkl} \leq u_{tk}, && \forall t \in [T],\ k, l \in [L] \\
& z_{tkl} \leq \theta_{kl}, && \forall t \in [T],\ k, l \in [L] \\
& z_{tkl} \geq u_{tk} + \theta_{kl} - 1, && \forall t \in [T],\ k, l \in [L] \\
& y_{tl} \leq u_{tl}, && \forall t \in [T],\ l \in [L] \\
& y_{tl} \leq \tau, && \forall t \in [T],\ l \in [L] \\
& y_{tl} \geq \tau + u_{tl} - 1, && \forall t \in [T],\ l \in [L] \\
& (\boldsymbol{\theta}, \mathbf{u}) \in \mathcal{T} \times \mathcal{U}, \quad \mathbf{v} \in \mathbb{R}_+^n, \quad \tau \geq 0, \quad \mathbf{z} \in (0,1)^{T\times L\times L}, \quad \mathbf{y} \in \mathbb{R}_+^{T\times L}.
\end{aligned}
$$

We can again strengthen our MIP using the reformulation-linearization technique, which in this case provides the valid equalities $\sum_{t\in[T]} y_{tl} = \tau$ and $\sum_{t\in[T]} z_{tkl} = \theta_{kl}$ for all $k, l \in [L]$.

7.2 Network models with independent component failures Finally, we turn our attention to an important general class of optimization problems with endogenous uncertainty, which includes stochastic network reliability and network interdiction problems. Problems in this class, as discussed in Haus et al. (2017), typically feature an underlying system represented by a graph whose edges and/or nodes are subject to random failures. Accordingly, the state of the system can be represented by binary vectors, where each coordinate corresponds to a network component. By convention, a value of 1 signifies the survival of the component, while a value of 0 indicates failure. These binary vectors can then be viewed as scenarios, with corresponding scenario probabilities determined by the (independent) survival probabilities of system components. Endogenous uncertainty arises when design decisions can be made to affect these survival probabilities, with the aim of improving the post-failure performance of the system. Performance is usually evaluated using an outcome function that quantifies network properties such as connectivity or shortest path lengths. An important assumption is that the outcome can be expressed solely as a function of the state of the system. Under this assumption, the objective function of a decision maker who is interested in minimizing the expected outcome (or, more generally, a risk measure of the outcome) will only depend on decisions through their effect on scenario probabilities. Before introducing a formal description of models in this problem class, we present an important application which is the primary motivation behind the developments in this section.

Stochastic pre-disaster investment planning problem (SPIPP) As discussed in the introduction, this problem (originally proposed by Peeta et al., 2010) has been receiving significant attention in the recent literature. The problem models a transportation network as an undirected graph, where the edges correspond to highway links, and the edge lengths correspond to traversal costs. In the event of a disaster (such as an earthquake), the links are subject to random failures, which are assumed to be independent. The failure probability of each link can be reduced by making an investment to strengthen it, and the goal is to use a limited budget to improve the post-disaster connectivity of the network. In the simplest case connectivity is quantified as the length of the shortest path between an origin-destination (O-D) pair. To obtain a more nuanced measure of connectivity, one can also consider an appropriately weighted sum of shortest path lengths between multiple O-D pairs.

While the above problem (SPIPP) will remain our main focus, we also briefly mention another well-known example of our problem class.

Example 7.5 *In the stochastic network interdiction problem (SNIP), introduced by Cormican et al. (1998), a defender attempts to block arcs in a capacitated network, using a limited budget, in order to diminish an attacker's ability to perform a task such as the distribution of nuclear weapons or illegal drugs. Blocking attempts are assumed to randomly succeed or fail in a binary fashion, independently for each arc. The attacker's goal is then to maximize a flow through the remaining network.*

7.2.1 Description of the base model Let us consider a system that consists of L components, which—keeping in mind the motivating pre-disaster planning problem—we will refer to as 'links'. Each link $\ell \in [L]$ has a baseline survival probability of $\sigma_\ell^0 \in [0,1]$, which increases to $\sigma_\ell^1 \in [\sigma_\ell^0, 1]$ if the link is strengthened (or "reinforced"). We introduce the binary decision variables $\mathbf{x} \in \{0,1\}^L$, where x_ℓ takes value 1 if and only if an investment is made to strengthen link $\ell \in [L]$. For the sake of our discussions we assume that the set $\mathcal{X} \subset \{0,1\}^L$ of feasible decisions is defined by a set of linear inequalities (in practice typically by a single budget constraint), although our formulations remain valid for the general case.

The post-failure state of the system can described via a binary vector of length L, whose ℓth component takes value 1 if and only if link $\ell \in [L]$ survives. In order to be consistent with the formalism of the preceding sections, we introduce the following notation. Observing that there are $n = 2^L$ possible system states ("scenarios"), let $\{\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^n\} = \{0,1\}^L$ be a list of all L-dimensional binary vectors. For $i \in [n]$, $\ell \in [L]$ we will interpret the equality $\xi_\ell^i = 1$ as "*link ℓ survives in scenario i*", and conversely interpret the equality $\xi_\ell^i = 0$ as "*link ℓ fails in scenario i*". Given the reinforcement decision $x_\ell \in \{0,1\}$, the survival probability of link $\ell \in [L]$ is $(1-x_\ell)\sigma_\ell^0 + x_\ell\sigma_\ell^1$, while its failure probability is $(1-x_\ell)(1-\sigma_\ell^0) + x_\ell(1-\sigma_\ell^1)$. Therefore, under the assumption that link failures are independent, the decision-dependent probability distribution $\mathbb{P}(\mathbf{x})$ is given by the following formula for the probability of scenario $i \in [n]$:

$$p^i(\mathbf{x}) = \prod_{\ell \in [L] \,:\, \xi_\ell^i = 1} \left[(1-x_\ell)\sigma_\ell^0 + x_\ell\sigma_\ell^1\right] \prod_{\ell \in [L] \,:\, \xi_\ell^i = 0} \left[(1-x_\ell)(1-\sigma_\ell^0) + x_\ell(1-\sigma_\ell^1)\right]. \tag{45}$$

Finally, recalling that under our assumption the outcome function is decision-independent, let us denote its value in scenario $i \in [n]$ by G^i. We can then formulate the risk-neutral version of our base problem as

$$\min_{\mathbf{x} \in \mathcal{X}} \sum_{i \in [n]} p^i(\mathbf{x}) G^i. \tag{46}$$

Model specifics for SPIPP. Let us assume that a highway network is modeled by the undirected graph (V, E) on vertex set V with edge set $E = \{e_1, \dots, e_L\}$, and edge lengths represent traversal costs. The interpretation of the reinforcement decision variables $\mathbf{x}$ and the survival probabilities σ_ℓ^0, σ_ℓ^1 is self-explanatory. If the cost of reinforcing the highway link $\ell \in [L]$ is c_ℓ, and the available budget is C, then the set of feasible decisions is given by $\mathcal{X} = \left\{\mathbf{x} \in \{0,1\}^L \;:\; \sum_{\ell \in [L]} c_\ell x_\ell \le C\right\}$. The post-disaster state of the network in scenario $i \in [n]$ can be modeled by the graph (V, E^i), where $E^i = \{e_\ell \;:\; \xi_\ell^i = 1\}$ is the set of edges that survive in scenario i. In the simplest case, when we consider only single O-D pair, the outcome G^i in scenario $i \in [n]$ is given by the length of the shortest O-D path in the surviving graph (V, E^i). However, in practice this value is modified: if the shortest path length exceeds a certain acceptability threshold, or if no O-D path exists in (V, E^i), then the value of G^i is set to a fixed penalty parameter. This parameter typically represents the traversal costs associated with an alternate mode of transportation (such as helicopter), which relief organizations can resort to when no acceptable routes survive in the network. When multiple O-D pairs are given, along with corresponding importance weights, the outcome G^i naturally becomes the weighted sum of the (modified) shortest O-D path lengths in (V, E^i).

