

Robust Chance-Constrained Optimization using a Continuous Parameter Space Wasserstein-2 Ambiguity Set of Gaussian Mixtures

Shibshankar Dey

Department of Industrial Engineering and Management Sciences, Northwestern University shibshankardey2025@u.northwestern.edu

Sanjay Mehrotra

Department of Industrial Engineering and Management Sciences, Northwestern University mehrotra@northwestern.edu

Abstract. We study distributionally robust linear chance constraint optimization models where coefficients follow a Gaussian mixture model (GMM). We develop a novel computable Wasserstein-2-type metric for GMMs in the parameter space supported over continuous sets. This significantly generalizes a previously known finite-set-based definition. We show the equivalence of our metric to the commonly used data space-based Wasserstein-2 metric between two distributions. For the ambiguity set based on our metric (CDR), under mild regularity conditions, we prove strong duality for the inner worst-case chance-constraint problem and derive its semi-infinite reformulation. We then develop an adaptive cutting-surface algorithm, which endogenously determines the locations of mixture components receiving mass, and the mean and covariances of the Gaussian distributions at these locations. We prove that this algorithm attains any prescribed optimality gap in finitely many iterations using finitely many linear segments used to approximate the non-convexity of the Gaussian cumulative distribution functions. A block-alternating local-search procedure is used to identify new Gaussian distributions added to the current pool.

An energy-allocation problem at an electric vehicle charging station is used as a case study to demonstrate the added value of continuous-support-based robustification developed in this paper. For chance probability targets $\theta \in \{0.95, 0.97, 0.99\}$ and Wasserstein radii $\rho \in \{0.001, 0.005, 0.01\}$, the results (using previously known data) show a superior out-of-sample chance constraint satisfaction, when compared with no robustification or a finite-support-based robustification (FDR) of a nominal GMM-based chance constraint. In the out-of-sample testing, FDR fails to attain the prescribed target probability for every tested (θ, ρ) pair. In contrast, CDR exhibits a consistent ρ -based improvement in out-of-sample chance constraint satisfaction (OSS) probability as ρ increases under $\pm 5\%$ or $\pm 10\%$ support uncertainty allowance from the mean of the nominal Gaussians. In particular, the $\theta = 0.95$ target is satisfied in every CDR specification even at $\rho = 0.001$, with OSS values of 95.04%–95.87%. For $\rho = 0.01$, the 97% target is attained across all continuous-support settings, with OSS values of 97.35%–98.45%; for $\theta = 0.99$, OSS is 98.49% for $\pm 5\%$ mean-support uncertainty, and 99.17% for $\pm 10\%$ mean-support uncertainty. As expected, the cost of the solution and solution time increase with the use of the CDR-based ambiguity set for GMMs. A detailed analysis of the generated solutions shows that the CDR model makes structural changes to the energy allocations, whereas the FDR model allocations are close to those from the nominal model. The results demonstrate that the proposed continuous support-based robust framework for Gaussian mixture should be considered when out-of-sample chance constraint satisfaction is important.

Funding: This research was fully funded by US National Science Foundation Grant DMS 2229410.

Key words: Gaussian mixture model (GMM), Wasserstein metric, pushforward measure, Gelbrich distance, optimal value function, piecewise linear approximation, semi-infinite program, cutting surface algorithm, block alternating minimization.

1. Introduction

We study a distributionally robust linear chance-constrained program with decision vector $\mathbf{x} \in \mathbb{R}^n$:

$$\min_{\mathbf{x} \in \mathcal{X}} \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \min_{\mathbb{P} \in \mathfrak{D}} \mathbb{P}[\boldsymbol{\xi}^\top \mathbf{x} \leq b] \geq \theta, \quad (\text{DR-C})$$

where $\mathbb{P}[\cdot]$ denotes the probability of an event, $\theta \in (0, 1)$ is the prescribed satisfaction probability of the uncertain requirement, and $\mathcal{X}$ is assumed to be a compact feasible set specified by deterministic constraints. $\mathfrak{D}$ denotes the ambiguity set of distributions representing the uncertainty of the random vector $\boldsymbol{\xi}$. In operational applications, the prescribed satisfaction probability θ is generally interpreted as a probability-based service-level target. This view is common in inventory and service operations, where service-level constraints are used to control stockout probabilities and fill rates, and in EV charging infrastructure planning, where service quality is measured using charging-request fulfillment, or lost demand [13, 44, 70].

We introduce a novel construction of $\mathfrak{D}$ by assuming that $\boldsymbol{\xi}$ follows a Gaussian mixture distribution. Our consideration of the Gaussian mixture-based uncertainty representation and the corresponding ambiguity set is motivated by the fact that Gaussian Mixture Models (GMMs) are dense in the space of probability distributions, meaning a mixture with enough components can approximate any probability distribution arbitrarily well [52, 54]. Thus, a GMM acts as an approximator for modeling multimodal and non-Gaussian data. Our ambiguity set allows the prescribed service level to be hedged against statistical errors in Gaussian mixture parameter estimation, particularly when limited data are used to estimate the nominal mixture. For probability distributions μ_1 and μ_2 , the Wasserstein-2 metric [81] is defined as:

$$W_2(\mu_1, \mu_2) := \left(\inf_{Y_1 \sim \mu_1, Y_2 \sim \mu_2} \mathbb{E}[\|Y_1 - Y_2\|^2] \right)^{1/2} = \left(\inf_{\pi \in \Pi(\mu_1, \mu_2)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\pi(\mathbf{y}_1, \mathbf{y}_2) \right)^{1/2}. \quad (1)$$

Here, μ_1 and μ_2 are assumed to be multivariate Gaussian mixture distributions. The objective in (1) minimizes the optimal transport cost obtained by minimizing the expected “Euclidean” distance of the transport cost over all joint probability distributions $\pi \in \Pi(\mu_1, \mu_2)$, called couplings, with marginals $Y_1 \sim \mu_1$ and $Y_2 \sim \mu_2$ [42, 43, 60, 82]. We call the metric in (1) the Wasserstein-2 metric in the data space, as it uses the space where samples (data) are observed in the metric definition. This metric can be used to describe a set $\{\mu_2 \mid W_2(\mu_1, \mu_2) \leq \rho\}$ of distributions called a Wasserstein-2 ball that are within ρ Wasserstein-2 distance away from the nominal distribution μ_1 .

1.1. Gaussian Mixture Model for Multi-Modal Uncertainty Representation

Multimodality and non-Gaussianity arise when the data reflect multiple clusters or regimes. Examples are electric-vehicle charging demand [63], wind power generation density [25, 88], and short-horizon financial returns modeled using mixture or regime-switching distributions [48, 66]. In such settings, population heterogeneity induces mixture distributions. Specifically, let $\mathbf{m}_k \in$

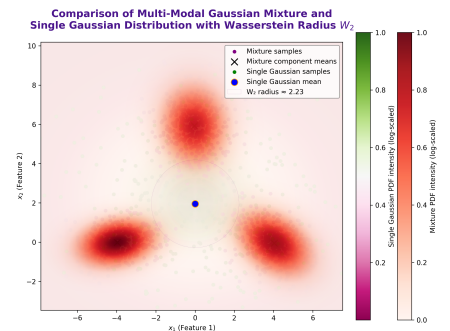


Figure 1: Any uncertainty ball with $W_2 < 2.23$ around a single-modal distribution fitted using the data (see footnote) would not contain the three-modal distribution.

$\mathbb{R}^n$, $\mathbf{Q}_k \in \mathbb{R}^{n \times n}$ for $k = 1, \dots, K$, be the mean vectors and covariance matrices of component Gaussian distributions $\mathcal{N}(\mathbf{m}_k, \mathbf{Q}_k)$.

Then the mixture distribution is given as $\sum_{k \in [K]} w_k \mathcal{N}(\mathbf{m}_k, \mathbf{Q}_k)$

where $\sum_{k \in [K]} w_k = 1$ and $w_k \geq 0$ for all $k \in [K]$. When the data-generating distribution is expected to have a mixture structure, a tighter ambiguity set can be obtained by modeling uncertainty over the mixing distribution on component parameters. For example, if a Gaussian distribution is used as a nominal distribution using the dataset¹, Fig. 1 shows that a Wasserstein ball of radius exceeding 2.23 units is required for a unimodal Gaussian surrogate to cover the fitted multimodal Gaussian mixture. In this case, the ambiguity set radius can be much smaller when a Gaussian mixture is used as the nominal distribution, yielding a more localized ambiguity set around the nominal multimodal structure for service-level analysis.

1.1.1. Wasserstein-2-type metric for finite Gaussian mixtures. Let $\widehat{\mathbb{GM}}_n$ denote the set of Gaussian mixture distributions in $\mathbb{R}^n$, and $\widehat{\mathbb{GM}}_n(K)$ those with K mixing terms. For $\mu_1 \in \widehat{\mathbb{GM}}_n(K)$ and $\mu_2 \in \widehat{\mathbb{GM}}_n(\hat{K})$, Delon and Desolneux [19] proposed a Wasserstein-type metric by restricting the coupling distributions $\Pi(\mu_1, \mu_2)$ to $\widehat{\mathbb{GM}}_{2n}$:

$$RW_2^2 := \inf_{\pi \in \Pi(\mu_1, \mu_2) \cap \widehat{\mathbb{GM}}_{2n}} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|y_1 - y_2\|^2 d\pi(\mathbf{y}_1, \mathbf{y}_2). \quad (2)$$

In particular, if μ_1, μ_2 follow the Gaussian mixture distributions $\sum_{k \in [\hat{K}]} \hat{w}_k \hat{\mathcal{N}}_k(\hat{\mathbf{m}}_k, \hat{\mathbf{Q}}_k)$ and $\sum_{l \in [K]} w_l \mathcal{N}_l(\mathbf{m}_l, \mathbf{Q}_l)$ respectively, then [19, Proposition 4] showed that $RW_2^2 = \inf_{q \in \Pi(\hat{\mathbf{w}}, \mathbf{w})} \sum_{k,l} q_{kl} d_{\text{BW}}^2(\hat{\mathcal{N}}_k, \mathcal{N}_l)$, where d_{BW} is the Burer-Wasserstein metric between two Gaussian laws (see equation (10)). Here $\Pi(\hat{\mathbf{w}}, \mathbf{w})$ is the standard coupling between two finitely supported probability distributions (represented by the transportation constraints). This result can be used to describe an ambiguity set around a nominal Gaussian-mixture distribution μ_1 as follows:

$$\mathfrak{D}_d := \left\{ (\mathbf{w}, \mathbf{q}) \left| \begin{array}{l} \sum_{k \in \hat{K}} \sum_{l \in K} q_{kl} d_{\text{BW}}^2(\hat{\mathcal{N}}_k, \mathcal{N}_l) \leq \rho \\ \sum_{l \in [K]} q_{kl} = \hat{w}_k, \forall k \in [\hat{K}] \\ \sum_{k \in [\hat{K}]} q_{kl} = w_l, \forall l \in [K] \\ \sum_{k \in [\hat{K}]} \hat{w}_k = 1, \hat{w}_k \geq 0 \forall k \in [\hat{K}] \end{array} \right. \right\}. \quad (3)$$

The solutions feasible to $\mathfrak{D}_d$ provide mixing weights $\mathbf{w}$ for a Gaussian mixture distribution $\sum_{k \in [K]} w_k \mathcal{N}_k(\mathbf{m}_k, \mathbf{Q}_k)$ in the ambiguity set. The ambiguity set in (3) is referred to as the finite-parameter ambiguity set, and the corresponding distributionally robust model from (DR-C) is called FDR.

Note that $d_{\text{BW}}^2(\mathcal{N}_k, \mathcal{N}_l)$ is pre-computable, hence, the ambiguity set in (3) is interesting in its own right. However, the distributions in this set are restricted either to those means and covariances selected a priori or exactly to those of the nominal finite Gaussian mixture distribution. For example, the approach of selecting

¹ We generated a synthetic 2D dataset (sample size = 1200, seed = 42) from an equally weighted 3-component Gaussian mixture with $w_k = 1/3$, $\mathbf{m}_1 = (-4, 0)^\top$, $\mathbf{m}_2 = (4, 0)^\top$, $\mathbf{m}_3 = (0, 6)^\top$, and full covariance matrices $\mathbf{Q}_1 = \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}$, $\mathbf{Q}_2 = \begin{pmatrix} 1.0 & -0.3 \\ -0.3 & 0.7 \end{pmatrix}$, and $\mathbf{Q}_3 = \begin{pmatrix} 0.8 & 0.0 \\ 0.0 & 1.2 \end{pmatrix}$. Using the `scikit-learn` `GaussianMixture` package in Python with full covariance, we fitted a 3-component GMM to represent the multimodal distribution and a 1-component Gaussian surrogate. The reported Wasserstein radius is the empirical W_2 distance between these two fitted distributions, computed with the `POT` `ot.emd2` package under squared Euclidean transport cost from Monte Carlo samples.

the unknown K shown in [89, Proposition 3] restricts the components' means and covariances to those obtained from GMM fitting to available data. Fixing the mixture component distribution parameters to those of the nominal mixture and optimizing only over the mixing weights is similar to a scenario-reweighting approach (see, e.g., [41]). Such support-restricted constructions may provide limited protection against realizations outside the observed support and underrepresent low-frequency (tail) regions unless the support is sufficiently rich [5, 24, 28, 56]. From the chance-constrained optimization perspective, the service-level assessment for any candidate decision x is therefore confined to adversarial distributions supported on a pre-selected finite set of component mean–covariance parameters. These limitations motivate us to study a generalization of (3) based on (1) in which the component Gaussian distributions are not restricted to a fixed finite set of parameters, and the mixing law has a continuous support. The proposed construction allows the worst-case mixing law to place probability mass over the continuous parameter support, with component mean-covariances selected endogenously.

1.2. Contributions

1.2.1. Contributions to optimal transport. Since the identification of, and computations with, the Gaussian mixture distributions in the Wasserstein-2 ball defined in the data space using (1) is not practical, in this paper, we show that under suitable conditions, this data-space Wasserstein-2 ball can be equivalently described using a Wasserstein-like metric representation in the parameter space (see Theorems 2 and 3). Although this paper focuses on the chance constraint optimization problem, from a distributional robust optimization and modeling perspective, this result applies to any optimization model that benefits from the ambiguity set representation based on a Gaussian mixture distribution.

1.2.2. Algorithmic development. For the ambiguity set developed in this paper, we establish a strong duality result for the robust chance constraint in (DR-C) by showing that the sufficient conditions in [8, Theorem 1(a)] are satisfied. This duality result allows us to reformulate (DR-C) as a nonlinear semi-infinite program. The τ -optimal convergence, after adding a finite number of cuts, of an adaptive cutting surface algorithm for this nonlinear semi-infinite program is proved. The solvability of the subproblem arising in the proposed cutting-surface algorithm is studied.

1.2.3. Computational contributions. We evaluate continuous- and finite-parameter distributionally robust formulations (CDR and FDR), along with their nominal counterparts, for a service-level-aware demand management problem in electric-vehicle charging station operations. To develop an implementation of the cutting-surface algorithm using the currently available commercial solvers, we develop a local-optimization heuristic to solve the cutting-surface identification problem (subproblem) efficiently. Across different experimental settings, the proposed adaptive cutting-surface algorithm's behavior indicates that the heuristic identifies the dominant violating mixture components in the first few iterations and then refines a reduced set. Beginning with five nominal cuts, the total cut count after four master iterations remains below 25.

1.2.4. Experimental findings and managerial implications. The computational study using a real dataset from repository [26] exhibits a quantitatively distinctive robustness frontier for the continuous-support ambiguity set. Table 1 summarizes the key messages from experimental findings in terms of out-of-sample satisfaction (OSS) probability over target $\theta \in \{0.95, 0.97, 0.99\}$, incremental objective overhead, and runtime. While FDR approach protects only over finite empirical support points, CDR approach additionally allows the nominal GMM means and covariances to vary within compact support sets, controlled by the Wasserstein radius $\rho \in \{0.001, 0.005, 0.01\}$, the mean-support expansion parameter $\varsigma \in \{0.05, 0.1\}$ that symmetrically expands the mixture component mean support around their respective nominal value, and the covariance-scaling bounds $(\underline{\lambda}_f, \bar{\lambda}_f) \in \{(1/2, 2), (1/3, 3)\}$.

Table 1 highlights the main gain of the proposed model with a clear demonstration of the added robustness by CDR. FDR does not achieve OSS under any settings. Whereas for mean-support uncertainty parameter value $\varsigma = 0.1$, CDR achieves OSS for all values of θ . The use of mean-covariance uncertainty beyond the Wasserstein radius in CDR expands the set of target-attaining solutions at a 1.5%-5% increase in the objective and a significant increase (from minutes to hours) in runtime overhead. Under the experimental settings considered, we also observe that the effect of changing the covariance-scaling multiplier is more limited compared to that of the mean-uncertainty hyperparameter.

Table 1 Comparative summary of prescribed target attainment, incremental objective, and runtime under FDR and CDR. The reported objective increase in % is calculated by $100 \times |\text{Obj}^{\text{CDR}} - \text{Obj}^{\text{FDR}}| / \text{Obj}^{\text{FDR}}$. FDR performance is investigated under both the stand-alone (22 hour) solve and the initial iteration budget used to initialize CDR, so that ambiguity-set effects can be distinguished from solve-time effects.

θ	DR Model	Mean-support uncertainty parameter ς	OSS probability range (%)	Target attained at ball radius ρ	Objective increase (%)	Total time (hr)
0.95	FDR	–	92.76–93.91	none	–	0.02–0.03
	CDR	0.05	95.04–96.41	all	1.49–2.31	4.64–12.01
	CDR	0.10	95.87–97.71	all	2.37–3.63	4.65–12.01
0.97	FDR	–	94.09–95.19	none	–	0.15–3.30
	CDR	0.05	96.21–97.36	0.01	1.64–2.48	15.70–22.00
	CDR	0.10	96.79–98.45	0.005, 0.01	2.53–3.99	16.08–20.00
0.99	FDR	–	95.57–95.60	none	–	22.00
	CDR	0.05	97.45–98.49	none	1.93–3.16	22.00
	CDR	0.10	98.10–99.17	0.01	2.82–5.00	22.00

The CDR model also provides decision-level guidance toward achieving higher service-level targets. The energy allocation diagnostics for the considered demand management case study further show that these gains are operationally substantive: FDR remains close to the nominal schedule as ρ varies, whereas CDR generates a distinct family of hourly energy allocations through targeted changes in a small set of recurring hours. Taken together, these findings indicate that allowing ambiguity over continuous mixture parameters, compared to only over empirical support points, provides complete control for service-level attainment, operating cost, and a meaningful structure of the charging schedule when the decision-maker can afford additional computational overhead.

1.3. Organization

Some basic notation used throughout the paper is given in Section 1.4. Existing literature on the use of Gaussian mixture in data representation, chance-constraint optimization, distributional robustness, and optimal transport is reviewed in Section 2. Sections 3-4 present our proposed generalized representation of the Wasserstein-2 metric and ambiguity set (2)-(3). In these sections, we develop the optimal transport (OT) results for this set. Section 5 provides strong duality results and reformulations resulting from its use for the distributionally robust GMM-based chance-constrained problem resulting from our proposed ambiguity set. Section 5 provides the resulting semi-infinite program to ensure the convergence of a cutting-surface (Cut-S) algorithm, followed by a comprehensive computational study focusing on an electric vehicle charging capacity planning problem in Section 7.

1.4. Notations and Basic Definitions

For any function $h(\cdot; \cdot)$, arguments before and after “;” respectively stand for variables and parameters of that function. $\mathcal{P}(\mathbb{R}^n)$ represents all sets of probability measures on $\mathbb{R}^n$, and $\mathcal{P}_2(\mathbb{R}^n)$ represents those having finite second moment. The notation $Y \sim \nu$ means that Y is a random variable with probability distribution ν . “a.e.” abbreviates almost everywhere. $\text{tr}(M)$ denotes the trace of a matrix M . For any non-negative integer n , we use $[n]$ to denote the set $\{1, 2, \dots, n\}$ and $[n]_0$ to denote $\{0, 1, 2, \dots, n\}$. $\mathcal{F}(\tau)$ represents a set defined by constraints, with the right-hand side of those constraints being τ . $\mathbb{S}_+^n$ represents the set of positive-definite matrices. $\|\cdot\|_F$ is used to denote the Frobenius norm of a matrix. $\mathbb{1}$ denotes an indicator function.

$\phi(\cdot)$ and $\Phi(\cdot)$ represent the standard univariate Gaussian probability density and cumulative distribution function (CDF), respectively. The derivative of ϕ is denoted by ϕ' . We let $\mathcal{S} := \mathbb{R}^n \times \mathbb{S}_+^n$ be the parameter space of n -variate Gaussian distribution $\mathcal{N}(\mathbf{m}, \mathbf{Q})$ with $\mathbf{m} \in \mathbb{R}^n$, $\mathbf{Q} \in \mathbb{S}_+^n$ be its mean vector and positive definite covariance matrix. $\mathcal{B}(\mathcal{S})$ is the Borel sigma algebra of $\mathcal{S}$. We let $\mathbf{s} := (\mathbf{m}, \mathbf{Q}) \in \mathcal{S}$, and $\mathcal{N}(\mathbf{s}) := \mathcal{N}(\mathbf{m}, \mathbf{Q})$. For 2n-variate Gaussian distributions, we use the notation $\mathcal{Z} := \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{S}_+^{2n}$; $\mathbf{z} := ((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q})$ and $\mathcal{N}(\mathbf{z}) = \mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q})$. We use the symbol $\#$ to represent a pushforward operator. pr_i denotes the coordinate projection onto the i -th component of a product space; specific domains will be clear from the context.

Definition 1 (Gaussian Mixture Measure). *We call a measure $\mu \in \mathcal{P}(\mathbb{R}^n)$ a Gaussian mixture measure if $\exists \lambda \in \mathcal{P}(\mathcal{S})$ such that for all $C \in \mathcal{B}(\mathbb{R}^n)$*

$$\mu(C) = \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(C) d\lambda(\mathbf{s}),$$

where $\mathcal{N}(\mathbf{s})(C) = \int_{\mathbb{R}^n} \mathbb{1}_C(\mathbf{y}) d\mathcal{N}(\mathbf{s})$ for $\mathbf{s} \in \mathcal{S}$.

Definition 2 (Mixing law via Borel sets). *Let $\mu \in \mathcal{P}(\mathbb{R}^n)$. We say that a probability measure π on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ is a mixing law for a Gaussian mixture distribution μ if, for every Borel set $C \in \mathcal{B}(\mathbb{R}^n)$,*

$$\mu(C) = \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(C) d\pi(\mathbf{s}). \quad (4)$$

Note that the mixing law for a Gaussian mixture distribution may not be unique, and the second moment of a Gaussian mixture distribution may not be bounded. We denote the set of all Gaussian mixture distributions with finite second moment as

$$\mathbb{GM}_n := \left\{ \nu \mid \exists \lambda \in \mathcal{P}(\mathcal{S}) \text{ s.t. } \nu = \int_{s \in \mathcal{S}} \mathcal{N}(s) d\lambda(s) \text{ and } \int_{s \in \mathcal{S}} (\|\mathbf{m}(s)\|^2 + \text{tr}(\mathbf{Q}(s))) d\lambda(s) < \infty \right\}. \quad (5)$$

2. Literature Review

This section summarizes three streams of relevant literature. The first stream provides the Wasserstein geometry, strong-duality reformulations, and finite-sample guarantees that justify data-driven radius calibration. We then cover distributionally robust chance-constrained optimization literature, emphasizing how the support structure of the random variables shapes the available deterministic reformulations and how the uncertainty law itself is represented, especially for multimodal data. The third identifies the related OT literature on Gaussian and mixture structure. Reviewing these three streams provides a clear, distinctive indication between sample-supported mixture models and the continuous-parameter mixture class adopted in this work.

2.1. Distributionally Robust Optimization with Wasserstein Set

Foundational Wasserstein distributionally robust optimization (WDRO) results show that worst-case expectations over Wasserstein balls admit strong dual representations and, under standard growth and Lipschitz conditions, reduce to finite-dimensional programs [8, 28, 56, 59, 90]. These results justify deterministic reformulations by viewing the Wasserstein-ball as a model-misspecification device and replacing the distributional supremum with a tractable dual program [8, 28, 56, 90]. In parallel, data-driven WDRO centers the ball at an empirical law, yields convex reformulations in broad settings, and links radius calibration to finite-sample reliability [59]. Wasserstein neighborhoods also underpin robust learning objectives based on adversarial perturbations of the data law, with population-level robustness guarantees and statistical consistency analyses [50, 77]. Algorithmically, the inner maximization has been handled in the literature via flow-based transport-map parameterizations [87], entropic smoothing and Sinkhorn-type discrepancies [83], and iterative minimax schemes over continuous densities with convergence guarantees [92].

2.2. Distributionally Robust Chance-Constrained Optimization (DRCCO)

A DRCCO problem replaces a nominal chance constraint under a fixed law by a worst-case chance constraint over an ambiguity set $\mathcal{P}$ of probability measures [74]. When the ambiguity-set size is calibrated from data, one can obtain finite-sample guarantees that the unknown data-generating law lies in $\mathcal{P}$ with prescribed confidence, which in turn supports out-of-sample reliability claims for decisions computed from the resulting DR model [51]. Most continuous-support DRCCO literature falls into three families—Wasserstein-metric, φ -divergence, and moment-based sets—each with distinct reformulation mechanisms and complexity tradeoffs.

