Data-Driven Two-Stage Conic Optimization with Zero-One Uncertainties

Anirudh Subramanyam[1], Mohamed El Tonbari[2], and Kibaek Kim[1]

[1]Argonne National Laboratory, Lemont, IL

[2]Georgia Institute of Technology, Atlanta, GA

July 16, 2021

Abstract

We address high-dimensional zero-one random parameters in two-stage convex conic optimization problems. Such parameters typically represent failures of network elements and constitute rare, high-impact random events in several applications. Given a sparse training dataset of the parameters, we motivate and study a distributionally robust formulation of the problem using a Wasserstein ambiguity set centered at the empirical distribution. We present a simple, tractable, and conservative approximation of this problem that can be efficiently computed and iteratively improved. Our method relies on a reformulation that optimizes over the convex hull of a mixed-integer conic programming representable set, followed by an approximation of this convex hull using lift-and-project techniques. We illustrate the practical viability and strong out-of-sample performance of our method on nonlinear optimal power flow and multi-commodity network design problems that are affected by random contingencies, and we report improvements of up to 20% over existing sample average approximation and two-stage robust optimization methods.

1 Motivation

This work is motivated by optimization problems arising in applications that are affected by an extremely large, yet finite, number of rare, high-impact random events. In particular, we are motivated by applications in which the decision-relevant random events consist of high-dimensional

binary outcomes. Such applications are ubiquitous in network optimization, where the uncertain events amount to failures of the nodes or edges of the underlying network. For example, in electric power networks, random node and edge failures have been used to model losses of physical components such as substations, transmission lines, generators, and transformers [13, 51]. Similarly, in natural gas [55], wireless communication [27], and transportation networks [10], they can be used to model failures of compressors and gas pipelines, antennas, and road links, respectively.

The challenges in modeling and solving such uncertainty-affected optimization problems are threefold. First, the number of possible failure states grows *exponentially* as the size of the input increases. For example, the number of failure states in a network with n failure-prone nodes is 2^n. Second, network failures are extremely *rare* but *critical*, and historical records are often not rich enough to include observations for every possible failure state. Third, failures of individual network elements are unlikely to be independent of each other. For example, transmission line failures in electric power systems often have a cascading effect that triggers the failure of other transmission lines. Because of this *high dimensionality*, the true distribution governing the random parameters is often unknown and difficult to estimate with a small number of historical observations.

In this high-dimensional context, we focus on *two-stage optimization* problems under uncertainty. Suppose that there are M uncertain parameters $\tilde{\xi}_1, \ldots, \tilde{\xi}_M$, and that $\Xi \subseteq \{0,1\}^M$ is the support of the underlying distribution of these parameters. If the distribution $\mathbb{P}$ is known, then the two-stage problem takes the form

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ c(\boldsymbol{x}) + \mathbb{E}_{\mathbb{P}}\left[\mathcal{Q}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right],$$

where $\boldsymbol{x}$ represents the *first-stage* decisions that must be made agnostically to the realization of the random parameters, $\mathcal{X} \subseteq \mathbb{R}^{N_1}$ is a convex compact set of feasible first-stage decisions, $c : \mathcal{X} \mapsto \mathbb{R}$ is a convex function representing the deterministic cost associated with these decisions, and $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is the *random loss* (or *second-stage cost*) corresponding to decisions $\boldsymbol{x}$ and a fixed realization $\boldsymbol{\xi} \in \Xi$ of the random parameters. We assume that the loss can be computed by solving a convex conic optimization problem of the form

$$\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) = \inf_{\boldsymbol{y}\in\mathcal{Y}} \left\{ \boldsymbol{q}(\boldsymbol{\xi})^\top \boldsymbol{y} : \boldsymbol{W}(\boldsymbol{\xi})\boldsymbol{y} \geq \boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi} + \boldsymbol{h}(\boldsymbol{x}) \right\}, \tag{1}$$

where $\boldsymbol{y}$ denotes the *second-stage* decisions that can be made after the realization $\boldsymbol{\xi}$ is known; $\mathcal{Y} \subseteq \mathbb{R}^{N_2}$ is a *proper* (closed, convex, pointed, and full-dimensional) cone; $\boldsymbol{q} : \Xi \mapsto \mathbb{R}^{N_2}$ and

$\boldsymbol{W} : \Xi \mapsto \mathbb{R}^{L \times N_2}$ are vector- and matrix-valued affine functions respectively, while $\boldsymbol{h} : \mathcal{X} \mapsto \mathbb{R}^L$ and $\boldsymbol{T} : \mathcal{X} \mapsto \mathbb{R}^{L \times M}$ are componentwise closed, proper, convex vector- and matrix-valued functions, respectively. We allow uncertainty to affect only the affine constraints of the problem and, similarly, the first- and second-stage decisions to interact only via the affine part.

Since the true underlying distribution $\mathbb{P}$ is unknown, this two-stage optimization formulation is ill-posed. Nevertheless, $\mathbb{P}$ is typically observable through a finite amount of historical data. We assume that we have access to N such independent and identically distributed observations, which we denote by $\{\hat{\boldsymbol{\xi}}^{(1)}, \ldots, \hat{\boldsymbol{\xi}}^{(N)}\}$. We also assume that generating additional data (e.g., via Monte Carlo computer simulations) is either costly or impossible. Thus, it is imperative to use the given data most efficiently.

A popular approach for solving the two-stage problem in such cases is to use *sample average approximation* [59]. This approach replaces the true distribution with the empirical distribution $\hat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{\boldsymbol{\xi}}^{(i)}}$, where $\delta_{\hat{\boldsymbol{\xi}}^{(i)}}$ denotes the Dirac distribution at $\hat{\boldsymbol{\xi}}^{(i)}$. In the context of rare events, however, obtaining accurate estimates of the true distribution and, hence, the optimal solution of the true two-stage problem may require unrealistically large amounts of data. The following example illustrates this phenomenon even if there are only $M = 3$ uncertain parameters.

Example 1. *Consider the network shown in Figure 1. Let A denote the set of arcs. The goal is to decide arc capacities $\boldsymbol{x} \in \mathbb{R}_+^{|A|}$ so as to route flow originating from the left layer of nodes to satisfy demand in the right layer of nodes. The per-unit cost $\boldsymbol{c} \in \mathbb{R}^{|A|}$ of installing capacity on each arc is denoted above the arc, and the first-stage deterministic cost is given by $\boldsymbol{c}^\top \boldsymbol{x}$. The middle layer of nodes can fail independently of each other, and the numbers above the nodes denote their failure probabilities. If node i fails (indicated by $\xi_i = 1$), then all arcs incident to that node become*

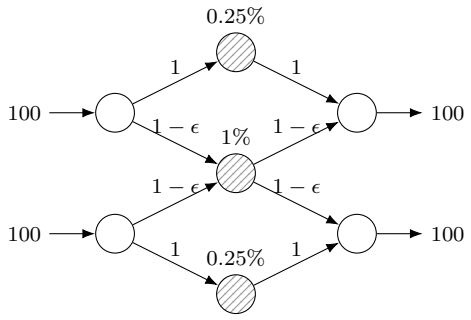
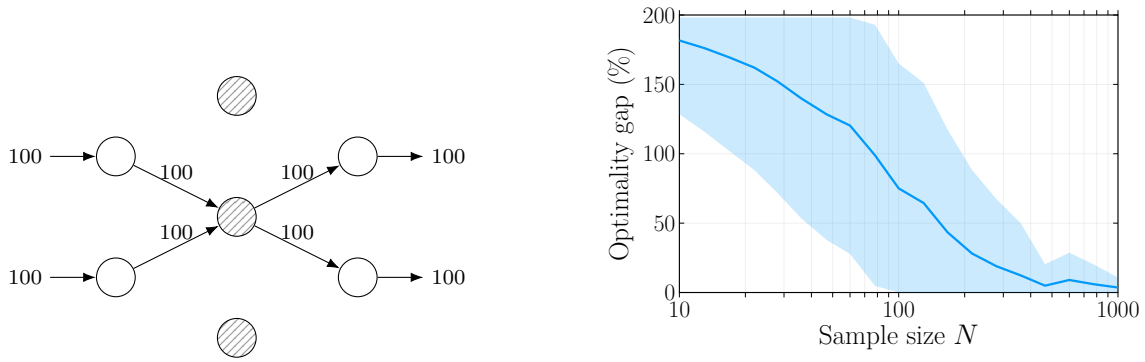


Figure 1. Illustrative network flow instance.

unusable, and any resulting supply shortfall is penalized at a cost of $1,000$ *per unit. For a given realization* $\boldsymbol{\xi} \in \{0,1\}^3$ *of the node failures, the loss function* $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ *is simply the total penalty cost under that realization and can be modeled as the optimal objective value of a linear program.*

We can show that the optimal solution assigns a capacity of 100 units to every arc (see Appendix A). In practice, however, the true failure probabilities are unknown. We therefore estimate the distribution using a sample average approximation. Figure 2 shows the performance of the solutions obtained by replacing the true distribution $\mathbb{P}$ *with the empirical distribution* $\hat{\mathbb{P}}_N$ *resulting from different sample sizes* N*. In particular, the figure shows the difference between the total expected costs (computed under the true distribution) of the sample average solution and the true optimal solution. We observe that despite the small dimensionality* ($M = 3$)*,* $N \geq 1000$ *samples are required for the sample average solution to estimate the true solution to an accuracy of 5%.*



(a) The optimal arc capacities determined by the sample average approximation when the empirical distribution $\hat{\mathbb{P}}_N$ puts all weight on the realization $\boldsymbol{\xi} = \mathbf{0}$, where no nodes fail (see Appendix A).

(b) The optimality gap of the sample average approximation with respect to true optimal objective for increasing sample size. The mean (solid line) and standard deviation (shaded) are estimated by using 1,000 statistically independent sets of samples.

Figure 2. Performance of the sample average approximation on the network flow instance from Figure 1.

2 Distributionally Robust Approach for Discrete Rare Events

The high dimensionality and rare occurrence of failure states render accurate estimation of the underlying distribution difficult. To remedy this situation, we adopt a *distributionally robust* approach, and construct an *ambiguity set* $\mathcal{P}$ of possible distributions that are consistent with the observed data. We then minimize the *worst-case* expected costs over all distributions in the ambi-

guity set. Specifically, we consider distributionally robust two-stage conic optimization problems, of the form

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ c(\boldsymbol{x}) + \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\mathcal{Q}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]. \tag{2}$$

The ambiguity set $\mathcal{P}$ must be chosen such that it contains the true distribution with high confidence or, at the very least, distributions that assign nonzero probability to the rare events. We focus on the Wasserstein ambiguity set, defined as the set of distributions that are close to the empirical distribution $\hat{\mathbb{P}}_N$ with respect to the Wasserstein distance d_W:

$$\mathcal{P} = \left\{\mathbb{Q}\in\mathcal{M}(\Xi) : d_W(\mathbb{Q},\hat{\mathbb{P}}_N)\leq\varepsilon\right\}. \tag{3}$$

Here, $\mathcal{M}(\Xi)$ denotes the set of all distributions supported on Ξ, and $\varepsilon\geq 0$ is the radius of the Wasserstein ball. Given any underlying metric $d(\cdot,\cdot)$ on the support set Ξ, the Wasserstein distance $d_W(\mathbb{P},\mathbb{P}')$ between two distributions $\mathbb{P},\mathbb{P}'$ can be defined as follows:

$$d_W(\mathbb{P},\mathbb{P}') = \min_{\Pi\in\mathcal{M}(\Xi\times\Xi)} \left\{ \sum_{\boldsymbol{\xi}\in\Xi}\sum_{\boldsymbol{\xi}'\in\Xi} d(\boldsymbol{\xi},\boldsymbol{\xi}')\Pi(\boldsymbol{\xi},\boldsymbol{\xi}') :\ \Pi \text{ is a coupling of } \mathbb{P} \text{ and } \mathbb{P}' \right\},$$

We motivate the choice of Wasserstein ambiguity sets in Section 2.1. The crucial role of the metric $d(\cdot,\cdot)$ and the radius ε of the Wasserstein ball $\mathcal{P}$ is discussed in Section 2.2

2.1 Advantages of Wasserstein ambiguity sets for discrete rare events

One can construct several ambiguity sets of distributions that are consistent with observed data. Broadly, these are either sets of distributions that satisfy constraints on their moments [25, 66] or those defined as balls centered on a reference distribution with respect to a metric such as the ϕ-divergence [7, 9] or the Wasserstein distance [30, 50, 67, 72]. Note that the high dimensionality and sparsity of the training data in the case of rare events prevent us from obtaining reliable estimates of moments, ruling out moment-based sets. In contrast, metric-based sets have the attractive feature of tying directly with available data; indeed, the empirical distribution corresponding to the training data is a natural choice for the reference distribution.

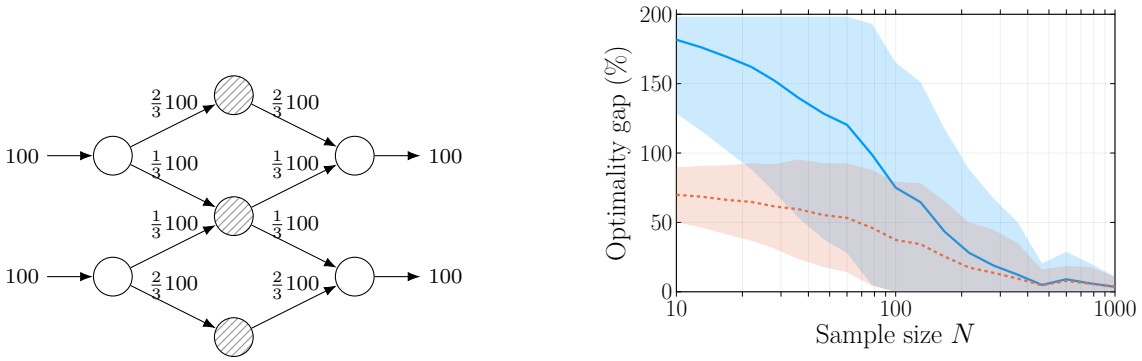
Ambiguity sets based on ϕ-divergence, especially Kullback-Liebler divergence, have certain shortcomings that are not shared by their Wasserstein counterparts. First, the former can exclude the true distribution while including pathological distributions; in fact, they can fail to represent confidence sets for the unknown true distribution [30]. In contrast, the latter contains the true

distribution with high confidence for an appropriate choice of ε, and hence the optimal value of the corresponding distributionally robust problem provides an upper confidence bound on the true out-of-sample cost [50]. Second, Kullback-Liebler and other ϕ-divergence based sets contain only those distributions that are absolutely continuous with respect to reference distribution; that is, these distributions can assign positive probability only to those realizations for which the empirical distribution also assigns positive mass [7]. This situation is undesirable for applications where uncertain events are rare, since the majority of possible uncertain states are unlikely to have been observed empirically. In contrast, the Wasserstein ball of any positive radius $\varepsilon > 0$ and corresponding to any finite-valued metric $d(\cdot, \cdot)$ includes distributions that assign nonzero probability to any arbitrary realization. Indeed, this property ensures that solutions of the distributionally robust problem (2) are robust to the occurrence of rare events. We illustrate this via Example 1.

Example 1 (continued)**.** *Consider again the network in Figure 1. In addition to the sample average approximation, Figure 3b now compares the performance of solutions computed by using the distributionally robust formulation (2) with a Wasserstein ambiguity set $\mathcal{P}$ of radius $\varepsilon = 10^{-3}$, centered around the empirical distribution $\hat{\mathbb{P}}_N$ resulting from different sample sizes N. Here, we use the metric $d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_1$ induced by the 1-norm. The figure shows the difference in total expected costs (computed under the true distribution) of the distributionally robust and sample average solutions with respect to the true optimal solution. We observe that the distributionally robust solution strongly outperforms the sample average solution while being more stable to changes in the training data (i.e., its performance has a smaller variance for a fixed N). The performance of the solutions computed using a ϕ-divergence ambiguity set are presented in the next subsection.*

2.2 Choice of the underlying metric and radius of the Wasserstein ball

The underlying metric $d(\cdot, \cdot)$ should ideally have the following properties: *(i)* if $d(\boldsymbol{\xi}', \boldsymbol{\xi}'')$ is small, then the probabilities of occurrence of the two realizations $\boldsymbol{\xi}', \boldsymbol{\xi}'' \in \Xi$ should be similar; and *(ii)* if $\boldsymbol{\xi}''$ is rarer than $\boldsymbol{\xi}'$, then for some fixed (say nominal) realization $\boldsymbol{\xi} \in \Xi$ (e.g., in a network, this could be the realization where none of the elements fail), we must have $d(\boldsymbol{\xi}, \boldsymbol{\xi}') \leq d(\boldsymbol{\xi}, \boldsymbol{\xi}'')$. These properties can be satisfied whenever $d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|$ is induced by a norm on $\mathbb{R}^M$ and by using an appropriate bit representation of the sample space. Throughout the paper, we will therefore assume that the metric $d(\cdot, \cdot)$ is induced by an arbitrary norm, although our results also apply when



(a) The optimal arc capacities determined by the distributionally robust formulation when the empirical distribution $\hat{\mathbb{P}}_N$ puts all its weight on $\boldsymbol{\xi} = \mathbf{0}$, where none of the nodes fail (see Appendix A).

(b) Comparison of the optimality gap of the distributionally robust formulation (dashed orange) with the sample average approximation (solid blue) for increasing sample size.

Figure 3. Performance of the distributionally robust formulation with a Wasserstein ambiguity set of radius $\varepsilon = 10^{-3}$ on the network flow instance from Figure 1.

$d(\cdot, \cdot)$ is any mixed-integer conic-programming-representable metric.

A noteworthy example of a metric that *does not* satisfy the above requirement is the *discrete metric*, defined as $d(\boldsymbol{\xi}', \boldsymbol{\xi}'') = 1$ whenever $\boldsymbol{\xi}' \neq \boldsymbol{\xi}''$ and 0 otherwise. In this case, the Wasserstein distance is equivalent to the *total variation distance*, and the distributions that solve the inner supremum in (2) can assign positive probability only to the training samples and the worst-case realization [36, 56]. In a network with rare failures, this means that only past observed realizations and the realization corresponding to complete network failure are taken into account, resulting in poor out-of-sample performance.

Example 1 (continued)**.** *Suppose that we define the metric $d(\cdot, \cdot)$ to be the discrete metric. Then, the performance of solutions of the distributionally robust problem (2) with a total variation ambiguity set $\mathcal{P}$ of* any *radius $\varepsilon \geq 0$ is equal to or worse than that of the sample average solution.*

For a given choice of the underlying metric, the radius ε of the ambiguity set allows us to control the level of conservatism of solutions of (2). Specifically, given a confidence level $\beta \in (0, 1)$, one can choose the radius as a function of β and the number of observations N such that the true distribution $\mathbb{P}$ is contained in the ambiguity set with high probability:

$$\mathbb{P}^N \left[d_W(\mathbb{P}, \hat{\mathbb{P}}_N) \leq \varepsilon_N(\beta) \right] \geq 1 - \beta. \tag{4}$$

Here, $\mathbb{P}^N$ is the product distribution that governs the observed data $\{\hat{\boldsymbol{\xi}}^{(1)}, \ldots, \hat{\boldsymbol{\xi}}^{(N)}\}$. It was shown in [50] that (4) holds if we select $\varepsilon_N(\beta) = c_0 \left(N^{-1} \log \beta^{-1}\right)^{1/M}$, where c_0 is a problem-dependent constant. Since this choice can lead to unnecessarily large values for the radius, the authors suggest solving the two-stage problem (2) for several *a priori* fixed choices of ε and then using cross-validation (e.g., k-fold cross-validation) to pick a good value. In the following, we show that Sanov's theorem can be used to obtain stronger finite sample guarantees; that is, less conservative values for $\varepsilon_N(\beta)$, by exploiting the finiteness of the support Ξ.

Theorem 1 (Finite sample guarantee)**.** *For every fixed sample size $N > 0$, and confidence level $\beta \in (0, 1)$, the probabilistic guarantee (4) holds whenever*

$$\varepsilon_N(\beta) \geq D\sqrt{(2N)^{-1}\left(|\Xi| \log(N+1) + \log \beta^{-1}\right)}, \tag{5}$$

where $D := \max_{\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ is the diameter of Ξ with respect to the metric d.

Proof. See Appendix G. □

The right-hand side of (5) indicates that for a fixed choice of the support Ξ and confidence level β, $\varepsilon_N(\beta)$ is roughly proportional to $\sqrt{N^{-1} \log(N+1)}$. For such choices, the optimal value of the corresponding distributionally robust two-stage problem (2) can be expected to provide an upper bound on the true (unknown) out-of-sample cost. We empirically verify this upper bound in Section 5, when we vary $\varepsilon = \nu\sqrt{N^{-1} \log(N+1)}$ as a function of a scalar ν. Theorem 1 also indicates that $\varepsilon_N(\beta) \to 0$ as the sample size becomes large, $N \to \infty$, and that $\varepsilon_N(\beta) \to \infty$ if we are overly conservative, $\beta \to 0$. This observation also elucidates that the distributionally robust formulation (2) generalizes both classical stochastic and robust optimization.

Remark 1 (Reduction to two-stage stochastic and robust optimization)**.** *The distributionally robust two-stage problem (2) reduces to the classical sample average approximation whenever the radius of the ambiguity set $\varepsilon = 0$, since $\mathcal{P} = \{\hat{\mathbb{P}}_N\}$ reduces to a singleton in this case. Similarly, it reduces to a classical two-stage robust optimization problem whenever $\varepsilon \geq \max_{\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ since Ξ is compact and $\mathcal{P}$ contains all Dirac distributions, $\delta_{\boldsymbol{\xi}}$, $\boldsymbol{\xi} \in \Xi$, in this case. Therefore, the worst-case expectation in (2) reduces to $\max_{\boldsymbol{\xi} \in \Xi} \mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$.*

2.3 Contributions

This paper addresses the relatively unexplored topic of rare high-impact uncertainties through the lens of data-driven distributionally robust optimization. Existing methods for addressing rare high-impact uncertainties [6, 17] are few, and they are all based on variants of Monte Carlo methods (e.g., importance sampling), which require the existence of a probability distribution that can be sampled to generate additional observations.

In contrast, our techniques fall within the scope of distributionally robust optimization [57, 58]. A surge in the popularity of Wasserstein ambiguity sets has occurred in this area because of recent results [11, 14, 30, 40, 50] that have established *(i)* strong finite sample and asymptotic guarantees of their formulations, *(ii)* connections with function regularization in machine learning, and *(iii)* tractable reformulations for various classes of loss functions $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ and metric spaces (Ξ, d). Most of the tractability results are for one-stage problems, however, with piecewise linear loss functions and continuous random parameters. Existing tractability results for two-stage problems are limited to the case where $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is the optimal value of a linear program, Ξ is a polyhedron, and d is induced by the 1-norm, see [35, 50]. Sufficient conditions for zero-one supports Ξ and type-∞ Wasserstein ambiguity sets have been established in [68]. In the absence of sufficient conditions that ensure tractability, one resorts either to iterative global optimization methods (e.g., see [46, 67, 71]) or tractable approximations. The latter commonly include discretization schemes (of which sample average approximation is a special case) [20, 45] and decision rule methods [8, 12, 32].

In this paper, we extend the state of the art in data-driven optimization by studying two-stage conic programs with a particular focus on high-dimensional zero-one uncertainties. This is crucial because existing reformulations and algorithms for distributionally robust optimization with finitely supported distributions [5, 9, 11, 53] scale with the size of the support set $|\Xi|$, which can grow exponentially large in such cases. We circumvent this exponential growth by utilizing tractable conservative approximations inspired by lift-and-project convexification techniques in global optimization [41, 43, 60].

Closest in spirit to our work are the papers of [2, 35, 69] who consider the case where the second-stage loss $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is the optimal value of a linear program with uncertain right-hand sides and the support set Ξ is a polytope. In this setting, [35, 69] reformulate (2) as a copositive cone program and then approximate this using semidefinite programming, whereas [2] provide approximations

by leveraging reformulation-linearization techniques from bilinear programming. Although some extensions of these approaches to the case of zero-one support sets Ξ have been made [37, 49], problems where $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is the optimal value of a conic program have not been addressed. This is partly because such extensions lead to so-called *generalized copsitive programs* or *set-semidefinite programs* (e.g., see [18, 28]), and relatively little is known about their tractable approximations. In contrast, generalizations of lift-and-project techniques to mixed zero-one conic problems are fairly well known (e.g., see [21, 61]), and we exploit these to derive tractable approximations for distributionally robust optimization. The relationship between convexification hierarchies based on copositive programming and lift-and-project techniques (for specific problem classes) has been explored in [19, 54].

We highlight the following main contributions:

1. By exploiting ideas from penalty methods and bilinear programming, we develop reformulations of the Wasserstein distributionally robust two-stage problem (2) that reduce its solution to optimization problems over the convex hulls of mixed-integer conic representable sets. Extensions to conditional value-at-risk are also presented.

2. We provide sufficient conditions for our convex hull reformulations and hence the distributionally robust two-stage problem (2) to be tractable. We also show that they are generically NP-hard, however, even if there are no first-stage decisions and the second-stage loss function is the optimal value of a two-dimensional linear program with uncertain objective coefficients.