Specifying the model components in accordance with the conventions we established is not always as straightforward as for SPIPP, as the next example shows.

Example 7.6 *Let us return to SNIP, as introduced in Example 7.5, and assume that a transportation network is modeled by a digraph (V, A) with arc set $A = \{a_1, \ldots, a_L\}$, specified source and sink nodes, and given arc capacities. We recall that under our assumptions the ordering of survival probabilities is fixed for every link: we have $\sigma_\ell^0 \leq \sigma_\ell^1$ for all $\ell \in [L]$. Since adherence to this convention will be necessary when applying the so-called* distribution shaping *methods discussed in the next section, we need to denote the baseline survival probability of arc a_ℓ by σ_ℓ^1. If an attempt is made to block this arc, its survival probability is reduced to σ_ℓ^0. Accordingly, our decision variables will have a somewhat counterintuitive interpretation, wherein $x_\ell = 1$ holds if and only if the defender does* ***not*** *attempt to block arc a_ℓ. If the cost of attempting to block arc a_ℓ is c_ℓ, and the defender's budget is C, then the set of feasible decisions is given by $\mathcal{X} = \left\{\mathbf{x} \in \{0,1\}^L \ : \ \sum_{\ell \in [L]} c_\ell (1 - x_\ell) \leq C\right\}$. Analogously to the case of SPIPP, the surviving digraph in scenario $i \in [n]$ is (V, A^i), with $A^i = \{a_\ell \ : \ \xi_\ell^i = 1\}$. The capacities of surviving arcs remain unchanged, and the outcome G^i in scenario $i \in [n]$ is given by the value of the maximum source-sink flow on (V, A^i).*

The base problem (46) presents two major challenges. First, the probability expression (45) is a highly non-linear function of the decision variables. Second, the number of scenarios is exponential in terms of the number of links, which leads to prohibitively large formulations. In the next sections we present two recently developed and closely related techniques that help tackle these issues.

7.2.2 Distribution shaping Peeta et al. (2010) have solved an approximate version of SPIPP which is obtained by replacing the highly polynomial objective of (46) with a multi-linear function. Significant effort has been made in the recent literature to develop efficient solution methods that improve on this rough approximation (Flach and Poggi, 2010; Laumanns et al., 2014; Schichl and Sellmann, 2015; Haus et al., 2017). Here, with the goal of a straightforward and intuitive presentation in mind, we closely follow the work of Laumanns et al. (2014), who provide efficient exact methods that will also remain applicable in our DRO context. A key element of their approach is the technique of *distribution shaping*, which enables one to characterize the decision-dependent scenario probabilities via a set of linear constraints.

Given a decision vector $\mathbf{x} \in \mathcal{X}$, we introduce for all $\ell \in [L]$ the truncated vector $\check{\mathbf{x}}^\ell$ given by $\check{x}_j^\ell = x_j$ if $1 \leq j \leq \ell$, and by $\check{x}_j^\ell = 0$ if $\ell < j \leq L$. Let us denote the corresponding scenario probabilities by $\pi_\ell^i = p^i(\check{\mathbf{x}}^\ell)$, and note that the trivial equality $\mathbf{x} = \check{\mathbf{x}}^L$ implies $p^i(\mathbf{x}) = \pi_L^i$ for all $i \in [n]$. The key observation behind scenario shaping is that, in accordance with Bayes' rule, the probability measures defined by successive truncations $\check{\mathbf{x}}^{\ell-1}$ and $\check{\mathbf{x}}^\ell$ have a linearly expressible relationship (see Laumanns et al., 2014, for the exact details). We can therefore formulate (46) as the MIP

$$\min \sum_{i \in [n]} \pi_L^i G^i \tag{47a}$$

$$\text{s.t.}\quad \pi_\ell^i \leq \frac{\sigma_\ell^1}{\sigma_\ell^0}\pi_{\ell-1}^i + 1 - x_\ell, \qquad \forall\ \ell \in [L],\ i \in [n]\ :\ \xi_\ell^i = 1 \tag{47b}$$

$$\pi_\ell^i \leq \frac{1-\sigma_\ell^1}{1-\sigma_\ell^0}\pi_{\ell-1}^i + 1 - x_\ell, \qquad \forall\ \ell \in [L],\ i \in [n]\ :\ \xi_\ell^i = 0 \tag{47c}$$

$$\pi_\ell^i \leq \pi_{\ell-1}^i + x_\ell, \qquad \forall\ \ell \in [L],\ i \in [n] \tag{47d}$$

$$\sum_{i\in[n]} \pi_\ell^i = 1, \qquad \ell \in [L] \tag{47e}$$

$$\boldsymbol{\pi} \in [0,1]^{n\times L}, \tag{47f}$$

$$\mathbf{x} \in \mathcal{X}, \tag{47g}$$

where $\pi_0^i = \prod_{\ell:\xi_\ell^i=1} \sigma_\ell^0 \prod_{\ell:\xi_\ell^i=0} (1-\sigma_\ell^0)$ denotes the baseline probability of scenario $i \in [n]$. It is straightforward to verify that, due to the distribution shaping relations (47b)–(47f), the defining equalities $\pi_\ell^i = p^i(\check{\mathbf{x}}^\ell)$ will hold for all $\ell \in [L]$, $i \in [n]$.

7.2.3 Scenario bundling We now turn our attention to the crucial problem of reducing the number of scenarios. Sampling methods are often used for this purpose; we will briefly discuss their applicability in our setting at the end of Section 7.2.7, but for the moment we keep our focus on exact methods. Taking advantage of the crucial assumption that outcomes are decision-independent, it is possible to group together scenarios that lead to the same outcome. Two important goals need to be kept in mind when attempting to implement this idea. First, suitable scenario groups should be identified without having to perform an excessive number of outcome function evaluations. Second, the highly effective distribution shaping approach should remain applicable. To accomplish these goals, we follow the scenario bundling technique described in Laumanns et al. (2014), which was demonstrated to work well on SPIPP.