A theme of Wasserstein-metric ambiguity set is deterministic reformulation of worst-case probability constraints through duality, leading to exact mixed-integer models in broad settings and MILP reductions under ℓ_1 or ℓ_∞ transport costs [15, 86]. Related developments treat individual and joint chance constraints and variants such as random right-hand sides, again relying on strong duality for the worst-case probability evaluation [34, 38]. Structural properties have also been identified under which Wasserstein robustification yields convex feasible regions [75].

φ -divergence-based ambiguity sets (e.g., KL, χ^2) have also been used to define confidence-region-based sets around a nominal law [6, 24]. For chance constraints, robustification can reduce to a nominal chance constraint with a modified chance satisfaction level, which can be calculated explicitly for KL divergence. It also has a relationship with Bernstein-type convex approximations [36, 62]. Divergence-based ambiguity set radii are calibrated for finite-sample coverage guarantees [24, 39, 51]. Related robust-counterpart views for uncertain probability vectors and application-driven MILP reformulations (e.g., χ^2 -robust surgery scheduling) appear in [5, 21]. Moment-based ambiguity sets hedge over all laws consistent with given moment (and possibly support) information. They typically yield conic reformulations when the first two moments are used in describing the ambiguity set formulation [18, 74]. Such models also lead to safe approximations for (distributionally robust) chance constraints, with possible extensions to generalized dispersion and broader admissible moment/support sets [23, 32].

2.2.1. Use of mixture models in uncertainty representation. The distributionally robust optimization models developed in the literature have focused on representing distribution ambiguity by using a finite-sample support of the empirical distributions. Unless the sample size or the ambiguity set is very large, such finite-support representations may miss regions with nontrivial probability mass (e.g., those arising in rare event regimes), and thus provide only a coarse description of multivariate dependence [3, 9, 46, 62, 74]. A possible solution is to use mixture laws. Using finite support for the mean and variance parameters, [41] use a Gaussian mixture law to represent distribution ambiguity in the context of a chemical engineering application. In the chance-constrained setting, finite Gaussian-mixture models have been studied in [25, 37, 85], where convex relaxations are derived under strong assumptions. In a recent paper [22], a finite Gaussian mixture was used for linear chance constraints and established its solvability by using piecewise-linear inner and outer approximations of the nonlinearity arising due to the Gaussian term in the mixture. The piecewise-linear approximation idea from [22] is further developed in the current paper in the analysis of the proposed cutting surface algorithm.

2.2.2. Service-level requirements through chance constraints. Beyond the illustrative case study of the electric vehicle charging center problem, the work in this paper has broad applicability. Chance constraint models arise in Operations Research from the service-level requirement perspective. In inventory and production systems, service-level constraints are often used to specify the probability of stockout, fill-rate, or multi-period demand-fulfillment targets. Example settings are minimum-service inventory models [13],

service-constrained stochastic lot-sizing formulations [79], and chance-constraint models that impose service targets jointly over the planning horizon [71], cycle-service, fill-rate, and α/β -service measures in stochastic inventory control [40, 70]. A more recent data-driven line of work emphasizes target attainment rather than only nominal service feasibility. In particular, empirical newsvendor decisions can undershoot prescribed service levels, motivating distributionally robust chance-constrained formulations to achieve on-target service levels [80]. Probability-based quality-of-service (QoS) constraints also appear in call-center staffing [17], charging-station sizing with QoS requirements [44], and peak charging-demand satisfaction constraints [12]. Surgery-assignment and operating-room bin-packing models provide another service-level interpretation: elective cases with uncertain durations are packed into operating room blocks so that overtime or capacity violation remains below a prescribed probability [76, 84]; related surgery-planning models also impose distributionally robust chance constraints on waiting and overtime [20]. Relative to these works, our focus is not to introduce a new service-level measure; rather, we study how a chance-constraint service target can be robustified when multimodal uncertainty is represented by a Gaussian mixture with continuous mean–covariance support.

2.3. Optimal Transport (OT) with Gaussian and Gaussian-Mixture Models

Wasserstein metric-based ambiguity sets are motivated from the theory of optimal transport. A key enabler for OT-based modeling with Gaussian uncertainty is that the 2-Wasserstein distance between Gaussian laws admits a closed-form expression in terms of means and covariances, which yields a computable geometry on Gaussian families [30, 78]. It introduces a Riemannian viewpoint on the manifold of Gaussian densities that supports geodesics and gradient-based constructions directly in mean–covariance variables [57]. There has been work on OT constructions with direct operations on Gaussian mixture models [14] and structured mixture-restricted OT distances that preserve the mixture form while yielding finite-dimensional transport problems over component couplings [19]. These results provide a principled finite-dimensional geometry for studying Gaussian-mixture uncertainty representations without reverting to empirical OT computations over samples in the data space.

Because mixture-structured OT still requires solving an OT problem whose cost matrix involves pairwise Gaussian Wasserstein terms, recent work emphasizes scalable surrogates and approximations that retain mixture geometry. These include the minimized aggregated Wasserstein distance, developed for efficient mixture barycenter computation and later used to integrate heterogeneous datasets summarized by mixtures [53], as well as sliced and multi-sliced constructions, which replace high-dimensional OT with collections of one-dimensional transport problems for mixture comparison and Wasserstein-based Gaussian mixture learning [47, 67].

These developments are related to robust modeling with heterogeneous data sources, where the empirical evidence indicates multiple regimes and a mixture representation is natural. In Wasserstein DRO with

heterogeneous data sources, distributional ambiguity is modeled through OT neighborhoods around source-specific empirical laws and their composition, providing a principled mechanism to combine biased or non-identically distributed samples [73]. Such multi-source OT developments also motivate tractable robust counterparts beyond finite support. Mean–covariance robust risk measurement uses Burer–Gelbrich balls in mean–covariance space to form conservative outer approximations of Wasserstein ambiguity sets that become exact under Gaussian structure and yield a reformulated model whose size does not scale with the nominal support [64]. However, these advancements remain fragmented across distribution comparison, barycenter computation, and learning, and do not yet have a unified robust-optimization framework for multimodal uncertainty distributions having a continuous-support parameter set.

3. Wasserstein-2 Metric for Gaussian Mixtures

This section develops a theoretical basis for a metric between two Gaussian mixture distributions by allowing mixing weights and parameters of the Gaussians to vary over an admissible parameter set. The key mechanism is an optimal-transport coupling that transports mass between Gaussian components indexed by their parameters. This construction preserves analytic structure while remaining computationally amenable.

3.1. Wasserstein-2 Metric between Two Gaussian Distributions in Data Space

The following proposition highlights that the second moment of a Gaussian distribution is bounded.

Proposition 1 (Second moment of a Joint Gaussian in $\mathbb{R}^{2n}$). *Let $Z \sim \mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q})$ in $\mathbb{R}^{2n}$ where $\mathbf{Q}$ is a covariance matrix. Then*

$$\mathbb{E}[\|Z\|^2] = \|(\mathbf{m}_1, \mathbf{m}_2)\|^2 + \text{tr}(\mathbf{Q}) = \|\mathbf{m}_1\|^2 + \|\mathbf{m}_2\|^2 + \text{tr}(\mathbf{Q}) < \infty. \quad (6)$$

Proof. See proof in Appendix B.1. $\square$

Definition 3. *Let μ_1, μ_2 be two n -variate probability distributions. Then, the Wasserstein-2 metric between μ_1 and μ_2 is defined as*

$$W_2^2(\mu_1, \mu_2) := \inf_{\nu \in \Pi(\mu_1, \mu_2)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2), \quad (7)$$

where

$$\Pi(\mu_1, \mu_2) = \{\nu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n) : \nu^{(1)} = \mu_1, \nu^{(2)} = \mu_2\}, \quad (8)$$

and $\nu^{(1)}$ and $\nu^{(2)}$ are the marginal distributions of ν on $\mathbb{R}^n$. A $2n$ -variate distribution attaining the optimal value in (7) is called the Wasserstein-2 optimal coupling.

Theorem 1. [30, 82] *For any $\mathbf{s}_1 = (\mathbf{m}_1, \mathbf{Q}_1) \in \mathcal{S}$ and $\mathbf{s}_2 = (\mathbf{m}_2, \mathbf{Q}_2) \in \mathcal{S}$, the Wasserstein-2 optimal coupling between Gaussian distributions $\mathcal{N}(\mathbf{s}_1)$ and $\mathcal{N}(\mathbf{s}_2)$ is given by the Gaussian distribution $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{P}_2(\mathbb{R}^n \times \mathbb{R}^n)$, where $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2) = \mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \bar{\mathbf{Q}}) \in \mathcal{P}_2(\mathbb{R}^n \times \mathbb{R}^n)$ with*

$$\bar{\mathbf{Q}} = \begin{pmatrix} \mathbf{Q}_1 & \Sigma \\ \Sigma^\top & \mathbf{Q}_2 \end{pmatrix} \in \mathbb{S}_+^{2n}, \quad \text{and} \quad \Sigma = \mathbf{Q}_1^{\frac{1}{2}} (\mathbf{Q}_1^{-\frac{1}{2}} \mathbf{Q}_2 \mathbf{Q}_1^{-\frac{1}{2}})^{\frac{1}{2}} \mathbf{Q}_1^{\frac{1}{2}}. \quad (9)$$

The matrix $\Sigma \in \mathbb{S}_+^n$. The distribution $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2)$ is called Gelbrich coupling. The Wasserstein-2 metric between $\mathcal{N}(\mathbf{s}_1)$ and $\mathcal{N}(\mathbf{s}_2)$, called the Burer–Gelbrich metric is given by:

$$d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) = \|\mathbf{m}_1 - \mathbf{m}_2\|^2 + \text{tr} \left(\mathbf{Q}_1 + \mathbf{Q}_2 - 2(\mathbf{Q}_1^{\frac{1}{2}} \mathbf{Q}_2 \mathbf{Q}_1^{\frac{1}{2}})^{\frac{1}{2}} \right). \quad (10)$$

We denote the optimal coupling between $\mathcal{N}(s_1)$ and $\mathcal{N}(s_2)$ by $\mathcal{G}(s_1, s_2)$ and refer to it as Gelbrich's coupling. Here and throughout, for any symmetric positive semidefinite matrix $Q \succeq 0$, $Q^{1/2}$ denotes the principal (symmetric PSD) square root, i.e., the unique matrix $Q^{1/2} \succeq 0$ such that $Q^{1/2}Q^{1/2} = Q$ [7]. Moreover, whenever an inverse square root appears, we adopt the convention followed in Theorem 1, i.e., $Q^{-1/2} \equiv Q^{\dagger/2} := (Q^{1/2})^\dagger$, where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse with $Q^{\dagger/2} = Q^{-1/2}$ when $Q \succ 0$ [61, 65]. In particular, $Q^\dagger$ is characterized as the unique matrix satisfying $QQ^\dagger Q = Q$, $Q^\dagger QQ^\dagger = Q^\dagger$, $(QQ^\dagger)^\top = QQ^\dagger$, and $(Q^\dagger Q)^\top = Q^\dagger Q$. Equivalently, if $Q = U \text{diag}(\lambda_1, \dots, \lambda_n)U^\top$ with $\lambda_i \geq 0$, then

$$Q^{\dagger/2} = U \text{diag}(\lambda_1^{-1/2} \mathbf{1}_{\{\lambda_1 > 0\}}, \dots, \lambda_n^{-1/2} \mathbf{1}_{\{\lambda_n > 0\}})U^\top, \quad \text{and} \quad (Q^{\dagger/2})^2 = Q^\dagger.$$

3.2. Construction of Transport Maps between Parameter Spaces

This section constructs two transport maps $\mathcal{T}$ and $\mathcal{M}$, respectively, for joint parameter construction and marginal parameter extraction.

3.2.1. Joint parameter construction map $\mathcal{T}$. For $s_i = (m_i, Q_i) \in \mathcal{S}$, define the induced parameter map

$$\mathcal{T} : \mathcal{S} \times \mathcal{S} \rightarrow \mathcal{Z}, \quad \mathcal{T}(s_1, s_2) := ((m_1, m_2), \bar{Q}), \quad (11)$$

where $\bar{Q}$ is defined in (9) under the convention $Q^{-1/2} := (Q^{1/2})^\dagger$ for all $Q \succeq 0$. Consequently, $\mathcal{T}$ is well-defined on $\mathcal{S} \times \mathcal{S}$. Moreover, since $Q \mapsto Q^{1/2}$ and $Q \mapsto (Q^{1/2})^\dagger$ are Borel measurable on $\mathbb{S}_+^n$, the map $\mathcal{T}$ is Borel measurable. Hence, for any $A \in \mathcal{B}(\mathcal{Z})$, we define the preimage $\mathcal{T}^{-1}(A) \in \mathcal{B}(\mathcal{S} \times \mathcal{S})$ as

$$\mathcal{T}^{-1}(A) := \{(s_1, s_2) \in \mathcal{S} \times \mathcal{S} \mid \mathcal{T}(s_1, s_2) \in A\}.$$

Let $\pi \in \mathcal{P}_2(\mathcal{S} \times \mathcal{S})$ and $\mathcal{T}_\# \pi \in \mathcal{P}(\mathcal{Z})$ denote the probability measure such that

$$(\mathcal{T}_\# \pi)(A) = \pi(\mathcal{T}^{-1}(A)) = \int_{\mathcal{S} \times \mathcal{S}} \mathbb{1}_A(\mathcal{T}(s_1, s_2)) d\pi(s_1, s_2) \quad \forall A \in \mathcal{B}(\mathcal{Z}).$$

Then we call $\lambda := \mathcal{T}_\# \pi \in \mathcal{P}(\mathcal{Z})$ a push-forward measure. Consequently, for any measurable Ψ that is either nonnegative or integrable (or, more strongly, bounded) [82],

$$\int_{\mathcal{S} \times \mathcal{S}} \Psi(\mathcal{T}(s_1, s_2)) d\pi(s_1, s_2) = \int_{\mathcal{Z}} \Psi(z) d(\mathcal{T}_\# \pi)(z) = \int_{\mathcal{Z}} \Psi(z) d\lambda(z). \quad (12)$$

The map $\mathcal{T}$ preserves the second moment of the probability measure π , as the following proposition states.

Proposition 2 (Moment preservation under $\mathcal{T}$). *Let $\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$ and set $\lambda := \mathcal{T}_\# \pi \in \mathcal{P}(\mathcal{Z})$. If*

$$\int_{\mathcal{S} \times \mathcal{S}} \left(\|m_1(s_1)\|^2 + \|m_2(s_2)\|^2 + \|Q_1(s_1)\|_F^2 + \|Q_2(s_2)\|_F^2 \right) d\pi(s_1, s_2) < \infty, \quad (13)$$

then $\lambda \in \mathcal{P}_2(\mathcal{Z})$ with respect to the metric $\|((m_1, m_2), \bar{Q})\|_{\mathcal{Z}}^2 := \|m_1\|^2 + \|m_2\|^2 + \|\bar{Q}\|_F^2$, i.e.,

$$\int_{\mathcal{Z}} \left(\|m_1(z)\|^2 + \|m_2(z)\|^2 + \|\bar{Q}(z)\|_F^2 \right) d\lambda(z) < \infty. \quad (14)$$

Proof. See proof in Appendix B.3. □

3.2.2. Marginal parameter extraction map $\mathcal{M}$. For $z = ((m_1, m_2), Q) \in \mathcal{Z}$ with

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \in \mathbb{S}_+^{2n}, \quad Q_{11}, Q_{22} \in \mathbb{S}_+^n, \quad (15)$$

define the marginal-parameter extraction map

$$\mathcal{M} : \mathcal{Z} \rightarrow \mathcal{S} \times \mathcal{S}, \quad \mathcal{M}(z) := ((m_1, Q_{11}), (m_2, Q_{22})). \quad (16)$$

The map $\mathcal{M}$ is continuous (hence Borel measurable; see [27, Theorem 1.6]). For $B \in \mathcal{B}(\mathcal{S} \times \mathcal{S})$, the preimage of $\mathcal{M}$ is

$$\mathcal{M}^{-1}(B) := \{z \in \mathcal{Z} \mid \mathcal{M}(z) \in B\} \in \mathcal{B}(\mathcal{Z}).$$

Given a probability measure $\lambda \in \mathcal{P}_2(\mathcal{Z})$, define the pushforward measure $\pi := \mathcal{M}_\# \lambda \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$ by

$$(\mathcal{M}_\# \lambda)(B) := \lambda(\mathcal{M}^{-1}(B)) = \int_{\mathcal{Z}} \mathbb{1}_B(\mathcal{M}(z)) d\lambda(z) \quad \forall B \in \mathcal{B}(\mathcal{S} \times \mathcal{S}).$$

Hence, for any measurable Ψ that is either nonnegative or integrable (or, more strongly, bounded) [82],

$$\int_{\mathcal{Z}} \Psi(\mathcal{M}(z)) d\lambda(z) = \int_{\mathcal{S} \times \mathcal{S}} \Psi(s_1, s_2) d(\mathcal{M}_\# \lambda)(s_1, s_2) = \int_{\mathcal{S} \times \mathcal{S}} \Psi(s_1, s_2) d\pi(s_1, s_2). \quad (17)$$

The transport map $\mathcal{M}$ is also moment-preserving: if λ has finite second moment, so does the pushforward measure π as stated in the following proposition.

Proposition 3 (Moment preservation under $\mathcal{M}$). *Let $\lambda \in \mathcal{P}(\mathcal{Z})$ and set $\pi := \mathcal{M}_\# \lambda \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$. If*

$$\int_{\mathcal{Z}} (\|\mathbf{m}_1(z)\|^2 + \|\mathbf{m}_2(z)\|^2 + \|\mathbf{Q}(z)\|_F^2) d\lambda(z) < \infty, \quad (18)$$

then $\pi \in \mathcal{P}_2(\mathcal{S} \times \mathcal{S})$ with respect to the product gauge $\|((\mathbf{m}_1, \mathbf{Q}_{11}), (\mathbf{m}_2, \mathbf{Q}_{22}))\|_{\mathcal{S} \times \mathcal{S}}^2 := \|\mathbf{m}_1\|^2 + \|\mathbf{m}_2\|^2 + \|\mathbf{Q}_{11}\|_F^2 + \|\mathbf{Q}_{22}\|_F^2$, i.e.,

$$\int_{\mathcal{S} \times \mathcal{S}} (\|\mathbf{m}_1(s_1)\|^2 + \|\mathbf{m}_2(s_2)\|^2 + \|\mathbf{Q}_{11}(s_1)\|_F^2 + \|\mathbf{Q}_{22}(s_2)\|_F^2) d\pi(s_1, s_2) < \infty. \quad (19)$$

Proof. See proof in Appendix B.3. □

3.2.3. Marginalization properties of Gaussian measures. For any fixed Gaussian parameter, the corresponding Gaussian measure *in the data space* satisfies the standard marginalization property. In particular, for $(Y_1, Y_2) \sim \mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q}) \in \mathbb{R}^n \times \mathbb{R}^n$, with block decomposition $\mathbf{Q}$ in (15), the marginals satisfy $Y_1 \sim \mathcal{N}(\mathbf{m}_1, \mathbf{Q}_{11})$ and $Y_2 \sim \mathcal{N}(\mathbf{m}_2, \mathbf{Q}_{22})$. Define the continuous coordinate projections $\text{pr}_i : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ by $\text{pr}_1(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{y}_1$ and $\text{pr}_2(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{y}_2$. Their preimages for any $C \in \mathcal{B}(\mathbb{R}^n)$ are

$$\text{pr}_1^{-1}(C) := \{(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{R}^n \times \mathbb{R}^n \mid \text{pr}_1(\mathbf{y}_1, \mathbf{y}_2) \in C\}, \quad \text{pr}_2^{-1}(C) := \{(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{R}^n \times \mathbb{R}^n \mid \text{pr}_2(\mathbf{y}_1, \mathbf{y}_2) \in C\}. \quad (20)$$

Also, denote $\mathcal{M}_1(z) := s_1(z) = (\mathbf{m}_1, \mathbf{Q}_{11})$, $\mathcal{M}_2(z) := s_2(z) = (\mathbf{m}_2, \mathbf{Q}_{22})$. Then, by the definition of pushforward measure, for any $C \in \mathcal{B}(\mathbb{R}^n)$,

$$\mathcal{N}(s_1(z))(C) = \mathcal{N}(\mathcal{M}_1(z))(C) = ((\text{pr}_1)_\# \mathcal{N}(z))(C) = \mathcal{N}(z)(\text{pr}_1^{-1}(C)), \quad (21)$$

$$\mathcal{N}(s_2(z))(C) = \mathcal{N}(\mathcal{M}_2(z))(C) = ((\text{pr}_2)_\# \mathcal{N}(z))(C) = \mathcal{N}(z)(\text{pr}_2^{-1}(C)).$$

Equivalently, $(\text{pr}_1)_\# \mathcal{N}(z)$ and $(\text{pr}_2)_\# \mathcal{N}(z)$ are the two marginals of $\mathcal{N}(z)$ (in data space).

3.3. Wasserstein-2-type Metric between Gaussian Mixtures

We now define a Wasserstein-2-type metric between two Gaussian mixture distributions. This metric is related to the Wasserstein-2 metric.

Definition 4. Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^n)$. Let $\Pi^{\text{mix}}(\mu_1, \mu_2) = \{\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}\}$. Consider the following optimization problem:

$$W_2^{\text{mix}}(\mu_1, \mu_2) := \inf_{\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2) \quad (22)$$

An optimal solution to (22) exists (see Proposition 12 in Appendix B). We call $W_2^{\text{mix}}(\mu_1, \mu_2)$ a Wasserstein-2-type metric between μ_1 and μ_2 . The corresponding optimal solution is called an optimal Wasserstein-2-type coupling between μ_1 and μ_2 . We will refer to $W_2^{\text{mix}}(\mu_1, \mu_2)$ and the corresponding coupling as the Wasserstein-2 metric and coupling defined in the data space.

Remark 1. Since $\{\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}\} \subseteq \Pi(\mu_1, \mu_2)$, $W_2(\mu_1, \mu_2) \leq W_2^{\text{mix}}(\mu_1, \mu_2)$.

4. Transport via Gaussian Mixture Distribution Parameters

We now define a metric that can be computed from an optimization model defined in the parameter space. For $\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$, let π_1, π_2 denote its first and second marginals: $\pi_1(C) := \pi(C \times \mathcal{S})$, $\pi_2(C) := \pi(\mathcal{S} \times C)$ for Borel $C \in \mathcal{B}(\mathcal{S})$. Let

$$\Gamma = \left\{ \pi \in \mathcal{P}_2(\mathcal{S} \times \mathcal{S}) \mid \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(A) d\pi_1(\mathbf{s}) = \mu_1(A), \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(B) d\pi_2(\mathbf{s}) = \mu_2(B), \forall A, B \in \mathcal{B}(\mathbb{R}^n) \right\}, \quad (23)$$

$$\text{and} \quad W_2^{\text{prm}} := \inf_{\pi \in \Gamma} \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2). \quad (24)$$

Theorem 2. Let $W_2^{\text{mix}}(\mu_1, \mu_2)$ and W_2^{prm} be defined as in (22) and (24), respectively. Then

$$W_2^{\text{mix}}(\mu_1, \mu_2) = W_2^{\text{prm}}. \quad (25)$$

The importance of Theorem 2 is that it allows us to compute the Wasserstein-2-type metric defined in (22) by setting up an optimization model using the parameters of the Gaussians in the mixture. W_2^{prm} is more amenable to algorithmic development, as will be seen in subsequent sections. We further show in Theorem 3 that the equivalence continues to hold when the mean–covariance parameter support is a compact subset of $\mathbb{R}^n \times \mathbb{S}_+^n$. We prove Theorem 2 below; the restricted-support case is proved in Appendix B.4.

4.1. Constructing Data-Space Coupling from Parameter Space Coupling

The following result shows that given any parameter space coupling from Γ , we can construct a data space coupling in the form of a Gaussian mixture distribution ν_π with marginals μ_1, μ_2 , i.e., $\nu_\pi \in \Pi^{\text{mix}}(\mu_1, \mu_2)$.

Lemma 1. Let $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}$, $\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)$ be the Gelbrich coupling, and $\pi \in \Gamma$. Then, for $\nu_\pi := \int_{\mathcal{S} \times \mathcal{S}} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2) d\pi(\mathbf{s}_1, \mathbf{s}_2)$

$$\nu_\pi^{(1)}(A) := \int_{\mathbb{R}^{2n}} \mathbb{1}_A(\mathbf{y}_1) d\nu_\pi(\mathbf{y}_1, \mathbf{y}_2) = \mu_1(A) \quad \forall \text{ Borel set } A \in \mathcal{B}(\mathbb{R}^n), \text{ therefore, } \nu_\pi^{(1)} = \mu_1, \quad (26)$$

$$\nu_\pi^{(2)}(B) := \int_{\mathbb{R}^{2n}} \mathbb{1}_B(\mathbf{y}_2) d\nu_\pi(\mathbf{y}_1, \mathbf{y}_2) = \mu_2(B) \quad \forall \text{ Borel set } B \in \mathcal{B}(\mathbb{R}^n), \text{ therefore, } \nu_\pi^{(2)} = \mu_2. \quad (27)$$

Moreover, $\nu_\pi \in \Pi^{\text{mix}}(\mu_1, \mu_2)$.