3. By using lift-and-project hierarchies to approximate the convex hull of the mixed-integer conic representable sets, we derive tractable conservative approximations of the distributionally robust two-stage problem (2), and we provide practical guidelines to compute them efficiently. The approximations are tractable irrespective of the aforementioned sufficient conditions, and they become exact as the Wasserstein radius ε shrinks to zero.

4. We demonstrate the practical viability of our method and its out-of-sample performance on challenging nonlinear optimal power flow and multi-commodity network design problems that are affected by rare network contingencies, and we study its behavior as a function of the rarity and impact of these contingencies, illustrating improvements over classical sample average and two-stage robust optimization formulations.

The rest of the paper is organized as follows. Section 3 derives the mixed-integer conic programming representation of interest, Section 4 derives their lift-and-project approximations, and Section 5 reports numerical results. For ease of exposition, the complexity analysis as well as proofs of all assertions is deferred to the appendix.

Notation. Vectors and matrices are printed in bold lowercase and bold uppercase letters, respectively, while scalars are printed in regular font. The set of non-negative integers and reals is denoted by $\mathbb{Z}_+$ and $\mathbb{R}_+$, respectively. For any positive integer N, we define $[N]$ as the index set $\{1,\ldots,N\}$. We use $\mathbf{e}_k$ to denote the k^{th} unit basis vector, $\mathbf{e}$ to denote the vector of ones, $\mathbf{I}$ to denote the identity matrix, and $\mathbf{0}$ to denote the vector or matrix of zeros, respectively; their dimensions will be clear from the context. For a matrix $\boldsymbol{A}$, we use $\mathrm{vec}(\boldsymbol{A})$ to denote the vector obtained by stacking the columns of $\boldsymbol{A}$ in order. The inner product between two matrices $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{m\times n}$ is denoted by $\langle \boldsymbol{A},\boldsymbol{B}\rangle \coloneqq \sum_{i\in[m]}\sum_{j\in[n]} A_{ij}B_{ij}$. We use $\mathcal{C}^n = \left\{(\boldsymbol{x},t)\in\mathbb{R}^{n-1}\times\mathbb{R} : \|\boldsymbol{x}\|\leq t\right\}$ to denote the norm cone associated with the norm $\|\cdot\|$. For a logical expression $\mathcal{E}$, we define $\mathbb{I}[\mathcal{E}]$ as the indicator function which takes a value of 1 if $\mathcal{E}$ is true and 0 otherwise. Throughout the paper, we refer to an optimization problem as *tractable* if it can be solved in polynomial time in the size of its input data, and *intractable* if it is NP-hard.

3 Mixed-Integer Conic Representations

Throughout the paper, we assume that the distributionally robust two-stage problem (2) satisfies the assumptions of *complete* and *sufficiently expensive recourse*.

(A1) For every realization $\boldsymbol{\xi}\in\Xi$, there exists $\boldsymbol{y}^+\in\mathrm{int}(\mathcal{Y})$ such that $\boldsymbol{W}(\boldsymbol{\xi})\boldsymbol{y}^+ > 0$.

(A2) For every first-stage decision $\boldsymbol{x}\in\mathcal{X}$ and every realization $\boldsymbol{\xi}\in\Xi$, the second-stage loss function $\mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi})$ is bounded.

A natural way to ensure these assumptions is to add slack variables in the formulation of the second-stage problem $\mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi})$ and penalize them in the objective function. Whenever the assumptions are satisfied, they imply that *(i)* $\mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi})$ is always strictly feasible and bounded, *(ii)* the dual of $\mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi})$, given in the following, is always feasible, and *(iii)* strong conic duality holds between the

second-stage problem and its dual, $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) = \mathcal{Q}_d(\boldsymbol{x}, \boldsymbol{\xi})$, where

$$\mathcal{Q}_d(\boldsymbol{x}, \boldsymbol{\xi}) := \sup_{\boldsymbol{\lambda} \in \mathbb{R}^L_+} \left\{ [\boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi} + \boldsymbol{h}(\boldsymbol{x})]^\top \boldsymbol{\lambda} : \boldsymbol{q}(\boldsymbol{\xi}) - \boldsymbol{W}(\boldsymbol{\xi})^\top \boldsymbol{\lambda} \in \mathcal{Y}^* \right\}. \tag{6}$$

Here, $\mathcal{Y}^*$ denotes the dual cone of $\mathcal{Y}$. We assume that the uncertain vectors and matrices in (1) are affine and can be represented as $\boldsymbol{q}(\boldsymbol{\xi}) = \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi}$ and $\boldsymbol{W}(\boldsymbol{\xi}) = \boldsymbol{W}_0 + \sum_{j \in [M]} \xi_j \boldsymbol{W}_j$, where $\boldsymbol{Q} \in \mathbb{R}^{N_2 \times M}$, and for each $j \in \{0, 1, \ldots, M\}$, we have $\boldsymbol{W}_j \in \mathbb{R}^{L \times N_2}$.

A consequence of the above assumptions is the following lemma, which states that computing the worst-case expectation in the two-stage problem (2) is equivalent to averaging N worst-case values of the loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ over $\boldsymbol{\xi} \in \Xi$, each regularized by one of the training samples. We omit the proof since it follows directly from the compactness of Ξ and the definition of the Wasserstein ambiguity set $\mathcal{P}$ in (3); see [15, 30] for proofs in much more general settings.

Lemma 1. *The distributionally robust two-stage problem (2) admits the following reformulation:*

$$\underset{\boldsymbol{x} \in \mathcal{X}, \alpha \geq 0}{\text{minimize}} \; c(\boldsymbol{x}) + \alpha\varepsilon + \frac{1}{N} \sum_{i=1}^{N} \max_{\boldsymbol{\xi} \in \Xi} \left\{ \mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) - \alpha d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}^{(i)}) \right\}. \tag{7}$$

The remainder of this section establishes that the inner maximization in (7) is equivalent to optimizing a linear function over the convex hull of the feasible region of a *mixed-integer conic program* (MICP). The following theorem is key to establishing this result.

Theorem 2 (Convex hull reformulation)**.** *The distributionally robust two-stage problem (2) admits the following convex hull reformulation:*

$$\underset{\boldsymbol{x} \in \mathcal{X}, \alpha \geq 0}{\text{minimize}} \; c(\boldsymbol{x}) + \alpha\varepsilon + \frac{1}{N} \sum_{i=1}^{N} Z_i(\boldsymbol{x}, \alpha), \tag{8}$$

where, for each $i \in [N]$, we define the function $Z_i : \mathcal{X} \times \mathbb{R}_+ \mapsto \mathbb{R}$ and the set $\mathcal{Z}_i$ as follows:

$$Z_i(\boldsymbol{x}, \alpha) = \underset{(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \tau) \in \operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)}{\text{maximize}} \left\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda} \rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \alpha\tau \right\} \tag{9a}$$

$$\mathcal{Z}_i = \left\{ (\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \tau) \in \Xi \times \mathbb{R}^L_+ \times \mathbb{R}^{L \times M} \times \mathbb{R}_+ : \begin{array}{l} \boldsymbol{\Lambda} = \boldsymbol{\lambda}\boldsymbol{\xi}^\top, \; (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}, \tau) \in \mathcal{C}^{M+1} \\ \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{W}_0^\top \boldsymbol{\lambda} - \sum_{j \in [M]} \boldsymbol{W}_j^\top \boldsymbol{\Lambda} \mathbf{e}_j \in \mathcal{Y}^* \end{array} \right\}. \tag{9b}$$

Proof. See Appendix G. □

The inner optimization problem (9a) is over the closed convex hull of the set $\mathcal{Z}_i$, which couples the binary uncertain parameters $\boldsymbol{\xi}$ with the continuous dual variables $\boldsymbol{\lambda}$ via the bilinear equation $\boldsymbol{\Lambda} = \boldsymbol{\lambda}\boldsymbol{\xi}^\top$. This set is, therefore, not the feasible region of an MICP. We propose two approaches to ensure MICP representability. The first is to linearize the bilinear equation $\boldsymbol{\Lambda} = \boldsymbol{\lambda}\boldsymbol{\xi}^\top$ using McCormick inequalities, which requires *a priori* upper bounds on the dual variables $\boldsymbol{\lambda}$; we briefly discuss this in Section 3.1. The second is to reformulate the loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ using ideas from penalty methods that circumvents any bilinear terms; this approach is the subject of Section 3.2. We compare the two approaches and summarize their merits in Section 3.3. We note that our results also apply if the risk-neutral expectation in the objective function of the two-stage problem (2) is replaced with the conditional value-at-risk; for ease of exposition, we defer this analysis to the appendix.

3.1 Linearized reformulation

The decision variables $\boldsymbol{\lambda}$ of the dual problem $\mathcal{Q}_d(\boldsymbol{x}, \boldsymbol{\xi})$ must be necessarily bounded for any fixed value of $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Xi$. Indeed, under assumptions (A1) and (A2), the value of $\mathcal{Q}_d(\boldsymbol{x}, \boldsymbol{\xi})$ is bounded for any fixed $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Xi$; and since $\mathcal{X}$ and Ξ are compact sets, the variables $\boldsymbol{\lambda}$ must also be necessarily bounded. Suppose that $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^L_+$ are *a priori* known upper bounds (independent of $\boldsymbol{x}$ and $\boldsymbol{\xi}$) on these variables. Such bounds may be analytically known whenever we explicitly add slack variables to ensure feasibility of the second-stage problem or if the latter has some structure (e.g., see [29]). Whenever such bounds are known, we can exactly linearize the bilinear equation $\boldsymbol{\Lambda} = \boldsymbol{\lambda}\boldsymbol{\xi}^\top$ using McCormick inequalities since $\boldsymbol{\xi}$ is binary-valued (e.g., see [33]), and reformulate the set $\mathcal{Z}_i$ in (9b) as the feasible region of an MICP:

$$\mathcal{Z}_i = \left\{ (\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \tau) \in \Xi \times \mathbb{R}^L_+ \times \mathbb{R}^{L\times M}_+ \times \mathbb{R}_+ : \begin{array}{l} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}, \tau) \in \mathcal{C}^{M+1} \\ \boldsymbol{\Lambda} - \boldsymbol{\lambda}\mathbf{e}^\top + \bar{\boldsymbol{\lambda}}(\mathbf{e} - \boldsymbol{\xi})^\top \in \mathbb{R}^{L\times M}_+ \\ \boldsymbol{\lambda}\mathbf{e}^\top - \boldsymbol{\Lambda} \in \mathbb{R}^{L\times M}_+,\ \bar{\boldsymbol{\lambda}}\boldsymbol{\xi}^\top - \boldsymbol{\Lambda} \in \mathbb{R}^{L\times M}_+ \\ \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{W}_0^\top \boldsymbol{\lambda} - \displaystyle\sum_{j\in[M]} \boldsymbol{W}_j^\top \boldsymbol{\Lambda}\mathbf{e}_j \in \mathcal{Y}^* \end{array} \right\}. \quad \text{(9b-}\ell\text{)}$$

The MICP representation (9b-ℓ) adds $O(ML)$ variables and constraints for each $\mathcal{Z}_i$, $i \in [N]$, which can be prohibitively large. Moreover, analytical upper bounds $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^L_+$ on the dual variables, which are independent of (and valid for all) $\boldsymbol{x}$ and $\boldsymbol{\xi}$, may be unavailable. The following section shows

that we can ensure MICP representability without *a priori* knowledge of these bounds, and at the expense of adding far fewer variables and constraints.

3.2 Penalty reformulation

Our goal in this section will be to move the uncertainty to the objective function of the second-stage problem, $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$. This approach is motivated by the following corollary to Theorem 2 that shows that MICP-representability is guaranteed if only the objective function of $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is uncertain.

Corollary 1 (Convex hull reformulation for objective uncertainty)**.** *Suppose that only the objective function of the second-stage problem $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is uncertain: $\boldsymbol{W}(\boldsymbol{\xi}) = \boldsymbol{W}_0$ and $\boldsymbol{T}(\boldsymbol{x}) = \mathbf{0}$. Then the distributionally robust two-stage problem (2) admits reformulation (8) where, for each $i \in [N]$, we define the function $Z_i : \mathcal{X} \times \mathbb{R}_+ \mapsto \mathbb{R}$ and the MICP-representable set $\mathcal{Z}_i$ as follows:*

$$Z_i(\boldsymbol{x}, \alpha) = \underset{(\boldsymbol{\xi}, \boldsymbol{\lambda}, \tau) \in \operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)}{\text{maximize}} \left\{ \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \alpha\tau \right\} \tag{10a}$$

$$\mathcal{Z}_i = \left\{ (\boldsymbol{\xi}, \boldsymbol{\lambda}, \tau) \in \Xi \times \mathbb{R}^L_+ \times \mathbb{R}_+ : (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}, \tau) \in \mathcal{C}^{M+1},\ \ \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{W}_0^\top \boldsymbol{\lambda} \in \mathcal{Y}^* \right\}. \tag{10b}$$

For the remainder of the paper, we shall make the following additional assumption of *fixed recourse* which will be key to achieving our goal.

(A3) For every realization $\boldsymbol{\xi} \in \Xi$ and every first-stage decision $\boldsymbol{x} \in \mathcal{X}$, we have $\boldsymbol{W}(\boldsymbol{\xi}) = \boldsymbol{W}_0$ and $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}_0$, respectively.

We note that since $\boldsymbol{\xi}$ is binary valued, this assumption is without loss of generality if the primal decision variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are bounded. Indeed, in such cases, we can exactly linearize any products of uncertain parameters and decisions in the second-stage problem $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ using McCormick inequalities [33], to ensure that it satisfies the assumption of fixed recourse. Under this additional assumption, the following theorem states that the second-stage problem $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ with constraint uncertainty can be equivalently reformulated as one with objective uncertainty.

Theorem 3 (Penalty reformulation of the loss function)**.** *There exists a sufficiently large, yet* finite*, penalty parameter $\rho > 0$ such that the second-stage loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ in (1) is equivalent to*

$$\mathcal{Q}^\rho(\boldsymbol{x}, \boldsymbol{\xi}) := \inf_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in [0,1]^M} \left\{ \boldsymbol{q}(\boldsymbol{\xi})^\top \boldsymbol{y} + \rho \left((\mathbf{e} - 2\boldsymbol{\xi})^\top \boldsymbol{z} + \mathbf{e}^\top \boldsymbol{\xi} \right) : \boldsymbol{W}_0 \boldsymbol{y} \geq \boldsymbol{T}_0 \boldsymbol{z} + \boldsymbol{h}(\boldsymbol{x}) \right\}. \tag{11}$$

Proof. See Appendix G. □

In conjunction with Corollary 1, Theorem 3 implies that the distributionally robust two-stage problem (2) admits a convex hull reformulation of the form (8), where the function $Z_i : \mathcal{X} \times \mathbb{R}_+ \mapsto \mathbb{R}$ and the MICP-representable set $\mathcal{Z}_i$ for each $i \in [N]$ are given as follows:

$$Z_i(\boldsymbol{x}, \alpha) = \underset{(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \tau) \in \operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)}{\text{maximize}} \left\{ \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} + \rho(\mathbf{e}^\top \boldsymbol{\xi}) - \mathbf{e}^\top \boldsymbol{\mu} - \alpha\tau \right\} \tag{9a-ρ}$$

$$\mathcal{Z}_i = \left\{ (\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \tau) \in \Xi \times \mathbb{R}^L_+ \times \mathbb{R}^M_+ \times \mathbb{R}_+ : \begin{array}{l} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}, \tau) \in \mathcal{C}^{M+1} \\ \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{W}_0^\top \boldsymbol{\lambda} \in \mathcal{Y}^* \\ \rho(\mathbf{e} - 2\boldsymbol{\xi}) + \boldsymbol{T}_0^\top \boldsymbol{\lambda} + \boldsymbol{\mu} \in \mathbb{R}^M_+ \end{array} \right\}. \tag{9b-ρ}$$

In contrast to the linearized reformulation (9b-ℓ), the MICP representation (9b-ρ) adds only $O(M+L)$ variables and constraints for each $\mathcal{Z}_i$, $i \in [N]$.

The reformulation (11) is reminiscent of penalty methods in nonlinear programming. However, in contrast to the latter, which suffer from numerical issues because the penalty parameter ρ must be driven to ∞, a finite value for ρ can be precomputed in our case. This process only requires solving the classical robust optimization formulation over any support $\Xi^0 \supseteq \Xi$; for example, some choices are $\Xi^0 = \Xi$ or $\Xi^0 = \{0,1\}^M$. The procedure is as follows. We compute the classical robust solution $\boldsymbol{x}^r$ and a corresponding worst-case realization, as per (12a) shown below, and then set ρ^r to be an optimal Lagrange multiplier of the last inequality in (12b).

$$\boldsymbol{x}^r \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \left\{ c(\boldsymbol{x}) + \max_{\boldsymbol{\xi} \in \Xi^0} \mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) \right\}, \ \boldsymbol{\xi}^r \in \arg\max_{\boldsymbol{\xi} \in \Xi^0} \ \mathcal{Q}(\boldsymbol{x}^r, \boldsymbol{\xi}), \tag{12a}$$

$$\inf_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in [0,1]^M} \left\{ \boldsymbol{q}(\boldsymbol{\xi}^r)^\top \boldsymbol{y} : \boldsymbol{W}_0 \boldsymbol{y} \geq \boldsymbol{T}_0 \boldsymbol{z} + \boldsymbol{h}(\boldsymbol{x}^r), \ (\mathbf{e} - 2\boldsymbol{\xi}^r)^\top \boldsymbol{z} + \mathbf{e}^\top \boldsymbol{\xi}^r \leq 0 \right\}. \tag{12b}$$

We next prove that our procedure is valid: the optimal Lagrange multiplier ρ^r is indeed an *exact value* for the penalty parameter.

Theorem 4 (Exact penalty reformulation)**.** *The optimal value of the distributionally robust two-stage problem* (2) *remains unchanged if we replace the second-stage loss function* $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ *with its penalty reformulation* $\mathcal{Q}^{\rho^r}(\boldsymbol{x}, \boldsymbol{\xi})$ *as defined in* (11) *with penalty parameter* ρ^r*.*

Proof. See Appendix G. □

The classical robust formulation (12a) is tractable if we choose $\Xi^0 = \{0,1\}^M$ and the loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ exhibits a down-monotone (or up-monotone) property with respect to the random parameters $\boldsymbol{\xi}$; that is, $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}') \geq \mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ whenever $\boldsymbol{\xi}' \geq \boldsymbol{\xi}$. In particular, this is the case for network

optimization problems where removing network elements is never advantageous. In such cases, the classical robust formulation reduces to a deterministic problem under the worst-case realization of the uncertain parameters. Moreover, this worst-case realization is often independent of the optimal robust solution $\boldsymbol{x}^r$. For example, suppose $\boldsymbol{\xi} \in \{0,1\}^M$ is a random vector denoting which of M arcs in a network have been "disrupted" and arc-flow variables $\boldsymbol{y}$ satisfy $0 \leq y_a \leq \xi_a \bar{y}_a$ (i.e., the flow on arc a is zero whenever it is disrupted $\xi_a = 0$, and is bounded between 0 and $\bar{y}_a$ otherwise). The worst-case realization is to disrupt all arcs in the network, $\boldsymbol{\xi}^r = \mathbf{0}$, independent of the optimal robust solution $\boldsymbol{x}^r$. Notably, this also implies that we can circumvent the computation of (12a) when determining the value of ρ^r. We generalize this observation in the following proposition.

Proposition 1. *Assume without loss of generality that $\mathcal{Y} \subseteq \mathbb{R}^{N_2}_+$. Suppose also that for all $j \in [M]$, we have either $\boldsymbol{T}_0\mathbf{e}_j, \boldsymbol{Q}\mathbf{e}_j \in \mathbb{R}^L_+$ or $-\boldsymbol{T}_0\mathbf{e}_j, -\boldsymbol{Q}\mathbf{e}_j \in \mathbb{R}^L_+$. Then, the classical robust optimization problem $\min_{\boldsymbol{x}\in\mathcal{X}} \left\{c(\boldsymbol{x}) + \max_{\boldsymbol{\xi}\in\{0,1\}^M} \mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi})\right\}$ reduces to a deterministic problem $\min_{\boldsymbol{x}\in\mathcal{X}} \{c(\boldsymbol{x}) + \mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi}^r)\}$, where for each $j \in [M]$ we have $\xi^r_j = 1$ if $\boldsymbol{T}_0\mathbf{e}_j, \boldsymbol{Q}\mathbf{e}_j \in \mathbb{R}^L_+$ and $\xi^r_j = 0$ otherwise.*

Proof. See Appendix G. □

3.2.1 Penalty reformulation for indicator constraints

The penalty reformulation $\mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi})$ in (11) introduces M additional variables $\boldsymbol{z} \in [0,1]^M$ in the second-stage loss function. This can be avoided by exploiting a structure that is common in network optimization problems. In several such applications, an uncertain parameter $\xi_j \in \{0,1\}$ may switch on and off a single constraint $f_j(\boldsymbol{y}) \geq 0$, leading to the following definition of the loss function:

$$\mathcal{Q}_{\text{ind}}(\boldsymbol{x},\boldsymbol{\xi}) = \inf_{\boldsymbol{y}\in\mathcal{Y}} \left\{ \boldsymbol{q}(\boldsymbol{\xi})^\top \boldsymbol{y} : \begin{array}{l} \boldsymbol{W}_0\boldsymbol{y} \geq \boldsymbol{h}(\boldsymbol{x}) \\ \xi_j = 1 \implies [f_j(\boldsymbol{y}) = 0]\,,\ j \in [M] \\ \xi_j = 0 \implies [f_j(\boldsymbol{y}) \geq 0]\,,\ j \in [M] \end{array} \right\}, \tag{13}$$

where $f_j : \mathcal{Y} \mapsto \mathbb{R}$ is an affine function for each $j \in [M]$. For example, in a network, $\xi_j = 1$ may indicate that link j has failed and the corresponding flow variable $f_j(\boldsymbol{y}) = y_j$ must be set to 0, whereas $\xi_j = 0$ may indicate that the flow variable y_j can take nonzero values. Such constraints are generally written as $f_j(\boldsymbol{y}) \leq \bar{f}(1-\xi_j)$, where $\bar{f}$ is a big-M upper bound on $f_j(\boldsymbol{y})$. Obtaining tight estimates on $\bar{f}$ may be non-trivial and at the same time, an overestimation of $\bar{f}$ can lead to

numerical issues. In such cases, we can avoid both *(i)* estimating $\bar{f}$, and *(ii)* introducing auxiliary variables $\boldsymbol{z}$ in the penalty reformulation (11), by simply retaining the constraint $f_j(\boldsymbol{y}) \geq 0$ and adding $+\rho\xi_j f(\boldsymbol{y}_j)$ to the objective, leading to a tighter penalty reformulation.

Corollary 2 (Penalty reformulation of the loss function with indicator constraints)**.** *There exists a sufficiently large, yet* finite*, penalty parameter $\rho > 0$ such that the second-stage loss function $\mathcal{Q}_{\text{ind}}(\boldsymbol{x}, \boldsymbol{\xi})$ in (13) is equivalent to*

$$\mathcal{Q}^{\rho}_{\text{ind}}(\boldsymbol{x}, \boldsymbol{\xi}) = \inf_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \boldsymbol{q}(\boldsymbol{\xi})^\top \boldsymbol{y} + \rho \boldsymbol{\xi}^\top \boldsymbol{f}(\boldsymbol{y}) : \boldsymbol{W}_0 \boldsymbol{y} \geq \boldsymbol{h}(\boldsymbol{x}),\ \boldsymbol{f}(\boldsymbol{y}) \geq \boldsymbol{0} \right\}. \tag{14}$$

The penalty term $+\rho\xi_j f(\boldsymbol{y}_j)$ ensures that $f_j(\boldsymbol{y})$ is driven to 0 whenever $\xi_j = 1$ for large values of ρ. Note that we no longer need to estimate the big-M upper bounds $\bar{f}$ since the constraint $f_j(\boldsymbol{y}) \leq \bar{f}(1 - \xi_j)$ is not required anymore.