We introduce the symbol $*$ to indicate that status of a link is unknown, and say that a ternary vector $\mathbf{s} \in \{0,1,*\}^L$ represents the scenario set $B_\mathbf{s} = \left\{i \in [n]\ :\ \xi_\ell^i = s_\ell \vee s_\ell = * \ \forall \ell \in [L]\right\}$. If we have $G^i = G^j$ for all $i, j \in B_\mathbf{s}$, that is, if the vector $\mathbf{s}$ represents a set of scenarios that all result in the same outcome, then we call $\mathbf{s}$ a *scenario bundle*, and denote this common outcome by $G^\mathbf{s}$. Given a decision $\mathbf{x} \in \mathcal{X}$ we define the probability of bundle $\mathbf{s}$ as $\bar{p}^\mathbf{s}(\mathbf{x}) = \sum_{i\in B_\mathbf{s}} p^i(\mathbf{x})$. If the elements of a family S of scenario bundles represent a partition $\bigcup^*_{\mathbf{s}\in S} B_\mathbf{s} = [n]$ of the scenario set, then we call S a *bundling*. We note if S is a bundling, then $\sum_{\mathbf{s}\in S} \bar{p}^\mathbf{s}(\mathbf{x}) = 1$ holds for any decision $\mathbf{x} \in \mathcal{X}$.

The distribution shaping method extends naturally to scenario bundlings. Analogously to Section 7.2.2, we introduce the notation $\phi_\ell^\mathbf{s} = \bar{p}^\mathbf{s}(\check{\mathbf{x}}^\ell)$ for bundle probabilities corresponding to truncated decisions, and observe that we have $\phi_L^\mathbf{s} = \bar{p}^\mathbf{s}(\mathbf{x})$. If S is a bundling, then with slight modification of the distribution shaping relations (47b)–(47f) we can now reformulate (47) as

$$\min \quad \sum_{i\in[n]} \phi_L^\mathbf{s} G^\mathbf{s} \tag{48a}$$

$$\text{s.t.}\quad \phi_\ell^\mathbf{s} \leq \frac{\sigma_\ell^1}{\sigma_\ell^0}\phi_{\ell-1}^\mathbf{s} + 1 - x_\ell, \qquad \forall\ \ell \in [L],\ s \in S\ :\ s_\ell = 1 \tag{48b}$$

$$\phi_\ell^\mathbf{s} \leq \frac{1-\sigma_\ell^1}{1-\sigma_\ell^0}\phi_{\ell-1}^\mathbf{s} + 1 - x_\ell, \qquad \forall\ \ell \in [L],\ s \in S\ :\ s_\ell = 0 \tag{48c}$$

$$\phi_\ell^\mathbf{s} \leq \phi_{\ell-1}^\mathbf{s}, \qquad \forall\ \ell \in [L],\ s \in S\ :\ s_\ell = * \tag{48d}$$

$$\phi_\ell^\mathbf{s} \leq \phi_{\ell-1}^\mathbf{s} + x_\ell, \qquad \forall\ \ell \in [L],\ s \in S \tag{48e}$$

$$\sum_{s\in S} \phi_\ell^\mathbf{s} = 1, \qquad \forall\ \ell \in [L] \tag{48f}$$

$$\phi \in [0,1]^{S\times L}, \tag{48g}$$

$$\mathbf{x} \in \mathcal{X}, \tag{48h}$$

where $\phi_0^{\mathbf{s}} = \prod_{l:s_l=1} \sigma_\ell^0 \prod_{l:s_l=0} (1-\sigma_\ell^0)$ denotes the baseline probability of bundle $\mathbf{s} \in S$.

If we can find a small bundling S, then the above formulation allows us to significantly reduce the problem size. Laumanns et al. (2014) show that, for the case of SPIPP with a single O-D pair, it is indeed possible to efficiently construct such small bundlings, which on a real-world instance can reduce 2^{30} scenarios to a few hundred bundles. We mention that, while for the base problem these advances translate readily to the case of multiple O-D pairs, this is no longer the case for DRO variants; see Section 7.2.7 for a more detailed discussion.

Remark 7.2 *A more refined version of the scenario bundling technique, shown to be effective both for SPIPP and SNIP, is developed in Haus et al. (2017). The key difference is that, instead of defining a bundle by specifying the statuses of some links, this scenario aggregation method utilizes binary decision diagrams. Our following results can be naturally adapted to this more general scheme. However, to keep our presentation straightforward, we do not discuss details of this approach.*

7.2.4 DRO problem variants We now proceed to introduce DRO variants of the underlying problem (46). To account for uncertainty about the scenario probabilities, we will use a discrete EMD ball of type (BALL-D) with radius κ around the decision-dependent nominal distribution $\mathbb{P}(\mathbf{x})$ as our ambiguity set. Noting that the assumptions of Section 6.1.1 are satisfied, we can formulate the arising instance of (DRO-RND) as the following simple variant of (33):

$$\min \quad \sum_{i\in[n]} p^i(\mathbf{x}) v^i + \kappa\tau \tag{49a}$$

$$\text{s.t.} \quad v^i \geq G^j - \delta^{ij}\tau, \qquad \forall i,j \in [n] \tag{49b}$$

$$\mathbf{v} \in \mathbb{R}^n, \ \tau \in \mathbb{R}_+, \ \mathbf{x} \in \mathcal{X}. \tag{49c}$$

The probabilities $p^i(\mathbf{x})$ in the objective are given by the highly non-linear formula (45). As seen in Section 7.2.2, distribution shaping allows us to eliminate this non-linearity. To this end, we can simply introduce the auxiliary variables π_ℓ^i for $\ell \in [L]$, $i \in [n]$, add the corresponding defining constraints (47b)–(47f) to the problem (49), and replace $p^i(\mathbf{x})$ in (49a) with π_L^i. While the resulting formulation will be valid for any choice of the "scenario distance" δ, from now on we will restrict ourselves to the special case where δ is the discrete metric. Accordingly, ambiguity sets will be based on the total variation distance, which has the downside of ignoring potentially meaningful information about degrees of similarity between various scenarios. However, this choice of δ will eventually enable us to use scenario bundling methods for our DRO problems in a straightforward fashion.

7.2.5 MIP formulations for the total variation ball When δ is the discrete metric, we can combine distribution shaping with the ideas of Section 6.1.2 to reformulate (49) as

$$\min \quad \sum_{i\in[n]} \pi_L^i v^i + \kappa\tau \tag{50a}$$

$$\text{s.t.} \quad v^i \geq G^i, \qquad \forall i \in [n] \tag{50b}$$

$$v^i \geq G^+ - \tau, \qquad \forall i \in [n] \tag{50c}$$

$$\text{(47b)–(47f)}, \tag{50d}$$

$$\mathbf{v} \in \mathbb{R}^n, \ \tau \in \mathbb{R}_+, \ \mathbf{x} \in \mathcal{X}. \tag{50e}$$

The constraints of the above mixed-integer program are linear, but the summation in the objective function features non-convex quadratic terms. Our next goal is to linearize these terms. According to Example 5.2, as we are using a total variation-based ambiguity set, the optimum of (50) is a convex combination of the worst-case outcome G^+ and the nominal $\mathrm{CVaR}_\kappa(G)$. Moreover, this optimum is attained when we have $\tau = G^+ - \mathrm{VaR}_\kappa(G)$. Keeping in mind the representation (5) of CVaR as a minimum taken over the finite set of realizations (interpreted as the possible VaR values), we can now solve our DRO problem as follows. If we fix the value of the variable τ in (50), then we can also fix the values v^i by setting $v^i = \max\{G^i, G^+ - \tau\}$ for all $i \in [n]$, and the problem becomes an MIP. Let us then separately solve the n different MIPs obtained from (50) by fixing $\tau = G^+ - G^j$ for some $j \in [n]$. The MIP with the smallest optimum value will also provide the solution for (50).