Proof. First, for each $(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{S} \times \mathcal{S}$ and for Borel sets $A \in \mathcal{B}(\mathbb{R}^n)$,

$$\int_{\mathbb{R}^{2n}} \mathbb{1}_A(\mathbf{y}_1) d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(A, \mathbb{R}^n) = \mathcal{N}(\mathbf{s}_1)(A), \quad (28)$$

where the first equality holds since $\mathbb{1}_A(\mathbf{y}_1)$ does not depend on $\mathbf{y}_2$, and the second equality holds since the marginal for the Gelbrich coupling is $\mathcal{N}(\mathbf{s}_1)$. Similarly, $\int_{\mathbb{R}^{2n}} \mathbb{1}_B(\mathbf{y}) d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2) =$

$\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbb{R}^n, B) = \mathcal{N}(\mathbf{s}_2)(B)$ for any Borel set $B \in \mathcal{B}(\mathbb{R}^n)$. Therefore, it is sufficient to prove (26). Let $\Psi(\mathbf{y}_1, \mathbf{y}_2) := \mathbb{1}_A(\mathbf{y}_1)$. Clearly, $\Psi(\mathbf{y}_1, \mathbf{y}_2)$ is measurable and bounded since $0 \leq \Psi \leq 1$. Then,

$$\begin{aligned}
\nu_\pi^{(1)}(A) &= \int_{\mathbb{R}^{2n}} \mathbb{1}_A(\mathbf{y}_1) d\nu_\pi(\mathbf{y}_1, \mathbf{y}_2) = \int_{\mathbb{R}^{2n}} \Psi(\mathbf{y}_1, \mathbf{y}_2) d\nu_\pi(\mathbf{y}_1, \mathbf{y}_2) \\
&= \int_{\mathbb{R}^{2n}} \Psi(\mathbf{y}_1, \mathbf{y}_2) d\left(\int_{S \times S} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2) d\pi(\mathbf{s}_1, \mathbf{s}_2)\right) \text{ (since } \nu_\pi(\mathbf{y}_1, \mathbf{y}_2) = \int_{S \times S} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2) d\pi(\mathbf{s}_1, \mathbf{s}_2)) \\
&= \int_{S \times S} \left(\int_{\mathbb{R}^{2n}} \Psi(\mathbf{y}_1, \mathbf{y}_2) d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2)\right) d\pi(\mathbf{s}_1, \mathbf{s}_2) \text{ (using Kernel Fubini-Tonelli Theorem 8)} \\
&= \int_{S \times S} \left(\int_{\mathbb{R}^{2n}} \mathbb{1}_A(\mathbf{y}_1) d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2)\right) d\pi(\mathbf{s}_1, \mathbf{s}_2) \text{ (using definition of } \Psi) \\
&= \int_{S \times S} \mathcal{N}(\mathbf{s}_1)(A) d\pi(\mathbf{s}_1, \mathbf{s}_2) \text{ (using (28))} \\
&= \int_S \mathcal{N}(\mathbf{s}_1)(A) d\pi_1(\mathbf{s}_1) \text{ (integrate out the } \mathbf{s}_2\text{-variable)} \\
&= \mu_1(A). \text{ (since by definition of } \Gamma, \pi \text{ is a mixing law for } \mu_1)
\end{aligned}$$

Since this identity holds for every Borel set $A \in \mathcal{B}(\mathbb{R}^n)$, the measures coincide, and $\nu_\pi^{(1)} = \mu_1$. Similarly, $\nu_\pi^{(2)}(B) = \mu_2(B)$. Hence, $\nu_\pi \in \Pi(\mu_1, \mu_2)$. We now prove that $\nu_\pi \in \mathbb{GM}_{2n}$. Since $\lambda := \mathcal{T}_\# \pi \in \mathcal{P}(\mathcal{Z})$, we obtain the following chain of equality:

$$\begin{aligned}
\nu_\pi &= \int_{S \times S} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2) d\pi(\mathbf{s}_1, \mathbf{s}_2) \stackrel{\text{Theorem 1}}{=} \int_{S \times S} \mathcal{N}((\mathbf{m}_1(\mathbf{s}_1), \mathbf{m}_2(\mathbf{s}_2)), \bar{\mathbf{Q}}(\mathbf{s}_1, \mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2) \\
&\stackrel{(11)}{=} \int_{S \times S} \mathcal{N}(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2) \stackrel{\text{by (12)}}{=} \int_{\mathcal{Z}} \mathcal{N}(\mathbf{z}) d(\mathcal{T}_\# \pi)(\mathbf{z}) \stackrel{\text{by (12)}}{=} \int_{\mathcal{Z}} \mathcal{N}(\mathbf{z}) d\lambda(\mathbf{z}).
\end{aligned}$$

We finally show finiteness of the second moment of ν_π leveraging Propositions 1 and 2 from Section 3.

First, since $\|\mathbf{y}\|^2 \geq 0$, by the kernel Fubini-Tonelli Theorem 8 under the ν_π construction we have

$$\int_{\mathbb{R}^{2n}} \|\mathbf{y}\|^2 d\nu(\mathbf{y}) = \int_{\mathcal{Z}} \left(\int_{\mathbb{R}^{2n}} \|\mathbf{y}\|^2 d\mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \bar{\mathbf{Q}})(\mathbf{y}) \right) d\lambda(\mathbf{z}).$$

Additionally, by Proposition 1, since

$$\int_{\mathbb{R}^{2n}} \|\mathbf{y}\|^2 d\mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \bar{\mathbf{Q}})(\mathbf{y}) = \|(\mathbf{m}_1, \mathbf{m}_2)\|^2 + \text{tr}(\bar{\mathbf{Q}}),$$

it implies that

$$\int_{\mathbb{R}^{2n}} \|\mathbf{y}\|^2 d\nu(\mathbf{y}) = \int_{\mathcal{Z}} \left(\|(\mathbf{m}_1, \mathbf{m}_2)\|^2 + \text{tr}(\bar{\mathbf{Q}}) \right) d\lambda(\mathbf{z}).$$

Now note that for any $\mathbf{Q} \succeq 0$, hence for $\bar{\mathbf{Q}} \succeq 0$, $\text{tr}(\bar{\mathbf{Q}}) \leq \sqrt{2n} \|\bar{\mathbf{Q}}\|_F$ and $\|\bar{\mathbf{Q}}\|_F \leq 1 + \|\bar{\mathbf{Q}}\|_F^2$ [7, 35]. Hence, $\text{tr}(\bar{\mathbf{Q}}) \leq \sqrt{2n} (1 + \|\bar{\mathbf{Q}}\|_F^2)$. Therefore,

$$\begin{aligned}
\|(\mathbf{m}_1, \mathbf{m}_2)\|^2 + \text{tr}(\bar{\mathbf{Q}}) &\leq \|\mathbf{m}_1\|^2 + \|\mathbf{m}_2\|^2 + \sqrt{2n} (1 + \|\bar{\mathbf{Q}}\|_F^2) \leq \sqrt{2n} (\|\mathbf{m}_1\|^2 + \|\mathbf{m}_2\|^2 + \|\bar{\mathbf{Q}}\|_F^2) + \sqrt{2n} \implies \\
&\int_{\mathcal{Z}} \left(\|(\mathbf{m}_1, \mathbf{m}_2)(\mathbf{z})\|^2 + \text{tr}(\bar{\mathbf{Q}}(\mathbf{z})) \right) d\lambda(\mathbf{z}) \leq \sqrt{2n} \left(\int_{\mathcal{Z}} (\|\mathbf{m}_1(\mathbf{z})\|^2 + \|\mathbf{m}_2(\mathbf{z})\|^2 + \|\bar{\mathbf{Q}}(\mathbf{z})\|_F^2) d\lambda(\mathbf{z}) + \int_{\mathcal{Z}} \lambda(\mathbf{z}) \right),
\end{aligned}$$

where the first integral on the right-hand side is finite by Proposition 2 and the second integral is 1 since for any probability measure $\lambda \in \mathcal{P}(\mathcal{Z})$, $\int_{\mathcal{Z}} d\lambda(\mathbf{z}) = 1$. Thus, we obtain $\int_{\mathbb{R}^{2n}} \|\mathbf{y}\|^2 d\nu(\mathbf{y}) < \infty$, i.e., $\nu \in \mathcal{P}_2(\mathbb{R}^{2n})$.

Hence, $\nu_\pi \in \mathbb{GM}_{2n}$. Therefore, $\nu_\pi \in \Pi^{\text{mix}}(\mu_1, \mu_2)$. $\square$

4.2. Construction of Parameter-space Coupling from Data Space Coupling

The following lemma shows that a parameter-space coupling $\pi_\nu \in \Gamma$ can be constructed from a given data-space coupling $\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)$.

Lemma 2. Let $\mathcal{M} : \mathcal{Z} \rightarrow \mathcal{S} \times \mathcal{S}$ be a Borel measurable map defined in (16). Assume $\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)$ is given with corresponding mixing law $\lambda \in \mathcal{P}(\mathcal{Z})$. Let $\pi_\nu = \mathcal{M}_\# \lambda$ and $\pi_\nu^{(1)}, \pi_\nu^{(2)}$ be the marginals of π_ν on $\mathcal{P}(\mathcal{S})$. Then (i) $\pi_\nu^{(1)}$ is a mixing law for μ_1 , (ii) $\pi_\nu^{(2)}$ is a mixing law for μ_2 . Therefore, $\pi_\nu \in \Gamma$.

Proof. Since $\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2) = \{\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}\}$, we have

$$\nu^{(1)} = \mu_1, \quad \nu^{(2)} = \mu_2, \quad \nu = \int_{\mathcal{Z}} \mathcal{N}(z) d\lambda(z). \quad (29)$$

Also, given a map $\mathcal{M}$, $\pi_\nu(E) = \mathcal{M}_\# \lambda(E)$ for all Borel set $E \in \mathcal{B}(\mathcal{S} \times \mathcal{S})$. It is sufficient to prove that the first marginal of π_ν is a mixing law for μ_1 . In particular, we show that $\mu_1(B) = \int_{\mathcal{S}} \mathcal{N}(s_1)(B) d\pi_\nu^{(1)}(s_1)$ for any Borel set $B \in \mathcal{B}(\mathbb{R}^n)$. We proceed by first expressing $\mu_1(B)$ via ν and then via λ through the following steps. Let $\Psi(\mathbf{y}_1, \mathbf{y}_2) := \mathbb{1}_B(\mathbf{y}_1)$. Clearly, $\Psi(\mathbf{y}_1, \mathbf{y}_2)$ is measurable and bounded since $0 \leq \Psi \leq 1$. Then,

$$\begin{aligned} \mu_1(B) &= \int_{\mathbb{R}^n} \mathbb{1}_B(\mathbf{y}_1) d\mu_1(\mathbf{y}_1) = \int_{\mathbb{R}^n} \mathbb{1}_B(\mathbf{y}_1) d\nu^{(1)}(\mathbf{y}_1) \quad (\text{since } \nu^{(1)} = \mu_1) \\ &= \int_{\mathbb{R}^n \times \mathbb{R}^n} \Psi(\mathbf{y}_1, \mathbf{y}_2) d\nu(\mathbf{y}_1, \mathbf{y}_2) \quad (\text{By the definition of the first marginal } \nu^{(1)}) \\ &= \int_{\mathbb{R}^n \times \mathbb{R}^n} \Psi(\mathbf{y}_1, \mathbf{y}_2) d\left[\int_{\mathcal{Z}} \mathcal{N}(z) d\lambda(z)\right](\mathbf{y}_1, \mathbf{y}_2) \quad (\text{Using the mixture representation } \nu = \int_{\mathcal{Z}} \mathcal{N}(z) d\lambda(z)) \\ &= \int_{\mathcal{Z}} \left[\int_{\mathbb{R}^n \times \mathbb{R}^n} \Psi(\mathbf{y}_1, \mathbf{y}_2) d\mathcal{N}(z)(\mathbf{y}_1, \mathbf{y}_2)\right] d\lambda(z) \quad (\text{By the kernel Fubini–Tonelli Theorem 8}) \\ &= \int_{\mathcal{Z}} \mathcal{N}(s_1(z))(B) d\lambda(z), \end{aligned} \quad (30)$$

where we obtain the last equality as follows: since $(Y_1, Y_2) \sim \mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q})$, $Y_1 \sim \mathcal{N}(\mathbf{m}_1, \mathbf{Q}_{11}) = \mathcal{N}(s_1(z))$. Then, using the coordinate projection operation, for any Borel set $B \in \mathcal{B}(\mathbb{R}^n)$,

$$\mathcal{N}(s_1(z))(B) \stackrel{\text{by (21)}}{=} \mathcal{N}(z)(\text{pr}_1^{-1}(B)) \stackrel{\text{by (20)}}{=} \int_{\mathbb{R}^n \times \mathbb{R}^n} \mathbb{1}_B(\mathbf{y}_1) d\mathcal{N}(z)(\mathbf{y}_1, \mathbf{y}_2).$$

Now, our goal is to replace λ by $\pi_\nu^{(1)}$ in (30). Since $\mathcal{M}_\# \lambda = \pi_\nu$ with $\mathcal{M}(z) = (s_1(z), s_2(z))$ and $\text{pr}_1(\mathcal{M}(z)) = s_1(z)$, by integration rule (17) and finally by definition of $\pi_\nu^{(1)}$,

$$\begin{aligned} \int_{\mathcal{Z}} \mathcal{N}(s_1(z))(B) d\lambda(z) &= \int_{\mathcal{Z}} \mathcal{N}(\text{pr}_1(\mathcal{M}(z)))(B) d\lambda(z) \stackrel{\text{by (17)}}{=} \int_{\mathcal{S} \times \mathcal{S}} \mathcal{N}(\text{pr}_1(s_1, s_2))(B) d(\mathcal{M}_\# \lambda)(s_1, s_2) \\ &\stackrel{\text{by (17)}}{=} \int_{\mathcal{S} \times \mathcal{S}} \mathcal{N}(s_1)(B) d\pi_\nu(s_1, s_2) = \int_{\mathcal{S}} \mathcal{N}(s_1)(B) d\pi_\nu^{(1)}(s_1). \end{aligned}$$

Combining this equality with (30), we obtain

$$\mu_1(B) = \int_{\mathcal{S}} \mathcal{N}(s_1)(B) d\pi_\nu^{(1)}(s_1), \quad \forall B \in \mathcal{B}(\mathbb{R}^n),$$

which proves that $\pi_\nu^{(1)}$ is a mixing law for μ_1 .

Additionally, by Proposition 3, for any given ν , the second moment of π_ν is finite. Therefore, $\pi_\nu \in \Gamma$. $\square$

4.3. Parameter Space Wasserstein-2 Metric Equality

We are now in a position to prove the main result. In the following, Proposition 4 shows that $W_2^{\text{mix}} \leq W_2^{\text{prm}}$, whereas Proposition 5 shows $W_2^{\text{mix}} \geq W_2^{\text{prm}}$ and thereby, proves Theorem 2.

Proposition 4. Let $\Pi(\mu_1, \mu_2)$, $\mathbb{GM}_{2n}$, Γ and W_2^2 be as defined in (8), (5), (23) and (10), respectively. Then

$$W_2^{\text{mix}} = \inf_{\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2) \leq \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(s_1), \mathcal{N}(s_2)) d\pi(s_1, s_2)$$

for any $\pi \in \Gamma$. Hence, $W_2^{\text{mix}} \leq \inf_{\pi \in \Gamma} \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(s_1), \mathcal{N}(s_2)) d\pi(s_1, s_2)$.

Proof. Since $\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \geq 0$, and our constructed probability measure $\nu_\pi := \int_{\mathcal{S} \times \mathcal{S}} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2) d\pi(\mathbf{s}_1, \mathbf{s}_2)$ for a given $\pi \in \Gamma$, using Theorem 8 (Kernel Fubini-Tonelli), we have

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_\pi = \int_{\mathcal{S} \times \mathcal{S}} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2) \right) d\pi(\mathbf{s}_1, \mathbf{s}_2).$$

Additionally, from Theorem 1, for each fixed pair $(\mathbf{s}_1, \mathbf{s}_2)$,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2) = d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)).$$

Hence,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_\pi = \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2). \quad (31)$$

Now, because $\nu_\pi \in \mathbb{GM}_{2n}$ by Lemma 1, ν_π has finite second moment. Hence, the above integral is finite. Additionally, considering all elements ν from $\Pi^{\text{mix}}(\mu_1, \mu_2)$, one of which is ν_π , we obtain $\inf_{\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)} \int_{\mathbb{R}^{2n}} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu \leq \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_\pi$. Therefore,

$$W_2^{\text{mix}} = \inf_{\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)} \int_{\mathbb{R}^{2n}} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu \leq \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_\pi \stackrel{\text{by (31)}}{=} \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2).$$

Since this inequality holds for all $\pi \in \Gamma$, $W_2^{\text{mix}} \leq W_2^{\text{prm}}$. $\square$

We now establish $W_2^{\text{mix}} \geq W_2^{\text{prm}}$.

Proposition 5. Let $\Pi(\mu_1, \mu_2)$, $\mathbb{GM}_{2n}$, Γ and W_2^2 be as defined in (8), (5), (23) and (10), respectively. Then

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2) \geq \inf_{\pi \in \Gamma} \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2) = W_2^{\text{prm}}$$

for any $\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)$. Hence, $W_2^{\text{mix}} \geq W_2^{\text{prm}}$.

Proof. Let $\nu_z := \mathcal{N}(z) \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$. We first show that the following inequality holds pointwise in $z \in \mathcal{Z}$

$$\underbrace{\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_z(\mathbf{y}_1, \mathbf{y}_2)}_{\text{cost of this coupling}} \geq \underbrace{d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z)))}_{\text{lowest possible cost}}. \quad (32)$$

By construction, the marginals of ν_z are $\nu_z^{(1)} = \mathcal{N}(\mathbf{s}_1(z))$, $\nu_z^{(2)} = \mathcal{N}(\mathbf{s}_2(z))$. Hence, $\nu_z \in \Pi(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z)))$. Then considering all the elements (coupling) of $\Pi(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z)))$,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_z(\mathbf{y}_1, \mathbf{y}_2) \geq \inf_{\nu \in \Pi(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z)))} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2),$$

where the right term is equal to $d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z)))$ by definition. Hence, inequality (32) follows.

We next obtain $\int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi_\nu(\mathbf{s}_1, \mathbf{s}_2)$ as an equality to an integral of the right-hand side of (32) over $\lambda(z)$ using the push-forward technique.

Let $f(\mathbf{s}_1(z), \mathbf{s}_2(z)) := d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z)))$. Then, by (16) and integration identity (17)

$$\begin{aligned} \int_{\mathcal{Z}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z))) d\lambda(z) &= \int_{\mathcal{Z}} f(\mathbf{s}_1(z), \mathbf{s}_2(z)) d\lambda(z) \stackrel{(16)}{=} \int_{\mathcal{Z}} f(\mathcal{M}(z)) d\lambda(z) \stackrel{\text{by (17)}}{=} \\ &= \int_{\mathcal{S} \times \mathcal{S}} f(\mathbf{s}_1, \mathbf{s}_2) d(\mathcal{M}_\# \lambda)(\mathbf{s}_1, \mathbf{s}_2) = \int_{\mathcal{S} \times \mathcal{S}} f(\mathbf{s}_1, \mathbf{s}_2) d\pi_\nu(\mathbf{s}_1, \mathbf{s}_2) = \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi_\nu(\mathbf{s}_1, \mathbf{s}_2). \end{aligned} \quad (33)$$

From (32) we can write

$$\int_{\mathcal{Z}} \left[\int_{\mathbb{R}^{2n}} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_z(\mathbf{y}_1, \mathbf{y}_2) \right] d\lambda(z) \geq \int_{\mathcal{Z}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1(z)), \mathcal{N}(\mathbf{s}_2(z))) d\lambda(z),$$

and using (33) implies that

$$\int_{\mathcal{Z}} \left[\int_{\mathbb{R}^{2n}} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_z(\mathbf{y}_1, \mathbf{y}_2) \right] d\lambda(z) \geq \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi_\nu(\mathbf{s}_1, \mathbf{s}_2).$$

Additionally, since $\|y_1 - y_2\|^2 \geq 0$, $\nu_z := \mathcal{N}(z)$ and $\nu(y_1, y_2) = \int_{\mathcal{Z}} \nu_z(y_1, y_2) d\lambda(z) = \int_{\mathcal{Z}} \mathcal{N}(z)(y_1, y_2) d\lambda(z)$,

$$\begin{aligned} \int_{\mathcal{Z}} \left[\int_{\mathbb{R}^{2n}} \|y_1 - y_2\|^2 d\mathcal{N}(z)(y_1, y_2) \right] d\lambda(z) &= \int_{\mathbb{R}^{2n}} \|y_1 - y_2\|^2 d \left[\int_{\mathcal{Z}} \mathcal{N}(z)(y_1, y_2) d\lambda(z) \right] \\ &= \int_{\mathbb{R}^{2n}} \|y_1 - y_2\|^2 d\nu(y_1, y_2). \quad \text{Thus,} \end{aligned}$$

$$I(\nu) := \int_{\mathbb{R}^{2n}} \|y_1 - y_2\|^2 d\nu(y_1, y_2) \geq \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(s_1), \mathcal{N}(s_2)) d\pi_\nu(s_1, s_2) \quad \forall \nu \in \Pi^{\text{mix}}(\mu_1, \mu_2).$$

Moreover, since π_ν is one admissible element of Γ (see Lemma (2)), considering all elements $\pi \in \Gamma$ we have

$$I(\nu) \geq \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(s_1), \mathcal{N}(s_2)) d\pi_\nu(s_1, s_2) \geq \inf_{\pi \in \Gamma} \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(s_1), \mathcal{N}(s_2)) d\pi(s_1, s_2) = W_2^{\text{prm}}.$$

The above inequality holds for all $\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)$. Therefore, $\inf_{\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)} I(\nu) = W_2^{\text{mix}} \geq W_2^{\text{prm}}$. $\square$

In the following sections, we use the Gaussian-mixture ambiguity set induced by the Wasserstein-2-type metric introduced above to formulate the distributionally robust chance-constrained problem and to solve its strong-duality-based semi-infinite reformulation. In these developments, the chance-constraint expression is viewed as a function of both the decision vector x and the Gaussian component parameters. To ensure that the resulting formulation is well-posed and for the algorithmic analysis, we require regularity conditions on the parameter space of the component distributions (see Assumption 1). Theorem 3 (provided in Appendix B.4) is a counterpart of Theorem 2 under the assumption that the covariance matrices of the Gaussians are positive definite and the support for mean–covariance is compact.

Theorem 3 (Wasserstein-type distance on compact parameter support). *Let $\mathcal{S} := \Theta \times \Xi \subset \mathbb{R}_+^n \times \mathbb{S}_{++}^n$ be nonempty and compact. Define*

$$\Gamma_{\mathcal{S}} := \left\{ \pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S}) : \int_{\mathcal{S}} \mathcal{N}(s)(A) d\pi_1(s) = \mu_1(A), \int_{\mathcal{S}} \mathcal{N}(s)(B) d\pi_2(s) = \mu_2(B), \forall A, B \in \mathcal{B}(\mathbb{R}^n) \right\}.$$

For this restricted support $\mathcal{S}$, let

$$W_{2,\mathcal{S}}^{\text{prm}}(\mu_1, \mu_2) := \inf_{\pi \in \Gamma_{\mathcal{S}}} \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(s_1), \mathcal{N}(s_2)) d\pi(s_1, s_2),$$

and

$$W_{2,\mathcal{S}}^{\text{mix}}(\mu_1, \mu_2) := \inf_{\nu \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^{\mathcal{S}}} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|y_1 - y_2\|_2^2 d\nu(y_1, y_2), \quad (34)$$

where $\mathbb{GM}_{2n}^{\mathcal{S}}$ is the class of Gaussian-mixture couplings whose component parameters are induced by elements of $\mathcal{S} \times \mathcal{S}$. Then $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^{\mathcal{S}}$ is non-empty and an optimal solution to (34) exists. Additionally, every $\pi \in \Gamma_{\mathcal{S}}$ has finite second moment, and

$$W_{2,\mathcal{S}}^{\text{mix}}(\mu_1, \mu_2) = W_{2,\mathcal{S}}^{\text{prm}}(\mu_1, \mu_2). \quad (35)$$

5. Strong Duality of Robust CC and Semi-infinite Formulation

We now return to the distributionally robust chance-constrained problem (DR-C) where the ambiguity set is defined using the Wasserstein-type parameter space coupling developed in the previous section. From this section onward, we focus on the compact support set $\mathcal{S} := \Theta \times \Xi \subset \mathbb{R}_+^n \times \mathbb{S}_{++}^n$ as introduced above in Theorem 3. We first derive a strong dual form of the resultant robust chance constraint and thereby

develop a semi-infinite program formulation to (DR-C). We assume that the nominal distribution estimated from data is a mixture of finitely many Gaussians. In particular, the nominal distribution $\hat{P} \in \mathbb{GM}_n(\hat{\mathcal{S}}) := \left\{ \sum_{k=1}^{\hat{K}} w_k^0 \mathcal{N}(\hat{\mathbf{s}}_k) \mid \hat{\mathbf{w}} \in \hat{\mathcal{W}}, \hat{\mathbf{s}} \in \hat{\mathcal{S}} \right\}$, where $\hat{\mathcal{W}} := \{\mathbf{w} \in \mathbb{R}_+^{\hat{K}} \mid \sum_{k=1}^{\hat{K}} w_k = 1\}$, $\hat{\mathcal{S}} := \{(\mathbf{m}, \mathbf{Q}) \in \mathcal{S} \mid |\hat{\mathcal{S}}| = \hat{K}\}$ for some $\hat{K} \in \mathbb{N}$. Note that the nominal distribution is continuous despite the mixing distribution being finite.