Our previous results continue to be valid as long as we replace each occurrence of the penalty term $(\mathbf{e} - 2\boldsymbol{\xi})^\top \boldsymbol{z} + \mathbf{e}^\top \boldsymbol{\xi}$ that multiples ρ in the objective function of (11) with this modification. For example, suppose that $\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{f}_0 + \boldsymbol{F}\boldsymbol{y}$. Then, Corollaries 1 and 2 imply that the distributionally robust two-stage problem (2) admits a convex hull reformulation of the form (8), where the function $Z_i : \mathcal{X} \times \mathbb{R}_+ \mapsto \mathbb{R}$ and the MICP-representable set $\mathcal{Z}_i$ for each $i \in [N]$ are given as follows:

$$Z_i(\boldsymbol{x}, \alpha) = \underset{(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \tau) \in \operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)}{\text{maximize}} \left\{ \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} + \rho \boldsymbol{f}_0^\top \boldsymbol{\xi} - \boldsymbol{f}_0^\top \boldsymbol{\mu} - \alpha\tau \right\}$$

$$\mathcal{Z}_i = \left\{ (\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \tau) \in \Xi \times \mathbb{R}^L_+ \times \mathbb{R}^M_+ \times \mathbb{R}_+ : \begin{array}{l} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}, \tau) \in \mathcal{C}^{M+1} \\ \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi} + \rho \boldsymbol{F}^\top \boldsymbol{\xi} - \boldsymbol{W}_0^\top \boldsymbol{\lambda} - \boldsymbol{F}^\top \boldsymbol{\mu} \in \mathcal{Y}^* \end{array} \right\}.$$

Similarly, a value for the penalty parameter ρ can be computed as follows. First, we compute the classical robust solution $\boldsymbol{x}^r$ and a corresponding worst-case realization by solving the problems:

$$\boldsymbol{x}^r \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \left\{ c(\boldsymbol{x}) + \max_{\boldsymbol{\xi} \in \Xi^0} \mathcal{Q}_{\text{ind}}(\boldsymbol{x}, \boldsymbol{\xi}) \right\},\ \boldsymbol{\xi}^r \in \arg\max_{\boldsymbol{\xi} \in \Xi^0} \mathcal{Q}_{\text{ind}}(\boldsymbol{x}^r, \boldsymbol{\xi}).$$

Next, we set ρ^r to be an optimal Lagrange multiplier of the last inequality of the following problem:

$$\inf_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \boldsymbol{q}(\boldsymbol{\xi}^r)^\top \boldsymbol{y} : \boldsymbol{W}_0 \boldsymbol{y} \geq \boldsymbol{h}(\boldsymbol{x}^r),\ \boldsymbol{f}(\boldsymbol{y}) \geq \boldsymbol{0},\ (\boldsymbol{\xi}^r)^\top \boldsymbol{f}(\boldsymbol{y}) \leq 0 \right\}$$

The proof for the validity of $\mathcal{Q}_{\text{ind}} = \mathcal{Q}^{\rho^r}_{\text{ind}}$ is similar to that of Theorem 4 and we omit it for the sake of brevity. Finally, note that one can avoid the computation of $\boldsymbol{x}^r$ and $\boldsymbol{\xi}^r$ (as before) by examining the structure of $\boldsymbol{f}(\boldsymbol{y})$ and $\boldsymbol{q}(\boldsymbol{\xi})$, see Proposition 1.

3.3 Summary and comparison

Table 1 summarizes the main differences between the linearized and penalty-based MICP reformulations of the sets $\mathcal{Z}_i$, $i \in [N]$ appearing in the convex hull reformulation (9a)–(9b). Notably, the penalty reformulation adds far fewer variables and constraints. However, it also requires additional assumptions and computations. In particular, it requires computing a value for the penalty parameter ρ, which may further entail the solution of a classical robust optimization problem; see (12a). We do not expect this to be a limitation, however, because the latter will likely reduce to a deterministic optimization problem for most practical applications.

Table 1. Summary of the MICP representations of $\mathcal{Z}_i$ based on the linearized and penalty reformulations.

Reformulation	Size	$\boldsymbol{W}(\boldsymbol{\xi})$	$\boldsymbol{T}(\boldsymbol{x})$	Requirements
Linearized (9b-ℓ)	$O(ML)$	affine	convex	*a priori* bounds on $\boldsymbol{\lambda}$ in $\mathcal{Q}_d(\boldsymbol{x}, \boldsymbol{\xi})$
Penalty (9b-ρ)	$O(M+L)^*$	constant†	constant†	computation of worst-case realization over any superset of $\Xi^\sharp$

*Can be further reduced; see Corollary 2

†Can be relaxed; see discussion following assumption (A3).

$^\sharp$Can be done in closed form; see Lemma 1 and preceding discussion.

4 Lift-and-Project Approximations

The key challenge in solving the convex hull reformulation (8) is the inner optimization (9a) over the convex hull of the MICP-representable set $\mathcal{Z}_i$, $i \in [N]$. Appendix B shows that although one can tractably compute these convex hulls in several cases, the problem remains NP-hard even in benign settings. Therefore, Appendix C presents a Benders scheme, similar to the ones proposed in [62, 71], to tackle the convex hull constraints. This scheme iteratively refines an *inner approximation* of the MICP representation of $\mathcal{Z}_i$. An alternative to solving the convex hull reformulation (8) is direct solution of the original reformulation (7) using a column-and-constraint generation scheme [70]. In contrast to the Benders scheme, the latter models the second-stage loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ via explicit second-stage variables and constraints by implicitly enumerating $\boldsymbol{\xi} \in \Xi$. In both schemes, however,

each iteration requires the solution of N global MICP problems and therefore, they can become computationally prohibitive. Moreover, intermediate solutions obtained from early termination provide no guarantees whatsoever since they bound the optimal value of the distributionally robust problem (2) from below, which itself is an upper bound on the true (unknown) optimal value.

These observations motivate the development of *tractable outer approximations* of the convex hulls of $\mathcal{Z}_i$, $i \in [N]$, that provide not only *(i)* upper bounds on the optimal value of (2) but also *(ii)* guarantees of polynomial time solvability. Our approximations are based on the following key observations. First, if we suppose that $\Xi = \{\boldsymbol{\xi} \in \mathbb{Z}_+^M : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{f}\}$ has a given outer description, then Section 3 establishes that each of the sets $\mathcal{Z}_i$, $i \in [N]$, can be represented as the feasible region of an MICP as follows:

$$\mathcal{Z} = \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{z} - \boldsymbol{b} \in \mathcal{K}, \;\; z_j \in \{0,1\}\; j \in [M]\}\,, \tag{15}$$

where, for ease of exposition, we have dropped the subscript i and included the bounds, $\boldsymbol{z} \geq \boldsymbol{0}$ and $1 \geq z_j \coloneqq \xi_j$, $j \in [M]$ in $\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b} \in \mathcal{K}$. For example, in case of (9b-ℓ), we have $\boldsymbol{z} = (\boldsymbol{\xi}, \boldsymbol{\lambda}, \mathrm{vec}(\boldsymbol{\Lambda}), \tau)$, $n = M + L + ML + 1$, $\mathcal{K} = \tilde{\mathcal{K}} \times \mathbb{R}^n_+$, where $\tilde{\mathcal{K}} = \mathbb{R}^F_+ \times_{i=1}^3 \mathbb{R}^{LM}_+ \times \mathcal{C}^{M+1} \times \mathcal{Y}^*$ and F is the dimension of $\boldsymbol{f} \in \mathbb{R}^F$, and $\boldsymbol{A} = [\tilde{\boldsymbol{A}}^\top \; \mathbf{I}]^\top$ and $\boldsymbol{b} = [\tilde{\boldsymbol{b}}^\top \; \boldsymbol{0}^\top]^\top$ model the right-hand half of (9b-ℓ). Similarly, in case of (9b-ρ), we have $\boldsymbol{z} = (\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \tau)$, $n = 2M + L + 1$, $\mathcal{K} = \tilde{\mathcal{K}} \times \mathbb{R}^n_+$, where $\tilde{\mathcal{K}} = \mathbb{R}^F_+ \times \mathbb{R}^M_+ \times \mathcal{C}^{M+1} \times \mathcal{Y}^*$, and $\boldsymbol{A} = [\tilde{\boldsymbol{A}}^\top \; \mathbf{I}]^\top$, $\boldsymbol{b} = [\tilde{\boldsymbol{b}}^\top \; \boldsymbol{0}^\top]^\top$ for suitable matrices $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{b}}$.

Second, given an MICP representation such as the above, its convex hull can be approximated by a hierarchy of increasingly tight *convex relaxations*,

$$\mathcal{Z}^0 \supseteq \mathcal{Z}^1 \supseteq \ldots \supseteq \mathcal{Z}^M = \mathrm{cl}\,\mathrm{conv}(\mathcal{Z}),$$

that converge to the convex hull in M iterations. Here, $\mathcal{Z}^0 \coloneqq \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{z} - \boldsymbol{b} \in \mathcal{K}\}$ is the continuous relaxation of $\mathcal{Z}$. Several such *sequential convexification* hierarchies are known, the most popular ones being those of [4, 41, 43, 60]. They are based on the concept of *lift-and-project* and represent $\mathrm{cl}\,\mathrm{conv}(\mathcal{Z})$ as the *projection* of another convex set lying in a higher-dimensional space. These hierarchies were originally proposed for (pure or mixed-) integer linear sets and later extended to mixed-integer convex sets in [21, 61]. Our proposal is to use an intermediate relaxation $\mathcal{Z}^t$ of any such hierarchy to outer approximate $\mathrm{cl}\,\mathrm{conv}(\mathcal{Z})$, which results in an outer approximation of the convex hull reformulation (8) and, hence, a conservative approximation of the distributionally robust two-stage problem (2). Notably, since we can optimize an objective function over the level-t

relaxation in time $n^{O(t)}$ (which is polynomial for fixed t), we can also obtain tight approximations of the original problem (2) in polynomial time. The approximation can be refined, if desired, by using higher values of t.

Third, the approximation of $\operatorname{cl}\operatorname{conv}(\mathcal{Z})$ when used in the convex hull reformulation allows us to dualize the inner optimization in (9a) using conic duality. The result is a single-stage convex conic optimization model that can be solved with off-the-shelf solvers. Notably, we can prove that the resulting approximations of the distributionally robust two-stage problem (2), obtained by replacing $\operatorname{cl}\operatorname{conv}(\mathcal{Z})$ with any of the relaxations $\mathcal{Z}^0, \ldots, \mathcal{Z}^M$, become exact if the radius ε of the Wasserstein ambiguity set $\mathcal{P}$ shrinks to zero with increasing sample size N.

Theorem 5 (Lift-and-project approximation quality)**.** *Suppose that $N\varepsilon_N \to 0$ as $N \to \infty$. Then, the optimal value of the distributionally robust two-stage problem (2) coincides with that of the convex hull reformulation (8) even if we approximate each $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$, $i \in [N]$, with $\mathcal{Z}_i^0$ in (9a).*

Proof. See Appendix G. □

We emphasize that our method is not tied to any particular convexification technique. This feature is important because each technique has its advantages and disadvantages. For example, in the linear case (i.e., $\mathcal{K} = \mathbb{R}^{n'}_+$), it is known [42] that the approximations in order of decreasing tightness are those of [41], [60], [43], and [4]; however, this ranking is reversed when they are ordered based on increasing computational complexity. For its simplicity and tradeoff between tightness and tractability, we focus on the Lovász-Schrijver approximation [43] in the remainder of this section. We show how it can be used to obtain a single-stage approximation of the distributionally robust two-stage problem (2), and we provide practical guidelines for its efficient computation.

4.1 Lovász-Schrijver approximation

The level-1 Lovász-Schrijver approximation $\mathcal{Z}^1$ is defined as a set-valued mapping, and the level-t approximation $\mathcal{Z}^t$ is defined as an iterated application of this mapping. For any $u \in \mathbb{R}_+$ and any conic representable set such as the continuous relaxation $\mathcal{Z}^0 = \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{z} - \boldsymbol{b} \in \mathcal{K}\}$, we denote by $\mathcal{Z}^0(u) = \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}u \in \mathcal{K}\}$ to be the homogenization of $\mathcal{Z}^0$ with respect to u. Next, we

define the following lifted set:

$$\mathcal{L}(\mathcal{Z}^0) = \left\{ \boldsymbol{z}, \{\boldsymbol{z}^{j0}\}_{j\in[M]}, \{\boldsymbol{z}^{j1}\}_{j\in[M]} \in \mathbb{R}^n : \begin{array}{ll} \exists u^{j0}, u^{j1} \geq 0,\ u^{j0} + u^{j1} = 1, & j \in [M] \\ \boldsymbol{z}^{j0} \in \mathcal{Z}^0(u^{j0}),\ \boldsymbol{z}^{j1} \in \mathcal{Z}^0(u^{j1}), & j \in [M] \\ \boldsymbol{z} = \boldsymbol{z}^{j0} + \boldsymbol{z}^{j1}, & j \in [M] \\ z_j^{j0} = 0,\ z_j^{j1} = u^{j1}, & j \in [M] \\ z_k^{j1} = z_j^{k1},\ \ k \in [M] : k > j, & j \in [M] \end{array} \right\}. \tag{16}$$

Consider now the following set-valued map, which is the projection of $\mathcal{L}(\mathcal{Z}^0)$ onto $\mathbb{R}^n$:

$$\mathcal{P}(\mathcal{Z}^0) = \left\{ \boldsymbol{z} \in \mathbb{R}^n : \exists \boldsymbol{z}^{j0}, \boldsymbol{z}^{j1},\ j \in [M] \text{ such that } \left(\boldsymbol{z}, \{\boldsymbol{z}^{j0}\}_{j\in[M]}, \{\boldsymbol{z}^{j1}\}_{j\in[M]}\right) \in \mathcal{L}(\mathcal{Z}^0) \right\}. \tag{17}$$

One can easily verify that $\mathcal{P}(\mathcal{Z}^0)$ is a convex relaxation of $\operatorname{cl}\operatorname{conv}(\mathcal{Z})$. In fact, we have the following relationship [21, Theorem 1]: $\operatorname{cl}\operatorname{conv}(\mathcal{Z}) \subseteq \mathcal{P}(\mathcal{Z}^0) \subseteq \bigcap_{j\in[M]} \operatorname{cl}\operatorname{conv}(\{\boldsymbol{z} \in \mathcal{Z}^0 : z_j \in \{0,1\}\}) \subseteq \mathcal{Z}^0$. The set $\mathcal{P}(\mathcal{Z}^0)$ corresponds to the level-1 relaxations of the Lovász-Schrijver hierarchy. For any $t \geq 1$, the level-t relaxation is given by $\mathcal{Z}^t = \mathcal{P}(\mathcal{Z}^{t-1})$, and one can show that $\mathcal{Z}^M = \operatorname{cl}\operatorname{conv}(\mathcal{Z})$. This is known as the *linear* Lovász-Schrijver hierarchy. One can obtain stronger relaxations by imposing positive semidefiniteness on the submatrix of $\{\boldsymbol{z}^{j1}\}_{j\in[M]}$ corresponding to the binary variables.

Remark 2 (Positive semidefinite Lovász-Schrijver hierarchy)**.** *Consider the following set:*

$$\mathcal{P}_+(\mathcal{Z}^0) = \left\{ \boldsymbol{z} \in \mathbb{R}^n : \begin{array}{l} \exists \boldsymbol{z}^{j0}, \boldsymbol{z}^{j1},\ j \in [M] \text{ such that } \left(\boldsymbol{z}, \{\boldsymbol{z}^{j0}\}_{j\in[M]}, \{\boldsymbol{z}^{j1}\}_{j\in[M]}\right) \in \mathcal{L}(\mathcal{Z}^0) \\ \text{and } \left[\boldsymbol{\xi}^{11} \ldots \boldsymbol{\xi}^{M1}\right] \succeq \boldsymbol{\xi}\boldsymbol{\xi}^\top, \text{where } \boldsymbol{\xi} = [z_1 \ldots z_M]^\top, \boldsymbol{\xi}^{j1} = [z_1^{j1} \ldots z_M^{j1}]^\top \end{array} \right\}.$$

This set corresponds to the level-1 relaxation of the positive semidefinite *Lovász-Schrijver hierarchy and is related to its linear counterpart as follows:* $\operatorname{cl}\operatorname{conv}(\mathcal{Z}) \subseteq \mathcal{P}_+(\mathcal{Z}^0) \subseteq \mathcal{P}(\mathcal{Z}^0)$. *For any* $t \geq 1$*, the level-*t *relaxation in this hierarchy is given by* $\mathcal{Z}^t = \mathcal{P}_+(\mathcal{Z}^{t-1})$. *As before, we have* $\mathcal{Z}^M = \operatorname{cl}\operatorname{conv}(\mathcal{Z})$.

For a given MICP representation of $\mathcal{Z}_i$ and any $t \geq 1$, we can use the level-t Lovász-Schrijver relaxation to approximate $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$ in (9a)–(9b). We can then convert the inner maximization (9a) to a minimization and embed it in the first-stage problem. The following lemma illustrates this for $t = 1$; we omit the proof since it follows from a straightforward application of conic duality. In this lemma, $\boldsymbol{\gamma}$ represents the objective function of the inner optimization (9a); in the linearized MICP representation (9b-ℓ) of $\mathcal{Z}$, it is given by $\boldsymbol{\gamma} = \left[\mathbf{0}^\top\ \boldsymbol{h}(\boldsymbol{x})^\top\ \operatorname{vec}(\boldsymbol{T}(\boldsymbol{x}))^\top\ -\alpha\right]^\top$, whereas in the penalty MICP representation (9b-ρ), we have $\boldsymbol{\gamma} = \left[\rho\mathbf{e}^\top\ \boldsymbol{h}(\boldsymbol{x})^\top\ -\mathbf{e}^\top\ -\alpha\right]^\top$.

Lemma 2. *Suppose that $\mathcal{Z}$ is defined as in (15) with $\mathcal{K} = \tilde{\mathcal{K}} \times \mathbb{R}^n_+$, $\boldsymbol{A} = [\tilde{\boldsymbol{A}}^\top \mathbf{I}]^\top$ and $\boldsymbol{b} = [\tilde{\boldsymbol{b}}^\top \mathbf{0}^\top]^\top$ for suitable matrices $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{b}}$. If $\mathcal{Z}^1$ denotes the level-1 Lovász-Schrijver relaxation of* $\operatorname{cl}\operatorname{conv}(\mathcal{Z})$*, then for any $\boldsymbol{\gamma} \in \mathbb{R}^n$, we have the following strong duality result:*

$$
\begin{array}{rcll}
\sup\limits_{\boldsymbol{z} \in \mathcal{Z}^1} \boldsymbol{\gamma}^\top \boldsymbol{z} & = & \text{minimize} & \sum\limits_{j \in [M]} \max\left\{ -\tilde{\boldsymbol{b}}^\top \tilde{\boldsymbol{\zeta}}^{j0}, -\tilde{\boldsymbol{b}}^\top \tilde{\boldsymbol{\zeta}}^{j1} - \beta^{j1} \right\} \\
& & \text{subject to} & \sum\limits_{j \in [M]} \boldsymbol{\delta}^j + \boldsymbol{\gamma} \leq \mathbf{0}, \\
& & & \left.\begin{array}{l} \tilde{\boldsymbol{A}}^\top \tilde{\boldsymbol{\zeta}}^{j\ell} + \beta^{j\ell} \mathbf{e}_j - \mathbb{I}[\ell = 1] \boldsymbol{Y} \mathbf{e}_j \leq \boldsymbol{\delta}^j \\ \boldsymbol{\delta}^j \in \mathbb{R}^n,\ \tilde{\boldsymbol{\zeta}}^{j\ell} \in \tilde{\mathcal{K}}^*,\ \beta^{j\ell} \in \mathbb{R} \end{array}\right\} \forall \ell \in \{0,1\}, j \in [M], \\
& & & \boldsymbol{Y} \in \mathbb{R}^{M \times M} : \boldsymbol{Y} = -\boldsymbol{Y}^\top.
\end{array}
\tag{18}
$$

Remark 3 (Relationship to approximations obtained by relaxing the support)**.** *An alternative outer approximation of the distributionally robust two-stage problem (2) can be obtained by simply relaxing the zero-one constraints on the support in reformulation (7); i.e., by replacing $\Xi = \{\boldsymbol{\xi} \in \mathbb{Z}^M_+ : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{f}\}$ in (7) with its continuous relaxation $\{\boldsymbol{\xi} \in \mathbb{R}^M_+ : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{f}\}$. The resulting approximation is intractable in general, unless the uncertainty $\boldsymbol{\xi}$ appears only in the objective of the loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ (see discussion in Section 3). In the latter case, it can be easily seen that the resulting approximation coincides precisely with that obtained by replacing* $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$ *in (10a)–(10b) with the continuous relaxation $\mathcal{Z}^0_i$. Therefore, our lift-and-project approximation technique can be viewed as a generalization of this approach. Unlike the former, however, a crucial difference is that this approach provides no formal mechanism to improve the quality of the final approximation; we illustrate this empirically in Section 5.*

4.2 Numerical considerations

Several factors can impact the numerical solution of the approximations obtained by replacing $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$, $i \in [N]$, in (9a), with their level-1 Lovász-Schrijver relaxations $\mathcal{Z}^1_i$. First, if we use the linearized MICP representation (9b-ℓ) to reformulate $\mathcal{Z}_i$, then formulation (18) has $O(M(F + N_2 + LM))$ variables, $O(M^2 L)$ linear constraints, and $O(M(M + N_2))$ conic constraints. If we use the penalty representation (9b-ρ), then these are reduced to $O(M(F + N_2 + L + M))$ variables, $O(M(L+M))$ linear constraints, and $O(M(M+N_2))$ conic constraints. In either case, the number of conic constraints can be reduced to $O(MN_2)$ whenever the metric $d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_1$ is induced by the 1-norm (see argument in proof of Proposition 3 in Appendix B).

Second, observe that setting $\boldsymbol{Y} = \boldsymbol{0}$ also reduces the number of variables in (18) for only a minor loss in approximation quality; indeed, the resulting relaxation is still equal to $\bigcap_{j\in[M]} \operatorname{cl}\operatorname{conv}(\{\boldsymbol{z} \in \mathcal{Z}_i^0 : z_j \in \{0,1\}\})$. In fact, it is equal to $\bigcap_{j\in\mathcal{J}_i} \operatorname{cl}\operatorname{conv}(\{\boldsymbol{z} \in \mathcal{Z}_i^0 : z_j \in \{0,1\}\})$, where $\mathcal{J}_i \subseteq [M]$ is the index set of binary parameters whose optimal values in the left-hand side of (18) are fractional. We expect $|\mathcal{J}_i|$ to be small since the optimal value of $\boldsymbol{\xi}$ in any convex relaxation of $\mathcal{Z}_i$ is unlikely to be far from the binary-valued $\hat{\boldsymbol{\xi}}^{(i)}$ (see argument in proof of Theorem 5). This motivates the following iterative heuristic to identify the index sets $\mathcal{J}_i$. Note that this procedure is independent of the MICP representation used for each $\mathcal{Z}_i$.

1. Select $\texttt{tol} \in (0, 0.5)$, $\texttt{niter} \in \mathbb{Z}_+$. Set $\texttt{iter} \leftarrow 1$. For each $i \in [N]$, set $\tilde{\mathcal{Z}}_i^1 \leftarrow \mathcal{Z}_i^0$ and $\mathcal{J}_i \leftarrow \emptyset$.

2. For each $i \in [N]$, replace $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$ with its current approximation $\tilde{\mathcal{Z}}_i^1$ and dualize the corresponding problem (9a). Solve the resulting convex hull approximation (8).

3. For each $i \in [N]$, let $\bar{\boldsymbol{\xi}}^{[i]}$ be the optimal value of $\boldsymbol{\xi}$ in (9a), recovered as scaled dual multipliers. For each $j \in [M] \setminus \mathcal{J}_i$, if $\bar{\xi}_j^{[i]} \in [\texttt{tol}, 1 - \texttt{tol}]$, update $\mathcal{J}_i \leftarrow \mathcal{J}_i \cup \{j\}$ and $\tilde{\mathcal{Z}}_i^1$ as follows:

$$\tilde{\mathcal{Z}}_i^1 \leftarrow \left\{ \boldsymbol{z} \in \mathbb{R}^n : \begin{array}{ll} \exists u^{j0}, u^{j1} \geq 0,\ u^{j0} + u^{j1} = 1, & j \in \mathcal{J}_i \\ \exists \boldsymbol{z}^{j0} \in \mathcal{Z}_i^0(u^{j0}),\ \boldsymbol{z}^{j1} \in \mathcal{Z}_i^0(u^{j1}), & j \in \mathcal{J}_i \\ \boldsymbol{z} = \boldsymbol{z}^{j0} + \boldsymbol{z}^{j1}, & j \in \mathcal{J}_i \\ z_j^{j0} = 0,\ z_j^{j1} = u^{j1}, & j \in \mathcal{J}_i \end{array} \right\}.$$

4. If none of the index sets $\mathcal{J}_1, \ldots, \mathcal{J}_N$ were updated or if $\texttt{iter} \geq \texttt{niter}$, stop. Otherwise, update $\texttt{iter} \leftarrow \texttt{iter} + 1$ and go to Step 2.

Note that the successive optimizations in Step 2 can benefit from an efficient initialization of their variables by using the optimal solution from the previous solve. Moreover, the size of these problems can be controlled by using smaller values of `niter` and larger values of `tol`, since they directly influence the size of $\mathcal{J}_i$ and $\tilde{\mathcal{Z}}_i^1$, albeit at the expense of coarser approximations. In our implementation, we found that a setting of $\texttt{iterlim} = 5$ and $\texttt{tol} = 10^{-2}$ achieved a good tradeoff between approximation quality and computational effort.

5 Computational Experiments

We illustrate the applicability of our method to operational problems in electric power systems in Section 5.1, and to design problems in multi-commodity flow networks in Section 5.2. Our goals are to: *(i)* study the lift-and-project approximations $\mathcal{Z}^0$ and $\mathcal{Z}^1$ in terms of their computational effort and ability to approximate $\operatorname{cl}\operatorname{conv}(\mathcal{Z})$; *(ii)* compare their out-of-sample performance with the standard sample average approximation and with classical two-stage robust optimization; and, *(iii)* elucidate the effect of two key parameters on the relative benefits of the distributionally robust two-stage problem (2) over these classical formulations: the "rareness" of network failures and the relative magnitude of "impact" when failures occur.