While the solution of the n MIPs can naturally be parallelized, we can also use an alternative disjunction-based approach to combine these subproblems into a single MIP formulation. Let us introduce for all $j \in [n]$ a binary variable β^j that takes value 1 if and only if the nominal κ-level Value-at-Risk of G equals the realization G^j. Then, recalling the condition $\tau = G^+ - \mathrm{VaR}_\kappa(G)$, at an the optimum solution of (50) we have $v^i = \max\{G^i, G^+ - \tau\} = \max\{G^i, \sum_{j\in[n]} G^j\beta^j\} = \sum_{j\in[n]} \max\{G^i, G^j\}\beta^j$. We can now use McCormick envelopes to linearize the quadratic terms $\pi_L^i v^i$ in the objective (50a) by introducing the auxiliary variables $z^{ij} = \pi_L^i \beta^j$, leading to the following MIP formulation of (50):

$$
\begin{aligned}
\min \quad & \sum_{i\in[n]}\sum_{j\in[n]} \max\{G^i, G^j\} z^{ij} + \kappa\left(G^+ - \sum_{j\in[n]} G^j\beta^j\right) && \text{(51a)}\\
\text{s.t.} \quad & \text{(47b)–(47f)}, && \text{(51b)}\\
& z^{ij} \le \pi_L^i, \qquad \forall i,\ j \in [n] && \text{(51c)}\\
& z^{ij} \le \beta^j, \qquad \forall i,\ j \in [n] && \text{(51d)}\\
& z^{ij} \ge \pi_L^i + \beta_j - 1, \qquad \forall i,\ j \in [n] && \text{(51e)}\\
& \sum_{j\in[n]} \beta^j = 1, && \text{(51f)}\\
& \boldsymbol{\beta} \in \{0,1\}^n, \quad \mathbf{z} \in [0,1]^{n\times n}. && \text{(51g)}
\end{aligned}
$$

Keeping in mind that the auxiliary variables $\mathbf{z}$ are non-negative, the constraints (51c)–(51e) will ensure that the defining equations $z^{ij} = \pi_L^i\beta^j$ hold for all $i, j \in [n]$.

We can significantly strengthen the above MIP. Introducing the notation $\mathbb{1} = (1,\dots,1) \in \mathbb{R}^n$, in any feasible solution of (51) the triple $(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\pi}_L)$ belongs to the mixed-integer set

$$T = \left\{(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\pi}_L) \in [0,1]^{n\times n} \times \{0,1\}^n \times [0,1]^n \ : \ \mathbf{z} = \boldsymbol{\pi}_L\boldsymbol{\beta}^\top, \quad \boldsymbol{\beta}^\top\mathbb{1} = 1, \quad \mathbb{1}^\top\boldsymbol{\pi}_L = 1\right\}.$$

Sets of similar structure appear in so-called *pooling problems*, as discussed, for example, by Gupte et al. (2017), who use the reformulation-linearization technique to describe the convex hull (see also Sherali et al., 1998). Adapting their result to our setting, we obtain

$$\mathrm{conv}(T) = \left\{(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\pi}_L) \in [0,1]^{n\times n} \times [0,1]^n \times [0,1]^n \ : \ \mathbb{1}^\top\mathbf{z} = \boldsymbol{\beta}^\top, \quad \mathbf{z}\boldsymbol{\beta} = \boldsymbol{\pi}_L, \quad \mathbb{1}^\top\boldsymbol{\pi}_L = 1\right\}.$$

Incorporating the arising valid inequalities into an MIP can dramatically improve computational performance, as observed for example by Liu et al. (2017) in a different context. We then arrive at the following formulation:

$$
\begin{aligned}
\min \quad & \sum_{i\in[n]}\sum_{j\in[n]} \max\{G^i, G^j\} z^{ij} + \kappa\left(G^+ - \sum_{j\in[n]} G^j\beta^j\right) && \text{(52a)}\\
\text{s.t.} \quad & \text{(47b)–(47f)}, && \text{(52b)}\\
& \sum_{i\in[n]} z^{ij} = \beta^j, \qquad \forall j \in [n] && \text{(52c)}\\
& \sum_{j\in[n]} z^{ij} = \pi_L^i, \qquad \forall i \in [n] && \text{(52d)}
\end{aligned}
$$

$$\sum_{j\in[n]} \beta^j = 1, \tag{52e}$$

$$\boldsymbol{\beta} \in \{0,1\}^n, \quad \mathbf{z} \in [0,1]^{n\times n}. \tag{52f}$$

Throughout this section we considered the DRO version of the risk-neutral underlying problem (46), which was observed to be equivalent to a (non-DRO) risk-averse variant of the underlying problem. Notably, we did not consider the DRO version of a risk-averse underlying problem. The reason behind this omission is the following: Let us assume $\kappa \leq 1-\alpha$. When δ is the discrete metric, we know from Section 5.2.2 that the κ-robustification of CVaR_α is a convex combination of the worst case outcome and $\mathrm{CVaR}_{\alpha+\kappa}$. When outcomes are scenario-independent, the worst-case outcome is constant. Therefore in this case minimizing the robustified risk measure $\mathrm{CVaR}_\alpha^\kappa$ is equivalent to minimizing $\mathrm{CVaR}_{\alpha+\kappa}$, which is in turn equivalent to minimizing the robustified expectation $\mathrm{CVaR}_0^{\alpha+\kappa} = \mathbb{E}^{\alpha+\kappa}$.

On a related note, while we derived the MIP formulations (51) and (52) from the starting point of robustifying a risk-neutral underlying problem, the above arguments show that equivalent formulations can be obtained starting from a non-DRO CVaR-minimization problem. Appendix A explores this perspective in more detail.