We now define the ambiguity set of interest $\mathfrak{D}$ using the coupling set Γ and metric W_2^{prn} :

$$\mathfrak{D} := \left\{ \pi^{(2)} \in \mathcal{P}_2(\mathcal{S}) \mid \exists \pi \in \Gamma \subseteq \mathcal{P}_2(\hat{\mathcal{S}} \times \mathcal{S}) \text{ with } \begin{aligned} & \sum_{k=1}^{\hat{K}} \int_{\mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) \leq \rho, \\ & \pi^{(1)} := \int_{\mathcal{S}} d\pi(\hat{\mathbf{s}}, \mathbf{s}) = \hat{\mathbf{w}} \end{aligned} \right\}, \quad (36)$$

where, each $\pi^{(2)} \in \mathcal{P}_2(\mathcal{S})$ induces a Gaussian mixture distribution on $\mathbb{R}^n$ via $\int \mathcal{N}(\mathbf{m}, \mathbf{Q}) d\pi^{(2)}(\mathbf{m}, \mathbf{Q})$. Observe that the mixing law $\pi^{(2)} \in \mathcal{P}_2(\mathcal{S})$ is allowed to freely vary over a continuum of measures on the parameter space $\mathcal{S}$, unlike the restriction of marginal $\pi^{(1)}$ of π to $\hat{\mathcal{W}}$. Moreover, $\pi^{(2)} \in \mathcal{P}_2(\mathcal{S})$ since $\Gamma \subset \mathcal{P}_2(\hat{\mathcal{S}} \times \mathcal{S})$. Hence, we can equivalently encode the Wasserstein-type ambiguity set (36) via feasible transport plans $\pi \in \mathcal{D}$ with

$$\mathcal{D} := \left\{ \pi \in \Gamma \mid \int_{\mathcal{S}} d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) = \hat{w}_k \forall k \in [\hat{K}], \int_{\hat{\mathcal{S}} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}), \mathcal{N}(\mathbf{s})) d\pi(\hat{\mathbf{s}}, \mathbf{s}) \leq \rho \right\}, \quad (37)$$

and represent distributionally robust chance-constrained program (DR-C) as (CDR):

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}} \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad & \sum_{k=1}^{\hat{K}} \int_{\mathbf{s} \in \mathcal{S}} G(\mathbf{s}, \mathbf{x}) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) \geq \theta \quad \forall \pi \in \mathcal{D}, \\ \text{where} \quad & G(\mathbf{s}, \mathbf{x}) := \begin{cases} \mathbb{1}_{\geq 0}(b), & \text{if } \mathbf{x} = \mathbf{0}, \\ \Phi\left(\frac{b - \mathbf{m}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q} \mathbf{x}}}\right) & \text{otherwise.} \end{cases} \end{aligned} \quad (\text{CDR}) \quad (38)$$

An additional restriction in the definition of $G(\mathbf{s}, \mathbf{x})$ is given in Assumption 1. Also, the isolated instance at $\mathbf{x} = \mathbf{0}$ with $\mathbb{1}_{\geq 0}(b)$ is handled in Propositions 6 and 7 below.

Remark 2. We emphasize that the formulation (CDR) does not take support in the data space $\mathbb{R}^n$ (also called observation or sample-value space) as in the classical Wasserstein definition. In the classical setting, the random element is $\boldsymbol{\xi} \in \mathbb{R}^n$. In contrast, the random element in (CDR) is the Gaussian component parameter $\mathbf{s} = (\mathbf{m}, \mathbf{Q})$, and therefore the relevant support set is $\mathcal{S} \subseteq \mathbb{R}^n \times \mathbb{S}_+^n$ with $\mathbf{s} \in \mathcal{S}$. In the finite case, the classical model fixes a finite set of support points $\{\boldsymbol{\xi}_k\}_{k=1}^{\hat{K}} \subset \mathbb{R}^n$ and the distributional decision reduces to choosing weights on these points. Likewise, a finite counterpart of our model fixes a finite set of parameter support points $\{\mathbf{s}_k\}_{k=1}^{\hat{K}}$ (typically the nominal component parameters $\{\hat{\mathbf{s}}_k\}$), so the distributional decision also reduces to selecting weights on these known support points. However, for the continuous case as this study focuses on, the support points $\mathbf{s} \in \mathcal{S}$ are not fixed a priori. Hence, the optimization needs to select: (i) where to place mass in $\mathcal{S}$ (i.e., which Gaussian parameter values to use) and (ii) how much mass to assign.

Theorem 4 below provides an equivalent formulation of (CDR) as a semi-infinite program (S-Inf) by establishing strong duality for the robust chance constraint. This result is developed under the assumption

(Assumption 1) that the Gaussian parameter set $\mathcal{S} := \Theta \times \Xi$ is compact, and restricted to strictly positive definite covariance matrices with a lower bound on its smallest eigenvalue. Therefore, the Gaussian distributions on affine subspaces of $\mathbb{R}^n$ are excluded from the mixture components. This strict positive definiteness is used in proving the continuity result in Lemma 5. Additionally, the compactness of the parameter set in Assumption 1 provides a condition needed for the (finite) convergence of the cutting surface algorithm we develop in Section 5 to solve CDR. Problem (CDR) involves infinitely many decision variables, whereas its dual counterpart (S-Inf) has infinitely many constraints.

Assumption 1. Covariance matrices $\mathbf{Q}$ appearing in the Gaussian mixture distribution belong to $\mathbb{S}_{\delta}^n := \{\mathbf{Q} \in \mathbb{S}_{++}^n : \mathbf{Q} \succeq \hat{\delta} \mathbf{I}\}$ for some $\hat{\delta} > 0$. Additionally, the Gaussian parameters $(\mathbf{m}, \mathbf{Q})$ belong to a compact set $\mathcal{S} := (\Theta \times \Xi) \subseteq \mathbb{R}^n \times \mathbb{S}_{\delta}^n$.

Theorem 4 (Semi-infinite program formulation). *Let Assumption 1 hold, and suppose Wasserstein radius $\rho > 0$. Then, for any $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}$ satisfies the robust constraint in (CDR) if and only if there exists $\beta \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$ such that $\sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta$ and $\beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) \leq G(\mathbf{s}, \mathbf{x})$, $\forall \mathbf{s} \in \mathcal{S}$, $\forall k \in [\hat{K}]$. Consequently, solving (CDR) is equivalent to solving the following semi-infinite program*

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}, \beta \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+} \quad & \mathbf{c}^\top \mathbf{x} \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta, \\ \text{s.t.} \quad & \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) \leq G(\mathbf{s}, \mathbf{x}), \quad \forall \mathbf{s} \in \mathcal{S}, \forall k \in [\hat{K}]. \end{aligned} \quad (\text{S-Inf})$$

In particular, the formulations (CDR) and (S-Inf) have the same optimal objective value, and their optimal $\mathbf{x}$ -solutions coincide whenever an optimal solution exists.

The (CDR) and (S-Inf) formulations are infeasible at $\mathbf{x} = \mathbf{0}$ whenever $b < 0$, as shown in Propositions 6 and 7 respectively below.

Proposition 6. *If $b < 0$, then $\mathbf{x} = \mathbf{0}$ is infeasible for (CDR). Additionally, if $b \geq 0$, then for $\mathbf{x} = \mathbf{0}$, (39) admits an optimal solution $\tilde{\pi}$ with $\mathbf{w} = \hat{\mathbf{w}}$, i.e., $\tilde{\pi}$ is a self-coupling of $\hat{\mathbf{w}}$.*

Proof. See proof in Appendix B.6. □

Proposition 7. *Let $\beta \in \mathbb{R}^{\hat{K}}$. Then the following hold:*

- i) *if $\|\beta\| = 0$, then (S-Inf) is infeasible.*
- ii) *If $b < 0$, then $\mathbf{x} = \mathbf{0}$ cannot be a feasible solution to (S-Inf).*
- iii) *if $b \geq 0$, for $\mathbf{x} = \mathbf{0}$ and $k \in [\hat{K}]$, $\hat{\mathbf{s}}_k$ is an optimal solution to the following problem:*

$$\max_{\mathbf{s} \in (\Theta \times \Xi)} \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) - G(\mathbf{s}; \mathbf{0}).$$

Proof. See proof in Appendix B.6. □

Theorem 4 follows from considering a dual of:

$$\min_{\pi \in \mathcal{D}} \sum_{k=1}^{\hat{K}} \int_{\mathbf{s} \in \Theta \times \Xi} G(\mathbf{s}, \mathbf{x}) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}). \quad (39)$$

Lemma 3 establishes this intermediate result.

Lemma 3 (Strong Duality). *Let us consider the following problem:*

$$\begin{aligned} \max_{\beta \in \mathbb{R}^{\hat{K}+1}} \quad & \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1}, \quad \beta_{\hat{K}+1} \geq 0, \\ \text{s.t.} \quad & \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{s}_k), \mathcal{N}(s)) \leq G(s; \mathbf{x}), \quad \forall s \in \mathcal{S}, \forall k \in [\hat{K}] \end{aligned} \quad (40)$$

where $G(s; \mathbf{x})$ is as defined in (38). Then, under Assumption 1, $\text{Val}(39) = \text{Val}(40)$.

We prove strong-duality claim in Theorem 4 using [8, Theorem 1(a)] (see Appendix A(vi)). To apply this result, we first show continuity of both $d_{\text{BW}}^2(\mathcal{N}(\hat{s}), \mathcal{N}(s))$ and $G(s; \mathbf{x})$ in s , respectively, in Lemma 4 and Lemma 5. We then complete the proof of Theorem 4 by establishing Lebesgue integrability of $G(s, \cdot)$, thereby verifying all the conditions of [8, Theorem 1(a)].

Lemma 4. *Let $(\Theta \times \Xi)$ be as defined in Assumption 1. Fix $(\hat{\mathbf{m}}, \hat{\mathbf{Q}}) \in (\Theta \times \Xi)$ and define $d_{\text{BW}}^2(\mathbf{m}, \mathbf{Q}; (\hat{\mathbf{m}}, \hat{\mathbf{Q}})) := \|\hat{\mathbf{m}} - \mathbf{m}\|^2 + \text{tr}(\hat{\mathbf{Q}} + \mathbf{Q} - 2(\mathbf{Q}^{1/2} \hat{\mathbf{Q}} \mathbf{Q}^{1/2})^{1/2})$. Then the map $(\mathbf{m}, \mathbf{Q}) \mapsto d_{\text{BW}}^2(\mathbf{m}, \mathbf{Q}; (\hat{\mathbf{m}}, \hat{\mathbf{Q}}))$ is convex on $\mathbb{R}^n \times \mathbb{S}_+^n$. Moreover, $d_{\text{BW}}^2(\mathbf{m}, \mathbf{Q}; (\hat{\mathbf{m}}, \hat{\mathbf{Q}}))$ is continuous on $(\Theta \times \Xi)$.*

Proof. See proof in Appendix B.5. $\square$

Note that the continuity statement in the following lemma is with respect to s , treating $\mathbf{x}$ as a parameter.

Lemma 5. *Let $s \mapsto G(s; \mathbf{x})$ be a mapping where $G(s; \mathbf{x})$ is defined in (38). Then under Assumptions 1, $G(s; \mathbf{x})$ is continuous on $(\Theta \times \Xi)$.*

Proof. Let us consider (case-1): $\mathbf{x} \neq 0$ and (case-2): $\mathbf{x} = 0$ with (case-2a): $b \neq 0$ and (case-2b): $b = 0$.

To show the claim in (case-1), let $g(\mathbf{x}) = \frac{b - \mathbf{m}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q} \mathbf{x}}}$ and $\{(\mathbf{m}_t, \mathbf{Q}_t)\}_{t=1}^\infty$ be any sequence in $(\Theta \times \Xi)$ converging to $(\check{\mathbf{m}}, \check{\mathbf{Q}})$. Since $\mathbf{Q}, \check{\mathbf{Q}} \in \mathbb{S}_\delta^n \subset \mathbb{S}_{++}^n$ and hence $\mathbf{x}^\top \mathbf{Q} \mathbf{x} \geq \delta \|\mathbf{x}\|^2 > 0$ for any $\mathbf{x} \neq 0$,

$$\lim_{t \rightarrow \infty} g(\mathbf{m}_t, \mathbf{Q}_t) = \lim_{t \rightarrow \infty} \frac{b - \mathbf{m}_t^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q}_t \mathbf{x}}} = \lim_{t \rightarrow \infty} \frac{b - \mathbf{m}_t^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q}_t \mathbf{x}}} = \frac{b - \check{\mathbf{m}}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \check{\mathbf{Q}} \mathbf{x}}} = g(\check{\mathbf{m}}, \check{\mathbf{Q}}).$$

The second equality holds by [72, Theorem 4.4] and the second-to-last equality follows due to the continuity of the linear $b - \mathbf{m}^\top \mathbf{x}$ and quadratic functions $\mathbf{x}^\top \mathbf{Q} \mathbf{x}$ in $(\mathbf{m}, \mathbf{Q})$, and since $\mathbf{x}^\top \mathbf{Q} \mathbf{x} > 0$ for any $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$ by Assumption 1. Thus $g(\mathbf{m}, \mathbf{Q})$ is continuous on $\Theta \times \Xi$. Additionally, since $\Phi(\cdot)$ is continuous on $\mathbb{R}$, by [72, Theorem 4.7], composition $\Phi(g(\mathbf{m}, \mathbf{Q}))$ is also continuous at all $(\mathbf{m}, \mathbf{Q}) \in (\Theta \times \Xi)$.

To prove in (case-2a), we use $\epsilon - \delta$ definition of continuity². Observe that, since $\mathbf{x} = 0$ with $b \neq 0$, $\Phi(\frac{b - \mathbf{m}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q} \mathbf{x}}})$ is defined and

$$|\Phi(\frac{b - \mathbf{m}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q} \mathbf{x}}}) - \Phi(\frac{b - \mathbf{m}_0^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x}}})| = 0 \quad \text{for every } (\mathbf{m}, \mathbf{Q}).$$

Thus the $\epsilon - \delta$ condition

$$\exists \delta > 0 : \|(\mathbf{m}, \mathbf{Q}) - (\mathbf{m}_0, \mathbf{Q}_0)\| < \delta \implies |\Phi(\frac{b - \mathbf{m}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q} \mathbf{x}}}) - \Phi(\frac{b - \mathbf{m}_0^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x}}})| < \epsilon$$

is satisfied by any $\delta > 0$.

Finally in (case-2b), $G(s; 0)$ is a constant function in $(\mathbf{m}, \mathbf{Q})$ with value 1 by definition (38). Hence, it is continuous. Therefore, $G(\mathbf{m}, \mathbf{Q}; \mathbf{x})$ is continuous at every point in $(\Theta \times \Xi)$. $\square$

² A function is continuous at any arbitrary point $(\mathbf{m}_0, \mathbf{Q}_0)$, if for any given $\epsilon > 0$, $\exists$ some $\delta(\epsilon) > 0$ such that for any $(\mathbf{m}, \mathbf{Q}) \in (\Theta \times \Xi)$ with $\|(\mathbf{m}, \mathbf{Q}) - (\mathbf{m}_0, \mathbf{Q}_0)\| < \delta(\epsilon)$ implies $|\Phi(\frac{b - \mathbf{m}^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q} \mathbf{x}}}) - \Phi(\frac{b - \mathbf{m}_0^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q}_0 \mathbf{x}}})| < \epsilon$.

Remark 3. We note that when $\mathbf{x} \neq \mathbf{0}$, if one allows either $(\mathbf{m}, \mathbf{Q}) \rightarrow (\mathbf{0}, \mathbf{0})$ with $b = 0$ or simultaneously $\mathbf{Q} \rightarrow \mathbf{0}$ and $\mathbf{m} \rightarrow \mathbf{m}_0 \neq \mathbf{0}$ so that $b - \mathbf{m}_0^\top \mathbf{x} = 0$, then $\lim_{(\mathbf{m}, \mathbf{Q}) \rightarrow (\mu_0, \mathbf{0})} g(\mathbf{m}, \mathbf{Q})$ is path-dependent and hence $\Phi(g(\mathbf{m}, \mathbf{Q}))$ may fail to be continuous at $(\mu_0, \mathbf{0})$ (see [72, Chapter 9] for counterexamples of continuity). We require regularity condition $\mathbf{x}^\top \mathbf{Q} \mathbf{x} \geq \|\delta\|^2 > 0$ for some $\delta > 0$ to overcome such an issue since $\mathbf{Q} \succeq \mathbf{0}, \mathbf{Q} \neq \mathbf{0}$ can still lead to $\mathbf{x}^\top \mathbf{Q} \mathbf{x} = 0$ if $\mathbf{x}$ lies entirely in the null-space of $\mathbf{Q}$.

Based on the continuity and Lebesgue integrability results, the proof of Theorem 4 is immediate:

Proof. (of Lemma 3) Proofs of Lemmas 4 and 5 respectively established the continuity of the function d_{BW}^2 and $G(\mathbf{s}; \mathbf{x})$ in $\mathbf{s}$. Moreover, for any given $\mathbf{x}$, $0 \leq G(\mathbf{s}; \mathbf{x}) \leq 1 \forall \mathbf{s} \in \Theta \times \Xi$. Then compactness of $\Theta \times \Xi$ (by Assumption 1) implies that

$$\int_{\mathbf{s} \in (\Theta \times \Xi)} |G(\mathbf{s}; \mathbf{x})| d\nu(\mathbf{s}) \leq \int_{\Theta \times \Xi} 1 d\nu = \nu(\Theta \times \Xi) < \infty.$$

Hence, $G(\mathbf{s}; \mathbf{x})$ is Lebesgue integrable with respect to Lebesgue measure ν . Therefore, strong duality immediately follows from [8, Theorem 1(a)]. $\square$

Proof. (of Theorem 4) See proof in Appendix B.5. $\square$

As a consequence of Theorem 4, we obtain a dual formulation of a finitely supported distributionally robust chance constraint model:

Corollary 1. Let $\mathcal{D}_d := \left\{ \boldsymbol{\pi} \in \mathbb{R}_+^{\hat{K} \times K} : \sum_{k=1}^{\hat{K}} \sum_{l=1}^K d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\hat{\mathbf{s}}_l)) \pi_{kl} \leq \rho \right\}$, $\sum_{l=1}^K \pi_{kl} = \hat{w}_k \forall k \in [\hat{K}]$. Consider the optimization problem

$$\min_{\mathbf{x} \in \mathcal{X}} \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \sum_{k=1}^{\hat{K}} \sum_{l=1}^K G(\mathbf{s}_l, \mathbf{x}) \pi_{kl} \geq \theta \quad \forall \boldsymbol{\pi} \in \mathcal{D}_d, \quad (\text{FDR})$$

Then, for any $\mathbf{x} \in \mathcal{X}$ satisfying the robust constraint in (FDR) if and only if there exists $\boldsymbol{\beta} \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$ such that $\sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta$ and $\beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\hat{\mathbf{s}}_l)) \leq G(\hat{\mathbf{s}}_l; \mathbf{x})$, $\forall k \in [\hat{K}], \forall l \in [K]$.

Consequently, solving (FDR) is equivalent to solving (41):

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}, \boldsymbol{\beta} \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+} \quad & \mathbf{c}^\top \mathbf{x} \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta, \\ \text{s.t.} \quad & \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s}_l)) \leq G(\mathbf{s}_l; \mathbf{x}), \quad \forall k \in [\hat{K}], \forall l \in [K]. \end{aligned} \quad (41)$$

In particular, (FDR) and (41) have the same optimal value, and their optimal $\mathbf{x}$ -solutions coincide whenever an optimal solution exists.

Note that, unlike Theorem 4, Corollary 1 does not require $\rho > 0$, because its inner worst-case problem, when feasible, is a finite-dimensional linear program with finite optimal value, and dual attainment follows from finite LP duality. We also note that a special case of (FDR) is when the Wasserstein-ball is supported on the same finite set of parameter values $\mathbf{s}_k$ that are used in defining the nominal distribution, with $K = \hat{K}$.

6. A Cutting Surface Algorithm and its Convergence

The cutting surface algorithm (CSA) framework is a general-purpose approach for solving semi-infinite programs [31, 33, 56, 58, 68]. This section focuses on developing CSA for the semi-infinite formulation (S-Inf). Algorithm 1 gives pseudo-code based on the main steps of the CSA, and its use to obtain a solution of the

user-desired tolerance $\hat{\tau}$ for (S-Inf). The user-desired tolerance $\hat{\tau}$ provides a parameter that is iteratively reduced to set the level of accuracy ϵ_j in developing inner and outer approximations of the subproblems. The convergence of the detailed algorithm depends on the solvability of the master problem (S-Inf) and the subproblems to this accuracy, which are studied in Sections 6.1 and 6.2, respectively. The accuracy ϵ_j is iteratively set because achieving a user-defined tolerance $\hat{\tau}$ depends on an unknown constant, which is known to exist; but computing its value is not practical.

6.1. Cutting-Surface Algorithm

For a given accuracy parameter ϵ_j , the CSA Algorithm 1 generates solutions to the increasingly tighter relaxations of (S-Inf). The master problem (42) is an outer approximation of (S-Inf) using finitely many constraints generated from a finite subset of elements in the set $(\Theta \times \Xi)$. Suppose, at a master iteration ℓ , set $\mathcal{S}_\ell$ contains l elements of $(\Theta \times \Xi)$ generated from previous ℓ -iterations. At an initialization step, these elements can be the parameters of the Gaussians in the nominal mixture distribution. Let a solution of the corresponding master problem:

$$\min_{\mathbf{x}, \boldsymbol{\beta}} \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{X}, \quad \boldsymbol{\beta} \in \mathfrak{B} := \left\{ \boldsymbol{\beta} \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+ \mid \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta \right\}, \quad (42)$$

$$J_k(\mathbf{s}_l) := \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s}_l)) \leq G(\mathbf{s}_l; \mathbf{x}), \quad \forall \mathbf{s}_l \in \mathcal{S}_\ell, k \in [\hat{K}]$$

be $(\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})$. Given this solution, CSA solves $[\hat{K}]$ subproblems (43):

$$\max_{\mathbf{s} \in (\Theta \times \Xi)} J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})) = \beta_k^{(\ell)} - \beta_{\hat{K}+1}^{(\ell)} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) - G(\mathbf{s}; \mathbf{x}^{(\ell)}) \quad \text{for each } k \in [\hat{K}] \quad (43)$$

for all $k \in [\hat{K}]$. Let the solutions to these subproblems be $\mathbf{s}_k^{(\ell+1)}$. We check the termination criteria based on sufficient constraint violation (i.e., $J_k(\mathbf{s}_k^{(\ell+1)}; (\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})) > \epsilon_j$). If violated constraints are identified, Step 2 in Algorithm 1 adds a set of new points $\{\mathbf{s}_k^{(\ell+1)}\}$ to the existing set $\mathcal{S}_\ell$, i.e., $\mathcal{S}_{\ell+1} \leftarrow \mathcal{S}_\ell \cup \{\mathbf{s}_k^{(\ell+1)}\}$ that provide these violated constraints. Note that prior to entering into Step 2, Algorithm 1 checks if the current iterate is small to handle the singularity at $\mathbf{x} = 0$. If one or more new violated constraints are found, then the algorithm adds these constraints to the master problem and proceeds to the next master iteration.

The following definitions are used in establishing the finite convergence of Algorithm 1 as stated in Theorem 5.

Definition 5 ($\hat{\tau}$ -accurate and $\hat{\tau}$ -optimal solutions). Assume that an optimization problem $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ with a nonempty feasible region $\mathcal{X}$ has a finite optimal objective value $f(\mathbf{x}^*)$ attained at a solution $\mathbf{x}^*$. A solution $\mathbf{x}_\tau^* \in \mathcal{X}$ is called $\hat{\tau}$ -accurate for some $\hat{\tau} > 0$ if it is $\hat{\tau}$ -feasible and its objective discrepancy from the true optimum satisfies $|f(\mathbf{x}_\tau^*) - f(\mathbf{x}^*)| \leq \hat{\tau}$. Moreover, $\mathbf{x}_\tau^*$ is called $\hat{\tau}$ -optimal if it is $\hat{\tau}$ -feasible and $f(\mathbf{x}_\tau^*) - f(\mathbf{x}^*) \leq \hat{\tau}$.