Our code was implemented in Julia 1.5.3, using JuMP 0.21.4. We used Mosek 9.2 for solving our lift-and-project approximations, and Gurobi 9.1.1 as the solver for the Benders and column-and-constraint generation schemes (which we compare in Sections 5.1 and 5.2 respectively), since the latter performed better than the former in solving the mixed-integer subproblems in those schemes; whereas Mosek performed better in solving the conic programming relaxations. All runs were conducted on an Intel Xeon 2.3 GHz computer, with a limit of four cores per run.

5.1 Optimal power flow

We use our method to address the security-constrained optimal power flow problem that is fundamental to the secure operation of electric power grids and solved every fifteen minutes or so by grid operators (e.g.,see [1, 22]). The goal is to determine voltages and generation levels of available generators so as to satisfy power demand in the network, while adhering to various physical and engineering constraints. For example, electric power between network nodes (also known as *buses*) can flow only along capacitated edges or transmission lines. As such, the latter are failure prone, and transmission line outages can lead to an unstable power network or even complete system failure, resulting in costly blackouts. However, such high-impact failure events are rare. For example, between the years 2000 and 2014, fewer than 1,500 power outages have occurred that affected 50,000 or more residents in the entire United States, which is fewer than 100 events per year [64]. This rarity complicates the accurate estimation of their underlying distribution.

Because electric power is governed by complex physical laws, optimal power flow is a highly nonlinear optimization problem. Nevertheless, the underlying physics can be approximated well by

using second-order cone or semidefinite programming relaxations [44]. Although our method generalizes to any convex cone relaxation, we focus on the standard second-order cone relaxation [38], where $\mathcal{X}$ is second-order cone representable and $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is the optimal value of a second-order cone program.

Our two-stage optimization model is inspired by [63] and presented in Appendix E. Conceptually, the first-stage problem determines minimum cost power generation levels assuming no line outages. Upon line failure, the second-stage model seeks to adjust the power generation levels subject to physical constraints where failed lines cannot be used, with a goal of minimizing the total penalty cost of violating power balances. This model satisfies assumptions (A1), (A2), and (A3) and allows the use of the penalty reformulation (9b-ρ), which also has the advantage of using fewer variables and constraints compared with the linearized reformulation (see Section 3.3).

The operational state of transmission lines is modeled as a random binary vector $\boldsymbol{\xi}$ with support $\Xi = \{0, 1\}^M$, where $\xi_l = 1$ indicates that line l has failed. In particular, since $\boldsymbol{\xi}$ represent on/off switches, we can use Corollary 2 to get not only a smaller MICP formulation but also tighter values of the penalty parameter $\rho = \rho^r$. The latter is computed by using Theorem 4, where the classical robust counterpart reduces to a deterministic problem (see Lemma 1); indeed, the second-stage loss function trivially attains its worst-case value when each component of $\boldsymbol{\xi}$ is one, that is, when all transmission lines fail.

5.1.1 Test instances

We conduct our experiments on the standard IEEE 14-, 30- and 118-bus test cases from the PGLib-OPF library [3]. In each case, the second-stage per-unit penalty cost for violating the power balance equations is set to be ϕ times the maximum per-unit first-stage generation cost. Note that this choice depends on the economic cost of failure to meet power demand. Since loss of power and blackouts tend to be costly and the associated penalty costs much larger than the cost of generation, we set $\phi = 100$ in Sections 5.1.2 and 5.1.3 and analyze its sensitivity in Section 5.1.4.

We generate empirical data using a Bernoulli model. Specifically, we model each component of $\tilde{\boldsymbol{\xi}}$ as independent and identically distributed Bernoulli random variables with parameter ψM^{-1}, where M denotes the number of transmission lines. Note that this choice reflects the rare nature of line failures; in particular, it implies that only $\psi \cdot 100\%$ of training samples record a line failure.

We set $\psi = 0.1$ in Sections 5.1.2 and 5.1.3 and analyze its impact in Section 5.1.4.

In all our experiments, for a fixed sample size N and radius ε, we report average results using 100 statistically independent sets of training samples, and we estimate the variance by reporting the standard deviation over these 100 runs. In Sections 5.1.3 and 5.1.4, the out-of-sample performances of a candidate solution are estimated by using 1,000 statistically independent sets of testing samples.

5.1.2 Approximation quality and computational effort

To study the quality of our lift-and-project approximations, we compute the following quantities for each sample size $N \in \{10, 100, 1000\}$ and radius $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$, where $\nu \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$: *(i)* the optimal value $v^\star$ of the convex hull reformulation (8) using the Benders scheme described in Appendix C, and *(ii)* the optimal values v^0, $\tilde{v}^1$ and v^1 of formulation (8) when the convex hulls $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$ in (9a)–(9b) are approximated by using the continuous relaxation $\mathcal{Z}^0$ and heuristically and exactly computed level-1 Lovász-Schrijver relaxations $\tilde{\mathcal{Z}}^1$ and $\mathcal{Z}^1$, respectively (see Section 4.2). The choice of the radius $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$ is motivated from Theorem 1, and we elaborate on it further in the next subsection.

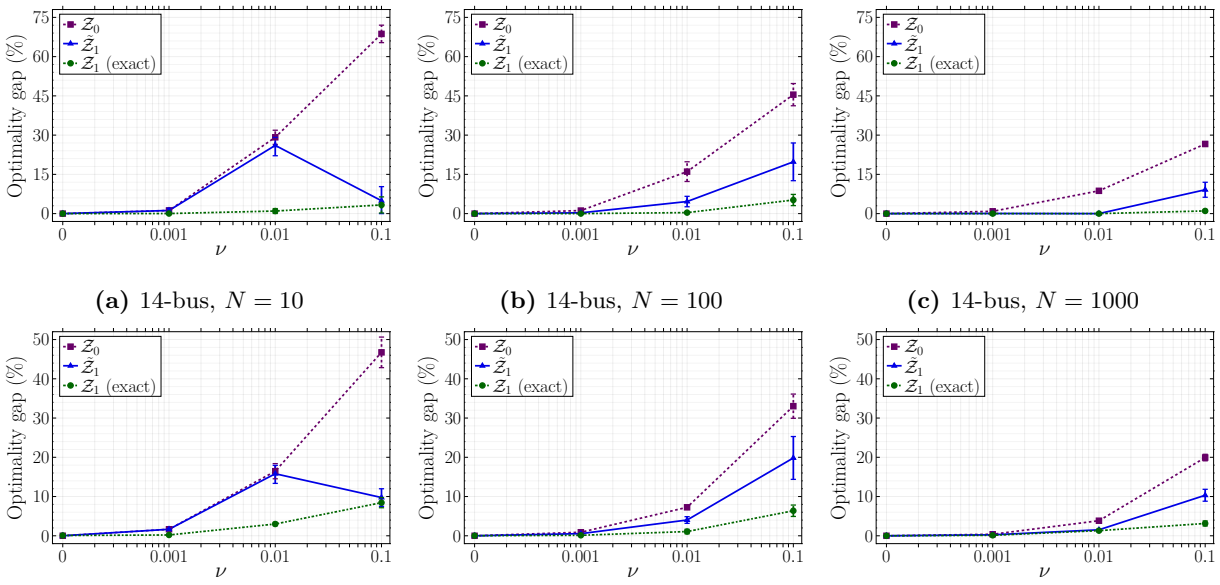
Figure 4 reports the average (line plot) and standard deviation (error bar) of the optimality gaps, defined as $(v - v^\star)/v^\star \times 100\%$, where $v \in \{v^0, \tilde{v}^1, v^1\}$, based on 100 statistically independent sets of training samples. We make the following observations.

- The exact level-1 relaxation $\mathcal{Z}^1$ is near optimal, with optimality gaps never exceeding 10%, whereas the continuous relaxation $\mathcal{Z}^0$ is less accurate, especially for larger radii (*e.g.*, 50% gap for $\nu = 0.1$). The heuristically computed level-1 relaxation $\tilde{\mathcal{Z}}^1$ is also near optimal for small and large radii but performs relatively poorly for intermediate values of ν.

- For a fixed sample size N and decreasing radius ν, the optimality gaps of all approximations decrease to 0. For increasing ν, the gaps of the level-1 relaxations increase far less rapidly than that of the continuous relaxation.

- For a fixed radius ν and increasing sample size N, the gaps of all approximations decrease.

Figure 5 reports the average computation time to solve the various approximations and compares them with that of the Benders decomposition scheme. We offer the following comments.



(a) 14-bus, $N = 10$ **(b)** 14-bus, $N = 100$ **(c)** 14-bus, $N = 1000$

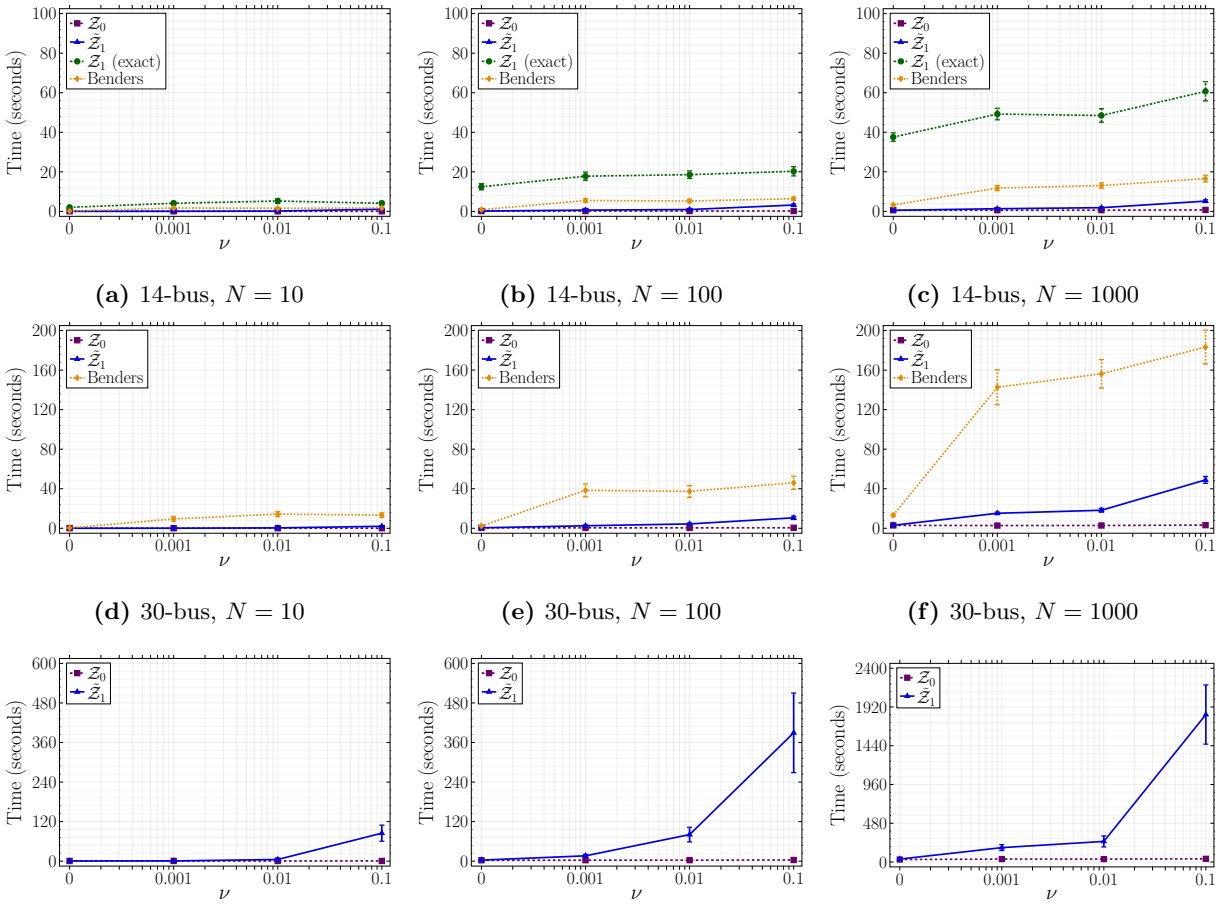
(d) 30-bus, $N = 10$ **(e)** 30-bus, $N = 100$ **(f)** 30-bus, $N = 1000$

Figure 4. Optimality gaps using the continuous relaxation $\mathcal{Z}^0$ and the heuristically and exactly computed level-1 Lovász-Schrijver relaxations $\tilde{\mathcal{Z}}^1$ and $\mathcal{Z}^1$ as a function of ν and N, where $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$.

- The continuous relaxation $\mathcal{Z}^0$ and heuristically computed level-1 relaxations $\tilde{\mathcal{Z}}^1$ have the smallest computation times, with the former being faster for larger values of N and ν and, in particular, for the larger 118-bus case where it is more than 10 times faster. When compared with the Benders scheme, the relative difference in their computation times is minor for small sample sizes N but increases significantly for large sample sizes. For $N = 1000$, both approximations run 10 times faster than the Benders scheme for the 14-bus case, while $\tilde{\mathcal{Z}}^1$ runs 4 times faster and $\mathcal{Z}^0$ runs almost 100 times faster for the 30-bus case.

- The exact level-1 relaxation $\mathcal{Z}^1$ and the Benders scheme appear to be the most difficult to solve. Although not shown, for the 30-bus case the former took about 1, 2 and 10 minutes for $N = 10, 100$, and $1,000$, respectively, whereas for the larger 118-bus case neither scheme terminated within 10 minutes for $N = 10, 100$ or within 1 hour for $N = 1000$. Moreover, some of the MICP subproblems within the Benders scheme can cause slow convergence (e.g., due to search tree enumeration). This is evidenced by the fact that about 1% of the Benders runs did not terminate within 10 minutes even for the smaller 14-bus and 30-bus cases, respectively.

- In conjunction with Figure 5, the heuristically computed level-1 relaxation $\tilde{\mathcal{Z}}^1$ appears to offer the best tradeoff in terms of approximation quality and computational effort.
- The run times of all approximations, and in particular $\tilde{\mathcal{Z}}^1$, can be significantly improved by using an efficient initialization of their variables (see Section 4.2), which we did not implement.



(a) 14-bus, $N = 10$ **(b)** 14-bus, $N = 100$ **(c)** 14-bus, $N = 1000$

(d) 30-bus, $N = 10$ **(e)** 30-bus, $N = 100$ **(f)** 30-bus, $N = 1000$

(g) 118-bus, $N = 10$ **(h)** 118-bus, $N = 100$ **(i)** 118-bus, $N = 1000$

Figure 5. Computation times for solving formulation (8) using the continuous relaxation $\mathcal{Z}^0$, the heuristically computed and exact level-1 Lovász-Schrijver relaxations $\tilde{\mathcal{Z}}^1$ and $\mathcal{Z}^1$, and the Benders decomposition scheme, as a function of ν and N, where $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$.

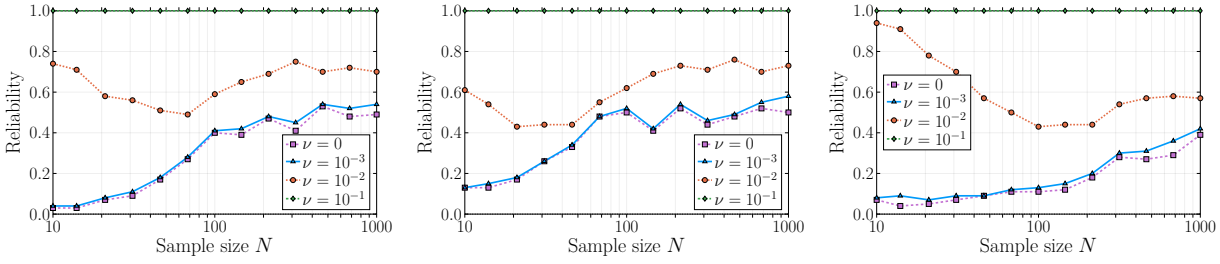
5.1.3 Out-of-sample performance and finite sample guarantee

To understand the potential benefits of a distributionally robust approach, we evaluate its out-of-sample performance. For a given training dataset of size N and a given choice of ν, we obtain

a candidate first-stage solution $\boldsymbol{x}^\nu$ by solving formulation (8) with the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$ and Wasserstein radius $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$. We then estimate the out-of-sample performance of $\boldsymbol{x}^\nu$ by recording $z^\nu = \boldsymbol{c}(\boldsymbol{x}^\nu) + 1000^{-1}\sum_{i=1}^{1000}\mathcal{Q}(\boldsymbol{x}^\nu, \hat{\boldsymbol{\xi}}^{(i)})$, where $\hat{\boldsymbol{\xi}}^{(1)}, \ldots, \hat{\boldsymbol{\xi}}^{(1000)}$ are 1,000 independently generated testing samples. This entire process is repeated 100 times for statistically independent sets of N training samples and 1,000 testing samples.

We first justify the choice of the radius $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$ as a function of the training sample size N. This dependence is motivated by inequality (5) in Theorem 1. However, the latter inequality can be loose, especially when accounting for problem-dependent constants such as the size and diameter of the support Ξ. Therefore, we empirically verify the finite sample guarantee of the first-stage solutions $\boldsymbol{x}^\nu$ under the tighter parameterization $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$ for several choices of the coefficient $\nu \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$. Figure 6 reports the *reliability* of $\boldsymbol{x}^\nu$, which we define as the empirical probability (over the 100 sets of training samples) that the optimal value $\tilde{v}^1$ of the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$ is an upper bound on the out-of-sample cost z^ν.

Figure 6 shows that, for fixed values of N, the reliability of $\boldsymbol{x}^\nu$ increases with increasing values of ν, and this can be used to guide the choice of ν. For example, depending on their risk level, decision-makers can select the smallest value of ν with sufficiently high reliability. In particular, for training datasets with small sample size N, observe that $\nu = 10^{-2}$ yields an upper bound on the out-of-sample cost with probability more than 0.5, whereas $\nu = 10^{-1}$ yields an upper bound with probability 1.0. However, note that we cannot always access the true out-of-sample cost (and hence, the true reliability). Nevertheless, one could estimate the out-of-sample cost by using cross validation or the holdout method (e.g, see [37, 50]), and then select a value for the coefficient ν that offers a good trade-off between low out-of-sample cost and high reliability.



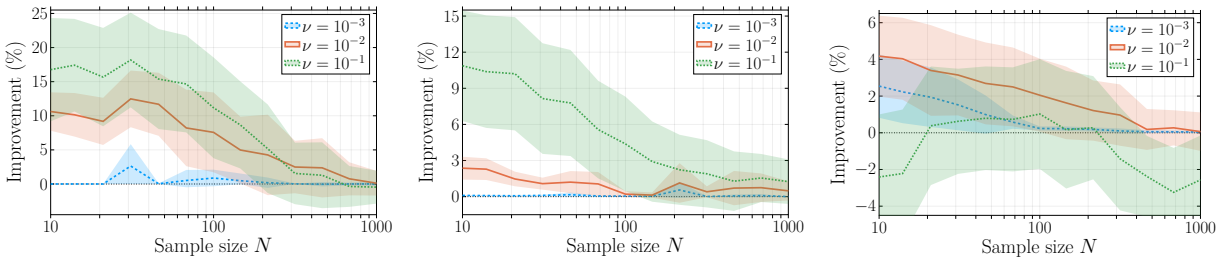
(a) 14-bus **(b)** 30-bus **(c)** 118-bus

Figure 6. Reliability of the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$, as a function of training sample size N.

We now evaluate the benefits of our distributionally robust model over the sample average approximation, by computing the relative improvement in out-of-sample cost, which we define as $(z^0 - z^\nu)/z^0 \times 100\%$. Figure 7 reports the mean (solid line) and standard deviation (shaded ribbon) of the relative improvement over the 100 independent sets of training samples. We make the following observations from Figure 7.

- The distributionally robust model (2) consistently outperforms the sample average approximation, particularly for small sample sizes N. The magnitude of the relative improvement is instance dependent (roughly 15%, 10%, and 5% for the 14-, 30-, and 118-bus cases, respectively) but consistently decreases for large values of N as expected. The magnitude of the radius that leads to the best possible improvement also is instance dependent.

- The larger variances in improvement for smaller N and for larger instances can be partially explained by the combinatorial growth in the number of truly distinct training datasets of size N (i.e., those that lead to distinct first-stage solutions) that are possible under the rare event model of line outages. The large variances for $\nu = 10^{-1}$ can also be similarly explained by the larger number of truly distinct first-stage solutions that can result from slight variations in the training dataset.



(a) 14-bus **(b)** 30-bus **(c)** 118-bus

Figure 7. Relative improvement in the out-of-sample performance of the distributionally robust two-stage model (2) when compared with the sample average approximation, as a function of sample size N.

We now compare the out-of-sample performance of the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$ corresponding to the best choice of ν (which is 10^{-1} for the 14- and 30-bus cases, and 10^{-2} for the 118-bus case) with *(i)* the continuous relaxation $\mathcal{Z}^0$ for the same ν, and *(ii)* a classical two-stage

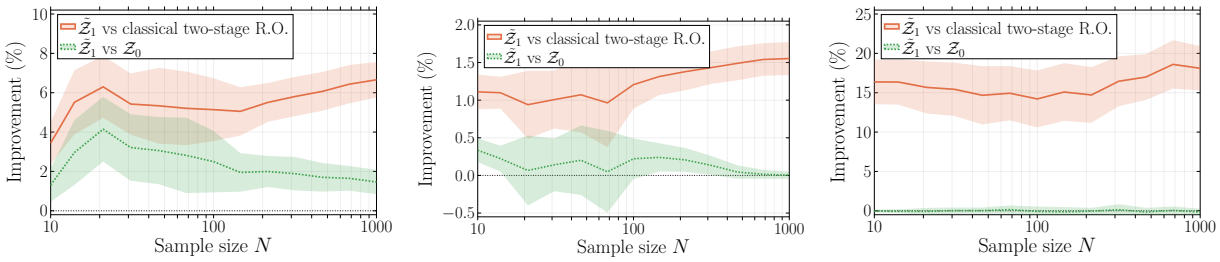
robust optimization model of the following form:

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ c(\boldsymbol{x}) + \max_{\boldsymbol{\xi}\in\Xi_K} \mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi}), \tag{19}$$

where $\Xi_K = \{\boldsymbol{\xi} \in \{0,1\}^M : \xi_1 + \ldots + \xi_M \leq K\}$ is the uncertainty set with K being the *budget* of uncertainty. In particular, for each instance, we let $K \in \{0,1,2,5,10\}$ and obtain optimal first-stage solutions $\boldsymbol{x}^K$ of problem (19) with the Benders decomposition scheme. As before, we estimate the out-of-sample performance of $\boldsymbol{x}^K$ as $z^K = \boldsymbol{c}(\boldsymbol{x}^K) + 1000^{-1}\sum_{i=1}^{1000} \mathcal{Q}(\boldsymbol{x}^K, \hat{\boldsymbol{\xi}}^{(i)})$, where $\hat{\boldsymbol{\xi}}^{(1)},\ldots,\hat{\boldsymbol{\xi}}^{(1000)}$ are the same 1,000 testing samples used to estimate z^ν. We then record the best possible K yielding the lowest z^K, and compute the relative improvement in out-of-sample cost, which we define as $(z^K - z^\nu)/z^\nu \times 100\%$. Figure 8 reports the mean (solid line) and standard deviation (shaded ribbon) of the relative improvement over the 100 independent sets of training samples. We make the following observations from Figure 8.

- The distributionally robust model strongly outperforms its classical robust counterpart, across all instances, with relative improvements of 5%, 1% and 15% for the 14-, 30- and 118-bus cases, respectively. In contrast to Figure 7, the relative improvements increase with increasing values of N; this is expected since the classical robust model (19) ignores all sample data and therefore, it becomes overly conservative in the presence of a moderate amount of data. Thus, we observe that for small to moderate values of N, the distributionally robust model improves upon both the sample average approximation and classical robust optimization. Finally, although not shown, solving the classical robust model with the Benders scheme took longer than solving the level-1 relaxation, especially for the larger 118-bus case.

- The relative improvement of the level-1 relaxation $\tilde{\mathcal{Z}}^1$ over the continuous relaxation $\mathcal{Z}^0$ is smaller, but can be as high as 4% as seen in the 14-bus case. It should be noted that these quantities are necessarily upper bounded by the improvements over the sample average approximations reported in Figure 7, and can be significant for applications including optimal power flow, which are executed several times each day of the year. Finally, we note that the tradeoff between tighter in-sample optimality gaps (see Figure 4) and hence stronger finite sample reliability guarantees (see Figure 6) offered by the level-1 relaxation $\tilde{\mathcal{Z}}^1$, with the faster computation times for solving the continuous relaxation $\mathcal{Z}^0$ (see Figure 5), can guide the design of an algorithmic scheme.



(a) 14-bus **(b)** 30-bus **(c)** 118-bus

Figure 8. Relative improvement in the out-of-sample performance of the distributionally robust two-stage model (2) solved using the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$, when compared with the classical two-stage robust optimization model (19), and the continuous relaxation $\mathcal{Z}^0$.