7.2.6 DRO and scenario bundling The DRO formulations we presented so far all rely on a full (exponential sized) scenario set, which usually makes it impossible to solve problem instances of practical interest. Our goal in this section is to outline how to ameliorate this situation by incorporating scenario bundling into the formulations (50), (51), and (52), in order to reduce problem sizes. The first natural step—recalling the definitions and notation from Section 7.2.3—is to everywhere replace the full scenario set $[n]$ with a bundling S, and the scenario-based indexing $i \in [n]$ with bundle-based indexing $\mathbf{s} \in S$. Accordingly, we also replace the scenario probabilities $\boldsymbol{\pi}$ with the bundle probabilities $\boldsymbol{\phi}$, the distribution shaping constraints (47b)–(47f) with their bundle-based counterparts (48b)–(48g), and the scenario outcomes G^i with bundle outcomes $G^{\mathbf{s}}$. The last step is to replace the scenario distances δ^{ij} with suitably defined *bundle distances.*

For a general distance δ this last step is highly non-trivial, and typically leads to undesired consequences, such as formulations whose optimal solution depends on the particular choice of bundling. In more detail: Let S be a bundling, as defined in Section 7.2.3. In line with the considerations of that section, any probability distribution $\mathbb{P}$ on the scenario set induces a distribution $\bar{\mathbb{P}}$ on the bundles, given by $\bar{p}^{\mathbf{s}} = \sum_{i\in B_{\mathbf{s}}} p^i$ for $\mathbf{s} \in S$. Let us now consider an EMD ball around $\mathbb{P}$, based on some distance δ. In order for the scenario bundling approach to work in a DRO context, we would need to characterize the family of probability distributions on S that are induced by the elements of this EMD ball. Furthermore, if we aim to achieve this goal via a straightforward adaptation of our previous MIP formulations, the characterization should be in the form of an EMD ball around the induced distribution $\bar{\mathbb{P}}$, with respect to some distance $\bar{\delta}$ among bundles. Unfortunately, there is no general scheme to define such a bundle distance. In particular, applying natural schemes (such as a Hausdorff distance-like maximin approach, or defining the distance between two bundles as the smallest distance between any two of their respective scenarios) to our problems can lead to results that are not only inexact, but depend strongly on the particular choice of bundling. However, as the next lemma shows, adopting the total variation distance (i.e., setting δ as the discrete metric) neatly sidesteps such issues.

Lemma 7.1 *Let us denote the discrete metric on $\{0,1\}^L$ by δ, and the discrete metric on a bundling S by $\bar{\delta}$. Then, denoting the identity map of S by $\bar{\boldsymbol{\xi}}$, for any radius $\kappa > 0$, we have*

$$\mathcal{B}^{\bar{\boldsymbol{\xi}}}_{\bar{\delta},\kappa}(\bar{\mathbb{P}}) = \left\{ \bar{\mathbb{Q}} \ : \ \mathbb{Q} \in \mathcal{B}^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P}) \right\}. \tag{53}$$

Proof. Let us first consider a distribution $\mathbb{Q} \in \mathcal{B}^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$. According to Lemma 5.1, there exists a corresponding solution $\boldsymbol{\gamma} \in \mathbb{R}^{n\times n}_{+}$ to system (26). Let us now define the aggregated solution $\bar{\boldsymbol{\gamma}} \in \mathbb{R}^{S\times S}_{+}$ by $\bar{\gamma}^{\mathbf{st}} = \sum_{i\in B_{\mathbf{s}}}\sum_{j\in B_{\mathbf{t}}} \gamma^{ij}$ for $\mathbf{s},\mathbf{t} \in S$, and observe that for all $i \in B_{\mathbf{s}}$, $j \in B_{\mathbf{t}}$ we have $\bar{\delta}^{\mathbf{st}} \leq \delta^{ij}$. Then, using Lemma 5.1 again, $\bar{\mathbb{Q}} \in \mathcal{B}^{\bar{\boldsymbol{\xi}}}_{\bar{\delta},\kappa}(\bar{\mathbb{P}})$ follows from the next set of inequalities:

$$\begin{aligned}
&\sum_{\mathbf{t}\in S}\bar{\gamma}^{\mathbf{st}} = \sum_{\mathbf{t}\in S}\sum_{i\in B_{\mathbf{s}}}\sum_{j\in B_{\mathbf{t}}}\gamma^{ij} = \sum_{i\in B_{\mathbf{s}}}\sum_{j\in[n]}\gamma^{ij} = \sum_{i\in B_{\mathbf{s}}} p^i = \bar{p}^{\mathbf{s}}, && \forall\, \mathbf{s}\in S\\
&\sum_{\mathbf{s}\in S}\bar{\gamma}^{\mathbf{st}} = \sum_{\mathbf{s}\in S}\sum_{i\in B_{\mathbf{s}}}\sum_{j\in B_{\mathbf{t}}}\gamma^{ij} = \sum_{j\in B_{\mathbf{t}}}\sum_{i\in[n]}\gamma^{ij} = \sum_{j\in B_{\mathbf{t}}} q^j = \bar{q}^{\mathbf{t}}, && \forall\, \mathbf{t}\in S\\
&\sum_{\mathbf{s}\in S}\sum_{\mathbf{t}\in S}\bar{\delta}^{\mathbf{st}}\bar{\gamma}^{\mathbf{st}} = \sum_{\mathbf{s}\in S}\sum_{\mathbf{t}\in S}\sum_{i\in B_{\mathbf{s}}}\sum_{j\in B_{\mathbf{t}}}\bar{\delta}^{\mathbf{st}}\gamma^{ij} \leq \sum_{\mathbf{s}\in S}\sum_{\mathbf{t}\in S}\sum_{i\in B_{\mathbf{s}}}\sum_{j\in B_{\mathbf{t}}}\delta^{ij}\gamma^{ij} = \sum_{i\in[n]}\sum_{j\in[n]}\delta^{ij}\gamma^{ij} \leq \kappa.
\end{aligned}$$

Now let us consider a distribution $\mathbb{T} \in \mathcal{B}^{\bar{\boldsymbol{\xi}}}_{\bar{\delta},\kappa}(\bar{\mathbb{P}})$, along with the corresponding $\bar{\gamma} \in \mathbb{R}^{S\times S}_{+}$ guaranteed by Lemma 5.1. As the bundling S represents a partition of $[n]$, for every $i \in [n]$ there exists a unique bundle $\mathbf{s}(i) \in S$ such that $i \in B_{\mathbf{s}(i)}$ holds. Using this notation, let us define the disaggregation $\boldsymbol{\gamma} \in \mathbb{R}^{n\times n}_{+}$ by

$$\gamma^{ij} = \begin{cases} \frac{p^i p^j}{\bar{p}^{\mathbf{s}(i)}\bar{p}^{\mathbf{s}(j)}}\bar{\gamma}^{\mathbf{s}(i)\mathbf{s}(j)} & \text{if } \mathbf{s}(i) \neq \mathbf{s}(j),\\ \frac{p^i}{\bar{p}^{\mathbf{s}(i)}}\bar{\gamma}^{\mathbf{s}(i)\mathbf{s}(i)} & \text{if } i = j,\\ 0 & \text{otherwise.} \end{cases}$$

If we define the probability distribution $\mathbb{Q}$ on the scenarios by $q^j = \sum_{i\in[n]}\gamma^{ij}$, then it is easy to verify that have $\mathbb{T} = \bar{\mathbb{Q}}$. Observing that $\gamma^{ij} \neq 0$ implies $\delta^{ij} = \bar{\delta}^{\mathbf{s}(i)\mathbf{s}(j)}$, it is also straightforward to verify that $\boldsymbol{\gamma}$ is a solution of the system (26). It follows that we have $\mathbb{Q} \in \mathcal{B}^{\boldsymbol{\xi}}_{\delta,\kappa}(\mathbb{P})$, which completes the proof. □