Definition 6. Let $\mathfrak{B} := \left\{ \boldsymbol{\beta} \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+ \mid \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta \right\}$ and, for some given $\hat{\mathbf{x}}$, $g(\mathbf{s}; \hat{\mathbf{x}}) := (b - \mathbf{m}^\top \hat{\mathbf{x}}) / \sqrt{\hat{\mathbf{x}}^\top \mathbf{Q} \hat{\mathbf{x}}}$. Further, let the constraint set of the master problem (42) and the semi-infinite problem (S-Inf) be respectively $\mathcal{F}_\ell(0) := \{(\mathbf{x}, \boldsymbol{\beta}) \mid \mathbf{x} \in \mathcal{X}, \boldsymbol{\beta} \in \mathfrak{B}, J_k(\mathbf{s}_l) \leq 0, \mathbf{s}_l \in \mathcal{S}_\ell, k \in [\hat{K}]\}$ and $\mathcal{F}(0) := \{(\mathbf{x}, \boldsymbol{\beta}) \mid \mathbf{x} \in \mathcal{X}, \boldsymbol{\beta} \in \mathfrak{B}, J_k(\mathbf{s}) \leq 0, \mathbf{s} \in \mathcal{S}, k \in [\hat{K}]\}$. Then we define the optimal objective

Algorithm 1 A cutting-surface algorithm (CSA) to solve (CDR).**Require:** Optimality Tolerance $\hat{\tau}$ **Ensure:** A $\hat{\tau}$ -optimal solution or infeasibility1: Set $\mathcal{S}_0 \leftarrow \{\hat{s}_1, \dots, \hat{s}_{\hat{K}}\}$, $\ell \leftarrow 0$, $j \leftarrow 1$, $\delta_o \leftarrow 0$ $lb \leftarrow -\infty$.2: **while true do**3: Set $\epsilon_j \leftarrow \hat{\tau}/2^j$ 4: Set $z_{\epsilon_j} \leftarrow \Phi^{-1}(1 - \min\{\epsilon_j, \frac{1}{2}\})$ 5: **Step 1. Solve Master Problem:** Solve (44) at ϵ_j -feasibility

$$\min \left\{ \mathbf{c}^\top \mathbf{x} : \mathbf{x} \in \mathcal{X}, \beta \in \mathfrak{B}, J_k(\mathbf{s}_l) \leq 0, \mathbf{s}_l \in \mathcal{S}_\ell, k \in [\hat{K}] \right\}. \quad (44)$$

6: **if** (44) is infeasible **then**7: **return** Infeasibility Status8: **end if**9: Denote $(\mathbf{x}^{(\ell)}, \beta^{(\ell)})$ as the ϵ_j -feasible solution of (44)10: Set $\delta_{\epsilon_j} \leftarrow \frac{b}{\bar{m}_\Theta + \sqrt{\bar{\lambda}_\Xi} z_{\epsilon_j}}$ for $\bar{m}_\Theta := \max_{\mathbf{m} \in \Theta} \|\mathbf{m}\|$, $\bar{\lambda}_\Xi := \max_{\mathbf{Q} \in \Xi} \lambda_{\max}(\mathbf{Q})$.11: **if** $\|\mathbf{x}^{(\ell)}\| \leq \delta_{\epsilon_j}$ **then**12: Set $\mathbf{s}^{(\ell+1)} \leftarrow \mathbf{s}^{(\ell)}$ and go to Step 3.13: **end if**14: **Step 2. Solve Subproblems:** Solve all $\hat{K}$ sub-problems (43) with ϵ_j -accuracy and define $\mathcal{S}_{\ell+1}$ and $\mathbf{s}^{(\ell+1)}$ as follows

$$\mathcal{S}_{\ell+1} := \bigcup_{k \in [\hat{K}]} \left\{ \mathbf{s}_k^{(\ell+1)} \mid \mathbf{s}_k^{(\ell+1)} \text{ is an } \epsilon_j\text{-accurate solution to } \max_{\mathbf{s} \in \mathcal{S}} J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})), J_k(\mathbf{s}_k^{(\ell+1)}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) > \epsilon_j \right\}$$

$$\mathbf{s}^{(\ell+1)} \in \arg \max_{\mathbf{s} \in \mathcal{S}_{\ell+1}} J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) \quad (\text{break tie arbitrarily}).$$

15: **Step 3. Check the Termination Criteria:**16: **if** $J(\mathbf{s}^{(\ell+1)}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) \leq \epsilon_j$ **then**17: Solve (44) at ϵ_j -optimality and obtain a feasible solution $(\tilde{\mathbf{x}}, \tilde{\beta})$ to (S-Inf)18: $lb \leftarrow \mathbf{c}^\top \tilde{\mathbf{x}}$ 19: **if** $\mathbf{c}^\top (\tilde{\mathbf{x}} - \mathbf{x}^{(\ell)}) \leq \hat{\tau}$ **then**20: **return** $\mathbf{x}^{(\ell)}$ 21: **else**22: Set $j \leftarrow j + 1$ 23: **end if**24: **end if**25: $\mathcal{S}_{\ell+1} \leftarrow \mathcal{S}_\ell \cup \mathcal{S}_{\ell+1}$, $\ell \leftarrow \ell + 1$.26: **end while**

value as a function of the right-hand side parameter $\tau \in \mathbb{R}$ for the master problem (42) and subproblem (43), respectively, as:

$$V_\ell^*(\tau) = \min \left\{ \mathbf{c}^\top \mathbf{x} \mid (\mathbf{x}, \beta) \in \mathcal{F}(\tau) \right\} = \left\{ (\mathbf{x}, \beta) \mid \mathbf{x} \in \mathcal{X}, \beta \in \mathfrak{B}, J_k(\mathbf{s}_l) \leq \tau, \mathbf{s}_l \in \mathcal{S}_\ell, k \in [\hat{K}] \right\}$$

$$V_{sub}^*(\tau) = \max \left\{ J(\mathbf{s}, \zeta'; \hat{\mathbf{x}}, \hat{\beta}) \mid (\mathbf{s}, \zeta') \in \mathcal{Q}(\tau) := \{(\mathbf{s}, \zeta') \mid \mathbf{s} \in \mathcal{S}, \zeta' \in [-1, 0], \tau \geq \Phi(g(\mathbf{s}; \hat{\mathbf{x}})) + \zeta'\} \right\},$$

with $(\mathbf{x}^*(\tau), \beta^*(\tau))$ and $(\mathbf{s}^*(\tau), \zeta^*(\tau))$ being the solution corresponding to $V_\ell^*(\tau)$ and $V_{sub}^*(\tau)$.

Theorem 5. Consider the adaptive cutting-surface Algorithm 1. Suppose that every master-problem and separation-subproblem call invoked by the algorithm is solved according to the accuracy requirement stated in the corresponding steps of the algorithm. Then given a user-specified tolerance $\hat{\tau} > 0$,

- i) Algorithm 1 attains a $\hat{\tau}/2$ -feasible solution $\hat{\mathbf{x}}$ of (S-Inf) with $\mathbf{c}^\top \hat{\mathbf{x}} \leq \text{val}(\text{S-Inf})$ after adding finitely many points from $\mathcal{S}$ to $\mathcal{S}_\ell$.
- ii) Let ℓ^* be the termination iteration of Algorithm 1. Assume that, for every $\varepsilon > 0$, there exists an oracle $\mathcal{O}$ returning $(\tilde{\mathbf{x}}, \tilde{\beta}) \in \arg \min \{ \mathbf{c}^\top \mathbf{x} : (\mathbf{x}, \beta) \in \mathcal{F}(-\varepsilon) \}$. Then Algorithm 1 terminates finitely, returns a $\hat{\tau}$ -accurate certificate $\mathbf{x}^{(\ell^*)}$, and yields a $\hat{\tau}$ -optimal feasible solution $\tilde{\mathbf{x}}$ to (S-Inf).

Theorem 5 requires solving both the master problem (42) and the subproblems (43) at each master iteration. The function $G(\mathbf{x}, \mathbf{s})$ defined in (38) appears in both the master problem and subproblems, and it involves the function Φ for which a closed analytical expression is not known. Hence, we first develop

approximate formulations of these problems together with an explicit mechanism to control the approximation accuracy of the resulting feasible set. To identify verifiable conditions under which insufficient constraint violation identified through the solutions of the subproblems translates to a $\hat{\tau}/2$ -feasible solution of (S-Inf), we study the master and subproblem optimal-value functions by analyzing $G(x, s)$ as a function of both x and s , and through properties of $\Phi(z)$ for $z \in \mathbb{R}$.

The univariate function $\Phi(z)$ is convex in z when $z \leq 0$ and concave when $z \geq 0$. Using convexity, concavity, and curvature properties of $\Phi(z)$, [22] recently developed outer and inner approximations of a finite Gaussian mixture-based chance constraint optimization problem. We restate results from [22] here for completeness. We consider a piecewise-linear outer (PWL-O) approximation-based formulation for (42) while the piecewise-linear inner approximation (PWL-I) is used in (43). Since (43) is a maximization problem, using its inner approximation provides an overall outer approximation to the feasible set of (S-Inf). In the PWL-O approximation of $\Phi(z)$, we approximate the convex part of $\Phi(z)$ with secant lines and its concave part with tangent lines. For the PWL-I approximation, they are respectively swapped to tangent and secant lines.

Let $\tilde{z} = (\tilde{z}_{-L}, \tilde{z}_{-L+1}, \dots, \tilde{z}_{-1}, \tilde{z}_0, \tilde{z}_1, \dots, \tilde{z}_{R-1}, \tilde{z}_R) \in \mathbb{R}^{L+R+1}$ be some ordered array of breakpoints on z -axis, parameterized by integers $L, R > 0$. Let us define the set of tangent and secant lines for PWL-O so that the curve $\bar{\Phi}(\cdot; \tilde{z})$, as defined below, always lies above that of Φ (see Figure 3a).

$$\bar{\Phi}(z; \tilde{z}) := \begin{cases} \min \left\{ 1, \min_{i \in [R]_0} \{g_i z + g_i^0\} \right\}, & \text{if } z \geq 0, \text{ } g_i := \phi(\tilde{z}_i), \text{ } g_i^0 = \Phi(\tilde{z}_i) - \phi(\tilde{z}_i)\tilde{z}_i, \text{ (Tangent)} \\ \max \left\{ \Phi(\tilde{z}_{-L}), \max_{i \in [L]} \{h_i z + h_i^0\} \right\}, & \text{otherwise where, } h_i := \frac{\Phi(\tilde{z}_{-i+1}) - \Phi(\tilde{z}_{-i})}{\tilde{z}_{-i+1} - \tilde{z}_{-i}}, \\ & h_i^0 = \Phi(\tilde{z}_{-i}) - \frac{\Phi(\tilde{z}_{-i+1}) - \Phi(\tilde{z}_{-i})}{\tilde{z}_{-i+1} - \tilde{z}_{-i}} \tilde{z}_{-i}, \text{ (Secant)} \end{cases} \quad (45)$$

Similarly, for PWL-I, the curve $\underline{\Phi}(\cdot, \tilde{z})$ defined in (46) lies below Φ :

$$\underline{\Phi}(z; \tilde{z}) = \begin{cases} \min \left\{ \Phi(\tilde{z}_R), \min_{i \in [R]} \{g_i z + g_i^0\} \right\}, & \text{if } z \geq 0, \text{ (Secant)} \\ \max \left\{ 0, \max_{i \in [L]_0} \{h_i z + h_i^0\} \right\}, & \text{otherwise, (Tangent)} \end{cases} \quad (46)$$

where

$$g_i := \frac{\Phi(\tilde{z}_i) - \Phi(\tilde{z}_{i-1})}{\tilde{z}_i - \tilde{z}_{i-1}}, \text{ } g_i^0 := \Phi(\tilde{z}_{i-1}) - \frac{\Phi(\tilde{z}_i) - \Phi(\tilde{z}_{i-1})}{\tilde{z}_i - \tilde{z}_{i-1}} \tilde{z}_{i-1}, \text{ } i \in [R],$$

and $h_i := \phi(\tilde{z}_i), \text{ } h_i^0 = \Phi(\tilde{z}_i) - \phi(\tilde{z}_i)\tilde{z}_i, \text{ } i \in [L]_0.$

We now provide a relaxation of the master problem (42), replacing Φ with $\bar{\Phi}$, followed by its equivalent mixed-integer reformulation. An approximate formulation of subproblem (43) uses $\underline{\Phi}$.

Proposition 8. *Let $\tilde{z}$ be an ordered array of breakpoints of z -axis and $\Phi(\cdot)$ be replaced with its PWL outer approximation $\bar{\Phi}(\cdot; \tilde{z})$ defined in (45). Then, the resulting formulation (47a) is a relaxation to master problem (42):*

$$\begin{aligned}
\min \quad & \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{X}, \quad \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \beta_{\hat{K}+1} \rho \geq \theta, \quad \boldsymbol{\beta} \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+ \\
& \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s}_l)) - \zeta_l \leq 0 \quad \forall k \in [\hat{K}] \\
& \bar{\Phi}(z_l; \tilde{\mathbf{z}}) \geq \zeta_l, \quad b - \mathbf{m}_l^\top \mathbf{x} \geq z_l \sigma_l, \quad \mathbf{x}^\top \mathbf{Q}_l \mathbf{x} = \sigma_l^2, \\
& \zeta_l \in [0, 1], \quad z_l \in \mathbb{R}, \quad \sigma_l \in \mathbb{R}_+ \quad \forall \mathbf{s}_l \in \mathcal{S}_\ell.
\end{aligned} \tag{47a}$$

Additionally (ignoring subscript l), $\exists(z, \zeta) \in \mathbb{R} \times \mathbb{R}$ satisfying $\zeta \leq \bar{\Phi}(z; \tilde{\mathbf{z}})$ if and only if $\mathcal{H}^O(\tilde{\mathbf{z}}) \neq \emptyset$,

$$\mathcal{H}^O(\tilde{\mathbf{z}}) := \left\{ \begin{array}{l} \boldsymbol{\alpha} \in \mathbb{R}_+^L, \\ \mathbf{t} \in \{0, 1\}^3, \\ \mathbf{y} \in \mathbb{R}^3, \\ z, \zeta \in \mathbb{R} \end{array} \left| \begin{array}{l} \Phi(\tilde{z}_{-L})t_1 + t_2 + \mathbf{g}_i y_3 + \mathbf{g}_i^0 t_3 \geq \zeta \quad i \in [R]_0, \quad \Phi(\tilde{z}_{-L})t_1 + \sum_{i=0}^L \alpha_i \Phi(\tilde{z}_{-i}) + t_3 \geq \zeta, \\ t_1 + t_2 + t_3 = 1, \quad z = y_1 + y_2 + y_3, \quad t_1 = 0 \implies y_1 = 0, \quad t_3 = 0 \implies y_3 = 0, \\ y_1 \leq t_1 \tilde{z}_{-L}, \quad y_2 = \sum_{i=0}^L \alpha_i \tilde{z}_{-i}, \quad t_2 = \sum_{i=0}^L \alpha_i, \quad \boldsymbol{\alpha} \in \text{SOS2} \end{array} \right. \right\}. \tag{47b}$$

Proof. See proof in Appendix B.6. $\square$

Theorem 6. ([22, Lemma 1, Theorem 1]) Let Algorithms 2a and 2b be used for generating breakpoint array $\tilde{\mathbf{z}}$. Then $O(\frac{1}{\sqrt{\tau}} \sqrt{\log(\frac{1}{\tau})})$ number of breakpoints are sufficient to approximate $\Phi(z)$ so that $0 \leq \bar{\Phi}(z; \tilde{\mathbf{z}}) - \Phi(z) \leq \tau$ for all $z \in \mathbb{R}$. Similarly, $O(\frac{1}{\sqrt{\tau}} \sqrt{\log(\frac{1}{\tau})})$ breakpoints are sufficient when $\Phi(z)$ is approximated as $\underline{\Phi}(z; \tilde{\mathbf{z}})$ so that $0 \leq \Phi(z) - \underline{\Phi}(z; \tilde{\mathbf{z}}) \leq \tau$ for all $z \in \mathbb{R}$.

6.2. Subproblem and its Properties

Since (43) is a maximization problem, to follow the same convention for inequality direction between $\bar{\Phi}(z_l; \tilde{\mathbf{z}})$ and ζ as used in (47a) and (47b), we use $\zeta' := -\zeta$ here. This leads to $-\underline{\Phi}(z; \tilde{\mathbf{z}}) \geq \zeta'$ in (48b) since $\underline{\Phi}(z; \tilde{\mathbf{z}}) \leq \zeta \iff -\underline{\Phi}(z; \tilde{\mathbf{z}}) \geq -\zeta = \zeta'$. Hence, the following proposition provides a reformulation of the relaxation of the subproblem (43) obtained by taking a piecewise linear approximation of $\Phi(\cdot)$.

Proposition 9. Let $\tilde{\mathbf{z}}$ be an ordered array of breakpoints of z -axis and $\Phi(\cdot)$ be replaced with its PWL inner approximation $\underline{\Phi}(\cdot; \tilde{\mathbf{z}})$ defined in (46). Then, for any given $\mathbf{x} \neq \mathbf{0}$ from $\mathcal{X}$, the resulting formulation (48a)-(48c) is a relaxation to subproblem (43).

$$\max \quad J(\mathbf{s}, \zeta'; (\mathbf{x}, \boldsymbol{\beta})) := \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{m}, \mathbf{Q})) + \zeta', \tag{48a}$$

$$\text{s.t.} \quad -\underline{\Phi}(z; \tilde{\mathbf{z}}) \geq \zeta', \quad b - \mathbf{m}^\top \mathbf{x} \leq z \sigma, \quad \mathbf{x}^\top \mathbf{Q} \mathbf{x} = \sigma^2, \tag{48b}$$

$$\mathbf{m} \in \Theta, \quad \mathbf{Q} \in \Xi, \quad \zeta' \in [-1, 0], \quad z \in \mathbb{R}, \quad \sigma \geq \sqrt{\hat{\delta}} \|\mathbf{x}\|. \tag{48c}$$

Additionally, $\exists(z, \zeta) \in \mathbb{R} \times \mathbb{R}$ satisfying $-\underline{\Phi}(z; \tilde{\mathbf{z}}) \geq \zeta'$ if and only if $\mathcal{H}^I(\tilde{\mathbf{z}}) \neq \emptyset$, where

$$\mathcal{H}^I(\tilde{\mathbf{z}}) := \left\{ \begin{array}{l} \boldsymbol{\alpha} \in \mathbb{R}_+^L, \\ \mathbf{t} \in \{0, 1\}^3, \\ \mathbf{y} \in \mathbb{R}^3, \\ z, \zeta' \in \mathbb{R} \end{array} \left| \begin{array}{l} -\mathbf{h}_i y_1 - \mathbf{h}_i^0 t_1 - t_2 - \Phi(\tilde{z}_R) t_3 \geq \zeta' \quad i \in [L]_0, \quad -t_1 - \sum_{i=0}^R \alpha_i \Phi(\tilde{z}_i) - \Phi(\tilde{z}_R) t_3 \geq \zeta', \\ t_1 + t_2 + t_3 = 1, \quad z = y_1 + y_2 + y_3, \quad t_1 = 0 \implies y_1 = 0, \quad t_3 = 0 \implies y_3 = 0, \\ y_2 = \sum_{i=0}^R \alpha_i \tilde{z}_i, \quad t_2 = \sum_{i=0}^R \alpha_i, \quad \boldsymbol{\alpha} \in \text{SOS2}, \quad y_3 \geq t_3 \tilde{z}_R \end{array} \right. \right\}. \tag{48d}$$

Proof. See proof in Appendix B.6. $\square$

Proposition 10. For a master iterate $\mathbf{x}^{(\ell)}$ at iteration ℓ , let $g(\mathbf{s}; \mathbf{x}^{(\ell)}) := \frac{b - \mathbf{m}^\top \mathbf{x}^{(\ell)}}{\sqrt{\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)}}}$. Define

$$\mathcal{Q} := \{(\mathbf{s}, \zeta') \mid \mathbf{s} \in \mathcal{S}, \zeta' \in [-1, 0], -\Phi(g(\mathbf{s}; \mathbf{x}^{(\ell)})) \geq \zeta'\}, \quad \tilde{\mathcal{Q}} := \{(\mathbf{s}, \zeta') \mid \mathbf{s} \in \mathcal{S}, \zeta' \in [-1, 0], -\underline{\Phi}(g(\mathbf{s}; \mathbf{x}^{(\ell)})) \geq \zeta'\},$$

and $v^* := \max_{(\mathbf{s}, \zeta') \in \mathcal{Q}} J(\mathbf{s}, \zeta')$, $\tilde{v}^* := \max_{(\mathbf{s}, \zeta') \in \tilde{\mathcal{Q}}} J(\mathbf{s}, \zeta')$. Then, if $0 \leq \Phi(z) - \underline{\Phi}(z) \leq \hat{\tau}/2$ for all $z \in \mathbb{R}$, (i) $\mathcal{Q} \subseteq \tilde{\mathcal{Q}}$ and (ii) $0 \leq \tilde{v}^* - v^* \leq \hat{\tau}/2$.

Proof. We begin by noting that $g(\mathbf{s}; \mathbf{x}^{(\ell)})$ is well defined since an iterate $\mathbf{x}^{(\ell)}$ is always considered for the subproblem after filtering it through the near-zero case through Lines 11-13 of Algorithm 1 and hence satisfies $\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)} > 0$.

First part of the claim is immediate since for any $\tau > 0$ and $z \in \mathbb{R}$, we know $0 \leq -\Phi(z) + \Phi(z) \leq \tau$. It implies $-\Phi(z) \leq -\underline{\Phi}(z) \leq \tau - \Phi(z)$. Thus, $\mathcal{Q} \subseteq \tilde{\mathcal{Q}}$. Now for *part-(ii)*, let $(\tilde{\mathbf{s}}, \tilde{\zeta}')$ be optimal for the approximate problem, and define

$$\Delta := \Phi(g(\tilde{\mathbf{s}}; \mathbf{x}^{(\ell)})) - \underline{\Phi}(g(\tilde{\mathbf{s}}; \mathbf{x}^{(\ell)})) \in [0, \hat{\tau}/2], \quad \zeta^\circ := \max\{-1, \tilde{\zeta}' - \Delta\}.$$

Since $-\Phi(g(\tilde{\mathbf{s}}; \mathbf{x}^{(\ell)})) \geq \tilde{\zeta}'$ and $-\Phi = -\underline{\Phi} - \Delta$, we have

$$-\Phi(g(\tilde{\mathbf{s}}; \mathbf{x}^{(\ell)})) = -\underline{\Phi}(g(\tilde{\mathbf{s}}; \mathbf{x}^{(\ell)})) - \Delta \geq \tilde{\zeta}' - \Delta \geq \zeta^\circ,$$

and clearly $\zeta^\circ \in [-1, 0]$. Hence, $(\tilde{\mathbf{s}}, \zeta^\circ) \in \mathcal{Q}$. Moreover, because only ζ' changes in $J(\mathbf{s}, \zeta')$, we obtain

$$J(\tilde{\mathbf{s}}, \zeta^\circ) \geq J(\tilde{\mathbf{s}}, \tilde{\zeta}') - \Delta \geq \tilde{v}^* - \hat{\tau}/2.$$

Since $\mathcal{Q} \subseteq \tilde{\mathcal{Q}}$, $v^* = \max_{(\mathbf{s}, \zeta') \in \mathcal{Q}} J(\mathbf{s}, \zeta') \geq J(\tilde{\mathbf{s}}, \zeta^\circ)$. Combining the two inequalities yields $0 \leq \tilde{v}^* - v^* \leq \hat{\tau}/2$. $\square$

6.3. Finite convergence of the Cutting-Surface Algorithm

We make some additional assumptions (Assumptions 2-4) required in our convergence analysis. The results for the master problem, along with those in Section 6.2 for subproblems, provide the tools required to prove Theorem 5. Assumption 2 excludes the optimal solution from being in a neighborhood of $\mathbf{0}$. Assumption 3: non-empty-interior is a regularity condition requiring that our distributionally robust chance constraint program is strictly feasible.