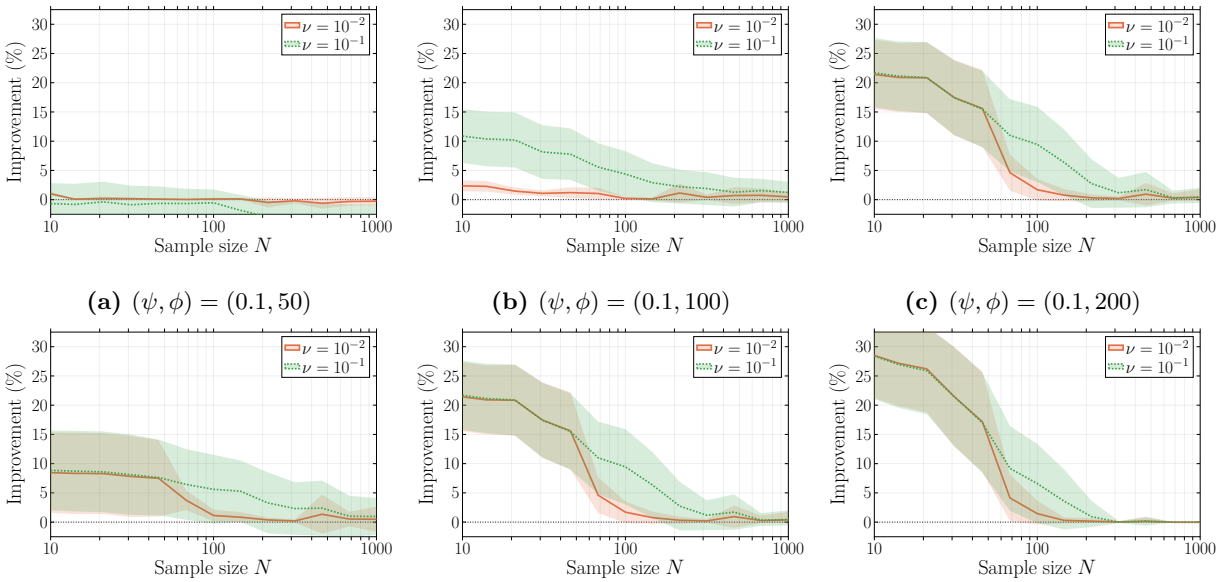
5.1.4 Sensitivity analysis

The instance-dependent behavior of the out-of-sample performance from the previous subsection suggests that it might also be influenced by other parameters. Here, we investigate the effect of the "rareness" ψ of transmission line failures and the relative magnitude ϕ of "impact" when failures occur. Recall from Section 5.1.1 that higher values of ψ increase the probability of line failures, whereas higher values of ϕ increase the penalty cost for failing to satisfy power demand due to transmission line failures. Figure 9 shows the relative improvement of the distributionally robust two-stage model (2) over the sample average approximation for various choices of ψ and ϕ. For brevity, we report results only for the 30-bus instance; the high-level insights do not change for other instances. We make the following observations from Figure 9.

- For fixed values of the line failure probability ψ, Figures 9a–9c show that as the impact due to failure ϕ increases, the relative benefits of a distributionally robust approach strongly increase. In other words, benefits increase with higher impacts of failures. Interestingly, Figure 9a also shows that if failures are rare but low impact, then ignoring them (as in the sample average approximation) may not incur high out-of-sample costs, even for small values of N.

- For fixed values of the magnitude of impact ϕ, Figures 9d–9f show that as the probability of failures ψ increases, the relative benefits of a distributionally robust approach increases. However, observe that this does not necessarily imply that the relative benefits are small when line failure probabilities are small. Indeed, we observe that the relative benefits remain

as high as 10% even when individual line failure probabilities are less than $0.05M^{-1}$.



(a) $(\psi, \phi) = (0.1, 50)$ **(b)** $(\psi, \phi) = (0.1, 100)$ **(c)** $(\psi, \phi) = (0.1, 200)$

(d) $(\psi, \phi) = (0.05, 200)$ **(e)** $(\psi, \phi) = (0.1, 200)$ **(f)** $(\psi, \phi) = (0.2, 200)$

Figure 9. Relative improvement in the out-of-sample performance of the distributionally robust two-stage model (2) when compared with the sample average approximation, for various values of (ψ, ϕ).

5.2 Multi-commodity network design

We now consider multi-commodity network design problems that have applications in telecommunications, transportation, logistics and production planning, among others (e.g., see [23, 48]). In several of these applications, it is required to send flows to satisfy known demands between multiple origin-destination pairs or commodities. The goal is to minimize the total cost, which is the sum of fixed costs of installing arc capacities and variable costs of routing flows for each commodity. As such, failures of network elements can lead to a reduction in its available flow capacity and a subsequent failure to meet demands. This is particularly true in telecommunication networks where the loss of even a single (typically high-capacity) fiber-optic cable or router equipment can cause a substantial fraction of the overall flow (e.g., internet traffic) to be lost, leading to potentially huge economic impacts [47]. Fortunately, these networks are typically well-engineered and therefore, such high-impact failures are rare. At the same time, this general lack of failure data in real networks complicates the accurate estimation of their underlying distribution.

The two-stage optimization model we consider is presented in Appendix F, and can be described as follows. The first-stage problem determines the arc capacities that can be used for routing flows. Upon failure, the second-stage model determines the routing of each commodity along the degraded network topology constrained by the first-stage arc capacities, and with the objective of minimizing the sum of variable routing costs and penalty costs for not satisfying demands.

For ease of exposition, we model only node failures, where $\boldsymbol{\xi}$ is supported on $\Xi = \{0,1\}^M$ and $\xi_i = 1$ indicates that node i has failed. Since $\boldsymbol{\xi}$ represent on/off switches, we can employ Corollary 2, and the penalty parameter $\rho = \rho^r$ can be computed using Theorem 4. Here, the classical robust counterpart reduces to a deterministic problem (see Lemma 1), since the second-stage loss function trivially attains its worst-case value when each component of $\boldsymbol{\xi}$ is one; that is, when all nodes fail.

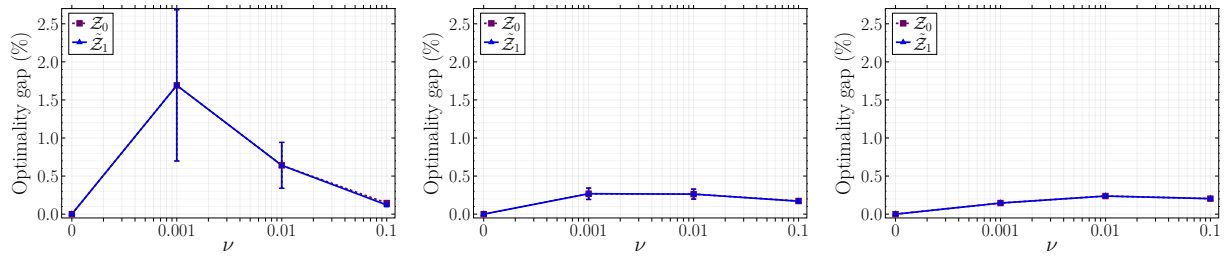
We conduct our experiments on the $(20, 230, 40, V, L)$ instance from the so-called Class I set of instances in [23]. As the name indicates, the instance has 20 nodes, 230 arcs, 40 commodities, and the letters V and L indicate that fixed-costs are relatively low compared to variable costs, and that the problem is loosely capacitated. We generate empirical data by modeling each component of $\tilde{\boldsymbol{\xi}}$ as independent and identically distributed Bernoulli random variables with parameter ψM^{-1}, where $\psi = 0.1$. As before, for a fixed sample size N and radius ε, we report average results using 100 statistically independent sets of training samples, and we estimate the variance by reporting the standard deviation over these 100 runs. The out-of-sample performances of candidate solutions are estimated by using 1,000 statistically independent sets of testing samples.

5.2.1 Approximation quality and computational effort

Similar to Section 5.1.2, we compute the optimality gaps of the convex hull reformulation (8) when the convex hulls $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$ in (9a)–(9b) are approximated by using the continuous relaxation $\mathcal{Z}^0$ and heuristically computed level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$, for sample sizes $N \in \{10, 100, 1000\}$ and radii $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$, $\nu \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$. In doing so, the true optimal value of each instance is computed by solving formulation (7) using the column-and-constraint generation scheme. The mixed-integer subproblems in this scheme are solved by dualizing the second-stage loss function as opposed to using its Karush-Kuhn-Tucker conditions, since it results in fewer additional variables and constraints, see [70]. Any bilinear expressions involving dual variables and uncertain parameters are reformulated using indicator constraints since

the lack of *a priori* known upper bounds on the dual variables prohibits direct linearization using McCormick inequalities.

Figure 10 reports the average (line plot) and standard deviation (error bar) of the optimality gaps, whereas Figure 11 reports the corresponding computation times, based on 100 statistically independent sets of training samples. Figure 10 shows that both the continuous and level-1 relaxations provide very similar and near-optimal approximations with optimality gaps never exceeding 3% for $N = 10$ and 0.5% for $N \geq 100$. Interestingly, the optimal first-stage decisions are different, and this can be seen from their out-of-sample performance that we present in the next subsection.



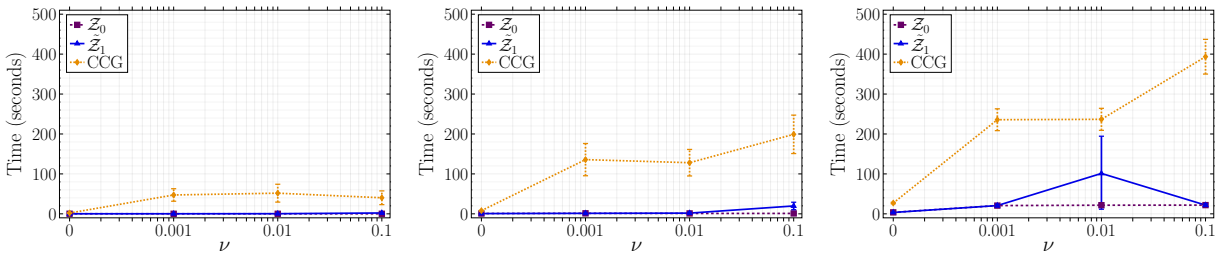
(a) $N = 10$ **(b)** $N = 100$ **(c)** $N = 1000$

Figure 10. Optimality gaps using the continuous relaxation $\mathcal{Z}^0$ and the heuristically computed level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$, as a function of ν and N where $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$.

Figure 11 shows that both relaxations have small computation times (< 60 seconds on average) across all N and ν. When compared with the column-and-constraint generation scheme, the relative difference in their computation times is minor for small sample sizes N but increases significantly for large sample sizes, where the column-and-constrain generation method can be slower by more than a factor of 10. Similar to the Benders scheme, the mixed-integer subproblems in the column-and-constraint generation scheme can cause slow convergence, and roughly 3% of its runs did not terminate within 10 minutes.

5.2.2 Out-of-sample performance and finite sample guarantee

Similar to Section 5.1.3, we estimate the out-of-sample performance of the first-stage solutions of the lift-and-project and sample average approximations for different sample sizes N and Wasserstein radii $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$, where $\nu \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$. The results are summarized in Figure 12.



(a) $N = 10$ **(b)** $N = 100$ **(c)** $N = 1000$

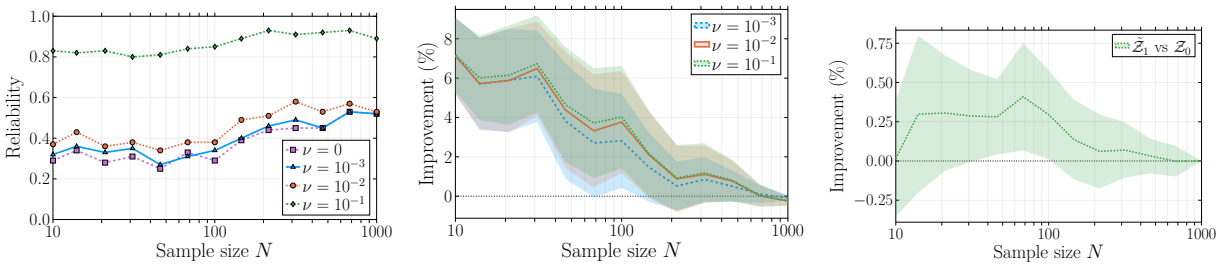
Figure 11. Computation times using the continuous $\mathcal{Z}^0$ and heuristically computed level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$ for solving formulation (8), and using the column-and-constraint generation scheme for solving formulation (7), as a function of ν and N, where $\varepsilon = \nu\sqrt{N^{-1}\log(N+1)}$.

First, Figure 12a reports the reliability of the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$, which is the empirical probability (over the 100 sets of training samples) that its optimal value is an upper bound on its out-of-sample cost. We observe that the reliability increases not only with increasing values of ν (for fixed values of N) but also with increasing values of N (for fixed values of ν). The choice $\nu = 10^{-1}$ is reliable with probability 0.8 for training datasets of small size N and this increases to > 0.9 for large values of N.

Figure 12b reports the relative improvement in out-of-sample cost of the distributionally robust model over the sample average approximation, over the 100 independent sets of training samples. As before, we find that the distributionally robust model consistently outperforms the sample average approximation, particularly for small sample sizes N, where the magnitude of the relative improvement can be as high as 7%, but decreases for large values of N.

Finally, Figure 12c reports the improvement in the out-of-sample performance of the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$ for the best choice of $\nu = 10^{-1}$, relative to the continuous relaxation $\mathcal{Z}^0$ for the same ν. The relative improvement of the level-1 relaxation $\tilde{\mathcal{Z}}^1$ over the continuous relaxation $\mathcal{Z}^0$ is small yet consistently non-negative. This is not unexpected since the latter already has tight in-sample optimality gaps, as can be seen from Figure 10. Nevertheless, we expect the relative improvements to be instance-dependent similar to optimal power flow, and even a few percentage points can result in long-term economic benefits.

We conclude this section by noting that [68] address problems where $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is the optimal value of a linear program, and the ambiguity set $\mathcal{P}$ is the type-∞ Wasserstein ball which is closely



(a) Reliability **(b)** Out-of-sample performance **(c)** Relative improvement

Figure 12. Reliability (left plot), and relative improvements in the out-of-sample performance of the level-1 Lovász-Schrijver relaxation $\tilde{\mathcal{Z}}^1$ when compared with the sample average approximation (middle plot), and the continuous relaxation $\mathcal{Z}^0$ (right plot), as a function of training sample size N.

related to yet distinct from (3). They provide formulations that can serve as an alternative means to generate first-stage decisions. Interestingly, when applied to the multi-commodity network design problem, they reduce to one of the following: *(i)* sample average approximation ($\varepsilon < 1$) or deterministic problem under the worst-case realization where all nodes have failed ($\varepsilon \geq 1$), when d is induced by the ∞-norm [68, Theorem 4], or *(ii)* classical robust optimization (19) with $K = 1$ when d is induced by the p-norm with $p \in [1, \infty)$ and $\varepsilon < \sqrt[p]{2}$ [68, Theorem 6].

6 Conclusions

Despite their ubiquity in real-world networks, optimization problems affected by rare high-impact uncertainties have not received much attention. This is partly because of the lack of available data given their rare nature, and partly because of the incapability of classical sample average approximations to address them effectively. This paper takes a step toward addressing these limitations by motivating a distributionally robust approach to the problem using Wasserstein ambiguity sets. Notably, we extend the state of the art in data-driven optimization by encompassing not only two-stage conic problems but also high-dimensional discrete uncertainties. By exploiting ideas from nonlinear penalty methods and lift-and-project techniques in global optimization, we present a simple, tractable, and tight approximation of the problem that can be efficiently computed and iteratively improved. We use our method to tackle optimal power flow problems with random transmission line outages and multi-commodity network design problems with random node failures. We find that the method can strongly outperform classical sample average and robust optimization approaches,

especially when failures are rare but can lead to high costs associated with loss of electric power or commodity flows.

Acknowledgments

This material is based upon work supported in part by the U.S. Department of Energy, Office of Science and Office of Electricity Delivery & Energy Reliability, Advanced Grid Research and Development (under contract number DE-AC02-06CH11357), and in part by the Office of Naval Research (under Award number N00014-18-1-2075).

References

[1] O. Alsac and B. Stott. "Optimal load flow with steady-state security". *IEEE Transactions on Power Apparatus and Systems* (1974), pp. 745–751.

[2] A. Ardestani-Jaafari and E. Delage. "Linearized Robust Counterparts of Two-stage Robust Optimization Problems with Applications in Operations Management". *Available at Optimization Online* (2017).

[3] S. Babaeinejadsarookolaee et al. "The power grid library for benchmarking ac optimal power flow algorithms". *arXiv preprint arXiv:1908.02788* (2019).

[4] E. Balas, S. Ceria, and G. Cornuéjols. "A lift-and-project cutting plane algorithm for mixed 0–1 programs". *Mathematical Programming* 58.1-3 (1993), pp. 295–324.

[5] M. Bansal, K.-L. Huang, and S. Mehrotra. "Decomposition algorithms for two-stage distributionally robust mixed binary programs". *SIAM Journal on Optimization* 28.3 (2018), pp. 2360–2383.

[6] J. Barrera et al. "Chance-constrained problems and rare events: an importance sampling approach". *Mathematical Programming* 157.1 (2016), pp. 153–189.

[7] G. Bayraksan and D. K. Love. "Data-driven stochastic programming using phi-divergences". *The Operations Research Revolution*. INFORMS, 2015, pp. 1–19.

[8] A. Ben-Tal et al. "Adjustable robust solutions of uncertain linear programs". *Mathematical Programming* 99.2 (2004), pp. 351–376.

[9] A. Ben-Tal et al. "Robust solutions of optimization problems affected by uncertain probabilities". *Management Science* 59.2 (2013), pp. 341–357.

[10] B. Berche et al. "Resilience of public transport networks against attacks". *The European Physical Journal B* 71.1 (2009), pp. 125–137.

[11] D. Bertsimas, V. Gupta, and N. Kallus. "Robust sample average approximation". *Mathematical Programming* 171.1-2 (2018), pp. 217–282.

[12] D. Bertsimas, S. Shtern, and B. Sturt. "Two-stage sample robust optimization". *arXiv preprint arXiv:1907.07142* (2019).

[13] D. Bienstock and A. Verma. "The $N-k$ Problem in Power Grids: New Models, Formulations, and Numerical Experiments". *SIAM Journal on Optimization* 20.5 (2010), pp. 2352–2380.

[14] J. Blanchet, Y. Kang, and K. Murthy. "Robust Wasserstein profile inference and applications to machine learning". *Journal of Applied Probability* 56.3 (2019), pp. 830–857.

[15] J. Blanchet and K. Murthy. "Quantifying distributional model risk via optimal transport". *Mathematics of Operations Research* 44.2 (2019), pp. 565–600.

[16] N. Boland et al. "Proximity Benders: a decomposition heuristic for stochastic programs". *Journal of Heuristics* 22.2 (2016), pp. 181–198.

[17] A. Budhiraja et al. "Minimization of a Class of Rare Event Probabilities and Buffer Probabilities of Exceedance". *arXiv preprint arXiv:1902.07829* (2019).

[18] S. Burer and H. Dong. "Representing quadratically constrained quadratic programs as generalized copositive programs". *Operations Research Letters* 40.3 (2012), pp. 203–206.

[19] S. Burer and A. N. Letchford. "Non-convex mixed-integer nonlinear programming: A survey". *Surveys in Operations Research and Management Science* 17.2 (2012), pp. 97–106.

[20] G. Calafiore and M. C. Campi. "Uncertain convex programs: randomized solutions and confidence levels". *Mathematical Programming* 102.1 (2005), pp. 25–46.

[21] M. Çezik and G. Iyengar. "Cuts for mixed 0-1 conic programming". *Mathematical Programming* 104.1 (2005), pp. 179–202.

[22] N. Chiang and A. Grothey. "Solving security constrained optimal power flow problems by a structure exploiting interior point method". *Optimization and Engineering* 16.1 (2015), pp. 49–71.

[23] T. G. Crainic, A. Frangioni, and B. Gendron. "Bundle-based relaxation methods for multicommodity capacitated fixed charge network design". *Discrete Applied Mathematics* 112.1 (2001), pp. 73–99.

[24] E. De Klerk. "The complexity of optimizing over a simplex, hypercube or sphere: a short survey". *Central European Journal of Operations Research* 16.2 (2008), pp. 111–125.

[25] E. Delage and Y. Ye. "Distributionally robust optimization under moment uncertainty with application to data-driven problems". *Operations research* 58.3 (2010), pp. 595–612.

[26] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg, 2010.

[27] B. T. Doshi et al. "Optical network design and restoration". *Bell Labs Technical Journal* 4.1 (1999), pp. 58–84.

[28] G. Eichfelder and J. Jahn. "Set-semidefinite optimization". *Journal of Convex Analysis* 15.4 (2008), pp. 767–801.

[29] V. Gabrel et al. "Robust location transportation problems under uncertain demands". *Discrete Applied Mathematics* 164 (2014), pp. 100–111.

[30] R. Gao and A. J. Kleywegt. "Distributionally robust stochastic optimization with Wasserstein distance". *arXiv preprint arXiv:1604.02199* (2016).

[31] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1979. ISBN: 0716710447.

[32] A. Georghiou, D. Kuhn, and W. Wiesemann. "The decision rule approach to optimization under uncertainty: methodology and applications". *Computational Management Science* (2018).

[33] F. Glover. "Improved linear integer programming formulations of nonlinear integer problems". *Management Science* 22.4 (1975), pp. 455–460.

[34] C. Ha. "A noncompact minimax theorem". *Pacific Journal of Mathematics* 97.1 (1981), pp. 115–117.

[35] G. A. Hanasusanto and D. Kuhn. "Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls". *Operations Research* 66.3 (2018), pp. 849–869.

[36] R. Jiang and Y. Guan. "Risk-Averse Two-Stage Stochastic Program with Distributional Ambiguity". *Operations Research* 66.5 (2018), pp. 1390–1405.

[37] R. Jiang, M. Ryu, and G. Xu. "Data-Driven Distributionally Robust Appointment Scheduling over Wasserstein Balls". *arXiv preprint arXiv:1907.03219* (2019).

[38] B. Kocuk, S. S. Dey, and X. A. Sun. "Strong SOCP relaxations for the optimal power flow problem". *Operations Research* 64.6 (2016), pp. 1177–1196.

[39] H. Konno. "Maximization of a convex quadratic function under linear constraints". *Mathematical Programming* 11.1 (1976), pp. 117–127.

[40] D. Kuhn et al. "Wasserstein distributionally robust optimization: Theory and applications in machine learning". *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019, pp. 130–166.

[41] J. B. Lasserre. "Global optimization with polynomials and the problem of moments". *SIAM Journal on optimization* 11.3 (2001), pp. 796–817.

[42] M. Laurent. "A comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre relaxations for 0–1 programming". *Mathematics of Operations Research* 28.3 (2003), pp. 470–496.

[43] L. Lovász and A. Schrijver. "Cones of matrices and set-functions and 0–1 optimization". *SIAM Journal on Optimization* 1.2 (1991), pp. 166–190.

[44] S. H. Low. "Convex relaxation of optimal power flow—Part I: Formulations and equivalence". *IEEE Transactions on Control of Network Systems* 1.1 (2014), pp. 15–27.

[45] J. Luedtke and S. Ahmed. "A sample approximation approach for optimization with probabilistic constraints". *SIAM Journal on Optimization* 19.2 (2008), pp. 674–699.

[46] F. Luo and S. Mehrotra. "Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models". *European Journal of Operational Research* 278.1 (2019), pp. 20–35.

[47] A. Markopoulou et al. "Characterization of Failures in an Operational IP Backbone Network". *IEEE/ACM Transactions on Networking* 16.4 (2008), pp. 749–762.

[48] M. Minoux. "Networks synthesis and optimum network design problems: Models, solution methods and applications". *Networks* 19.3 (1989), pp. 313–360.

[49] A. Mittal, C. Gokalp, and G. A. Hanasusanto. "Robust quadratic programming with mixed-integer uncertainty". *INFORMS Journal on Computing* 32.2 (2020), pp. 201–218.

[50] P. Mohajerin Esfahani and D. Kuhn. "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations". *Mathematical Programming* 171.1-2 (2018), pp. 115–166.

[51] North American Electric Reliability Corporation. "Transmission System Planning Performance Requirements". *TPL-001-4* (2017).

[52] M. Padberg. "Approximating separable nonlinear functions via mixed zero-one programs". *Operations Research Letters* 27.1 (2000), pp. 1–5.

[53] K. Postek, D. den Hertog, and B. Melenberg. "Computationally tractable counterparts of distributionally robust constraints on risk measures". *SIAM Review* 58.4 (2016), pp. 603–650.

[54] J. Povh and F. Rendl. "Copositive and semidefinite relaxations of the quadratic assignment problem". *Discrete Optimization* 6.3 (2009), pp. 231–241.

[55] P. Praks, V. Kopustinskas, and M. Masera. "Monte-Carlo-based reliability and vulnerability assessment of a natural gas transmission system due to random network component failures". *Sustainable and Resilient Infrastructure* 2.3 (2017), pp. 97–107.

[56] H. Rahimian, G. Bayraksan, and T. Homem-de-Mello. "Identifying effective scenarios in distributionally robust stochastic programs with total variation distance". *Mathematical Programming* 173.1-2 (2019), pp. 393–430.

[57] H. Rahimian and S. Mehrotra. "Distributionally robust optimization: A review". *arXiv preprint arXiv:1908.05659* (2019).

[58] A. Shapiro. "Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming". *Available at Optimization Online* (2018).