Remark 7.3 *The above lemma remains valid if we replace the discrete metric with a reflexive scenario distance δ that depends only on the outcomes. More precisely, let us assume that for all $i,j \in [n]$ we have $\delta^{ij} = d(G^i, G^j)$, for some mapping $d : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_{+}$ that satisfies $d(G,G) = 0$ for all $G \in \mathbb{R}$. Then, given two scenario bundles $\mathbf{s},\mathbf{t} \in S$, for any $i_1, i_2 \in B_{\mathbf{s}}$ and $j_1, j_2 \in B_{\mathbf{t}}$ we have $\delta^{i_1 j_1} = d(G^{i_1}, G^{j_1}) = d(G^{i_2}, G^{j_2}) = \delta^{i_2 j_2}$. Let us denote this common value by $\bar{\delta}^{\mathbf{st}}$ to define a reflexive bundle distance $\bar{\delta} : S \times S \to \mathbb{R}_{+}$. The equality (53) then follows by using the same proof as before.*

7.2.7 Limitations of scenario bundling

The problem originally introduced in Peeta et al. (2010) features multiple O-D pairs with corresponding weights. The formulations (49)–(52) can naturally be adapted to this case by having the random outcome G represent the weighted sum of shortest path lengths between these pairs. However, bundling methods can no longer be directly applied here, as links that are irrelevant to the length of a shortest path for one O-D pair will typically not be irrelevant to the lengths of shortest paths between other pairs. As proposed in Laumanns et al. (2014), it is possible to instead perform bundling separately for each O-D pair, and again use distribution shaping to express the marginal distribution of the shortest path length for each O-D pair. When solving the risk-neutral underlying problem, these marginal distributions will suffice, because—due to the linearity of expectation—the expected shortest path lengths for the individual O-D pairs can be aggregated into a global objective.

This approach is unfortunately no longer viable when working either in a risk-averse or in a DRO context. Expressing CVaR, or any more complex risk measure, of the global objective would require knowledge of the joint distribution of shortest path lengths, because the risk of a weighted sum is in general not equal to the weighted sum of individual risks. Similarly, the obstacle for DRO problems is that the ambiguity set around the joint distribution cannot be reduced to ambiguity sets around the marginal distributions. Therefore, while we might be able to find the worst-case distribution for each

O-D pair, this does not give us the global worst-case distribution. The scope of the methods outlined in this section is thus limited to cases where either scenario bundling can effectively be applied in terms of the global objective function, or the overall number of scenarios is relatively small. Fortunately, in addition to the single O-D pair problem we explored here, the "global bundling" approach also produces good results for other problems such as stochastic network interdiction (Haus et al., 2017).

8. Further avenues of research One of the main distinguishing features of our approach is that the nominal distribution at the center of the ambiguity set is decision-dependent. In this regard, it would be essential to investigate possible ways to describe this dependence, and in particular to develop meaningful and tractable characterizations of decision-dependent nominal parameter realizations and/or scenario probabilities (akin to the distribution shaping equations) for practical applications.

We have mentioned in Section 7.2.6 that while most EMDs are not compatible with the scenario bundling approach, the total variation metric is a notable exception. Remark 7.3 presents another, potentially more informative, class of outcome-based scenario distances, which give rise to EMDs that can be used in conjunction with bundling. However, as many of our developments strongly depend on the structural properties of the variation metric, incorporating this new class of distances into our formulations would require additional work.

As noted in Section 7.2.7, there are problems of practical interest where bundling methods do not appear to be applicable. In such cases one might instead consider using sampling methods to reduce the number of scenarios. The use of sampling, however, comes with two important caveats. First, sampling typically sacrifices exact solutions in exchange for tractability. Second, even in cases where a particular sampling method (such as importance sampling) is known to work well for the underlying problem, this does not automatically translate to a performance guarantee for the DRO variant. For example, when using an ambiguity set based on the discrete metric, the worst-case distribution will be highly sensitive to the worst scenario included in a sample. However, studies such as Bardou et al. (2009) indicate that sampling approaches could be better suited for CVaR-based and other risk-averse formulations which do not explicitly feature the ambiguity set. Along similar lines, the use of scenario reduction techniques could also be explored.

In line with the majority of the DRO literature, we have adopted a risk-averse pessimistic viewpoint, focusing on the worst outcome in the ambiguity set. In contrast, optimistic robust optimization has recently been suggested to be meaningful and relevant for certain application areas, including machine learning (see, e.g., Norton et al., 2017). Taking an optimistic view in our problems would lead us to replace maximization over the ambiguity set with minimization, typically making the resulting problems significantly more straightforward. Therefore, this seems to be a worthwhile avenue to explore whenever an optimistic view is warranted by a particular application.

References

Acerbi, C. (2002). Spectral measures of risk: a coherent representation of subjective risk aversion. *Journal of Banking and Finance*, 26(7):1505–1518.

Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Bardou, O., Frikha, N., and Pages, G. (2009). Computing var and cvar using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods and Applications*, 15(3):173–210.

Bayraksan, G. and Love, D. (2015). *TutORials in Operations Research*, chapter Data-Driven Stochastic Programming using Phi-Divergences, pages 1–15. INFORMS.

Bertsimas, D. and Vayanos, P. (2014). Data-driven learning in dynamic pricing using adaptive optimization. Available at Optimization Online: http://www.optimization-online.org/DB_FILE/2014/10/4595.pdf.

Blanchet, J., Kang, Y., Zhang, F., He, F., and Hu, Z. (2017). Doubly robust data-driven distributionally robust optimization. Technical report. Technical Report 1705.07168, ArXiv.

Calafiore, G. C. (2007). Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization*, 18(3):853–877.

Cormican, K. J., Morton, D. P., and Wood, R. K. (1998). Stochastic network interdiction. *Operations Research*, 46(2):184–197.

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

Dupacova, J. (2006). Optimization under exogenous and endogenous uncertainty. University of West Bohemia in Pilsen. Unpublished, http://www.karlin.mff.cuni.cz/~dupacova/papers/MME06sty.pdf.

Erdoğan, E. and Iyengar, G. (2006). Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61.

Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*. Online first https://link.springer.com/article/10.1007/s10107-017-1172-1.

Flach, B. and Poggi, M. (2010). On a class of stochastic programs with endogenous uncertainty: theory, algorithm and application. Technical report. Monografias em Ciencia da Computacao No 05/10, PUC, Rio.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447.

Gao, R., Chen, X., and Kleywegt, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. *CoRR*, abs/1712.06050.

Gao, R. and Kleywegt, A. (2016). Distributionally robust stochastic optimization with Wasserstein distance. Technical report. Technical Report 1604.02199, ArXiv.

Gao, R. and Kleywegt, A. (2017). Distributionally robust stochastic optimization with dependence structure. Available at Optimization Online: http://www.optimization-online.org/DB_HTML/2017/01/5817.html.

Goel, V. and Grossmann, I. E. (2004). A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers & Chemical Engineering*, 28(8):1409 – 1429.

Goel, V. and Grossmann, I. E. (2006). A class of stochastic programs with decision dependent uncertainty. *Mathematical Programming*, 108(2):355–394.

Goh, J. and Sim, M. (2010). Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1):902–917.