Assumption 2 (Exclusion of a zero-neighborhood). *Let $\mathcal{F}(\mathbf{0})$ be as in Definition 6. For some $\hat{\delta}$, define $\mathcal{F}_{\hat{\delta}}(\mathbf{0}) := \mathcal{F}(\mathbf{0}) \cap \{(\mathbf{x}, \boldsymbol{\beta}) : \|\mathbf{x}\| \geq \hat{\delta}\}$. Then for $b = 0$ and $(\mathbf{0}, \boldsymbol{\beta}) \in \mathcal{F}(\mathbf{0})$, there exists $\hat{\delta} > 0$ such that $\min_{(\mathbf{x}, \boldsymbol{\beta}) \in \mathcal{F}_{\hat{\delta}}(\mathbf{0})} \mathbf{c}^\top \mathbf{x} = \min_{(\mathbf{x}, \boldsymbol{\beta}) \in \mathcal{F}(\mathbf{0})} \mathbf{c}^\top \mathbf{x}$.*

Assumption 3 (Local nonemptiness). *There exists $\varepsilon > 0$ such that $\mathcal{F}(\tau) \neq \emptyset$, $\forall \tau \in [-\varepsilon, \varepsilon]$.*

Assumption 4 (LICQ). *Let $(\mathbf{x}^*(0), \boldsymbol{\beta}^*(0))$ be an optimal solution to $\min\{\mathbf{c}^\top \mathbf{x} \mid (\mathbf{x}, \boldsymbol{\beta}) \in \mathcal{F}(\mathbf{0})\}$. Linear independence constraint qualification (LICQ) holds at $(\mathbf{x}^*(0), \boldsymbol{\beta}^*(0))$.*

Remark 4. *Assumption 2 is needed because of the continuity and differentiability requirements in our analysis (see Lemma 7). Together, Assumptions 3 and 4 provide the regularity conditions commonly assumed in nonlinear optimization. Under Assumption 3, one can ensure the existence of the oracle $\mathcal{O}$ assumed in part-(ii) of Theorem 5 by taking a PWL-inner approximation for the master problem and a PWL-outer approximation for the subproblem.*

Recall that $\mathcal{X}$ is assumed to be compact. Lemma 6 ensures that when considering the master problem (42), we are dealing with compact feasible sets.

Lemma 6. Suppose that in the nominal Gaussian mixture distribution, $\hat{w}_k > 0$ for all $k \in [\hat{K}]$, $\sum_{k=1}^{\hat{K}} \hat{w}_k = 1$, Wasserstein radius $\rho > 0$, and $\theta \in (0, 1]$. Suppose also that $\hat{s}_k \in \mathcal{S}$ for all $k \in [\hat{K}]$ and that, at master iteration ℓ , $\{\hat{s}_1, \dots, \hat{s}_{\hat{K}}\} \subseteq \mathcal{S}_\ell$. Then every feasible solution $(\mathbf{x}, \beta)$ of the master problem (42) satisfies $\beta \in \overline{\mathfrak{B}}$, where $\overline{\mathfrak{B}} := \prod_{k=1}^{\hat{K}} \left[1 - \frac{1-\theta}{\hat{w}_k}, 1\right] \times \left[0, \frac{1-\theta}{\rho}\right]$ is compact.

Proof. Fix any feasible solution $(\mathbf{x}, \beta)$ of the finite master problem at iteration ℓ . Since $\{\hat{s}_1, \dots, \hat{s}_{\hat{K}}\} \subseteq \mathcal{S}_\ell$, the master problem contains, for each $k \in [\hat{K}]$, the constraint generated by $\mathbf{s}_l = \hat{s}_k$. Hence,

$$\beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{s}_k), \mathcal{N}(\hat{s}_k)) \leq G(\hat{s}_k; \mathbf{x}).$$

Because $d_{\text{BW}}^2(\mathcal{N}(\hat{s}_k), \mathcal{N}(\hat{s}_k)) = 0$, we get $\beta_k \leq G(\hat{s}_k; \mathbf{x})$. Additionally, since $G(\mathbf{s}; \mathbf{x})$ corresponds to a cumulative probability term, $0 \leq G(\mathbf{s}; \mathbf{x}) \leq 1$. Together they imply $\beta_k \leq 1$, $\forall k \in [\hat{K}]$.

Next, because $\beta \in \mathfrak{B}$, $\sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta$. Using $\beta_k \leq 1$ for all k and $\sum_{k=1}^{\hat{K}} \hat{w}_k = 1$, we obtain

$$\theta \leq \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \leq 1 - \rho \beta_{\hat{K}+1}.$$

Hence,

$$0 \leq \beta_{\hat{K}+1} \leq \frac{1-\theta}{\rho},$$

where the lower bound follows from $\beta \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$.

It remains to lower-bound each β_k . Since $\beta_{\hat{K}+1} \geq 0$ and $\sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta$, we have $-\rho \beta_{\hat{K}+1} \leq 0$. Fix any $i \in [\hat{K}]$. Also, for every $k \neq i$, $\beta_k \leq 1$. Therefore,

$$\theta \leq \hat{w}_i \beta_i + \sum_{k \neq i} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \leq \hat{w}_i \beta_i + \sum_{k \neq i} \hat{w}_k.$$

Because $\sum_{k \neq i} \hat{w}_k = 1 - \hat{w}_i$, we get $\theta \leq \hat{w}_i \beta_i + 1 - \hat{w}_i$. Since $\hat{w}_i > 0$, $\beta_i \geq \frac{\theta - (1 - \hat{w}_i)}{\hat{w}_i}$. As $i \in [\hat{K}]$ was arbitrary, the bound holds for all $k \in [\hat{K}]$.

Combining these inequalities gives

$$\frac{\theta - (1 - \hat{w}_k)}{\hat{w}_k} \leq \beta_k \leq 1, \quad k \in [\hat{K}], \quad \text{and} \quad 0 \leq \beta_{\hat{K}+1} \leq \frac{1-\theta}{\rho}.$$

Thus, $\beta \in \overline{\mathfrak{B}}$. Since $\overline{\mathfrak{B}}$ is a finite product of closed and bounded intervals, it is compact. $\square$

Lemma 7. ([22, Lemma 2-3]) Suppose Assumption 2 holds. Then given any $\mathbf{s} \in \mathcal{S}$, $G(\mathbf{x}, \beta; \mathbf{s})$ is continuously differentiable on $\mathcal{X} \times \overline{\mathfrak{B}}$.

Lemma 8. Let Assumptions 1-4 hold, and at some master iteration ℓ , $(\mathbf{x}_*^{(\ell)}, \beta_*^{(\ell)})$ and $(\mathbf{x}^{(\ell)}, \beta^{(\ell)})$ respectively be the exact optimal solutions to (44) and (47a). Then there exists a finite PWL-O approximation using $O(\frac{1}{\sqrt{\epsilon_j}} \sqrt{\log(\frac{1}{\epsilon_j})})$ breakpoints so that $(\mathbf{x}^{(\ell)}, \beta^{(\ell)})$ is ϵ_j -feasible to (44) for some tolerance $\epsilon_j > 0$, corresponding optimal value function $V_\ell^*(\epsilon_j)$ is continuous at 0 and $\mathbf{c}^\top \mathbf{x}_*^{(\ell)} - \mathbf{c}^\top \mathbf{x}^{(\ell)} \leq \epsilon_j$.

Proof. A proof is given in Appendix B.6. It adapts the arguments in [22, Lemma 4, Theorem 3]. $\square$

When $b > 0$, a master iterate close to $\mathbf{0}$ can make the normal-CDF term approach one. In numerical computation, this may appear as $\Phi\left(\frac{b - \mathbf{m}^\top \mathbf{x}^{(\ell)}}{\sqrt{\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)}}}\right) \approx 1$ for all points in $\mathcal{S}_\ell$. This numerical situation, however, is not used as a standalone termination certificate. Instead, we treat it as a *near-zero case* prior to Step 2 in Algorithm 1 and use a deterministic norm-based certificate. Lemma 9 shows that for a feasibility tolerance

ϵ , if the iterate $\mathbf{x}^{(\ell)}$ is in the ϵ -neighborhood of $\mathbf{0}$, then it is already ϵ -feasible for all the constraints of the semi-infinite constraint set in (42). Hence the subproblem separation step is unnecessary in this branch, and the algorithm proceeds directly to the termination criteria checking step (Step 3).

Lemma 9 (Near-zero feasibility certificate). *Suppose $b > 0$, $\mathcal{S} = \Theta \times \Xi$ is compact, and define*

$$\bar{m}_\Theta := \max_{\mathbf{m} \in \Theta} \|\mathbf{m}\|, \quad \bar{\lambda}_\Xi := \max_{\mathbf{Q} \in \Xi} \lambda_{\max}(\mathbf{Q}).$$

Let $\epsilon \in (0, \min\{\hat{\tau}/2, \frac{1}{2}\})$, and define $z_\epsilon := \Phi^{-1}(1 - \epsilon)$, $\delta_\epsilon := \frac{b}{\bar{m}_\Theta + \sqrt{\bar{\lambda}_\Xi} z_\epsilon}$. If $(\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})$ is feasible for the finite master problem at iteration ℓ and $\|\mathbf{x}^{(\ell)}\| \leq \delta_\epsilon$, then $J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})) \leq \epsilon$, $\forall \mathbf{s} \in \mathcal{S}$, $k \in [\hat{K}]$. Equivalently, $(\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}) \in \mathcal{F}(\epsilon)$.

Proof. Fix any support point $\mathbf{s} = (\mathbf{m}, \mathbf{Q}) \in \mathcal{S}$. Since $\mathcal{S} = \Theta \times \Xi$ is compact, the quantities

$$\bar{m}_\Theta := \max_{\mathbf{m} \in \Theta} \|\mathbf{m}\|, \quad \bar{\lambda}_\Xi := \max_{\mathbf{Q} \in \Xi} \lambda_{\max}(\mathbf{Q})$$

are finite. For $\mathbf{x}^{(\ell)} \neq \mathbf{0}$, the definiteness Assumption 1 on $\mathbf{Q} \in \Xi$ gives $\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)} > 0$. Moreover, $b - \mathbf{m}^\top \mathbf{x}^{(\ell)} \geq b - \bar{m}_\Theta \|\mathbf{x}^{(\ell)}\|$, and $\sqrt{\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)}} \leq \sqrt{\bar{\lambda}_\Xi} \|\mathbf{x}^{(\ell)}\|$. Since

$$\begin{aligned} \|\mathbf{x}^{(\ell)}\| &\leq \delta_\epsilon := \frac{b}{\bar{m}_\Theta + \sqrt{\bar{\lambda}_\Xi} z_\epsilon}, \quad z_\epsilon := \Phi^{-1}(1 - \epsilon), \\ b - \bar{m}_\Theta \|\mathbf{x}^{(\ell)}\| &\geq z_\epsilon \sqrt{\bar{\lambda}_\Xi} \|\mathbf{x}^{(\ell)}\|. \end{aligned}$$

Therefore, $\frac{b - \mathbf{m}^\top \mathbf{x}^{(\ell)}}{\sqrt{\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)}}} \geq z_\epsilon$. Hence,

$$G(\mathbf{s}; \mathbf{x}^{(\ell)}) = \Phi\left(\frac{b - \mathbf{m}^\top \mathbf{x}^{(\ell)}}{\sqrt{\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)}}}\right) \geq \Phi(z_\epsilon) = 1 - \epsilon.$$

The same conclusion holds for $\mathbf{x}^{(\ell)} = \mathbf{0}$ since $G(\mathbf{s}; \mathbf{0}) = 1$ for $b > 0$ by definition (38).

Next, since the algorithm is initialized with $\mathcal{S}_0 = \hat{\mathcal{S}} := \{\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_{\hat{K}}\}$, we have $\hat{\mathcal{S}} \subseteq \mathcal{S}_\ell$ for every iteration ℓ . Therefore, for each $k \in [\hat{K}]$, the finite master problem contains the cut generated by the nominal support point $\hat{\mathbf{s}}_k$: $\beta_k^{(\ell)} - \beta_{\hat{K}+1}^{(\ell)} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\hat{\mathbf{s}}_k)) \leq G(\hat{\mathbf{s}}_k; \mathbf{x}^{(\ell)})$. Since $d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\hat{\mathbf{s}}_k)) = 0$ and $G(\hat{\mathbf{s}}_k; \mathbf{x}^{(\ell)}) \leq 1$, it follows that $\beta_k^{(\ell)} \leq 1$, $\forall k \in [\hat{K}]$.

Now fix arbitrary $k \in [\hat{K}]$ and $\mathbf{s} \in \mathcal{S}$. Then

$$\begin{aligned} J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})) &= \beta_k^{(\ell)} - \beta_{\hat{K}+1}^{(\ell)} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) - G(\mathbf{s}; \mathbf{x}^{(\ell)}) \\ &\leq \beta_k^{(\ell)} - G(\mathbf{s}; \mathbf{x}^{(\ell)}) \leq 1 - (1 - \epsilon) = \epsilon, \end{aligned}$$

where we used $\beta_{\hat{K}+1}^{(\ell)} \geq 0$ and $d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) \geq 0$. Since k and $\mathbf{s}$ were arbitrary,

$$J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})) \leq \epsilon, \quad \forall \mathbf{s} \in \mathcal{S}, k \in [\hat{K}].$$

Thus, $(\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}) \in \mathcal{F}(\epsilon)$. □

6.3.1. Proof of the optimality and finite termination claim of Theorem 5.

Proof. Part-(i). Proof arguments for this part are established in the following order by setting $j = 1$ and hence feasibility tolerance $\hat{\tau}/2$:

- claim that satisfying the Step-3 condition $J(\mathbf{s}^{(\ell+1)}; (\mathbf{x}^{(\ell)}, \boldsymbol{\beta}^{(\ell)})) \leq \hat{\tau}/2$ in Line 16 of Algorithm 1 is equivalent to obtaining $\hat{\tau}/2$ -feasible solution to the semi-infinite program (S-Inf). This is proved by separately considering the regular case and the near-zero case in Line 11 of Algorithm 1

- show the objective value obtained from the current iterate $\mathbf{x}^{(\ell)}$ is bounded above by $\text{Val}(\text{S-Inf})$ based on the master relaxation relation
- use compactness and continuity property to conclude that only finitely many points from $\mathcal{S}$ are needed

First suppose that, at iteration ℓ , Step 2 of Algorithm 1 is executed. For fixed $(\mathbf{x}^{(\ell)}, \beta^{(\ell)})$, define

$$\check{J}_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) := \beta_k^{(\ell)} - \beta_{\hat{K}+1}^{(\ell)} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) - \Phi(g(\mathbf{s}; \mathbf{x}^{(\ell)})),$$

and

$$J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) := \beta_k^{(\ell)} - \beta_{\hat{K}+1}^{(\ell)} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) - \Phi(g(\mathbf{s}; \mathbf{x}^{(\ell)})).$$

Because $\Phi \leq \bar{\Phi}$, we have $J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) \leq \check{J}_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)}))$ for every $\mathbf{s} \in \mathcal{S}$ and $k \in [\hat{K}]$. Let $(\bar{k}, \mathbf{s}^{(\ell+1)})$ solve $\max_{k \in [\hat{K}]} \max_{\mathbf{s} \in \mathcal{S}} \check{J}_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)}))$. If the condition in Line 16 is satisfied, then $\check{J}_{\bar{k}}(\mathbf{s}^{(\ell+1)}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) \leq \hat{\tau}/2$. Since $(\bar{k}, \mathbf{s}^{(\ell+1)})$ maximizes $\check{J}_k$ over $k \in [\hat{K}]$ and $\mathbf{s} \in \mathcal{S}$, it follows that $\check{J}_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) \leq \hat{\tau}/2$ for all $\mathbf{s} \in \mathcal{S}$ and $k \in [\hat{K}]$. Combining this with $J_k \leq \check{J}_k$ gives $J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) \leq \hat{\tau}/2$ for all $\mathbf{s} \in \mathcal{S}$ and $k \in [\hat{K}]$. Hence $(\mathbf{x}^{(\ell)}, \beta^{(\ell)}) \in \mathcal{F}(\hat{\tau}/2)$.

Now suppose that Line 11 is invoked and the subproblem step is skipped. Let $\delta_{\hat{\tau}/2}$ be the radius in Lemma 9 with $\epsilon = \hat{\tau}/2$. If $b > 0$ and $\|\mathbf{x}^{(\ell)}\| \leq \delta_{\hat{\tau}/2}$, then Lemma 9 gives $J_k(\mathbf{s}; (\mathbf{x}^{(\ell)}, \beta^{(\ell)})) \leq \hat{\tau}/2$ for all $\mathbf{s} \in \mathcal{S}$ and $k \in [\hat{K}]$. Thus $(\mathbf{x}^{(\ell)}, \beta^{(\ell)}) \in \mathcal{F}(\hat{\tau}/2)$ also in this case. This proves that the condition used to enter Step 3 is a valid feasibility certificate for the exact semi-infinite constrained problem (S-Inf).

We next prove the objective bound. Let $\mathcal{F}_\ell(0)$ and $\mathcal{F}_\ell(\hat{\tau}/2)$ be the finite-cut feasible sets in Definition 6, and let $\tilde{\mathcal{F}}_\ell(0)$ be the master feasible set obtained by replacing Φ with $\bar{\Phi}$. By construction, $\mathcal{F}_\ell(0) \subseteq \tilde{\mathcal{F}}_\ell(0) \subseteq \mathcal{F}_\ell(\hat{\tau}/2)$. At iteration ℓ , the master problem solves $\min\{\mathbf{c}^\top \mathbf{x} : (\mathbf{x}, \beta) \in \tilde{\mathcal{F}}_\ell(0)\}$. Therefore, for an optimal master solution $(\mathbf{x}^{(\ell)}, \beta^{(\ell)})$, $\mathbf{c}^\top \mathbf{x}^{(\ell)} \leq \min\{\mathbf{c}^\top \mathbf{x} : (\mathbf{x}, \beta) \in \mathcal{F}_\ell(0)\}$. Since $\mathcal{S}_\ell \subseteq \mathcal{S}$, the finite-cut feasible set $\mathcal{F}_\ell(0)$ is a relaxation of the exact feasible set $\mathcal{F}(0)$. Hence $\min\{\mathbf{c}^\top \mathbf{x} : (\mathbf{x}, \beta) \in \mathcal{F}_\ell(0)\} \leq \text{Val}(\text{S-Inf})$, and consequently $\mathbf{c}^\top \mathbf{x}^{(\ell)} \leq \text{Val}(\text{S-Inf})$.

It remains to show that only finitely many cuts are needed to reach $\mathcal{F}(\hat{\tau}/2)$. By Lemma 6, because all nominal cuts are included in the initial set $\mathcal{S}_0$, every feasible master solution satisfies $\beta \in \bar{\mathfrak{B}}$, where $\bar{\mathfrak{B}}$ is compact. Thus the master variables lie in the compact set

$$\mathcal{Y} = \begin{cases} \mathcal{X} \times \bar{\mathfrak{B}}, & b > 0, \\ \mathcal{X}_\delta \times \bar{\mathfrak{B}}, & b = 0, \end{cases}$$

where the second case uses the zero-excluded domain in Assumption 2. By Lemma 4, Lemma 5, and Lemma 7, $J_k(\mathbf{s}, \mathbf{x}, \beta)$ is continuous on $\mathcal{S} \times \mathcal{Y}$ for every $k \in [\hat{K}]$. Since $\mathcal{S}$ is compact and $[\hat{K}]$ is finite, the semi-infinite system is a finite collection of compactly indexed continuous constraints. Therefore, $\mathcal{F}(0)$ is a closed subset of the compact set $\mathcal{Y}$, and each finite-cut feasible set $\mathcal{F}_\ell(0)$ is compact. The hypotheses of [33, Theorem 7.2] hold. Applying that result with tolerance $\hat{\tau}/2$, Algorithm 1 reaches $\mathcal{F}(\hat{\tau}/2)$ after adding finitely many points from $\mathcal{S}$. This proves part-(i).

Proof of Part-(ii). Let $\epsilon_j := \hat{\tau}/2^j$, $j \geq 1$. The proof of part-(i) does not rely on any condition on $\hat{\tau}/2$ except that $\hat{\tau}/2 > 0$, so the same argument applies with ϵ_j in place of $\hat{\tau}/2$. Thus, for each fixed j , only finitely many cuts are added before the algorithm obtains an ϵ_j -feasible point for the exact semi-infinite constrained problem (S-Inf).

Now, let $\mathbf{x}^*$ be an optimal solution of (S-Inf), and $\mathbf{x}_*^{(\ell)}$ be an exact optimal solution of the finite master problem (44) at iteration ℓ . Since the finite master problem contains only a subset of the constraints of (S-Inf), $\mathbf{c}^\top \mathbf{x}_*^{(\ell)} \leq \mathbf{c}^\top \mathbf{x}^* = \text{Val}(\text{S-Inf})$. Let $V_\ell^*(\epsilon_j)$ denote the optimal value over $\mathcal{F}_\ell(\epsilon_j)$, and let $V^*(0)$ be the optimal value of (S-Inf) over $\mathcal{F}(0)$. The master step gives $V_\ell^*(\epsilon_j) \leq \mathbf{c}^\top \mathbf{x}^{(\ell)} \leq \mathbf{c}^\top \mathbf{x}_*^{(\ell)} \leq \mathbf{c}^\top \mathbf{x}^* = V^*(0)$.

Once an ϵ_j -feasible solution $\mathbf{x}^{(\ell)}$ to (S-Inf) is obtained (via *part-(i)*), Line 17 uses the oracle $\mathcal{O}$ to return a feasible solution $(\tilde{\mathbf{x}}, \tilde{\beta})$ of (S-Inf) by solving (44) at ϵ_j optimality. Lemma 8 guarantees achieving such a desired ϵ_j -optimal solution to (44) under finite approximation accuracy through establishing continuity of the optimal value function $V^*(\cdot)$ at the origin. Hence, for every $\epsilon_j > 0$, there exists $\tau^+(\epsilon_j) > 0$ such that $0 < \tau \leq \tau^+(\epsilon_j)$ implies $|V^*(0) - V^*(\tau)| \leq \epsilon_j$ and $|V^*(0) - V^*(-\tau)| \leq \epsilon_j$. Therefore,

$$|V^*(-\tau) - V^*(\tau)| \leq |V^*(-\tau) - V^*(0)| + |V^*(0) - V^*(\tau)| \leq 2\epsilon_j = 2\hat{\tau}/2^j \leq \hat{\tau}.$$

Because ϵ_j is halved whenever $j \leftarrow j + 1$, a finite j is reached for which the corresponding tolerance lies inside the required continuity neighborhood and $2\epsilon_j \leq \hat{\tau}$. For this j , the feasible solution $\tilde{\mathbf{x}}$ and the certificate iterate $\mathbf{x}^{(\ell^*)}$ satisfy $|\mathbf{c}^\top \tilde{\mathbf{x}} - \mathbf{c}^\top \mathbf{x}^{(\ell^*)}| \leq \hat{\tau}$ at some terminating iteration ℓ^* .

For each fixed j , the preceding paragraph and *part-(i)* imply finite termination of the cut-generation step. Since only finitely many updates of j are required, the total number of master iterations is finite. Hence $j < \infty$, $\ell^* < \infty$, and Algorithm 1 returns a $\hat{\tau}$ -accurate solution in finitely many iterations. $\square$

7. Case Study

We now discuss our computational experience with continuous support-based robustification of the chance constraint, specified using the Gaussian mixture distribution, as studied in the earlier sections. We discuss this in the context of a service-level planning case study for EV charging stations, developed using a public dataset from [26]. Our case study is motivated by the broader EV charging literature on the problem of aggregate charging-profile control and planning over time, as well as service-level-based charging infrastructure planning; see, for example, [1, 55]. We introduce the model in Section 7.1 and describe our computational settings in Section 7.2. This section also specifies Wasserstein-2 ambiguity set radii and the mean-covariance support sets used in our experiments. Section 7.3 presents the implementation details for both the finite-support DR (FDR) and continuous-support DR (CDR) formulations of the EV model. Section 7.4 compares properties of the solutions generated by the two DR approaches along with the nominal distribution in terms of objective value, out-of-sample performance, and computation time. Section 7.5 compares the energy-allocation profile in the nominal, FDR, and CDR solutions using different metrics and visual tools to provide insights into how robustness reshapes the energy profile in these models.

7.1. An Aggregate EV Charging Service-Level Planning Model

We consider a least-cost EV charging-station energy allocation model subject to a minimum daily demand-fulfillment requirement as follows:

$$\min_{\mathbf{x} \in [0,1]^T} \sum_{t=1}^T \left(c_t^f x_t + C_t^o(x_t) \right) \quad \text{s.t.} \quad \mathbb{P} \left(\sum_{t=1}^T \xi_t x_t \geq E_{\min} \right) \geq \theta. \quad (49)$$

Here ξ_t denotes the random charging demand in hour t , and $x_t \in [0, 1]$ denotes the fraction of that demand served in that hour. $\xi_t x_t$ is the served demand in hour t , and $\sum_{t=1}^T \xi_t x_t$ is the total demand served over the day. The parameter $E_{\min}$ specifies a minimum daily served-energy requirement, and the chance constraint requires this requirement to be met with probability at least θ . The objective consists of a baseline linear cost $c_t^f x_t$, where c_t^f captures the hour- t time-of-use cost, and an additional operational cost term $C_t^o(x_t)$ which allows the cost of serving demand to vary nonlinearly with the demand fulfillment fraction. The convex piecewise-linear form provides a regime-dependent penalty across service levels.

7.1.1. Instance Generation. To instantiate the EV demand-management model (49) from the data in [26], each charging session was partitioned across clock hours in proportion to the time spent in each hour and then aggregated into a daily 24-dimensional demand vector. The entire dataset was divided for training and out-of-sample testing using a 60/40 split. The setting of the cost parameters and the procedure for estimating necessary parameters for the Gaussian distribution model (49) from the source data and its training/testing split are elaborated in Appendix D.1.