[59] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[60] H. D. Sherali and W. P. Adams. "A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems". *SIAM Journal on Discrete Mathematics* 3.3 (1990), pp. 411–430.

[61] R. A. Stubbs and S. Mehrotra. "A branch-and-cut method for 0-1 mixed convex programming". *Mathematical Programming* 86 (1999), pp. 515–532.

[62] A. Thiele, T. Terry, and M. Epelman. *Robust linear optimization with recourse*. Tech. rep. Bethlehem, PA: Lehigh University, 2009.

[63] U.S. Department of Energy Advanced Research Projects Agency-Energy. "SCOPF Problem Formulation: Challenge 1". *Grid Optimization Competition* (2019).

[64] U.S. Energy Information Administration. "Annual Summaries". *Electric Disturbance Events (OE-417)* (2017). See also http://insideenergy.org/2014/08/18/data-explore-15-years-of-power-outages/.

[65] J. Vielma. "Mixed Integer Linear Programming Formulation Techniques". *SIAM Review* 57.1 (2015), pp. 3–57.

[66] W. Wiesemann, D. Kuhn, and M. Sim. "Distributionally robust convex optimization". *Operations Research* 62.6 (2014), pp. 1358–1376.

[67] D. Wozabal. "A framework for optimization under ambiguity". *Annals of Operations Research* 193.1 (2012), pp. 21–47.

[68] W. Xie. "Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball". *Operations Research Letters* 48.4 (2020), pp. 513–523.

[69] G. Xu and S. Burer. "A copositive approach for two-stage adjustable robust optimization with uncertain right-hand sides". *Computational Optimization and Applications* 70.1 (2018), pp. 33–59.

[70] B. Zeng and L. Zhao. "Solving two-stage robust optimization problems using a column-and-constraint generation method". *Operations Research Letters* 41.5 (2013), pp. 457–461.

[71] C. Zhao. "Data-driven risk-averse stochastic program and renewable energy integration". PhD thesis. University of Florida, 2014.

[72] C. Zhao and Y. Guan. "Data-driven risk-averse stochastic optimization with Wasserstein metric". *Operations Research Letters* 46.2 (2018), pp. 262–267.

[73] R. D. Zimmerman and C. E. Murillo-Sánchez. "MATPOWER 6.0 user's manual". *PSERC: Tempe, AZ, USA* (2016).

Appendix A Extended discussion of Example 1

In Section A.1, we analyze the optimal solution of the network optimization problem in Example 1 under the true failure distribution. In Section A.2, we analyze the solutions of the sample average and distributionally robust formulations when the empirical distribution puts all its weight on the realization $\boldsymbol{\xi} = \mathbf{0}$, where none of the nodes fail.

A.1 Optimal solution under the true distribution

Problem symmetry implies that the optimal arc capacities have the structure shown in Figure 13. Table 2 presents a calculation of the second-stage loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ for each $\boldsymbol{\xi} \in \{0,1\}^3$. The optimal arc capacities can then be determined by solving the two-dimensional piecewise linear convex optimization problem (note that $0 < \epsilon \ll 1$ can be any small positive constant):

$$\underset{\boldsymbol{x} \in \mathbb{R}^2_+}{\text{minimize}} \; 4x_1 + (4-\epsilon)x_2 + \sum_{\boldsymbol{\xi} \in \{0,1\}^3} \mathbb{P}[\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}] \mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$$

A straightforward calculation shows that the objective function minimized at $(x_1, x_2) = (100, 100)$.

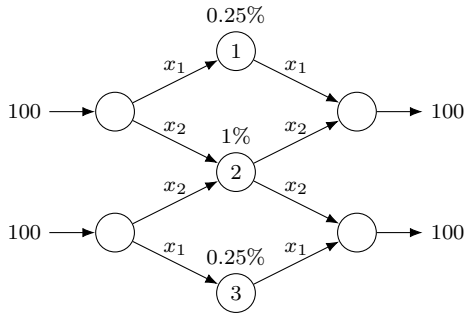


Figure 13. The structure of the optimal solution, with arc capacities x_1 and x_2 indicated above each arc.

A.2 Optimal solutions using the sample average and distributionally robust formulations

Suppose now that the empirical distribution $\hat{\mathbb{P}}_N = \delta_{\mathbf{0}}$; that is, it puts all its weight on the realization $\boldsymbol{\xi} = (0,0,0)$, where none of the nodes fail. The problem symmetry carries forth to the sample average approximation, and its optimal solution has the same symmetry as Figure 13. The solution

Table 2. The second-stage loss function $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ for each $\boldsymbol{\xi} \in \{0,1\}^3$. Here, we use $q = 1000$ to denote the penalty cost per unit of supply shortfall, and we use $[\cdot]_+ := \max\{\cdot, 0\}$.

$\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)$	$\|\boldsymbol{\xi}\|_1$	$\mathbb{P}[\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}]$	$\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$
$(1,1,1)$	3	6.3×10^{-8}	$200q$
$(1,0,1)$	2	6.2×10^{-6}	$2[100 - x_2]_+ q$
$(1,1,0)$	2	2.5×10^{-5}	$(100 + [100 - x_1]_+)\, q$
$(0,1,1)$	2	2.5×10^{-5}	$(100 + [100 - x_1]_+)\, q$
$(1,0,0)$	1	2.5×10^{-3}	$(100 - x_2 + [100 - x_1 - x_2]_+)\, q$
$(0,0,1)$	1	2.5×10^{-3}	$(100 - x_2 + [100 - x_1 - x_2]_+)\, q$
$(0,1,0)$	1	1.0×10^{-2}	$2[100 - x_1]_+ q$
$(0,0,0)$	0	9.6×10^{-1}	$2\,[100 - x_1 - x_2]_+\, q$

can be determined by solving the following problem:

$$\underset{\boldsymbol{x} \in \mathbb{R}^2_+}{\text{minimize}}\ 4x_1 + (4-\epsilon)x_2 + \mathcal{Q}(\boldsymbol{x}, \mathbf{0}) \;=\; \underset{\boldsymbol{x} \in \mathbb{R}^2_+}{\text{minimize}}\ 4x_1 + (4-\epsilon)x_2 + 2\,[100 - x_1 - x_2]_+\, q$$

Since $\epsilon > 0$, the objective function is minimized at $(x_1, x_2) = (0, 100)$; this is plotted in Figure 2a.

Consider now the distributionally robust formulation (2) using a Wasserstein ambiguity set (3) of radius $\varepsilon > 0$, and defined with respect to the metric $d(\boldsymbol{\xi}, \boldsymbol{\xi'}) = \|\boldsymbol{\xi} - \boldsymbol{\xi'}\|_1$. Again, problem symmetry leads to the same optimal solution structure as Figure 13. The optimal solution can be determined by solving formulation (7):

$$\underset{(x_1, x_2, \alpha) \in \mathbb{R}^3_+}{\text{minimize}}\ 4x_1 + (4-\epsilon)x_2 + \alpha\varepsilon + \max_{\boldsymbol{\xi} \in \{0,1\}^3} \{\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) - \alpha\|\boldsymbol{\xi}\|_1\}$$

For any $\varepsilon \in (0, 3)$, the minimizer is $(x_1, x_2, \alpha) = (\frac{2}{3}100, \frac{1}{3}100, \frac{2}{3}100q)$; this is plotted in Figure 3a.

If we change the definition of the metric to be $d(\boldsymbol{\xi'}, \boldsymbol{\xi''}) = 1$ whenever $\boldsymbol{\xi'} \neq \boldsymbol{\xi''}$ and 0 otherwise, then the ambiguity set is equivalent to the ϕ-divergence ambiguity set based on the total variation distance. The corresponding optimal solution can be determined by solving formulation (7):

$$\underset{(x_1, x_2, \alpha) \in \mathbb{R}^3_+}{\text{minimize}}\ 4x_1 + (4-\epsilon)x_2 + \alpha\varepsilon + \max\left\{\mathcal{Q}(\boldsymbol{x}, \mathbf{0}), \max_{\boldsymbol{\xi} \in \{0,1\}^3 \setminus \{\mathbf{0}\}} \mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) - \alpha\right\}$$

The innermost maximum is attained at $\boldsymbol{\xi} = \mathbf{e} = (1,1,1)$ irrespective of $\varepsilon, \boldsymbol{x}, \alpha$. Therefore, the outer maximum becomes $\max\{\mathcal{Q}(\boldsymbol{x}, \mathbf{0}), \mathcal{Q}(\boldsymbol{x}, \mathbf{e}) - \alpha\}$ and it is minimized at its break-point where

$\mathcal{Q}(\boldsymbol{x}, \mathbf{0}) = \mathcal{Q}(\boldsymbol{x}, \mathbf{e}) - \alpha$. The problem then becomes equivalent to the sample average approximation

$$\underset{(x_1,x_2,\alpha)\in\mathbb{R}^3_+}{\text{minimize}} \;\; 4x_1 + (4-\epsilon)x_2 + \alpha\varepsilon + \mathcal{Q}(\boldsymbol{x}, \mathbf{0})$$

and its minimum is attained at $(x_1, x_2, \alpha) = (0, 100, 0)$ for any $\varepsilon \geq 0$.

Appendix B Complexity analysis

The feasible region of the inner optimization problem in Theorem 2 is the convex hull of an MICP-representable set. This allows us to exploit existing tools on characterizing the convex hulls of MICP sets. Section B.1 highlights several problem instances (i.e., sufficient conditions) where this convex hull is efficiently computable and hence the distributionally robust two-stage problem (2) is tractable. In general, however, the presence of discrete random parameters makes the problem intractable even when the second-stage problem $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is benign. This is proved in Section B.2.

B.1 Tractable cases

We preface this subsection by noting that any notion of tractability of the distributionally robust two-stage problem (2) requires that optimizing $c(\boldsymbol{x})$ over $\boldsymbol{x} \in \mathcal{X}$ can be done in a tractable manner. We therefore assume this to be the case throughout this subsection.

Suppose that $\Xi = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(K)}\}$ has a known inner description, where the number of vertices K of Ξ grows polynomially with the dimension M. For example, this is the case of budget supports of the form $\Xi = \{\boldsymbol{\xi} \in \{0,1\}^M : \boldsymbol{e}^\top \boldsymbol{\xi} \leq k\}$, where k is a small, fixed input (say, ≤ 3) and $K = O(M^k)$. Now, if $\boldsymbol{\xi}$ is fixed to one of $\boldsymbol{v}^{(k)}$, $k \in [K]$, then the set $\mathcal{Z}_i$ in (9b) becomes a closed convex set. In other words, $\mathcal{Z}_i$ is the union of K closed convex sets, and its convex hull can be described compactly.

Proposition 2. *Assume $\Xi = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(K)}\}$, where K is a polynomial function of M. Then, the distributionally robust two-stage problem (2) is equivalent to the following tractable problem:*

$$\begin{array}{ll} \textit{minimize} & c(\boldsymbol{x}) + \alpha\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \sigma_i \\ \textit{subject to} & \boldsymbol{x} \in \mathcal{X},\ \alpha \geq 0 \\ & \left.\begin{array}{l} \sigma_i \in \mathbb{R},\ \boldsymbol{y}^{(k)} \in \mathcal{Y} \\ \sigma_i \geq \boldsymbol{q}(\boldsymbol{v}^{(k)})^\top \boldsymbol{y}^{(k)} - \alpha d(\boldsymbol{v}^{(k)}, \hat{\boldsymbol{\xi}}^{(i)}) \\ \boldsymbol{W}(\boldsymbol{v}^{(k)})\boldsymbol{y}^{(k)} \geq \boldsymbol{T}(\boldsymbol{x})\boldsymbol{v}^{(k)} + \boldsymbol{h}(\boldsymbol{x}) \end{array}\right\} \forall k \in [K], i \in [N]. \end{array}$$

Proof. From Lemma 1 and the stated hypothesis, problem (2) is equivalent to

$$\underset{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0}{\text{minimize}} \quad c(\boldsymbol{x}) + \alpha\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \max_{k\in[K]} \left\{ \mathcal{Q}(\boldsymbol{x}, \boldsymbol{v}^{(k)}) - \alpha d(\boldsymbol{v}^{(k)}, \hat{\boldsymbol{\xi}}^{(i)}) \right\}.$$

We can then introduce the epigraphical variables σ_i, $i \in [N]$ to model the ith inner maximization in the above sum. The stated reformulation then follows after introducing second-stage variables $\boldsymbol{y}^{(k)}$ to represent a minimizer of $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{v}^{(k)})$, $k \in [K]$. □

Suppose now that we have an outer description of $\Xi = \{\boldsymbol{\xi} \in \mathbb{Z}^M : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{f}\}$ and the second-stage problem $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is a linear program $\mathcal{Y} = \mathbb{R}^{N_2}_+$ with objective uncertainty so that Corollary 1 is applicable. The next observation relies on the concept of an *ideal formulation* [52, 65]. A mixed-integer linear set $\{\boldsymbol{x} \in \mathbb{Z}^p \times \mathbb{R}^{n-p} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ is said to be *ideal* if its convex hull has at least one vertex and it is equal to its linear relaxation. We show that the convex hull reformulation (8) becomes tractable if the set $\mathcal{Z}_i$ defined in (10b) is mixed-integer linear and ideal.

Proposition 3. *Suppose that $\Xi = \{\xi \in \mathbb{Z}^M : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{f}\}$ and $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is the optimal value of a linear program with uncertain objective coefficients: $\boldsymbol{W}(\boldsymbol{\xi}) = \boldsymbol{W}_0$ and $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{0}$. Suppose also that the mixed-integer linear formulation $\{(\boldsymbol{\xi}, \boldsymbol{\lambda}) \in \mathbb{Z}^M \times \mathbb{R}^L_+ : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{f},\ \boldsymbol{W}_0^\top\boldsymbol{\lambda} - \boldsymbol{Q}\boldsymbol{\xi} \leq \boldsymbol{q}_0\}$ is ideal.*

(i) If the reference metric $d(\boldsymbol{\xi}, \boldsymbol{\xi'}) = \|\boldsymbol{\xi} - \boldsymbol{\xi'}\|_1$ is the 1-norm, then the two-stage problem (2) is equivalent to the following tractable optimization problem:

$$\begin{array}{ll} \textit{minimize} & c(\boldsymbol{x}) + \alpha\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \left[\boldsymbol{q}_0^\top \boldsymbol{y}^{(i)} + \boldsymbol{f}^\top \boldsymbol{\eta}^{(i)} - \alpha \mathbf{e}^\top \hat{\boldsymbol{\xi}}^{(i)}\right] \\ \textit{subject to} & \boldsymbol{x} \in \mathcal{X},\ \alpha \geq 0 \\ & \left.\begin{array}{l} \boldsymbol{y}^{(i)} \in \mathcal{Y},\ \boldsymbol{\eta}^{(i)} \in \mathbb{R}^F_+ \\ -\boldsymbol{Q}^\top \boldsymbol{y}^{(i)} + \boldsymbol{E}^\top \boldsymbol{\eta}^{(i)} \geq -\alpha(\mathbf{e} - 2\hat{\boldsymbol{\xi}}^{(i)}) \\ \boldsymbol{W}_0 \boldsymbol{y}^{(i)} \geq \boldsymbol{h}(\boldsymbol{x}) \end{array}\right\} \ \forall i \in [N]. \end{array}$$

(ii) If $\varepsilon \geq \max_{\boldsymbol{\xi},\boldsymbol{\xi'}\in\Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'})$, then the two-stage (classical robust optimization) problem (2) is equivalent to the following tractable optimization problem:

$$\begin{array}{ll} \textit{minimize} & c(\boldsymbol{x}) + \boldsymbol{q}_0^\top \boldsymbol{y} + \boldsymbol{f}^\top \boldsymbol{\eta} \\ \textit{subject to} & \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{y} \in \mathcal{Y},\ \boldsymbol{\eta} \in \mathbb{R}^F_+ \\ & -\boldsymbol{Q}^\top \boldsymbol{y} + \boldsymbol{E}^\top \boldsymbol{\eta} \geq \boldsymbol{0} \\ & \boldsymbol{W}_0 \boldsymbol{y} \geq \boldsymbol{h}(\boldsymbol{x}) \end{array}$$

Proof. The conditions of this proposition allow us to use the result of Corollary 1.

(i) When the reference metric $d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_1$ is the 1-norm, the condition $(\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}, \tau) \in \mathcal{C}^{M+1}$ appearing in (10b) is equivalent to $\tau \geq \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}\|_1$. If we exploit the fact that $\boldsymbol{\xi}$ and $\hat{\boldsymbol{\xi}}^{(i)}$ are both binary-valued, then the right-hand side of this inequality is equal to $(\mathbf{e} - 2\hat{\boldsymbol{\xi}}^{(i)})^\top \boldsymbol{\xi} + \mathbf{e}^\top \hat{\boldsymbol{\xi}}^{(i)}$ and hence, linear in $\boldsymbol{\xi}$. Since the objective function of (10a) is linear in τ and τ only appears in a single constraint in $\mathcal{Z}_i$ that is linear in $\boldsymbol{\xi}$, we can reformulate (10a)–(10b) as follows:

$$
\begin{aligned}
Z_i(\boldsymbol{x}, \alpha) = \underset{(\boldsymbol{\xi}, \boldsymbol{\lambda}) \in \operatorname{cl}\operatorname{conv}(\hat{\mathcal{Z}})}{\text{maximize}} & \left\{ \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \alpha \left((\mathbf{e} - 2\hat{\boldsymbol{\xi}}^{(i)})^\top \boldsymbol{\xi} + \mathbf{e}^\top \hat{\boldsymbol{\xi}}^{(i)} \right) \right\}, \\
\hat{\mathcal{Z}} = & \left\{ (\boldsymbol{\xi}, \boldsymbol{\lambda}) \in \mathbb{Z}^M \times \mathbb{R}^L_+ : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{d},\ \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi} - \boldsymbol{W}_0^\top \boldsymbol{\lambda} \geq \mathbf{0} \right\},
\end{aligned}
$$

where $\Xi = \{\xi \in \mathbb{Z}^M : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{d}\}$ and $\mathcal{Y} = \mathbb{R}^{N_2}_+$. By hypothesis, this formulation of $\hat{\mathcal{Z}}$ is ideal, and hence $\operatorname{cl}\operatorname{conv}(\hat{\mathcal{Z}}) = \{(\boldsymbol{\xi}, \boldsymbol{\lambda}) \in \mathbb{R}^M \times \mathbb{R}^L_+ : \boldsymbol{E}\boldsymbol{\xi} \leq \boldsymbol{f},\ \boldsymbol{W}_0^\top \boldsymbol{\lambda} - \boldsymbol{Q}\boldsymbol{\xi} \leq \boldsymbol{q}_0\}$. We then obtain the stated result by exploiting strong linear programming duality.

(ii) When $\varepsilon \geq \max_{\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}')$, problem (2) reduces to a classical two-stage robust optimization as per Remark 1. In this case, the optimal value of α in the convex hull reformulation (8) is 0 and the two-stage problem (2) is equivalent to

$$
\underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}} \quad c(\boldsymbol{x}) + \hat{Z}(\boldsymbol{x}),
$$

where the set $\hat{\mathcal{Z}}$ is defined as above and we define the function $\hat{Z} : \mathcal{X} \mapsto \mathbb{R}$ as follows:

$$
\hat{Z}(\boldsymbol{x}) = \underset{(\boldsymbol{\xi}, \boldsymbol{\lambda}) \in \operatorname{cl}\operatorname{conv}(\hat{\mathcal{Z}})}{\text{maximize}} \ \boldsymbol{h}_0(\boldsymbol{x})^\top \boldsymbol{\lambda}.
$$

By the same argument as in part *(i)*, the above formulation of $\hat{\mathcal{Z}}$ is ideal, and we obtain the stated result using strong linear programming duality. □

A well-known sufficient condition that guarantees idealness of a mixed-integer linear formulation is *total unimodularity* of the constraint matrices. In particular, if $\boldsymbol{W}_0$, $\boldsymbol{Q}$ and $\boldsymbol{E}$ are totally unimodular (e.g., they are network matrices), then the conditions of Proposition 3 can be satisfied. We refer to [65] for a more general overview of ideal formulations.

B.2 Computational complexity

Even though the preceding subsection presented some sufficient conditions for tractability, the distributionally robust two-stage problem (2) is intractable even in benign settings.

Theorem 6 (NP-hardness)**.** *The distributionally robust two-stage problem* (2) *is strongly NP-hard even if there are no first-stage decisions* ($N_1 = 0$)*, the support* $\Xi = \{0,1\}^M$ *is the zero-one hypercube,* $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ *is the optimal value of a linear program* ($\mathcal{Y} = \mathbb{R}^{N_2}_+$)*, and either*

(i) uncertainty affects only the objective function, $\boldsymbol{W}(\boldsymbol{\xi}) = \boldsymbol{W}_0$ *and* $\boldsymbol{T}(\boldsymbol{x}) = \mathbf{0}$*, or*

(ii) uncertainty affects only the constraint right-hand sides: $\boldsymbol{W}(\boldsymbol{\xi}) = \boldsymbol{W}_0$*,* $\boldsymbol{Q} = \mathbf{0}$ *and* $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}_0$*.*

It remains NP-hard even if $N_2 = 2$ *in case (i).*

Proof. We prove part *(i)* by describing a polynomial reduction of the strongly NP-hard integer programming feasibility problem [31]:

> Given $\boldsymbol{Q} \in \mathbb{Z}^{N_2 \times M}$ and $\boldsymbol{q}_0 \in \mathbb{Z}^{N_2}$, is there a vector $\boldsymbol{\xi} \in \{0,1\}^M$ such that $\boldsymbol{Q}\boldsymbol{\xi} \geq -\boldsymbol{q}_0$?

We show that this problem has an affirmative answer if and only if the problem

$$\max_{\boldsymbol{\xi} \in \{0,1\}^M} \min_{\boldsymbol{y} \in \mathbb{R}^{N_2}_+} \left\{ (\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_0)^\top \boldsymbol{y} : \mathbf{e}^\top \boldsymbol{y} = 1 \right\}$$

has an optimal value greater than or equal to 0. This can be viewed as an instance of the two-stage problem (2) without first-stage decisions ($N_1 = 0$), see Remark 1. For fixed $\boldsymbol{\xi}$, the inner minimization evaluates to $\min\left\{\mathbf{e}_1^\top(\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_0), \ldots, \mathbf{e}_{N_2}^\top(\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_0)\right\}$, where $\mathbf{e}_j$ denotes the jth unit canonical vector in $\mathbb{R}^{N_2}$. Therefore, the optimal value of the above problem is greater than or equal to 0 if and only if there exists $\boldsymbol{\xi} \in \{0,1\}^M$ such that $\min\left\{\mathbf{e}_1^\top(\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_0), \ldots, \mathbf{e}_{N_2}^\top(\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_0)\right\} \geq 0$, that is, if and only if $\mathbf{e}_j^\top(\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_0) \geq 0$ for all $j \in [N_2]$, that is, if and only if $\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_0 \geq \mathbf{0}$.

To prove part *(ii)*, we recall the strongly NP-hard problem of convex quadratic maximization over the unit hypercube [24]:

> Given a symmetric positive semidefinite matrix $\boldsymbol{T}_0 \in \mathbb{R}^{M \times M}$ and a scalar $t_0 \geq 0$, is there a vector $\boldsymbol{\xi} \in [0,1]^M$ such that $\boldsymbol{\xi}^\top \boldsymbol{T}_0 \boldsymbol{\xi} \geq t_0$?

We show that this problem has an affirmative answer if and only if the problem

$$\max_{\boldsymbol{\xi} \in \{0,1\}^M} \min_{\boldsymbol{y} \in \mathbb{R}^M_+} \left\{ \mathbf{e}^\top \boldsymbol{y} : \boldsymbol{y} \geq \boldsymbol{T}_0 \boldsymbol{\xi} \right\}$$

has an optimal value greater than or equal to h_0. As before, this problem is an instance of the two-stage problem (2) with $\boldsymbol{W}_0 = \mathbf{I}$ and where the uncertainty affects only the right-hand sides of

the second-stage problem. By strong linear programming duality, it is equivalent to

$$\max_{\substack{\boldsymbol{\xi}\in\{0,1\}^M \\ \boldsymbol{\lambda}\in[0,1]^M}} \boldsymbol{\xi}^\top \boldsymbol{T}_0 \boldsymbol{\lambda} = \max_{\substack{\boldsymbol{\xi}\in[0,1]^M \\ \boldsymbol{\lambda}\in[0,1]^M}} \boldsymbol{\xi}^\top \boldsymbol{T}_0 \boldsymbol{\lambda},$$

where the equality follows from the fact that there always exists an optimal vertex solution of the bilinear program on the right-hand side of the equality [39, Theorem 2.1]. The claim now follows from the fact that the optimal value of the right-hand side bilinear program is greater than or equal to t_0 if and only if there exists a vector $\boldsymbol{\xi} \in [0,1]^M$ such that $\boldsymbol{\xi}^\top \boldsymbol{T}_0 \boldsymbol{\xi} \geq t_0$ [39, Theorem 2.2].

To show NP-hardness in part *(i)* even when $N_2 = 2$, we describe a polynomial reduction from the weakly NP-hard subset sum problem [31]:

> Given $\boldsymbol{q} \in \mathbb{Z}^M$ and $q_0 \in \mathbb{Z}$, is there a subset $J \subseteq [M]$ such that $\sum_{j\in J} q_j = q_0$?