Gupte, A., Ahmed, S., Dey, S. S., and Cheon, M. S. (2017). Relaxations and discretizations for the pooling problem. *Journal of Global Optimization*, 67(3):631–669.

Hansen, P., Jaumard, B., and Lu, S.-H. (1992). Global optimization of univariate lipschitz functions: Ii. new algorithms and computational comparison. *Mathematical Programming*, 55(1):273–292.

Haus, U.-U., Michini, C., and Laumanns, M. (2017). Scenario aggregation using binary decision diagrams for stochastic programs with endogenous uncertainty. Unpublished, https://arxiv.org/pdf/1701.04055.pdf.

Hellemo, L. (2016). Managing uncertainty in design and operation of natural gas infrastructure. Ph.D. thesis, Norwegian University of Science and Technology.

Hellemo, L., Barton, P. I., and Tomasgard, A. (2014). Stochastic programming with decision-dependent probabilities. Unpublished, http://strato.impa.br/videos/2014-festival-incerteza/09_AsgeirTomasgard.pdf.

Hu, Z. and Hong, L. J. (2012). Kullback-Leibler divergence constrained distributionally robust optimization. Available at Optimization Online: http://www.optimization-online.org/DB_FILE/2012/11/3677.pdf.

Ji, R. and Lejeune, M. (2017). Data-driven optimization of reward-risk ratio measures. Available at Optimization Online: http://www.optimization-online.org/DB_HTML/2017/01/5819.html.

Jiang, R. and Guan, Y. (2015). Data-driven chance constrained stochastic program. *Mathematical Programming*, 158:291–327.

Jiang, R. and Guan, Y. (2018). Risk-averse two-stage stochastic program with distributional ambiguity. *Operations Research*. Online first https://pubsonline.informs.org/doi/10.1287/opre.2018.1729.

Jonsbråten, T., Wets, R.-B., and Woodruff, D. (1998). A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82(0):83–106.

Kantorovich, L. V. and Rubinshtein, G. S. (1958). On a space of totally additive functions. Vestnik Leningradskogo Universiteta, 13, pp. 52-59.

Keha, A. B., Khowala, K., and Fowler, J. W. (2009). Mixed integer programming formulations for single machine scheduling problems. *Computers and Industrial Engineering*, 56(1):357–367.

Khaligh, F. H. and MirHassani, S. (2016). A mathematical model for vehicle routing problem under endogenous uncertainty. *International Journal of Production Research*, 54(2):579–590.

Kusuoka, S. (2001). On law invariant coherent risk measures. *Advances in Mathematical Economics*, 3:83–95.

Kyparisis, J. and Fiacco, A. V. (1987). Generalized convexity and concavity of the optimal value function in nonlinear programming. *Mathematical Programming*, 39(3):285–304.

Lam, H. (2016). Recovering best statistical guarantees via the empirical divergence-based distributionally. Technical report. Technical Report 1605.09349, ArXiv.

Lappas, N. H. and Gounaris, C. E. (2018). Robust optimization for decision-making under endogenous uncertainty. *Computers & Chemical Engineering*. Online first https://www.sciencedirect.com/science/article/pii/S0098135418300152.

Laumanns, M., Prestwich, S., and Kawas, B. (2014). Distribution shaping and scenario bundling for stochastic programs with endogenous uncertainty. Unpublished, https://edoc.hu-berlin.de/bitstream/handle/18452/9095/5.pdf.

Lejeune, M. A. and Shen, S. (2016). Multi-objective probabilistically constrained programs with variable risk: Models for multi-portfolio financial optimization. *European Journal of Operational Research*, 252(2):522 – 539.

Lindvall, T. (1992). *Lectures on the Coupling Method*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley.

Liu, X., Kucukyavuz, S., and Noyan, N. (2017). Robust multicriteria risk-averse stochastic programming models. *Annals of Operations Research*, 259(1):259–294.

Luo, F. and Mehrotra, S. (2017). Decomposition algorithms for distributionally robust optimization using wasserstein metric. http://www.optimization-online.org/DB_HTML/2017/04/5946.html.

McCormick, G. (1976). Computability of global solutions to factorable nonconvex programs: Part I – convex underestimating problems. *Mathematical Programming*, 10(1):147–175.

Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. John Wiley & Sons, Chichester.

Nohadani, O. and Sharma, K. (2016). Optimization under decision-dependent uncertainty. Technical report. Technical Report 1611.07992, ArXiv.

Norton, M., Takeda, A., and Mafusalov, A. (2017). Optimistic Robust Optimization With Applications To Machine Learning. *ArXiv e-prints*.

Noyan, N. and Rudolf, G. (2013). Optimization with multivariate conditional value-at-risk-constraints. *Operations Research*, 61(4):990–1013.

Noyan, N. and Rudolf, G. (2015). Kusuoka representations of coherent risk measures in general probability spaces. *Annals OR*, 229(1):591–605.

Noyan, N. and Rudolf, G. (2018). Optimization with stochastic preferences based on a general class of scalarization functions. *Operations Research*, 66(2):463–486.

Peeta, S., Salman, F., Gunnec, D., and Viswanath, K. (2010). Pre-disaster investment decisions for strengthening a highway network. *Computers and Operations Research*, 37:1708–1719.

Pflug, G. and Wozabal, D. (2007). Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442.

Pflug, G. C. and Pichler, A. (2011). *Approximations for Probability Distributions and Stochastic Optimization Problems*, pages 343–387. Springer New York, New York, NY.

Pflug, G. C., Pichler, A., and Wozabal, D. (2012). The 1/N investment strategy is optimal under high model ambiguity. *Journal of Banking & Finance*, 36(2):410–417.

Pflug, G. C. and Römisch, W. (2007). *Modelling, managing and measuring risk.* World Scientific publishing, Singapore.

Pinedo, M. (2008). *Scheduling: Theory, Algorithms, and Systems.* Springer, 3rd edition.

Postek, K., Den Hertog, D., and Melenberg, B. (2016). Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650.

Rahimian, H., Bayraksan, G., and Homem-de Mello, T. (2018). Identifying effective scenarios in distributionally robust stochastic programs with total variation distance. *Mathematical Programming.*

Rockafellar, R. (2007). Coherent approaches to risk in optimization under uncertainty. In *Tutorials in Operations Research*, pages 38–61. INFORMS.

Rockafellar, R. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41.

Royset, J. O. and Wets, R. J.-B. (2017). Variational theory for optimization under stochastic ambiguity. *SIAM Journal on Optimization*, 27(2):1118–1149.

Rubner, Y., Tomasi, C., and Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66.

Schichl, H. and Sellmann, M. (2015). Predisaster preparation of transportation networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 709–715. AAAI Press.

Shabtay, D. and Steiner, G. (2007). A survey of scheduling with controllable processing times. *Discrete Applied Mathematics*, 155(13):1643 – 1666.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory.* The society for industrial and applied mathematics and the mathematical programming society, Philadelphia, USA.

Sherali, H. D. and Adams, W. P. (1994). A hierarchy of relaxations and convex hull representations for mixed-integer zero-one programming problems. *Discrete Appl. Math.*, 52(1):83–106.