The nominal Gaussian mixture model fitted to the training data had $\hat{K} = 5$ components. It was determined as follows. Gaussian mixtures with $k = 1, \dots, 20$ components were fitted using the `GaussianMixture` class in the `scikit-learn` package. In the package-specific configuration, the covariance type was set to `full` and diagonal regularization was set to 10^{-6} . For each fixed order k , the native expectation-maximization (EM) algorithm was run from ten different initializations, and the fit with the largest log-likelihood value was retained. This entire order-selection procedure was then repeated over ten independent random seeds. In each replicate, the selected order was the value of k minimizing the Bayesian information criterion (BIC). Sorting these ten replicate-specific selections, we took the lower of the two middle values as the final choice, yielding $\hat{K} = 5$ for the fitted demand data. The resulting nominal distribution had mixture weights $\hat{\mathbf{w}} = (0.167, 0.167, 0.204, 0.241, 0.222)$.

7.1.2. Data preparation for out-of-sample performance evaluation. For the out-of-sample performance analysis, energy allocation solutions obtained from the FDR and CDR models were evaluated against demand samples generated from a Gaussian mixture model fitted on the holdout data. Gaussian mixture models with orders $k = 1, \dots, 20$ were fitted using the EM algorithm, with 15 initializations per order and up to 100 EM iterations per fit; for each k , the best fit among the 15 initializations was retained. Let K^* correspond to the one with the smallest BIC over $k = 1, \dots, 20$. out-of-sample satisfaction (OSS) probability was then computed from samples drawn for K^* , $K^* - 1$ and $K^* + 1$, using Monte Carlo sample sizes of 500 for each, and reporting the proportion of sampled demand vectors ξ for which $\xi^\top \mathbf{x} \geq E_{\min}$ with candidate allocation $\mathbf{x}$. Such an approach is statistically motivated: finite-mixture likelihoods are non-convex,

EM estimates may depend on initialization, and adjacent mixture orders often have nearly indistinguishable information-criterion values on moderate sample sizes. Evaluating OSS at $K^* - 1$, K^* , and $K^* + 1$, therefore, serves as a robustness check against model-order and local-optimum sensitivity rather than relying on a single fitted mixture.

7.2. Experimental Settings

All computations were performed on a server with Intel Xeon 2.80 GHz CPUs using Python 3.11. Gurobi 11.0.1 was used to solve master and subproblems that do not involve positive semi-definite (PSD) constraints; Mosek 11.1.6 was used otherwise. A thread count of 2 for the master problem and each subproblem was set. For both the finite- and continuous-support formulations, a violation tolerance of 10^{-4} was used as the stopping threshold for cut generation. The target satisfaction probability θ , representing the desired service level, was set to take values in $\{0.95, 0.97, 0.99\}$. The cutting-surface algorithm used to solve CDR was run for four master iterations. For each of the five master problems, Gurobi's `MIPGap` parameter was set to 0.1%. Since solving nonlinear mixed integer programs generated in the cutting surface algorithm to optimality is time-consuming (and not required), an iteration-dependent time limit, as specified in implementation detail Section 7.3.2, was followed to keep the total computation time ≤ 22 hours for each instance. All other Gurobi and Mosek settings were kept to their default values.

7.2.1. Wasserstein-2 set. We experimented with Wasserstein-radius parameters $\rho \in \{0.001, 0.005, 0.01\}$. For the CDR case, the compact support $\mathcal{S} := \Theta \times \Xi$ was generated as follows. The mean support set Θ was specified by componentwise bounds $\underline{m} \leq \mathbf{m}_k \leq \bar{\mathbf{m}}$ for all $k \in [K]$ which along each coordinate $j \in [n]$ (for the case study j and $[n]$ respectively correspond to t and $[T]$) are obtained from the nominal component means as:

$$\underline{m}_j = \min_{k \in \hat{K}} m_{jk} - \varsigma \left| \min_{k \in \hat{K}} m_{jk} \right|, \quad \bar{m}_j = \max_{k \in \hat{K}} m_{jk} + \varsigma \left| \max_{k \in \hat{K}} m_{jk} \right|.$$

Similarly, the covariance support set Ξ was specified by imposing multiplicative semidefinite bounds around the nominal covariance matrices $\hat{\mathbf{Q}}_k$, namely, $\underline{\lambda}_f \hat{\mathbf{Q}}_k \preceq \mathbf{Q}_k \preceq \bar{\lambda}_f \hat{\mathbf{Q}}_k$, $\forall k \in [K]$. We considered two settings for the mean and covariance bounds: $\varsigma \in \{0.05, 0.1\}$ and $(\underline{\lambda}_f, \bar{\lambda}_f) \in \{(1/2, 2), (1/3, 3)\}$.

7.3. Implementation Details of Nominal Model and DR variants

7.3.1. Nominal model. Given the GMM fitted to the EV-station demand data, the nominal counterpart of model (49) is $\min_{\mathbf{x} \in [0,1]^T} \sum_{t=1}^T (c_t^f x_t + C_t^o(x_t))$ s.t. $\sum_{k=1}^{\hat{K}} \hat{w}_k \Phi\left(\frac{E_{\min} - \mathbf{m}_k^\top \mathbf{x}}{\sqrt{\mathbf{x}^\top \mathbf{Q}_k \mathbf{x}}}\right) \geq \theta$, which we solved via a piecewise-linear approximation of $\Phi(\cdot)$ and the corresponding chance-constraint reformulation in (47b). Note that obtaining the array of breakpoints for the piecewise-linear approximation of $\Phi(\cdot)$ from Algorithm 2a-2b requires the approximation accuracy tolerance τ ; the same approximation tolerance $\hat{\tau} := \tau = 10^{-4}$ was used to decide the number of points in the piecewise linear approximations of the univariate Gaussian for the nominal model and the DR instances. The outermost loop in Algorithm 1 that iteratively reduces the approximation accuracy and tolerances was not performed.

7.3.2. Distributionally robust models.

We solve (41) as the FDR benchmark with the unknown support set restricted to the nominal GMM components. Since formulation (41) contains only finitely many constraints, it can be handled directly by an off-the-shelf solver such as Gurobi and does not require iterative subproblem separation. In contrast, the CDR cutting-surface algorithm is initialized with the cuts generated from $\mathcal{S}_0 = \{\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_{\hat{K}}\}$, constructed from the means and covariance matrices of the $\hat{K} = 5$ nominal components. By Corollary 1 in Section 5, the iteration-0 master problem in the cutting-surface algorithm corresponds to the FDR model (41). However, because the iteration-adaptive schedule described below allocates only up to 2 hours to iteration 0, which may be insufficient to obtain an FDR solution with $\text{MIPGap}=0.1\%$, we solved (41) as a standalone benchmark run with a time limit set to 22 hours, which equals the time allocated to the first four CDR iterations.

Iteration-dependent master iteration time limits. For the CDR model, one separation problem is solved for each $k \in [\hat{K}]$. These $\hat{K}$ subproblem solutions may generate multiple violated cuts, and each admitted cut introduces a nonconvex constraint together with a new block of binary variables. As a result, the master problem becomes increasingly computationally demanding along iterations. To manage this computational burden, rather than solving each master problem to a prescribed optimality gap, we impose iteration-dependent time limits of 2, 2, 4, 6, and 8 hours for iterations 0 through 4, respectively. This practical adaptation does not compromise the CSA algorithmic framework, as an optimal solution of the master problem in this framework is required when proving optimality. Sub-optimal solutions to the master problem are acceptable for generating the cutting surfaces. Since we have FDR solutions available from two separate contexts: under iteration-0 of the cutting surface algorithm as well as from its stand-alone run of 22 hours, we report both for performance comparisons.

Separation problem. Let $\mathbf{x}^{(\ell)} \neq \mathbf{0}$ and $\beta^{(\ell)}$ be the master solution at some iteration ℓ . For each nominal component $k \in [\hat{K}]$, the separation problem searches over a support point $\mathbf{s} = (\mathbf{m}, \mathbf{Q}) \in \mathcal{S}$ to identify a violated semi-infinite constraint. We suppress the nominal component subscript k in the following description. Thus, $\hat{\mathbf{m}}$, $\hat{\mathbf{Q}}$, and $\beta^{(\ell)}$ denote $\hat{\mathbf{m}}_k$, $\hat{\mathbf{Q}}_k$, and $\beta_k^{(\ell)}$, respectively. Note that $\beta_{K+1}^{(\ell)}$ does not depend on $k \in [K]$. Additionally, after replacing $\Phi(\cdot)$ by its PWL inner approximation and using $\zeta := -\zeta'$, and the mean-covariance support set $\Theta \times \Xi$ specified in Section 7.2.1, the separation problem (48a)-(48c) can be written in minimization form as

$$\begin{aligned} \min_{\mathbf{m}, \mathbf{Q}, \sigma, z, \zeta, \alpha, \mathbf{t}, \mathbf{y}} \quad & -\beta^{(\ell)} + \beta_{K+1}^{(\ell)} \left[\|\mathbf{m} - \hat{\mathbf{m}}\|_2^2 + \text{tr}(\mathbf{Q}) + \text{tr}(\hat{\mathbf{Q}}) - 2 \text{tr} \left((\mathbf{Q}^{1/2} \hat{\mathbf{Q}} \mathbf{Q}^{1/2})^{1/2} \right) \right] + \zeta \\ \text{s.t.} \quad & \underline{\mathbf{m}} \leq \mathbf{m} \leq \bar{\mathbf{m}}, \quad \underline{\lambda}_f \hat{\mathbf{Q}} \preceq \mathbf{Q} \preceq \bar{\lambda}_f \hat{\mathbf{Q}}, \end{aligned} \quad (50)$$

$\mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)} = \sigma^2$, $b - \mathbf{m}^\top \mathbf{x}^{(\ell)} \leq z\sigma$, $(z, \zeta, \alpha, \mathbf{t}, \mathbf{y}) \in \mathcal{H}^I(\tilde{\mathbf{z}})_+$, where, $\mathcal{H}^I(\tilde{\mathbf{z}})_+$ denotes the set $\mathcal{H}^I(\tilde{\mathbf{z}})$ except using $\zeta := -\zeta'$ for $\zeta \in [0, 1]$.

The above model is not directly representable using the algebraic modeling language of the off-the-shelf solver Mosek. Although the mean-distance term and the trace terms $\text{tr}(\mathbf{Q}) + \text{tr}(\hat{\mathbf{Q}})$ are directly representable, the difficulty arises from the matrix square-root term $\text{tr}((\mathbf{Q}^{1/2} \hat{\mathbf{Q}} \mathbf{Q}^{1/2})^{1/2})$, which couples the

variable covariance matrix $\mathbf{Q}$ with the nominal covariance matrix $\hat{\mathbf{Q}}$. To express this term through semidefinite constraints, we use the identity

$$\text{tr}\left(\left(\mathbf{Q}^{1/2}\hat{\mathbf{Q}}\mathbf{Q}^{1/2}\right)^{1/2}\right) = \max_{\mathbf{W}} \left\{ \text{tr}(\mathbf{W}) : \begin{bmatrix} \mathbf{Q} & \mathbf{W} \\ \mathbf{W}^\top & \hat{\mathbf{Q}} \end{bmatrix} \succeq 0 \right\}.$$

Substituting this representation into (50) gives the following $\mathbf{W}$ -based form:

$$\begin{aligned} \min_{\mathbf{m}, \mathbf{Q}, \mathbf{W}, \sigma, z, \zeta, \boldsymbol{\alpha}, \mathbf{t}, \mathbf{y}} \quad & -\beta^{(\ell)} + \beta_{K+1}^{(\ell)} \left[d_{m,k}(\mathbf{m}) + d_{Q,k}(\mathbf{Q}, \mathbf{W}) \right] + \zeta \\ \text{s.t.} \quad & \underline{\mathbf{m}} \leq \mathbf{m} \leq \bar{\mathbf{m}}, \quad \underline{\lambda}_f \hat{\mathbf{Q}} \preceq \mathbf{Q} \preceq \bar{\lambda}_f \hat{\mathbf{Q}}, \\ & \begin{bmatrix} \mathbf{Q} & \mathbf{W} \\ \mathbf{W}^\top & \hat{\mathbf{Q}} \end{bmatrix} \succeq 0, \\ & \mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)} = \sigma^2, \quad b - \mathbf{m}^\top \mathbf{x}^{(\ell)} \leq z\sigma, \quad (z, \zeta, \boldsymbol{\alpha}, \mathbf{t}, \mathbf{y}) \in \mathcal{H}^I(\tilde{\mathbf{z}})_+, \end{aligned} \quad (51)$$

where

$$d_m(\mathbf{m}) := \|\mathbf{m} - \hat{\mathbf{m}}\|_2^2, \quad d_Q(\mathbf{Q}, \mathbf{W}) := \text{tr}(\mathbf{Q}) + \text{tr}(\hat{\mathbf{Q}}) - 2\text{tr}(\mathbf{W}).$$

The resultant model (51) is a reformulation to (50). Indeed, for a fixed $\mathbf{Q}$, minimizing $d_Q(\mathbf{Q}, \mathbf{W})$ is equivalent to maximizing $\text{tr}(\mathbf{W})$ over the block PSD constraint. Since $\beta_{K+1}^{(\ell)} \geq 0$, the objective selects such a maximizer whenever the covariance-distance term is active. If $\beta_{K+1}^{(\ell)} = 0$, the Bures–Wasserstein distance term is absent from the objective and the choice of $\mathbf{W}$ does not affect the separation objective.

The separation problem reformulation (51) is a mixed-binary problem with PSD constraints, quadratic equalities ($\sigma^2 = \mathbf{x}^\top \mathbf{Q} \mathbf{x}$), and bilinear terms. Despite the simplification through reformulation (51), Gurobi and Mosek (and other commercial mixed-integer solvers, to our knowledge) are not advanced enough to accept this formulation in its native form. Therefore, we developed an alternating block minimization heuristic to obtain solutions to (51). It is acceptable (possibly desirable) to use a heuristic in the cut generation subproblems at an earlier stage of CSA since its optimal solution is required only in proving optimality.

An alternating block minimization heuristic. We propose an alternating block minimization heuristic by leveraging the fact that the cross-block coupling arising in the subproblem (51) only couples through the link constraint $b(\mathbf{m}; \mathbf{x}^{(\ell)}) := b - \mathbf{m}^\top \mathbf{x}^{(\ell)} \leq z\sigma$, which decomposes in two blocks of variables: $\mathbf{m}$ -block variables $(\mathbf{m}, z, \zeta)$ and $\mathbf{Q}$ -block variables $(\mathbf{Q}, \mathbf{W}, \sigma)$. All other constraints and objectives are block-separable, and the interaction between the two blocks is confined to the product of the scalar variables z and σ . At iteration r of this alternating block minimization heuristic, we present the formulations below for both blocks:

$\mathbf{m}$ -block at sub-iteration r :

$$\begin{aligned} \min_{\mathbf{m}, z, \zeta, \boldsymbol{\alpha}, \mathbf{t}, \mathbf{y}} \quad & -\beta^{(\ell)} + \beta_{K+1}^{(\ell)} d_m(\mathbf{m}) + \beta_{K+1}^{(\ell)} \bar{c}_Q^{(r)} + \zeta \\ \text{s.t.} \quad & \underline{\mathbf{m}} \leq \mathbf{m} \leq \bar{\mathbf{m}}, \\ & b(\mathbf{m}; \mathbf{x}^{(\ell)}) \leq z\sigma^{(r)}, \\ & (z, \zeta, \boldsymbol{\alpha}, \mathbf{t}, \mathbf{y}) \in \mathcal{H}^I(\tilde{\mathbf{z}})_+, \\ \text{where,} \quad & \bar{c}_Q^{(r)} := d_Q(\mathbf{Q}^{(r)}, \mathbf{W}^{(r)}), \\ & \sigma^{(r)} := \sqrt{\mathbf{x}^{(\ell)\top} \mathbf{Q}^{(r)} \mathbf{x}^{(\ell)}}. \end{aligned}$$

$\mathbf{Q}$ -block at sub-iteration $r+1$:

$$\begin{aligned} \min_{\mathbf{Q}, \mathbf{W}, \sigma} \quad & -\beta^{(\ell)} + \beta_{K+1}^{(\ell)} \bar{c}_m^{(r+1)} + \beta_{K+1}^{(\ell)} d_Q(\mathbf{Q}, \mathbf{W}) + \zeta^{(r+1)} \\ \text{s.t.} \quad & \underline{\lambda}_f \hat{\mathbf{Q}} \preceq \mathbf{Q} \preceq \bar{\lambda}_f \hat{\mathbf{Q}}, \\ & \begin{bmatrix} \mathbf{Q} & \mathbf{W} \\ \mathbf{W}^\top & \hat{\mathbf{Q}} \end{bmatrix} \succeq 0, \\ & \sigma^2 \leq \mathbf{x}^{(\ell)\top} \mathbf{Q} \mathbf{x}^{(\ell)}, \\ & b(\mathbf{m}^{(r+1)}; \mathbf{x}^{(\ell)}) \leq z^{(r+1)}\sigma, \\ \text{where,} \quad & \bar{c}_m^{(r+1)} := d_m(\mathbf{m}^{(r+1)}). \end{aligned}$$

In our implementation, the heuristic was initialized on the covariance side by setting $Q^{(0)} = \hat{Q}$ and $W^{(0)} = \hat{Q}$, that is, at the nominal covariance pair of the Gaussian component under consideration. With $(Q^{(0)}, W^{(0)})$ fixed, the algorithm started by solving the m -block to obtain $(m^{(1)}, z^{(1)}, \zeta^{(1)})$. Fixing these coefficients it then solves the Q -block problem to obtain $(Q^{(1)}, W^{(1)})$, and iterates from thereon.

Mosek was used to solve the Q -block of the subproblem containing PSD constraints with a tolerance of 10^{-6} . The alternating block minimization heuristic was run for up to $r = 15$ iterations. A resulting candidate solution was added only if its violation exceeded the threshold 10^{-4} .

7.4. Numerical Findings

This section reports computational insights from the CDR and FDR experiments. Across all parameter settings of the CDR model, cumulative solution times and cut counts over the master iterations of the cutting-surface approach indicate that the computational burden is driven primarily by repeated master-problem reoptimization rather than by separation. For both mean-support expansion values $\varsigma = 0.05$ and $\varsigma = 0.1$, the separation heuristic accounts for less than 1% of total computation time. By contrast, the master problems become increasingly difficult as the algorithm proceeds, because each admitted cut adds a nonconvex constraint together with an additional block of binary variables. This effect becomes more pronounced at larger target satisfaction levels θ . The number of admitted cuts remains relatively stable across all the tested settings. After four CDR iterations, the total number of cuts is observed to be between 21 and 23.

This behavior is also reflected in the iteration-wise runtimes. Early CDR iterations are solved within minutes, whereas later iterations may require hours and, in the most difficult cases, the cumulative runtime reaches the 22-hour budget. Detailed iteration-wise breakdown of solution times and cut counts are reported in Tables 2a and 2b separately for $\varsigma = 0.1$ and $\varsigma = 0.05$. Setting $\varsigma = 0.1$ generally requires more computation time than $\varsigma = 0.05$, which is noticeable for $\theta \geq 0.97$. This additional expense is mainly attributable to the larger cut space, which makes identifying highly violated cuts more costly. Runtime differences, however, are hardly distinguishable for different covariance-scaling parameter λ_f values across all these test settings.

The FDR model is evaluated and compared under two computational contexts. The first is a benchmark setting that solves the stand-alone FDR formulation to the desired optimality gap of 10^{-4} within the same 22-hour time budget. The second uses the CDR iteration-0 master problem as an FDR reference, where runtime is capped below 2 hours by the iteration-adaptive schedule. This distinction is important because the stand-alone FDR solve may still consume the full 22-hour budget toward the desired optimality gap, whereas the iteration-0 CDR master is deliberately restricted to a shorter runtime. For FDR, the effect of a larger ambiguity radius ρ on computation time is visible for higher satisfaction targets $\theta \geq 0.97$. For CDR, this effect is further amplified by larger mean-support uncertainty. These two factors together make master reoptimization more computationally demanding, while separation remains a negligible component of total runtime.

Table 2 Finite-support DR (FDR) solution time, cumulative completion time, and cuts across master iterations of the cutting surface algorithm under continuous-support DR (CDR) for all prescribed service-level targets $\theta \in \{0.95, 0.97, 0.99\}$ and Wasserstein ambiguity radius $\rho \in \{0.001, 0.005, 0.01\}$, shown for mean-support expansion parameters $\varsigma = 0.10$ and $\varsigma = 0.05$.

(a) $\varsigma = 0.10$															
		FDR		CDR											
				Covariance-scaling factor $\underline{\lambda}_f = 0.333$								Covariance-scaling factor $\underline{\lambda}_f = 0.5$			
		Benchmark	From CDR Iter 0	Iter 2		Iter 3		Iter 4		Iter 2		Iter 3		Iter 4	
θ	ρ	Time (hr)	Time (hr)	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut
0.95	0.001	0.022	0.009	0.015	15	0.700	19	4.662	22	0.014	15	1.077	19	3.637	22
	0.005	0.026	0.006	3.968	14	9.968	17	17.968	21	2.770	15	8.770	18	16.770	22
	0.010	0.024	0.009	6.009	15	12.009	18	20.009	21	6.009	14	12.009	18	20.009	22
0.97	0.001	0.669	0.705	6.705	15	12.705	19	20.705	22	6.696	15	12.696	19	20.696	22
	0.005	0.154	0.149	6.149	15	12.149	19	20.149	23	6.142	15	12.142	19	20.142	23
	0.010	3.300	2.000	8.000	15	14.000	18	22.000	22	8.000	15	14.000	19	22.000	23
0.99	0.001	22.00	2.000	8.000	15	14.000	19	22.000	22	8.000	15	14.000	19	22.000	22
	0.005	22.00	2.000	8.000	15	14.000	19	22.000	23	8.000	15	14.000	19	22.000	23
	0.010	22.02	2.000	8.000	15	14.000	19	22.000	23	8.000	15	14.000	19	22.000	23

(b) $\varsigma = 0.05$															
		FDR		CDR											
				Covariance-scaling factor $\underline{\lambda}_f = 0.333$								Covariance-scaling factor $\underline{\lambda}_f = 0.5$			
		Benchmark	From CDR Iter 0	Iter 2		Iter 3		Iter 4		Iter 2		Iter 3		Iter 4	
θ	ρ	Time (hr)	Time (hr)	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut	Time (hr)	Cut
0.95	0.001	0.022	0.008	0.849	15	2.112	18	10.112	21	0.558	15	3.043	18	5.789	22
	0.005	0.026	0.008	6.007	15	12.007	19	20.007	23	6.006	15	12.006	19	20.006	23
	0.010	0.024	0.010	6.010	14	12.010	18	20.010	22	6.010	14	12.010	18	20.010	22
0.97	0.001	0.669	0.077	5.184	15	10.418	18	18.418	21	5.024	15	11.024	18	19.024	22
	0.005	0.154	0.139	6.139	15	12.139	18	20.139	22	6.138	15	12.138	18	20.138	22
	0.010	3.300	2.000	8.000	15	14.000	19	22.000	23	7.245	15	13.245	19	21.245	23
0.99	0.001	22.00	2.000	8.000	15	14.000	18	22.000	22	8.000	15	14.000	18	22.000	22
	0.005	22.00	2.000	8.000	14	14.000	18	22.000	21	8.000	14	14.000	18	22.000	21
	0.010	22.02	2.000	8.000	14	14.000	18	22.000	22	8.000	14	14.000	18	22.000	22

7.4.1. Out-of-Sample Performance and the Cost-Reliability Tradeoff. Table 3 quantifies the associated cost–reliability tradeoff. The nominal objective is constant across ρ for each fixed θ , because ρ enters only through the ambiguity set. Its out-of-sample satisfaction (OSS) probability in the hold-out data is 92.52%, 94.02%, and 95.20%, which is significantly below the desired probabilities at $\theta = 0.95, 0.97$, and 0.99 , respectively. The FDR model changes this baseline only modestly. Across all nine (θ, ρ) combinations, the FDR objective increase remains 0.06% to 0.77%, and the corresponding OSS probability gain over the nominal model is at most 1.39%. In none of the reported cases does FDR increase the empirical satisfaction probability to the target level irrespective of runtime settings. Also note that increasing the FDR runtime beyond the iteration-0 budget minimally improves OSS probability: OSS values under both runtime settings are identical in five of the nine (θ, ρ) , with small differences in the cases $\theta = 0.99$. Particularly at $\theta = 0.99$, although the additional computational time was about 20 hours, the OSS improvement was only 0.28% to 0.33% points. Hence, increasing the provable accuracy of a solution to the FDR model has some but incremental value.

Table 3 Nominal objective values and out-of-sample satisfaction probabilities (OSS, %) are reported as baselines for all $\rho \in \{0.001, 0.005, 0.01\}$ and $\theta \in \{0.95, 0.97, 0.99\}$. For FDR, OSS values are reported both from a dedicated benchmark FDR run with $\text{MIPGap} = 0.1\%$ and the same 22-hour runtime budget as CDR, and from iteration 0 of the cutting surface approach. Relative objective difference with respect to nominal are calculated as $100 \frac{|\text{Obj}^{\text{nominal}} - \text{Obj}^{\text{CDR}}|}{\text{Obj}^{\text{nominal}}}$ and represented under Obj Diff (%) column. For the CDR approach, OSS probability and relative objective differences are reported for four different hyperparameter (mean perturbation, covariance scaling factor) = $(\varsigma, \underline{\lambda}_f)$ settings.