We show that this problem has an affirmative answer if and only if the problem

$$\max_{\boldsymbol{\xi}\in\{0,1\}^M} \min_{\boldsymbol{y}\in\mathbb{R}^2_+} \left\{ (\boldsymbol{q}^\top \boldsymbol{\xi} - q_0)(y_1 - y_2) : y_1 + y_2 = 1 \right\}$$

has an optimal value greater than or equal to 0. As before, this problem can be viewed as a special case of the two-stage problem (2) with objective uncertainty. For fixed $\boldsymbol{\xi}$, the inner minimization evaluates to $-|\boldsymbol{q}^\top \boldsymbol{\xi} - q_0|$ which is always non-positive. Therefore, the optimal value is greater than or equal to 0 if and only if there exists a vector $\boldsymbol{\xi} \in \{0,1\}^M$ such that $|\boldsymbol{q}^\top \boldsymbol{\xi} - q_0| = 0$, that is, if and only if there exists a subset $J = \{j \in [M] : \xi_j = 1\}$ such that $\sum_{j\in J} q_j = q_0$. □

Appendix C Benders decomposition

The central idea in Benders decomposition is to solve the convex hull reformulation (8) by iteratively refining an inner approximation of the value function $Z_i(\boldsymbol{x}, \alpha)$, for each $i \in [N]$. We generically write the latter as $Z_i(\boldsymbol{x}, \alpha) = \max_{\boldsymbol{z}\in\mathcal{Z}_i} \boldsymbol{\gamma}(\boldsymbol{x}, \alpha)^\top \boldsymbol{z}$. Recall that the latter optimization problem can be formulated as an MICP, in one of two equivalent forms, by using the linearized reformulation (see Section 3.1) or the penalty reformulation (see Section 3.2). To present the Benders decomposition algorithm, we re-write (8) as a semi-infinite program:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathcal{X},\ \alpha\geq 0}{\text{minimize}} \quad & c(\boldsymbol{x}) + \alpha\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \sigma_i \\ \text{subject to} \quad & \sigma_i \geq \boldsymbol{\gamma}(\boldsymbol{x}, \alpha)^\top \boldsymbol{z}, \quad \forall \boldsymbol{z} \in \mathcal{Z}_i, i \in [N]. \end{aligned}$$

Observe that, in both cases where an MICP representation is possible, $\boldsymbol{\gamma}(\boldsymbol{x}, \alpha)$ is componentwise convex and $\mathcal{Z}_i \subseteq \mathbb{R}^n_+$; therefore, each of the semi-infinite constraints in the above problem defines a convex feasible region (in $\boldsymbol{x}$, α and $\boldsymbol{\sigma}$). We present the algorithm next.

1. Initialize $\hat{\mathcal{Z}}_i = \emptyset$, for each $i \in [N]$.

2. Solve the following *master problem*.

$$\begin{array}{ll} \underset{\boldsymbol{x}\in\mathcal{X},\, \alpha\geq 0,\, \boldsymbol{\sigma}\in\mathbb{R}^N}{\text{minimize}} & c(\boldsymbol{x}) + \alpha\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \sigma_i \\ \text{subject to} & \sigma_i \geq \max\left\{\mathcal{Q}(\boldsymbol{x}, \hat{\boldsymbol{\xi}}^{(i)}), \boldsymbol{\gamma}(\boldsymbol{x}, \alpha)^\top \boldsymbol{z}\right\}, \quad \forall \boldsymbol{z} \in \hat{\mathcal{Z}}_i, i \in [N]. \end{array} \tag{20}$$

   Let $(\boldsymbol{x}^\star, \alpha^\star, \boldsymbol{\sigma}^\star)$ denote an optimal solution.

3. Solve the following *subproblem*, for each $i \in [N]$:

$$\underset{\boldsymbol{z}\in\mathcal{Z}_i}{\text{maximize}}\ \boldsymbol{\gamma}(\boldsymbol{x}^\star, \alpha^\star)^\top \boldsymbol{z} \tag{21}$$

   Let $\boldsymbol{z}^{\star,i}$ denote an optimal solution.

4. For each $i \in [N]$, if $\boldsymbol{\gamma}(\boldsymbol{x}^\star, \alpha^\star)^\top \boldsymbol{z}^{\star,i} > \sigma_i^\star$, add $\boldsymbol{z}^{\star,i}$ to $\hat{\mathcal{Z}}_i$.

   If $\hat{\mathcal{Z}}_i$ was not updated for any $i \in [N]$, stop. Otherwise, go to Step 2.

We make some remarks about the algorithm next.

- The master problem (20) is always feasible. Indeed, its optimal value always constitutes a lower bound to the optimal value of the distributionally robust two-stage problem, which always exists and is finite (see Section 3). Similarly, the term $\mathcal{Q}(\boldsymbol{x}, \hat{\boldsymbol{\xi}}^{(i)})$ in the constraint ensures that it is also always bounded.

- The optimal value of the subproblem (21) also always exists and is finite, because of the assumption of complete recourse.

- The subproblem (21) can be solved as an MICP by using either the McCormick linearization or the penalty-based formulation from Section 3.

- The computational efficiency of the algorithm can be improved in several ways. First, the solution of the subproblems in Step 3 can be carried out in parallel, if desired. Second, we

don't need to solve the subproblems to global optimality; we can stop as soon as we find a solution $\boldsymbol{z}^{\dagger,i}$ that satisfies $\boldsymbol{\gamma}(\boldsymbol{x}^\star,\alpha^\star)^\top \boldsymbol{z}^{\dagger,i} > \sigma_i^\star$. Third, for the same reason, we can employ $\sigma^\star$ as a lower bound when solving the subproblem for $i \in [N]$.

We note that convergence of the algorithm is guaranteed only asymptotically in general. Finite convergence is guaranteed only if the feasible region of the second-stage problem is a linear program, that is, $\mathcal{Y} = \mathbb{R}^{N_2}_+$; and Step 3 solves the resulting mixed-integer linear programs to global optimality. The reason is the finite number of extreme points of the sets $\mathcal{Z}_i$, $i \in [N]$, in this case.

Appendix D Extension to risk-averse objective functions

The objective function of the distributionally robust two-stage problem (2) minimizes the worst-case expectation of the loss and reflects a risk-neutral approach. In the context of rare high-impact events, where the loss function increases sharply with extreme realizations of the uncertainty, it might be preferable to adopt a risk-averse approach and minimize the tail of the distribution of the random loss. A natural risk measure in this case is the conditional value-at-risk, which is the conditional expectation above the $(1-p)$-quantile of the random loss function $\mathcal{Q}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$:

$$\mathbb{P}\text{-CVaR}_p\left[\mathcal{Q}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] = \inf_{w\in\mathbb{R}} w + \frac{1}{p}\mathbb{E}_{\mathbb{P}}\left[\max\left\{\mathcal{Q}(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) - w, 0\right\}\right]. \tag{22}$$

In this subsection, we show that convex hull reformulation of Theorem 2 also extends to this setting. Specifically, we study the following distributionally robust two-stage risk-averse stochastic program:

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ c(\boldsymbol{x}) + \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\text{-CVaR}_p\left[\mathcal{Q}(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] \tag{23}$$

Theorem 7 (Convex hull reformulation for conditional-value-at-risk)**.** *For any $p \in (0,1]$, the distributionally robust two-stage risk-averse stochastic program (23) admits the following reformulation,*

$$\underset{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0,w\in\mathbb{R}}{\textit{minimize}}\ c(\boldsymbol{x}) + \alpha\varepsilon + w + \frac{1}{N}\sum_{i=1}^{N} Z_i(\boldsymbol{x},\alpha,w), \tag{24}$$

where, for each $i \in [N]$, we define the function $Z_i : \mathcal{X} \times \mathbb{R}_+ \times \mathbb{R} \mapsto \mathbb{R}$ and the set $\mathcal{Z}_i$ as follows:

$$Z_i(\boldsymbol{x}, \alpha, w) = \underset{(\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\Lambda},\tau,\theta)\in \operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)}{\text{maximize}} \left\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda} \rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} - w\theta - \alpha\tau \right\} \tag{25a}$$

$$\mathcal{Z}_i = \left\{ (\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \tau, \theta) \in \Xi \times \mathbb{R}^L_+ \times \mathbb{R}^{L\times M} \times \mathbb{R}_+ \times [0, p^{-1}] : \begin{array}{l} \boldsymbol{\Lambda} = \boldsymbol{\lambda}\boldsymbol{\xi}^\top,\ (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}, \tau) \in \mathcal{C}^{M+1} \\ (\boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi})\theta - \boldsymbol{W}_0^\top \boldsymbol{\lambda} - \sum_{j\in[M]} \boldsymbol{W}_j^\top \boldsymbol{\Lambda} \mathbf{e}_j \in \mathcal{Y}^* \end{array} \right\}. \tag{25b}$$

Proof. Substituting (22) into (23), we obtain

$$\underset{\boldsymbol{x}\in\mathcal{X}, w\in\mathbb{R}}{\text{minimize}}\ c(\boldsymbol{x}) + w + \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\frac{1}{p} \max\left\{ \mathcal{Q}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - w, 0 \right\} \right].$$

We can thus interpret the above problem as an instance of the two-stage problem (2) where the second-stage loss function is given by $\frac{1}{p}\max\{\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) - w, 0\}$. This, in turn, is equivalent to $\max_{\theta\in[0,p^{-1}]} \theta \cdot \{\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) - w\} = \max_{\theta\in[0,p^{-1}]} \{\theta\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) - \theta w\}$. By definition of the loss function, the first term is equal to $\inf_{\boldsymbol{y}\in\mathcal{Y}} \{\theta \boldsymbol{q}(\boldsymbol{\xi})^\top \boldsymbol{y} : \boldsymbol{W}(\boldsymbol{\xi})\boldsymbol{y} \geq \boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi} + \boldsymbol{h}(\boldsymbol{x})\}$. An application of Theorem 2 to this instance of problem (2) leads to the stated result. Details are omitted for brevity. □

We note that the sets $\mathcal{Z}_i$, $i \in [N]$ appearing in the statement of Theorem 7 are similar to those appearing in Theorem 2 except for the additional variable θ that multiplies the objective coefficients $(\boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi})$. This can be exactly linearized by using McCormick inequalities since $\boldsymbol{\xi}$ is binary-valued and θ is bounded. The other considerations remain exactly the same as in Sections 3.1 and 3.2. Therefore, we obtain MICP-representable sets $\mathcal{Z}_i$ even in the risk-averse setting.

Appendix E Two-stage optimal power flow model

Our presentation of the first-stage model closely follows [38], whereas the second-stage model is inspired by [63]. Conceptually, the first-stage problem determines power generation levels in the so-called *base case*, where there are no line outages. After transmission lines fail, the second-stage model may adjust the first-stage power generation levels subject to physical and engineering constraints where failed lines cannot be used.

Let $\mathcal{G}, \mathcal{B}$, and $\mathcal{M}$ be the set of generators, buses, and transmission lines, respectively, and let $\mathcal{G}_i$ be the set of generators associated with bus $i \in \mathcal{B}$. We define $\delta(i) := \{j \in \mathcal{B} : (i, j) \in \mathcal{M} \text{ or } (j, i) \in \mathcal{M}\}$ to be the set of neighbors of bus $i \in \mathcal{B}$. Let p^g_k and q^g_k be the real and reactive power output of

generator $k \in \mathcal{G}$, respectively, with lower and upper bounds denoted by $p_k^{\min}, p_k^{\max}$ and $q_k^{\min}, q_k{}^{max}$. We assume a linear cost c_k of power generation for generator $k \in \mathcal{G}$. Real load and reactive load at bus $i \in \mathcal{B}$ are denoted by p_i^d and q_i^d, respectively, and are known data. Let p_{ij}^F and q_{ij}^F be the real and reactive power flow on line (i,j), respectively, defined for $(i,j) \in \mathcal{M}$ and $(j,i) \in \mathcal{M}$, with line rating limit $f_{ij}^{\max}$ (note that $f_{ij}^{\max} = f_{ji}^{\max}$). Let $\boldsymbol{Y}$ be the $|B| \times |B|$ complex-valued nodal admittance matrix, whose components are $Y_{ij} = G_{ij} + \mathrm{i}B_{ij}$, $\mathrm{i} = \sqrt{-1}$, where G_{ij} and B_{ij} are the conductance and susceptance of line $(i,j) \in \mathcal{M}$, respectively (see [73] for details on computing $\boldsymbol{Y}$). We denote the real and imaginary parts of the complex voltage by e_i and f_i, respectively. As in [38], we define new variables such that $c_{ii} = e_i^2 + f_i^2$, $c_{ij} = e_i e_j + f_i f_j$ and $s_{ij} = e_i f_j - e_j f_i$. We define $\tilde{\boldsymbol{\xi}}$ to be a random binary vector with support $\Xi = \{0,1\}^{|\mathcal{M}|}$, where $\tilde{\xi}_{ij} = 1$ if line $(i,j) \in \mathcal{M}$ fails and 0 otherwise. We have $\boldsymbol{x} = \left(\boldsymbol{p^g},\ \boldsymbol{q^g},\ \boldsymbol{p^F},\ \boldsymbol{q^F},\ \boldsymbol{c},\ \boldsymbol{s},\ \boldsymbol{\sigma^{p+}},\ \boldsymbol{\sigma^{p-}},\ \boldsymbol{\sigma^{q+}},\ \boldsymbol{\sigma^{q-}}\right)$ as first-stage variables and $\boldsymbol{y} = \left(\tilde{\boldsymbol{\delta}},\ \tilde{\boldsymbol{p}}^{\boldsymbol{g}},\ \tilde{\boldsymbol{q}}^{\boldsymbol{g}},\ \tilde{\boldsymbol{p}}^{\boldsymbol{F}},\ \tilde{\boldsymbol{q}}^{\boldsymbol{F}},\ \tilde{\boldsymbol{c}},\ \tilde{\boldsymbol{s}},\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{p+}},\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{p-}},\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{q+}},\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{q-}},\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{pF}},\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{qF}}\right)$ as second-stage variables. The two-stage model can be written as follows:

$$\underset{\substack{\boldsymbol{p^g},\boldsymbol{q^g},\boldsymbol{p^F},\boldsymbol{q^F},\boldsymbol{c},\boldsymbol{s}\\ \boldsymbol{\sigma^{p+}},\boldsymbol{\sigma^{p-}},\boldsymbol{\sigma^{q+}},\boldsymbol{\sigma^{q-}}}}{\text{minimize}} \quad \sum_{k\in\mathcal{G}} c_k p_k^g + \sum_{i\in\mathcal{B}} g_i\left(\sigma_i^{p+} + \sigma_i^{p-} + \sigma_i^{q+} + \sigma_i^{q-}\right) + \mathbb{E}_{\mathbb{P}}\left[\mathcal{Q}(\boldsymbol{p^g}, \tilde{\boldsymbol{\xi}})\right]$$

$$\text{subject to} \quad \sum_{k\in\mathcal{G}_i} p_k^g - p_i^d + \sigma_i^{p+} - \sigma_i^{p-} = g_{ii}c_{ii} + \sum_{j\in\delta(i)} p_{ij}^F, \qquad \forall i \in \mathcal{B}, \tag{26a}$$

$$\sum_{k\in\mathcal{G}_i} q_k^g - q_i^d + \sigma_i^{q+} - \sigma_i^{q-} = -b_{ii}c_{ii} + \sum_{j\in\delta(i)} q_{ij}^F, \qquad \forall i \in \mathcal{B}, \tag{26b}$$

$$p_{ij}^F = -G_{ij}c_{ii} + G_{ij}c_{ij} + B_{ij}s_{ij}, \qquad \forall (i,j),(j,i) \in \mathcal{M}, \tag{26c}$$

$$q_{ij}^F = B_{ij}c_{ii} - B_{ij}c_{ij} + G_{ij}s_{ij}, \qquad \forall (i,j),(j,i) \in \mathcal{M}, \tag{26d}$$

$$c_{ij} = c_{ji},\ s_{ij} = -s_{ji}, \qquad \forall (i,j) \in \mathcal{M}, \tag{26e}$$

$$c_{ij}^2 + s_{ij}^2 + \left(\frac{c_{ii} - c_{jj}}{2}\right)^2 \le \left(\frac{c_{ii} + c_{jj}}{2}\right)^2, \qquad \forall (i,j) \in \mathcal{M}, \tag{26f}$$

$$\underline{V}_i^2 \le c_{ii} \le \bar{V}_i^2, \qquad \forall i \in \mathcal{B}, \tag{26g}$$

$$p_k^{\min} \le p_k^g \le p_k^{\max}, \qquad \forall k \in \mathcal{G}, \tag{26h}$$

$$q_k^{\min} \le q_k^g \le q_k^{\max}, \qquad \forall k \in \mathcal{G}, \tag{26i}$$

$$(p_{ij}^F)^2 + (q_{ij}^F)^2 \le (f_{ij}^{\max})^2, \qquad \forall (i,j),(j,i) \in \mathcal{M}, \tag{26j}$$

$$\sigma_i^{p+}, \sigma_i^{p-}, \sigma_i^{q+}, \sigma_i^{q-} \ge 0, \qquad \forall i \in \mathcal{B}, \tag{26k}$$

where $\mathcal{Q}(\boldsymbol{p^g}, \tilde{\boldsymbol{\xi}})$ is the optimal value of

$$\begin{aligned}
\underset{\substack{\tilde{\boldsymbol{\delta}},\tilde{\boldsymbol{p}}^{\boldsymbol{g}},\tilde{\boldsymbol{q}}^{\boldsymbol{g}},\tilde{\boldsymbol{p}}^{\boldsymbol{F}},\tilde{\boldsymbol{q}}^{\boldsymbol{F}},\tilde{\boldsymbol{c}},\tilde{\boldsymbol{s}} \\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{p+}},\tilde{\boldsymbol{\sigma}}^{\boldsymbol{p-}},\tilde{\boldsymbol{\sigma}}^{\boldsymbol{q+}},\tilde{\boldsymbol{\sigma}}^{\boldsymbol{q-}} \\ \tilde{\boldsymbol{\sigma}}^{\boldsymbol{pF}},\tilde{\boldsymbol{\sigma}}^{\boldsymbol{qF}}}}{\text{minimize}} \quad & \sum_{i\in\mathcal{B}} g_i \left(\tilde{\sigma}_i^{p+} + \tilde{\sigma}_i^{p-} + \tilde{\sigma}_i^{q+} + \tilde{\sigma}_i^{q-}\right) \\
\text{subject to} \quad & \tilde{p}_k^g = p_k^g + \Delta_k \tilde{\delta} && \forall k \in \mathcal{G}, && \text{(27a)} \\
& \sum_{k\in\mathcal{G}_i} \tilde{p}_k^g - p_i^d + \tilde{\sigma}_i^{p+} - \tilde{\sigma}_i^{p-} = g_{ii}\tilde{c}_{ii} + \sum_{j\in\delta(i)} \tilde{p}_{ij}^F, && \forall i \in \mathcal{B}, && \text{(27b)} \\
& \sum_{k\in\mathcal{G}_i} \tilde{q}_k^g - q_i^d + \tilde{\sigma}_i^{q+} - \tilde{\sigma}_i^{q-} = -b_{ii}\tilde{c}_{ii} + \sum_{j\in\delta(i)} \tilde{q}_{ij}^F, && \forall i \in \mathcal{B}, && \text{(27c)} \\
& \tilde{p}_{ij}^F = -G_{ij}\tilde{c}_{ii} + G_{ij}\tilde{c}_{ij} + B_{ij}\tilde{s}_{ij} + \tilde{\sigma}_{ij}^{pF}, && \forall (i,j),(j,i) \in \mathcal{M}, && \text{(27d)} \\
& \tilde{q}_{ij}^F = B_{ij}\tilde{c}_{ii} - B_{ij}\tilde{c}_{ij} + G_{ij}\tilde{s}_{ij} + \tilde{\sigma}_{ij}^{qF}, && \forall (i,j),(j,i) \in \mathcal{M}, && \text{(27e)} \\
& \tilde{c}_{ij} = \tilde{c}_{ji}, \;\; \tilde{s}_{ij} = -\tilde{s}_{ji}, && \forall (i,j) \in \mathcal{M}, \\
& \tilde{c}_{ij}^2 + \tilde{s}_{ij}^2 + \left(\frac{\tilde{c}_{ii} - \tilde{c}_{jj}}{2}\right)^2 \leq \left(\frac{\tilde{c}_{ii} + \tilde{c}_{jj}}{2}\right)^2, && \forall (i,j) \in \mathcal{M}, \\
& \underline{V}_i^2 \leq \tilde{c}_{ii} \leq \bar{V}_i^2, && \forall i \in \mathcal{B}, \\
& p_k^{\min} \leq \tilde{p}_k^g \leq p_k^{\max}, && \forall k \in \mathcal{G}, \\
& q_k^{\min} \leq \tilde{q}_k^g \leq q_k^{\max}, && \forall k \in \mathcal{G}, \\
& (\tilde{p}_{ij}^F)^2 + (\tilde{q}_{ij}^F)^2 \leq (f_{ij}^{\max})^2, && \forall (i,j),(j,i) \in \mathcal{M}, \\
& \tilde{\xi}_{ij} = 1 \implies \left[\tilde{p}_{ij}^F = 0, \; \tilde{q}_{ij}^F = 0\right], && \forall (i,j),(j,i) \in \mathcal{M}, && \text{(27f)} \\
& \tilde{\xi}_{ij} = 0 \implies \left[\tilde{\sigma}_{ij}^{pF} = 0, \; \tilde{\sigma}_{ij}^{qF} = 0\right], && \forall (i,j),(j,i) \in \mathcal{M}, && \text{(27g)} \\
& \tilde{\sigma}_i^{p+}, \tilde{\sigma}_i^{p-}, \tilde{\sigma}_i^{q+}, \tilde{\sigma}_i^{q-} \geq 0, && \forall i \in \mathcal{B}.
\end{aligned}$$

Constraints (26a) and (26b) are the real and reactive power balance equations, respectively. Constraints (26c) and (26d) define the real and reactive power flow in both directions of all lines, respectively. Constraints (26e) and (26f) model the change of variables (see [38] for details), where the latter is the result of convexifying the original constraint $c_{ij}^2 + s_{ij}^2 = c_{ii}c_{jj}$. Constraints (26g), (26h) and (26i) enforce bounds on the voltage magnitude, and the real and reactive power generation, respectively. In the first stage, each generator $k \in \mathcal{G}$ has an associated generation cost c_k, and we penalize violating constraints (26a) and (26b) by g_i.

The second stage involves the same constraints with a few modifications. Namely, constraint (27a) adjusts the first-stage real power generation, where all generators are adjusted by a constant $\tilde{\delta}$, scaled by their predefined automatic generation control participation factor Δ_k, also known as the droop control policy. We set participation factors Δ_k to be proportional to the generation

capacity for each generator $k \in \mathcal{G}$. Constraints (27f) ensure that no power can flow through lines under contingency ($\xi_{ij} = 1$). Note that if line (i, j) fails, then we must have $\tilde{p}^F_{ij} = \tilde{p}^F_{ji} = 0$ and $\tilde{q}^F_{ij} = \tilde{q}^F_{ji} = 0$, but variables $\tilde{c}_{ii}, \tilde{c}_{ij}$ and $\tilde{s}_{ij}$ should not be affected. Thus, unlike in the first stage, slacks $\tilde{\sigma}^{pF}_{ij}$ and $\tilde{\sigma}^{qF}_{ij}$ are added in constraints (27d) and (27e), respectively, so that (27d) and (27e) become redundant for (i, j) and (j, i) if line (i, j) has failed. These slacks are active only if $\xi_{ij} = 1$, enforced by constraints (27g). As in the first stage, the absolute values of the real and reactive power balance violation $\tilde{\sigma}^{p+}_i + \tilde{\sigma}^{p-}_i$ and $\tilde{\sigma}^{q+}_i + \tilde{\sigma}^{q-}_i$ for bus $i \in \mathcal{B}$, respectively, are penalized by g_i. Note that there is not cost on power generation in the second stage. We set the cost g_i of violating the balance equations to be $\phi \cdot \max_{k\in\mathcal{G}} c_k$ (see Section 5.1.1).

Appendix F Two-stage multi-commodity network design model

Our model is based on the fixed-charge multi-commodity network design problem which has been extensively studied in the literature, such as in [16] and [23]. We are given a directed network with nodes $\mathcal{V}$, arcs $\mathcal{A}$, and a set of commodities $\mathcal{K}$ with known demands d^k. For commodity $k \in \mathcal{K}$, let O_k be the origin node and D_k the destination node. Each arc (i, j) has an associated maximum capacity u_{ij}, per-unit cost of installing capacity f_{ij} and per-unit cost of flow c_{ij}. For each node $i \in \mathcal{V}$, we define the set of neighboring nodes incident to outgoing and incoming arcs as $\mathcal{V}^+(i) = \{j \mid (i, j) \in \mathcal{A}\}$ and $\mathcal{V}^-(i) = \{j \mid (j, i) \in \mathcal{A}\}$, respectively.