Sherali, H. D., Adams, W. P., and Driscoll, P. J. (1998). Exploiting special structures in constructing a hierarchy of relaxations for 0-1 mixed integer problems. *Operations Research*, 46(3):396–405.

Van Parys, B., Estafahani, M., and Kuhn, D. (2017). From data to decisions: Distributionally robust optimization is optimal. Technical report. Technical Report 1704.04118, ArXiv.

Wang, Z., Glynn, P. W., and Ye, Y. (2016). Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261.

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

Wozabal, D. (2014). Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research*, 62(6):1302–1315.

Zhang, J., Xu, H., and Zhang, L. (2016). Quantitative stability analysis for distributionally robust optimization with moment constraints. *SIAM Journal on Optimization*, 26(3):1855–1882.

Zhao, C. and Guan, Y. (2015). Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. Available at Optimization Online: http://www.optimization-online.org/DB_HTML/2015/07/5014.html.

Zhao, C. and Guan, Y. (2018). Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267.

Zymler, S., Kuhn, D., and Rustem, B. (2013). Worst-case value at risk of nonlinear portfolios. *Management Science*, 59(1):172–188.

Appendix A. Non-DRO risk-averse optimization perspective In this appendix we provide an alternative view of the developments in Section 7.2.5. Accordingly, we use the notation and assumptions from that section. Let us recall from Example 5.2 that, when we use a total variation-based ball of radius $\kappa \in [0,1]$, the robustified expectation is a convex combination of the worst-case outcome (with weight κ) and the κ-level CVaR of the outcome. Accordingly, under our assumptions (DRO-RND) is equivalent to the following problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \quad \kappa G^+ + (1-\kappa)\,\mathrm{CVaR}_\kappa\left([\mathbb{P}(\mathbf{x}), G]\right). \tag{54}$$

We remark that (54) is equivalent to a non-DRO CVaR minimization problem, because the worst-case outcome G^+ is not decision-dependent. In settings without endogenous uncertainty it is possible to obtain a linear formulation for CVaR minimization problems by using representation (5). A direct implementation of this approach in the presence of decision-dependent probabilities gives rise to a highly non-linear model. As we saw in Section 7.2.2, distribution shaping can be used to get rid of the nonlinearity in probability expressions of the form (45), and reformulate (54) as follows.

$$\begin{aligned}
\min \quad & \kappa G^+ + (1-\kappa)\sum_{j\in[n]}\left(G^j + \frac{1}{1-\kappa}\sum_{i\in[n]}\pi_L^i[G^i - G^j]_+\right)\beta^j && \text{(55a)}\\
\text{s.t.} \quad & \text{(47b)–(47g)}, && \text{(55b)}\\
& \sum_{j\in[n]}\beta^j = 1, && \text{(55c)}\\
& \boldsymbol{\beta}\in\{0,1\}^n. && \text{(55d)}
\end{aligned}$$

It is well-known that in settings without endogenous uncertainty CVaR minimization can be expressed in a linear fashion. In contrast, the formulation (55) still has a non-convex quadratic objective. Analogously to the case of problem (50), one way to obtain an optimal solution of the problem (55) is to separately solve the n MIP formulations that arise when we fix $\beta^j = 1$ for $j \in [n]$. Alternatively, we can again use McCormick envelopes to linearize the objective function, introducing the auxiliary variables $z^{ij} = \pi_L^i\beta^j$. We omit the details here, as the arising formulation will be essentially equivalent to (51). To see that this equivalence holds, it is sufficient to verify that the objective functions (51a) and (55a) are equal, which in turn follows by straightforward calculation from the simple observation that we have $[G^i - G^j]_+ = \max\{G^i, G^j\} - G_j$ for $i, j \in [n]$.

It is also possible to obtain an MIP formulation of (54) without linearizing the quadratic terms $\pi_L^i\beta^j$. Let us introduce the auxiliary variable μ to represent $\mathrm{CVaR}_\kappa\left([\mathbb{P}(\mathbf{x}), G]\right)$. Then, using a disjunction-based representation of the finite minimum in the CVaR representation (5), we arrive at

$$\begin{aligned}
\min \quad & \kappa G^+ + (1-\kappa)\mu && (56a)\\
\text{s.t.} \quad & (47b)\text{–}(47g), && (56b)\\
& \mu \le G^j + \frac{1}{1-\kappa}\sum_{i\in[n]} \pi_L^i [G^i - G^j]_+, \qquad \forall j \in [n] && (56c)\\
& \mu \ge G^j + \frac{1}{1-\kappa}\sum_{i\in[n]} \pi_L^i [G^i - G^j]_+ - (1-\beta^j)M, \qquad \forall j \in [n] && (56d)\\
& (55c) - (55d). && (56e)
\end{aligned}$$

Here, if $M \in \mathbb{R}_+$ is a suitably large constant, the constraints (56c)–(56e) ensure the validity of the defining equation

$$\mu = \mathrm{CVaR}_\kappa\left([\mathbb{P}(\mathbf{x}), \boldsymbol{\xi}]\right) = \min_{j\in[n]} G^j + \frac{1}{1-\kappa}\sum_{i\in[n]} \pi_L^i [G^i - G^j]_+. \tag{57}$$

We could also have arrived at an equivalent form of the MIP (56) by applying a similar disjunctive approach, but with the formulation (50) as our starting point. As discussed in Section 7.2.5, the optimum in (50) can be expressed as a finite minimum. More precisely, the optimum is attained when we have $\tau = G^+ - G^j$ for some $j \in [n]$, with corresponding values $v^i = \max\{G^i, G^j\}$. Recalling that we have $\max\{G^i, G^j\} = [G^i - G^j]_+ + G^j$ for $i, j \in [n]$, we can now express the optimum value of (50) as the minimum of the following expression, taken over all $(\mathbf{x}, \boldsymbol{\pi})$ satisfying (47b)–(47g):

$$\begin{aligned}
& \min_{j\in[n]} \sum_{i\in[n]} \pi_L^i \max\{G^i, G^j\} + \kappa(G^+ - G^j) && (58a)\\
&= \kappa G^+ + \min_{j\in[n]} \sum_{i\in[n]} \pi_L^i \max\{G^i, G^j\} - \kappa G^j && (58b)\\
&= \kappa G^+ + \min_{j\in[n]} \sum_{i\in[n]} \pi_L^i ([G^i - G^j]_+ + G^j) - \kappa G^j && (58c)\\
&= \kappa G^+ + \min_{j\in[n]} \sum_{i\in[n]} \pi_L^i [G^i - G^j]_+ + \sum_{i\in[n]} \pi_L^i G^j - \kappa G^j && (58d)\\
&= \kappa G^+ + \min_{j\in[n]} \sum_{i\in[n]} \pi_L^i [G^i - G^j]_+ + (1-\kappa) G^j && (58e)\\
&= \kappa G^+ + (1-\kappa) \min_{j\in[n]} G^j + \frac{1}{1-\kappa}\sum_{i\in[n]} \pi_L^i [G^i - G^j]_+ && (58f)\\
&= \kappa G^+ + (1-\kappa)\mu, && (58g)
\end{aligned}$$

where μ is given by the expression in (57).