In the first stage, let x_{ij} be a variable between 0 and 1 which determines the fraction of capacity of arc (i, j) that can be used in the second stage. Note that this differs from a typical fixed-charge multi-commodity network design model, where x_{ij} is a binary variable determining whether arc (i, j) can be used, but becomes the same problem considered in Example 1. In the second stage, let y^k_{ij} be the amount of flow of commodity k on arc (i, j), and let σ_k be the unsatisfied demand of commodity k which we penalize by g_k. We consider node failures, and define $\boldsymbol{\xi}$ to be a binary vector where $\xi_i = 1$ indicates that node $i \in \mathcal{V}$ has failed. As in Example 1, if a node fails, all arcs incident to it cannot be used. The model can be as written as follows:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \sum_{(i,j)\in\mathcal{A}} f_{ij}x_{ij} + \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\mathcal{Q}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \\ \text{subject to} \quad & x_{ij} \in [0, 1], \ \forall (i, j) \in \mathcal{A}. \end{aligned}$$

where $\mathcal{Q}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is defined as the optimal objective value of the following linear program:

$$\underset{y,\sigma}{\text{minimize}} \quad \sum_{(i,j)\in\mathcal{A}} \sum_{k\in\mathcal{K}} c_{ij} y_{ij}^k + \sum_{k\in\mathcal{K}} g_k \sigma_k$$

$$\text{subject to} \quad \sum_{j\in\mathcal{V}^+(i)} y_{ij}^k - \sum_{j\in\mathcal{V}^-(i)} y_{ji}^k = \begin{cases} d^k - \sigma_k, & \text{if } i = O_k \\ -d^k + \sigma_k, & \text{if } i = D_k \\ 0, & \text{otherwise} \end{cases}, \ \forall k \in \mathcal{K}, \ \forall i \in \mathcal{V}, \tag{28a}$$

$$\sum_{k\in\mathcal{K}} y_{ij}^k \le u_{ij} x_{ij}, \qquad \forall (i,j) \in \mathcal{A}, \tag{28b}$$

$$y_{ij}^k \le d^k (1 - \tilde{\xi}_i), \qquad \forall (i,j) \in \mathcal{A}, \ \forall k \in \mathcal{K}, \tag{28c}$$

$$y_{ij}^k \le d^k (1 - \tilde{\xi}_j), \qquad \forall (i,j) \in \mathcal{A}, \ \forall k \in \mathcal{K}, \tag{28d}$$

$$y_{ij}^k \ge 0 \qquad \forall (i,j) \in \mathcal{A}, \ \forall k \in \mathcal{K},$$

$$\sigma_k \ge 0 \qquad \forall k \in \mathcal{K}.$$

In the second stage, constraints (28a) are typical network flow constraints, with added non-negative variables σ_k which model unsatisfied demand. If we have a positive amount σ_k of unsatisfied demand for commodity k, then an amount of $d^k - \sigma_k$ would leave the origin node and enter the destination node. Note that variables σ_k ensure feasibility in the second stage for any $\boldsymbol{x}$ and $\boldsymbol{\xi}$. For each commodity k, unsatisfied demand σ_k is penalized by g_k. In our experiments, g_k is constant and set to $\phi \cdot \max_{(i,j)\in A} c_{ij}$, where $\phi = 1000$ is a pre-defined non-negative parameter. Constraint (28b) ensures the total flow on any arc (i,j) does not exceed the available capacity, which is determined by the first-stage decision x_{ij}. Finally, constraints (28c) and (28d) ensure there is no flow through arcs incident to a failed node.

Appendix G Proofs

Proof of Theorem 1. First, note that for any $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\Xi)$, we have

$$d_W(\mathbb{Q}, \mathbb{P}) \le D d_{TV}(\mathbb{Q}, \mathbb{P}) \le D\sqrt{d_{KL}(\mathbb{Q}, \mathbb{P})/2},$$

where $d_{TV}(\mathbb{Q}, \mathbb{P}) \coloneqq \frac{1}{2} \sup_{\boldsymbol{\xi}\in\Xi} |\mathbb{P}(\boldsymbol{\xi}) - \mathbb{Q}(\boldsymbol{\xi})|$ and $d_{KL}(\mathbb{Q}, \mathbb{P}) \coloneqq \sum_{\boldsymbol{\xi}\in\Xi} \mathbb{Q}(\boldsymbol{\xi}) \log\left(\mathbb{Q}(\boldsymbol{\xi})/\mathbb{P}(\boldsymbol{\xi})\right)$ are the *total variation* and *Kullback-Liebler* distance between $\mathbb{Q}$ and $\mathbb{P}$, respectively. The first inequality follows from $d(\boldsymbol{\xi}', \boldsymbol{\xi}'') \le D\mathbb{I}[\boldsymbol{\xi}' \ne \boldsymbol{\xi}'']$, whereas the second inequality is Pinkser's inequality. Now

define $\Pi := \{\mathbb{Q} \in \mathcal{M}(\Xi) : d_W(\mathbb{Q}, \mathbb{P}) > \varepsilon_N(\beta)\}$, where $\mathbb{P}$ is the true (unknown) distribution, and $\varepsilon_N(\beta)$ satisfies (5). Since all $\mathbb{Q} \in \Pi$ satisfy $d_W(\mathbb{Q}, \mathbb{P}) > D\sqrt{(2N)^{-1}\big(|\Xi| \log(N+1) + \log \beta^{-1}\big)}$ by construction, we have:

$$\begin{aligned} & d_{KL}(\mathbb{Q}, \mathbb{P}) > N^{-1}\big(|\Xi| \log(N+1) + \log \beta^{-1}\big) \quad \forall \mathbb{Q} \in \Pi, \\ \iff & \inf_{\mathbb{Q} \in \Pi} d_{KL}(\mathbb{Q}, \mathbb{P}) > N^{-1}\big(|\Xi| \log(N+1) + \log \beta^{-1}\big) \\ \iff & (N+1)^{|\Xi|} \exp\left[-N \inf_{\mathbb{Q} \in \Pi} d_{KL}(\mathbb{Q}, \mathbb{P})\right] < \beta \end{aligned} \tag{29}$$

An application of Sanov's theorem [26, inequality 2.1.12] gives:

$$\mathbb{P}^N\left[\hat{\mathbb{P}}_N \in \Pi\right] \leq (N+1)^{|\Xi|} \exp\left[-N \inf_{\mathbb{Q} \in \Pi} d_{KL}(\mathbb{Q}, \mathbb{P})\right] \tag{30}$$

Combining inequalities (29) and (30) along with the definition of Π, we have:

$$\mathbb{P}^N\left[d_W(\hat{\mathbb{P}}_N, \mathbb{P}) > \varepsilon_N(\beta)\right] < \beta,$$

which is equivalent to the probabilistic guarantee (4). □

Proof of Theorem 2. Under the stated assumptions of (A1) complete and (A2) sufficiently expensive recourse, strong duality holds between $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ and its dual $\mathcal{Q}_d(\boldsymbol{x}, \boldsymbol{\xi})$. Along with the fact that $d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}^{(i)}) = \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}\|$ is induced by a norm, the result from Lemma 1 allows us to equivalently reformulate the distributionally robust two-stage problem (2) in the form (8), where

$$Z_i(\boldsymbol{x}, \alpha) = \sup_{\boldsymbol{\xi} \in \Xi} \left\{\mathcal{Q}_d(\boldsymbol{x}, \alpha) - \alpha \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}\|\right\}.$$

By substituting the expression for $\mathcal{Q}_d(\boldsymbol{x}, \boldsymbol{\xi})$ from (6) and introducing the epigraphical variable τ, we obtain

$$Z_i(\boldsymbol{x}, \alpha) = \underset{\boldsymbol{\xi} \in \Xi, \boldsymbol{\lambda} \in \mathbb{R}^L_+, \tau \in \mathbb{R}_+}{\text{maximize}} \left\{[\boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi} + \boldsymbol{h}(\boldsymbol{x})]^\top \boldsymbol{\lambda} - \alpha\tau : \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^{(i)}\| \leq \tau,\ \boldsymbol{q}(\boldsymbol{\xi}) - \boldsymbol{W}(\boldsymbol{\xi})^\top \boldsymbol{\lambda} \in \mathcal{Y}^*\right\}.$$

Next, we *(i)* use the affinity of $\boldsymbol{q}$ and $\boldsymbol{W}$: $\boldsymbol{q}(\boldsymbol{\xi}) = \boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi}$ and $\boldsymbol{W}(\boldsymbol{\xi}) = \boldsymbol{W}_0 + \sum_{j \in [M]} \xi_j \boldsymbol{W}_j$, *(ii)* linearize the products $\boldsymbol{\lambda}\boldsymbol{\xi}^\top$ by setting them equal to (the new variable) $\boldsymbol{\Lambda}$, and *(iii)* use the definition of the norm cone $\mathcal{C}^{M+1} = \{(\boldsymbol{\xi}, \tau) \in \mathbb{R}^M \times \mathbb{R} : \|\boldsymbol{\xi}\| \leq \tau\}$ to obtain

$$Z_i(\boldsymbol{x}, \alpha) = \underset{(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \tau) \in \mathcal{Z}_i}{\text{maximize}} \left\{\langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda} \rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \alpha\tau\right\},$$

where $\mathcal{Z}_i$ is defined in (9b). The objective function of this maximization problem is linear in its decision variables. Therefore, we can equivalently replace the feasible region with the closure of its convex hull to obtain the stated reformulation. □

Proof of Theorem 3. First, we observe that under Assumption (A3), the second-stage loss function, $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$, can be equivalently reformulated as follows:

$$\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) = \inf_{\boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in [0,1]^M} \left\{ \boldsymbol{q}(\boldsymbol{\xi})^\top \boldsymbol{y} : \boldsymbol{W}_0 \boldsymbol{y} \geq \boldsymbol{T}_0 \boldsymbol{z} + \boldsymbol{h}(\boldsymbol{x}),\ (\mathbf{e} - 2\boldsymbol{\xi})^\top \boldsymbol{z} + \mathbf{e}^\top \boldsymbol{\xi} \leq 0 \right\}. \tag{31}$$

To see this, observe that satisfaction of the last inequality is equivalent to satisfying $\boldsymbol{z} = \boldsymbol{\xi}$ since

$$\begin{aligned}
(\mathbf{e} - 2\boldsymbol{\xi})^\top \boldsymbol{z} + \mathbf{e}^\top \boldsymbol{\xi} \leq 0 \iff & \sum_{i \in [M]} [\xi_i (1 - z_i) + (1 - \xi_i) z_i] \leq 0 \\
\overset{(\#)}{\iff} & \sum_{i \in [M]} |z_i - \xi_i| \leq 0 \\
\iff & \|\boldsymbol{z} - \boldsymbol{\xi}\|_1 \leq 0 \\
\iff & \boldsymbol{z} = \boldsymbol{\xi},
\end{aligned}$$

where the equivalence (#) follows from the fact that $\boldsymbol{z} \in [0,1]^M$ and $\boldsymbol{\xi} \in \Xi \subseteq \{0,1\}^M$.

Next, we construct the Lagrangian dual of the problem (31) with respect to the last inequality. Strong duality holds since the second-stage problem $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ is strictly feasible and convex, under Assumption (A2):

$$\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}) = \sup_{\rho \geq 0} \mathcal{Q}^\rho(\boldsymbol{x}, \boldsymbol{\xi}).$$

As a function of the penalty parameter ρ, $\mathcal{Q}^\rho(\boldsymbol{x}, \boldsymbol{\xi})$ is concave and nondecreasing since $(\mathbf{e} - 2\boldsymbol{\xi})^\top \boldsymbol{z} + \mathbf{e}^\top \boldsymbol{\xi} \geq 0$ whenever $\boldsymbol{z} \in [0,1]^M$ and $\boldsymbol{\xi} \in \{0,1\}^M$. Therefore, for a fixed choice of $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Xi$, it suffices to choose any value of ρ that is greater than or equal to the optimal Lagrange multiplier of the last constraint in (31). The claim then follows from the compactness of $\mathcal{X}$ and Ξ. □

The proof of Theorem 4 relies on the following technical lemma.

Lemma 3. *For each $i \in [N]$, let $f_i : \mathbb{R}_+ \mapsto \mathbb{R}$ be a concave and non-decreasing function such that the supremum $\sup_{\rho \geq 0} f_i(\rho)$ is achieved for some finite ρ. Then, the following equality holds:*

$$\sum_{i \in [N]} \max_{\rho \geq 0} f_i(\rho) = \max_{\rho \geq 0} \sum_{i \in [N]} f_i(\rho). \tag{32}$$

Proof. The inequality $\geq$ is trivially true, and it implies that the maximization on the right-hand side is attained. Next, we show that the inequality $\leq$ also holds. Let $\rho^* \in \arg\max_{\rho \geq 0} \sum_{i \in [N]} f_i(\rho)$. If $\rho^* \notin \arg\max_{\rho \geq 0} f_{i'}(\rho)$ for some $i' \in [N]$, then there exists $\hat{\rho} > \rho^*$ such that $f_{i'}(\hat{\rho}) > f_{i'}(\rho^*)$; and it follows from their monotonicity that $f_j(\hat{\rho}) \geq f_j(\rho^*)$ for all $j \in [N] \setminus \{i'\}$. This implies that $\sum_{i \in [N]} f_i(\hat{\rho}) > \sum_{i \in [N]} f_i(\rho^*)$, contradicting that ρ^* is a maximizer of the right-hand side. □

Proof of Theorem 4. The proof of Theorem 3 established that $\mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi}) = \max_{\rho\geq 0} \mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi})$. In conjunction with Lemma 1, we can conclude that the distributionally robust two-stage problem (2) is equivalent to

$$\begin{aligned}
&\min_{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0}\; c(\boldsymbol{x}) + \alpha\varepsilon + \sum_{i\in[N]} \max_{\rho\geq 0} \underbrace{\max_{\boldsymbol{\xi}\in\Xi} \frac{1}{N}\left\{\mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi}) - \alpha\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}^{(i)}\|\right\}}_{:= f_i(\boldsymbol{x},\alpha,\rho)} \\
&= \min_{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0}\; c(\boldsymbol{x}) + \alpha\varepsilon + \max_{\rho\geq 0} \sum_{i\in[N]} f_i(\boldsymbol{x},\alpha,\rho), \qquad (33a)
\end{aligned}$$

where the equality follows by applying Lemma 3 to $f_i(\boldsymbol{x},\alpha,\cdot)$, $i\in[N]$: indeed, the mapping $\rho\mapsto\mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi})$ is concave and nondecreasing (see proof of Theorem 3), and hence so is $f_i(\boldsymbol{x},\alpha,\cdot)$.

Theorem 3 also shows that there exists a finite $\bar{\rho}>0$ such that $\mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi}) = \mathcal{Q}(\boldsymbol{x},\boldsymbol{\xi})$ for all $\rho\geq\bar{\rho}$ and all $\boldsymbol{x}$ and $\boldsymbol{\xi}$. This implies that $\sum_{i\in[N]} f_i(\boldsymbol{x},\alpha,\cdot)$ is maximized at $\bar{\rho}$, and thus we have

$$\begin{aligned}
&(33a) \\
&\leq \min_{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0}\; c(\boldsymbol{x}) + \alpha\varepsilon + \sum_{i\in[N]} f_i(\boldsymbol{x},\alpha,\bar{\rho}) \qquad (33b)\\
&\leq \max_{\rho\geq 0} \min_{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0}\; c(\boldsymbol{x}) + \alpha\varepsilon + \sum_{i\in[N]} f_i(\boldsymbol{x},\alpha,\rho) \qquad (33c)\\
&= \max_{\rho\geq 0} \min_{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0}\; c(\boldsymbol{x}) + \alpha\varepsilon + \frac{1}{N}\sum_{i\in[N]} \max_{\boldsymbol{\xi}\in\Xi}\left\{\mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi}) - \alpha\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}^{(i)}\|\right\}. \qquad (33d)
\end{aligned}$$

The inequality (33b) $\leq$ (33c) follows by treating (33b) as a function that is evaluated at $\rho=\bar{\rho}$ and (33c) as maximizing this function. The max-min inequality implies (33c) $\leq$ (33a), and therefore, we have (33a) = (33b) = (33c) = (33d). We point out that, unlike the classical minimax theorem, we did not exploit convexity of $\mathcal{X}$. Indeed, we only exploited the fact that each $f_i(\boldsymbol{x},\alpha,\rho)$ is monotone in ρ and the feasible region of ρ is essentially compact because of Theorem 3.

We now show that it suffices to choose $\bar{\rho}$ as per the statement of the theorem. The key observation is that for any $\varepsilon\geq 0$, the expression inside $\max_{\rho\geq 0}$ in (33d) is bounded from above by the optimal value of the classical robust optimization problem with any uncertainty set $\Xi^0\supseteq\Xi$:

$$\begin{aligned}
&\max_{\rho\geq 0}\min_{\boldsymbol{x}\in\mathcal{X}}\; c(\boldsymbol{x}) + \max_{\boldsymbol{\xi}\in\Xi^0} \mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi}) \\
&\geq \max_{\rho\geq 0} \min_{\boldsymbol{x}\in\mathcal{X},\alpha\geq 0}\; c(\boldsymbol{x}) + \alpha\varepsilon + \frac{1}{N}\sum_{i\in[N]} \max_{\boldsymbol{\xi}\in\Xi}\left\{\mathcal{Q}^\rho(\boldsymbol{x},\boldsymbol{\xi}) - \alpha\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}^{(i)}\|\right\} \qquad \forall\varepsilon'\geq\varepsilon\geq 0.
\end{aligned}$$

By a similar argument as before, the nondecreasing nature of the mapping $\rho \mapsto \mathcal{Q}^\rho(\boldsymbol{x}, \boldsymbol{\xi})$ guarantees that any ρ of maximizing the left-hand side (i.e., the classical robust problem) must also maximize the right-hand side (i.e., the distributionally robust problem). The proof of Theorem 3 then shows that it suffices to choose a value that is at least as large as the optimal Lagrange multiplier ρ^r. □

Proof of Proposition 1. Observe that $\mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi}^r) \geq \mathcal{Q}(\boldsymbol{x}, \boldsymbol{\xi})$ for all $\boldsymbol{\xi} \in \Xi$ and all $\boldsymbol{x} \in \mathcal{X}$. Indeed, the objective function of the problem on the left-hand side is greater than that on the right-hand side: $(\boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi}^r)^\top \boldsymbol{y} \geq (\boldsymbol{q}_0 + \boldsymbol{Q}\boldsymbol{\xi})^\top \boldsymbol{y}$ for all $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^{N_2}_+$. Also, the feasible region of the problem on the left-hand side is a superset of the one on the right: $\boldsymbol{W}_0\boldsymbol{y} \geq \boldsymbol{T}_0\boldsymbol{\xi}^r + \boldsymbol{h}(\boldsymbol{x}) \geq \boldsymbol{T}_0\boldsymbol{\xi} + \boldsymbol{h}(\boldsymbol{x})$. Therefore, $\boldsymbol{\xi}^r$ is a worst-case realization of the parameters independent of the first-stage decision $\boldsymbol{x} \in \mathcal{X}$. □

Proof of Theorem 5. Let $v^\star_{\varepsilon,N}$ denote the optimal value of the distributionally robust two-stage problem (2) for a given sample $\{\hat{\boldsymbol{\xi}}^{(1)}, \ldots, \hat{\boldsymbol{\xi}}^{(N)}\}$ of size $N > 0$, and radius $\varepsilon \geq 0$. Let $v^0_{\varepsilon,N}$ denote the optimal value of the convex hull reformulation (8) when the convex hulls $\operatorname{cl}\operatorname{conv}(\mathcal{Z}_i)$ in (9a)–(9b) are approximated using the continuous relaxation $\mathcal{Z}^0_i$. For simplicity, denote $\boldsymbol{z} = (\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \tau)$ in (9a)–(9b). Then, observe that:

$$
\begin{aligned}
v^0_{\varepsilon_N,N} &= \min_{\boldsymbol{x}\in\mathcal{X}} c(\boldsymbol{x}) + \inf_{\alpha\geq 0} \alpha\varepsilon_N + \frac{1}{N}\sum_{i=1}^N \sup_{\boldsymbol{z}\in\mathcal{Z}^0_i} \Big\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda}\rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda} - \alpha\tau \Big\} \\
&= \min_{\boldsymbol{x}\in\mathcal{X}} c(\boldsymbol{x}) + \inf_{\alpha\geq 0} \alpha\varepsilon_N + \sup_{(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N)\in\mathcal{Z}^0_1\times\ldots\times\mathcal{Z}^0_N} \frac{1}{N}\sum_{i=1}^N \Big\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda}_i\rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda}_i - \alpha\tau_i \Big\} \\
&= \min_{\boldsymbol{x}\in\mathcal{X}} c(\boldsymbol{x}) + \sup_{(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N)\in\mathcal{Z}^0_1\times\ldots\times\mathcal{Z}^0_N} \frac{1}{N}\sum_{i=1}^N \Big\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda}_i\rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda}_i \Big\} + \inf_{\alpha\geq 0} \alpha \left[\varepsilon_N - \frac{1}{N}\sum_{i=1}^N \tau_i \right] \\
&= \min_{\boldsymbol{x}\in\mathcal{X}} c(\boldsymbol{x}) + \sup_{(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N)\in\mathcal{Z}^0_1\times\ldots\times\mathcal{Z}^0_N} \left\{ \frac{1}{N}\sum_{i=1}^N \Big\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda}_i\rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda}_i \Big\} : \sum_{i=1}^N \tau_i \leq N\varepsilon_N \right\},
\end{aligned}
$$

where the equality on the first line follows by definition of $v^0_{\varepsilon,N}$; and the second equation is obtained by interchanging the order of the summation and the supremum. The third equality follows from a non-compact variant of Sion's minimax theorem [34, Theorem 2], which is applicable since *(i)* the objective function is linear in both $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$ and α, *(ii)* their corresponding feasible regions are convex, and *(iii)* the term $\sup_{(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N)\in\mathcal{Z}^0_1\times\ldots\times\mathcal{Z}^0_N} \inf_{\alpha\geq 0}$ is bounded from below by $\sup_{(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N)\in\mathcal{K}^0_1\times\ldots\times\mathcal{K}^0_N} \frac{1}{N}\sum_{i=1}^N \big\{\langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda}_i\rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda}_i\big\} + \inf_{\alpha\in\{0\}} \alpha\big[\varepsilon_N - \frac{1}{N}\sum_{i=1}^N \tau_i\big] = \frac{1}{N}\sum_{i=1}^N \mathcal{Q}_d(\boldsymbol{x}, \hat{\boldsymbol{\xi}}^{(i)})$, where $\mathcal{K}^0_i := \{(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \tau) \in \mathcal{Z}^0_i : \boldsymbol{\xi} = \hat{\boldsymbol{\xi}}^{(i)}, \tau = 0\}$ is a convex compact subset of

$\mathcal{Z}_i^0$. The last equation is due to $v_{\varepsilon_N,N}^0$ being finite; see Assumptions **(A1)** and **(A2)** from Section 3. Therefore, we have

$$\begin{aligned}
\limsup_{N\to\infty} v_{\varepsilon_N,N}^0 &= \limsup_{N\to\infty} \left\{ \min_{\boldsymbol{x}\in\mathcal{X}} c(\boldsymbol{x}) + \sup_{(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N)\in\mathcal{Z}_1^0\times\ldots\times\mathcal{Z}_N^0} \left\{ \frac{1}{N}\sum_{i=1}^N \left\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda}_i\rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda}_i \right\} : \sum_{i=1}^N \tau_i \le N\varepsilon_N \right\} \right\} \\
&= \limsup_{N\to\infty} \left\{ \min_{\boldsymbol{x}\in\mathcal{X}} c(\boldsymbol{x}) + \sup_{(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N)\in\mathcal{K}_1^0\times\ldots\times\mathcal{K}_N^0} \frac{1}{N}\sum_{i=1}^N \left\{ \langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\Lambda}_i\rangle + \boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\lambda}_i \right\} \right\} \\
&= \lim_{N\to\infty} v_{0,N}^0 = v_{0,\infty}^\star,
\end{aligned}$$

where $v_{0,N}^\star$ is the optimal value of the sample average approximation, and $v_{0,\infty}^\star \coloneqq \lim_{N\to\infty} v_{0,N}^\star$ is the optimal value of the original two-stage stochastic problem, which is guaranteed to exist because of [59, Theorem 5.4] along with Assumptions **(A1)** and **(A2)** from Section 3. Note that the second equation follows from the assumption that $N\varepsilon_N \to 0$ as $N\to\infty$, which further implies that $\sum_{i=1}^N \tau_i \le 0$, leading to $\boldsymbol{\xi}_i = \hat{\boldsymbol{\xi}}^{(i)}$ for all $i \in [N]$.

Moreover, for any $N > 0$ and any $\varepsilon \ge 0$, it holds that $v_{\varepsilon,N}^0 \ge v_{\varepsilon,N}^\star \ge v_{0,N}^\star$. Therefore, we have

$$\liminf_{N\to\infty} v_{\varepsilon,N}^0 \ge \liminf_{N\to\infty} v_{\varepsilon,N}^\star \ge \lim_{N\to\infty} v_{0,N}^\star = v_{0,\infty}^\star.$$

Combining these observations, we obtain $\lim_{N\to\infty} v_{\varepsilon_N,N}^0 = \lim_{N\to\infty} v_{\varepsilon_N,N}^\star = v_{0,\infty}^\star$, which proves the statement of the theorem. □