θ	ρ	Nominal		FDR			CDR							
		Obj	OSS (%)	OSS (%)			$\varsigma = 0.05, \underline{\lambda}_f = 0.333$		$\varsigma = 0.05, \underline{\lambda}_f = 0.5$		$\varsigma = 0.10, \underline{\lambda}_f = 0.333$		$\varsigma = 0.10, \underline{\lambda}_f = 0.5$	
				Obj ^a Diff (%)	Benchmark	From CDR Iter 0	Obj Diff (%) (FDR) ^b	OSS (%)	Obj Diff (%) (FDR) ^b	OSS (%)	Obj Diff (%) (FDR) ^b	OSS (%)	Obj Diff (%) (FDR) ^b	OSS (%)
0.95	0.001	7011.376	92.52	0.10	92.76	92.76	1.59 (1.49)	95.040	1.59 (1.49)	95.060	2.47 (2.37)	95.867	2.47 (2.37)	95.867
	0.005	7011.376	92.52	0.43	93.45	93.45	2.42 (1.98)	95.940	2.42 (1.98)	95.940	3.72 (3.28)	97.187	3.74 (3.29)	97.173
	0.010	7011.376	92.52	0.77	93.91	93.91	3.10 (2.31)	96.413	3.10 (2.31)	96.413	4.43 (3.63)	97.673	4.43 (3.63)	97.713
0.97	0.001	7073.915	94.02	0.08	94.09	94.09	1.73 (1.65)	96.207	1.72 (1.64)	96.207	2.62 (2.53)	96.787	2.62 (2.53)	96.787
	0.005	7073.915	94.02	0.37	94.53	94.53	2.49 (2.12)	96.793	2.49 (2.12)	96.793	3.98 (3.60)	98.107	3.98 (3.60)	98.107
	0.010	7073.915	94.02	0.65	95.19	94.69	3.14 (2.48)	97.360	3.14 (2.48)	97.353	4.66 (3.99)	98.453	4.66 (3.99)	98.440
0.99	0.001	7176.816	95.20	0.06	95.57	95.27	1.99 (1.93)	97.447	1.99 (1.93)	97.447	2.89 (2.82)	98.100	2.89 (2.82)	98.100
	0.005	7176.816	95.20	0.25	95.60	95.32	2.86 (2.60)	98.160	2.86 (2.60)	98.173	4.51 (4.25)	98.833	4.51 (4.25)	98.833
	0.010	7176.816	95.20	0.41	95.59	95.26	3.58 (3.16)	98.487	3.59 (3.16)	98.487	5.44 (5.00)	99.173	5.44 (5.00)	99.173

^a Identical across the two FDR settings up to two decimals and hence reported once.

^b The non-parenthesized value is relative to the nominal solution, while the parenthesized value is relative to FDR.

By contrast, the CDR results exhibit a systematic progression toward the prescribed service-level targets as ρ increases within each fixed continuous-support setting $(\varsigma, \underline{\lambda}_f)$. For $\theta = 0.95$, all CDR settings satisfy the target already at $\rho = 0.001$, with OSS values ranging from 0.9504 to 0.9587. For $\theta = 0.97$, the $\varsigma = 0.05$ settings improve OSS from 0.9621 to 0.9735–0.9736, crossing the target at $\rho = 0.01$, while the $\varsigma = 0.10$ settings always exceed the target from $\rho = 0.005$ onward. For $\theta = 0.99$, with increasing ρ the $\varsigma = 0.05$ settings increases OSS from 0.9745 to 0.9849, and the $\varsigma = 0.10$ setting reaches 0.9917 at $\rho = 0.01$. These reliability gains are obtained with CDR solution objective value increases of 1.59%–5.44% relative to nominal and 1.49%–5.00% relative to FDR. This increase in the objective values under CDR is accompanied by substantial OSS gains: CDR exceeds the stand-alone FDR OSS by 1.88–3.80 percentage points overall, and by 2.50–3.80, 2.16–3.26, and 2.90–3.58 percentage points at $\rho = 0.01$ for $\theta = 0.95, 0.97$, and 0.99, respectively. Therefore, CDR not only produces higher OSS values than FDR, but also exhibits a consistent response to the ambiguity set radii ρ : under the same $(\varsigma, \underline{\lambda}_f)$ setting, OSS moves closer to the prescribed service level and often exceeds it, whereas FDR remains below the targets throughout the reported settings. The effect is particularly evident for the larger mean-support with parameter $\varsigma = 0.10$, under which CDR either attains the target or remains closest to it across all three service levels.

Within the continuous-support family of solutions, two features stand out. First, increasing ρ raises both cost and OSS monotonically. Second, ς is the economically meaningful hyperparameter, whereas $\underline{\lambda}_f$ is secondary across our settings. For example, at $(\theta, \rho) = (0.99, 0.01)$, moving from $\varsigma = 0.05$ to $\varsigma = 0.10$ increases the objective difference from about 3.58% to 5.44% and OSS from 0.9849 to 0.9917. By contrast, changing $\underline{\lambda}_f$ from 0.333 to 0.5 leaves the reported objective and OSS values almost unchanged. In these EV instances, the extent of uncertainty in the mean matters much more than the uncertainty in the covariance.

7.5. Solution Attributes and Managerial Insights

The out-of-sample performance and cost–reliability trade-off of the CDR solutions provide aggregate evidence of improved robustness, but they do not reveal how robustness is implemented in the energy allocation profile. This distinction is important in the EV demand planning problem because two solutions can have similar objective values or OSS probabilities while reallocating energy across different hours. In order to understand the source of improvement exhibited in the CDR solutions, we compared the nominal, FDR, and CDR allocation vectors directly. Let $\mathbf{x}^N$ denote the nominal allocation and let $\mathbf{x}^{\text{DR}}$ denote the allocation returned by a DR formulation. For each hour t , define the signed allocation difference $\Delta_t := x_t^{\text{DR}} - x_t^N$.

We use three diagnostic plots, each designed to capture and explain Δ from different aspects. The first diagnostic is a violin plot of the absolute deviations $\{|\Delta_t|\}_t$, which displays the empirical distribution of hourly change magnitudes; it is used to distinguish whether robustness is achieved through broad small-scale adjustments or through a few large reallocations. The second diagnostic is a heatmap of the signed deviations Δ_t , which preserves the hour index and the direction of change; it is used to identify when the DR solution increases or decreases allocation relative to the nominal schedule. The third diagnostic is a JS–cosine scatter plot, which compares each DR allocation profile with the nominal profile using two full-vector similarity measures: cosine similarity captures directional alignment between $\mathbf{x}^N$ and $\mathbf{x}^{\text{DR}}$, while Jensen–Shannon divergence, computed after normalizing the nonnegative allocation profiles, captures distributional separation in the allocation pattern. Together, the three diagnostics connect aggregate robustness outcomes to operational changes in the charging schedule: the violin plot measures how large the hourly changes are, the heatmap identifies where and in which direction they occur through hour-by-hour signed differences Δ_t , and the JS–cosine plot summarizes the overall distance of the DR profile from the nominal profile. Figure 2 reports these diagnostics for $\rho = 0.01$, while Figure 5 in Appendix D.3 reports the corresponding results for $\rho = 0.005$ and $\rho = 0.001$ in top-down order.

7.5.1. Discussion on violin plot, thermal map, and scatter plot. Across all nine (θ, ρ) settings, the FDR solution remains much more similar to the nominal benchmark than any CDR solution. Its violin plots have the tightest spread, its heatmap column stays close to zero in most hours, and its JS–cosine point lies nearest to the upper-left corner of the scatter plots. Quantitatively, the FDR solution never moves beyond about 0.0045 in Jensen–Shannon divergence and remains between about 0.994 and 1.000 in cosine similarity. The continuous-support solutions, on the other hand, are uniformly farther away, with JS divergences roughly staying between 0.003 and 0.018 and cosine similarities between about 0.973 and 0.998. The cost and reliability differences in Table 3 therefore correspond to genuinely different allocation policies rather than to negligible perturbations of the nominal schedule.

The heatmaps show that these allocation-profile differences are highly structured. The robust reallocations are concentrated in a small number of recurring time blocks. Positive shifts appear repeatedly around hours 11–12, 17, and 19–21, often extending to hour 22; in the more stringent $\theta = 0.99$ cases, these increases

are frequently offset by reductions at around hour 10. The continuous-support model thus improves reliability primarily through temporal redistribution. It shifts energy allocation away from hours whose nominal protection is less contributive and toward hours that repeatedly emerge as critical under distributional perturbation. This is also why the violin plots broaden mainly through their upper tails. Thus, CDR policy need not be uniformly different from the nominal schedule, but selectively different where protection matters most.

The similarity plots further indicate that a larger ambiguity radius does not necessarily correspond to the greatest global dissimilarity from the nominal allocation. The major difference in the allocation profile appears between the nominal and CDR. Within the continuous-support family, however, the ordering in JS–cosine space is not uniform. In particular, at $\theta = 0.99$ and $\rho = 0.005$, the $\varsigma = 0.10$ solutions are closer to the nominal profile than the $\varsigma = 0.05$ solutions under both similarity measures, even though both of them are more conservative in objective value and attain higher OSS. The heatmaps help explain this distinction: a more conservative model may obtain its gain through concentrated changes in a small set of influential hours, whereas a less conservative model may spread smaller adjustments over a broader portion of the day. Allocation profile similarity and robustness cost, therefore, measure related but distinct aspects of the solution.

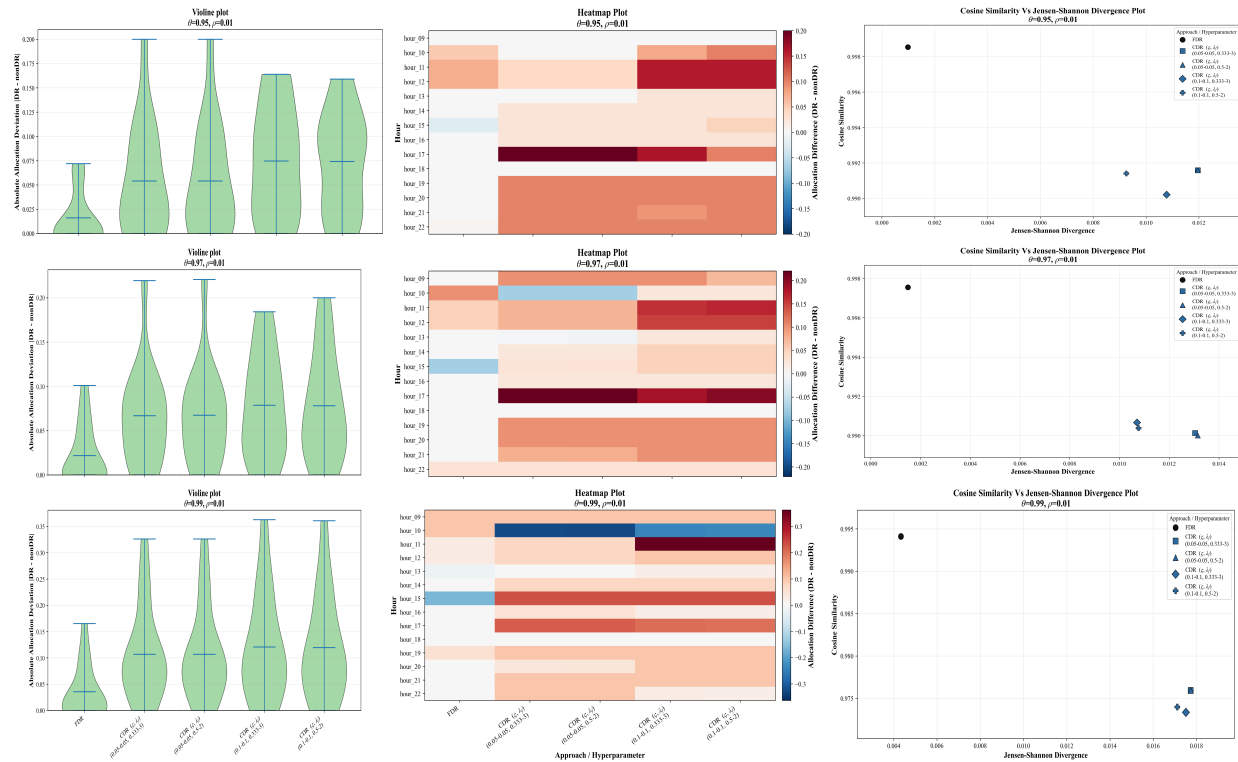


Figure 2 Comparison of finite- and continuous-support solutions with $(\varsigma, \lambda_{\min}) \in \{(0.05, 0.333), (0.05, 0.5), (0.1, 0.333), (0.1, 0.5)\}$ for $\rho = 0.01$, with rows ordered from top to bottom by $\theta = 0.95, 0.97, 0.99$. Within each row, the left panel shows the violin plot, i.e., the distribution of absolute deviations from the nominal solution; the middle panel shows the heatmap of allocation differences relative to the nominal solution, and the right panel shows cosine similarity versus Jensen-Shannon divergence.

7.5.2. Managerial insights. For the EV demand planning problem, a clear managerial message is that robustness is fundamentally a timing decision. The CDR model does not only recommend “more” protection; it identifies when protection is most valuable. Since the main changes are concentrated in a limited subset of hours, higher service reliability need not be pursued through a uniform expansion of hourly allocations. It can instead be achieved by redesigning the charging schedule toward the hours that matter most for maintaining feasibility under distributional mis-specification.

The results also clarify the distinct roles of the two DR formulations. FDR is best viewed as a low-cost hedge around the empirical support of the fitted model. It is computationally attractive and remains close to the nominal solution, but its gains in empirical reliability are limited. CDR is the relevant model when the planner is also concerned about mis-specification in the fitted mixture parameters themselves. In that case, robustness protects not only against variation in realized demand but also against error in the estimated distributional representation used to construct the plan.

Within the continuous-support model, the mean perturbation parameter ς acts as a practical robustness control. In the examined instances, $\varsigma = 0.05$ produces a moderate-robustness regime: it increases OSS probability materially above FDR while keeping the additional objective premium over FDR between about 1.5% and 3.2%. The setting $\varsigma = 0.10$ defines a high-protection regime: it yields the strongest empirical service levels, including the only cases that exceed 99% OSS when $\theta = 0.99$, but it can require up to about a 5% cost premium over FDR. The computational results thus suggest a corresponding workflow: FDR can serve as a fast screening model, while CDR is most valuable when the benefit of higher reliability justifies the additional solve time and cost premium.

Taken together, the numerical evidence shows that CDR adds a distinct robustness layer beyond FDR. It improves empirical satisfaction markedly; it does so through interpretable changes in the timing of the charging plan, and it provides a transparent cost-reliability tradeoff through the design of the parameter-support set.

8. Conclusion

We studied a distributionally robust chance-constrained framework with continuous support in Gaussian-mixture parameters and developed a cutting-surface-based solution approach for its implementation. Relative to the nominal and finite-support DR models, the continuous-support formulation delivers stronger out-of-sample reliability and can induce a meaningfully different allocation pattern, showing that protection against broader mean–covariance misspecification may require more than a mild perturbation of the nominal plan. The computational results also indicate that the implemented approach is practically meaningful, with early iterations identifying the dominant violating distributions and later iterations performing refinement.

An important direction for future work is computational acceleration. In the current implementation, the master problem relies on off-the-shelf solvers, with limited control over the piecewise-linear approximation

and the associated branching behavior on binary variables, and it is re-solved from scratch after each newly added cut. These features create a significant runtime burden. Developing warm-started master reoptimization, cut bundling, adaptive management of the PWL approximation, and more tailored branch-and-cut or decomposition strategies would substantially improve scalability and strengthen the practical applicability of the proposed framework. Overall, the study shows that continuous-support robustification is both practically relevant and computationally approachable, while also highlighting the need for more specialized algorithms to fully unlock its benefits.

References

- [1] Feras Al Taha, Tyrone L. Vincent, and Eilyan H. Bitar. An efficient method for quantifying the aggregate flexibility of plug-in electric vehicle populations. *IEEE Transactions on Smart Grid*, 2024.
- [2] Tsuyoshi Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979.
- [3] Javiera Barrera, Tito Homem-de Mello, Eduardo Moreno, Bernardo K. Pagnoncelli, and Gianpiero Canessa. Chance-constrained problems and rare events: An importance sampling approach. *Mathematical Programming*, 157(1):153–189, 2016.
- [4] Mokhtar S. Bazaraa, Hanif D. Sherali, and Chitharanjan M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, Hoboken, NJ, 3 edition, 2006.
- [5] Aharon Ben-Tal, Dick den Hertog, and Anja De Waegenare. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- [6] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.
- [7] Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, Princeton, NJ, 2007.
- [8] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.
- [9] Jose Blanchet, Duc K. Nam, and Bert Zwart. Efficient scenario generation for heavy-tailed chance-constrained optimization. *Stochastic Systems*, 14(1):22–46, 2024.
- [10] Vladimir I. Bogachev. *Measure Theory*, volume 1–2. Springer, Berlin, 2007. See Theorem 10.4.1 for integration with respect to a measurable kernel.
- [11] J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, New York, 2000.
- [12] Marco Casini, Antonio Vicino, and Giovanni Gino Zanvetto. A chance constraint approach to peak mitigation in electric vehicle charging stations. *Automatica*, 131:109746, 2021.
- [13] Frank Y. Chen and Dmitry Krass. Inventory models with minimal service level constraints. *European Journal of Operational Research*, 134(1):120–140, 2001.
- [14] Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018.
- [15] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Technical note—data-driven chance constrained programs over Wasserstein balls. *Operations Research*, 72(1):410–424, 2024.
- [16] Erhan Çinlar. *Probability and Stochastics*. Springer, New York, 2011.
- [17] Tien Thanh Dam, Thuy Anh Ta, and Tien Mai. Joint chance-constrained staffing optimization in multi-skill call centers. *Journal of Combinatorial Optimization*, 44(1):354–378, 2022.
- [18] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.
- [19] Julie Delon and Agnès Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.
- [20] Yan Deng, Siqian Shen, and Brian Denton. Chance-constrained surgery planning under conditions of limited and ambiguous data. *INFORMS Journal on Computing*, 31(3):559–575, 2019.
- [21] Yan Deng, Siqian Shen, and Brian T. Denton. Chance-constrained surgery planning under conditions of limited and ambiguous data. *INFORMS Journal on Computing*, 31(3):559–575, 2019.
- [22] Shibshankar Dey, Sanjay Mehrotra, and Anirudh Subramanyam. On solving chance-constrained models with gaussian mixture distribution, 2025.
- [23] Ke-wei Ding, Nan-jing Huang, and Lei Wang. Distributionally robust joint chance constrained problem under moment uncertainty. *Journal of Applied Mathematics*, 2014:487178, 2014. Replaces an unvalidated placeholder entry for a Ding et al. moment-information DR chance-constraint reference.

- [24] John C. Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–73, 2019.
- [25] Abolhassan Mohammadi Fathabad, Jianqiang Cheng, Kai Pan, and Boshi Yang. Asymptotically tight conic approximations for chance-constrained AC optimal power flow. *European Journal of Operational Research*, 305(2):738–753, 2023.
- [26] Figshare. A high-resolution electric vehicle charging transaction dataset with multidimensional features in china. Figshare dataset repository, 2025. Companion dataset repository for Scientific Data article.
- [27] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, 2 edition, 1999.
- [28] Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.
- [29] Jacques Gauvin and Jon W. Tolle. Differential stability in nonlinear programming. *SIAM Journal on Control and Optimization*, 15(2):294–311, 1977.
- [30] Matthias Gelbrich. On a formula for the l^2 -wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.
- [31] Miguel Ángel Goberna and Marco A. López, editors. *Semi-Infinite Programming: Recent Advances*, volume 57 of *Nonconvex Optimization and Its Applications*. Kluwer Academic Publishers, Dordrecht, 2001.
- [32] Grani A. Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. Ambiguous joint chance constraints under mean and dispersion information. *Operations Research*, 65(3):751–767, 2017.
- [33] Rainer Hettich and Kenneth O. Kortanek. Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, 35(3):380–429, 1993.
- [34] Nam Ho-Nguyen, Fatma Kılınç-Karzan, Simge Küçükyavuz, and Dabeen Lee. Distributionally robust chance-constrained programs with right-hand side uncertainty under Wasserstein ambiguity, 2020.
- [35] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2 edition, 2012.
- [36] Jian Hu and L. Jeff Hong. Kullback–Leibler divergence constrained distributionally robust optimization. *Operations Research*, 61(4):1073–1086, 2013.
- [37] Zhaolin Hu, Wenjie Sun, and Shushang Zhu. Chance constrained programs with gaussian mixture models. *IIE Transactions*, 54(12):1117–1130, 2022.
- [38] Ran Ji and Miguel A. Lejeune. Data-driven distributionally robust chance-constrained optimization with Wasserstein metric. *Journal of Global Optimization*, 79(4):779–811, 2021.
- [39] Rui Jiang and Yongpei Guan. Data-driven chance-constrained stochastic program. *Mathematical Programming*, 158(1–2):291–327, 2016.
- [40] Yuchen Jiang, Cong Shi, and Siqian Shen. Service level constrained inventory systems. *Production and Operations Management*, 28(9):2365–2389, 2019.
- [41] Sanjula Kammammettu, Shu-Bo Yang, and Zukui Li. Distributionally robust chance-constrained optimization with gaussian mixture ambiguity set. *Computers & Chemical Engineering*, 186:108703, 2024.
- [42] L. V. Kantorovich. On the translocation of masses. *Doklady Akademii Nauk SSSR*, 37:227–229, 1942.
- [43] L. V. Kantorovich and G. Sh. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad University*, 13(7):52–59, 1958. Often cited as an origin of the Kantorovich–Rubinstein metric; transliteration and bibliographic details vary.
- [44] Alireza Khaksari, Georgios Tsaousoglou, Prodromos Makris, Konstantinos Steriotis, Nikolaos Efthymiopoulos, and Emmanouel Varvarigos. Sizing of electric vehicle charging stations with smart charging capabilities and quality of service requirements. *Sustainable Cities and Society*, 70:102872, 2021.
- [45] Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, London, 2008.
- [46] Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.
- [47] Soheil Kolouri, Gustavo K. Rohde, and Heiko Hoffmann. Sliced Wasserstein distance for learning Gaussian mixture models, 2018.
- [48] Stanley J. Kon. Models of stock returns—a comparison. *The Journal of Finance*, 39(1):147–165, 1984.
- [49] Fumio Kubo and Tsuyoshi Ando. Means of positive linear operators. *Mathematische Annalen*, 246:205–224, 1980.
- [50] Junhyeok Kwon, Dongha Kim, Joong-Ho Won, and Myunghee Cho Paik. A principled learning method for wasserstein distributionally robust optimization with local perturbations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5567–5576. PMLR, 2020.
- [51] Henry Lam. Recovering the best-case distribution in robust stochastic optimization. *Operations Research*, 67(5):1328–1352, 2019.
- [52] Jonathan Li and Andrew Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems*, volume 12, 1999.
- [53] Yi Lin, Jing Mu, Jun Zhou, and Hongtu Zhou. Multisource single-cell data integration by MAW barycenter for gaussian mixture models. *Biometrics*, 79(2):866–877, 2023.
- [54] Bruce G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, CA, 1995.

- [55] Jiayan Liu, Gang Lin, Christian Rehtanz, Sunhua Huang, Yang Zhou, and Yong Li. Data-driven intelligent EV charging operating with limited chargers considering the charging demand forecasting. *International Journal of Electrical Power & Energy Systems*, 141:108218, 2022.
- [56] Fengqiao Luo and Sanjay 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):20–35, 2019.
- [57] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein riemannian geometry of gaussian densities. *Information Geometry*, 1:137–179, 2018.
- [58] Sanjay Mehrotra and Dávid Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimization*, 24(4):1670–1697, 2014.
- [59] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.
- [60] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.
- [61] Eliakim Hastings Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26(9):394–395, 1920.
- [62] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance-constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.
- [63] Christopher Neuman, Andrew Meintz, and Myungsoo Jun. Workplace charging data collection and behavior. Technical Report NREL/TP-5400-78983, National Renewable Energy Laboratory, Golden, CO, 2021.
- [64] Viet Anh Nguyen, Soroosh Shafiee, Damir Filipović, and Daniel Kuhn. Mean-covariance robust risk measurement, 2025. arXiv:2112.09959v3, 1 Oct 2025.
- [65] Roger Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.
- [66] Gabriel Pérez-Quiros and Allan Timmermann. Firm size and cyclical variations in stock returns. *The Journal of Finance*, 55(3):1229–1262, 2000.
- [67] Niklas Piening, Daniel Reich, and Tim Schulze. Double-sliced mixture Wasserstein distances: Parallelizable one-dimensional optimal transport for Gaussian mixtures, 2025.
- [68] R. Reemtsen and J.-J. Rückmann, editors. *Semi-Infinite Programming*. Kluwer Academic Publishers, Boston, 1998.
- [69] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 1998.
- [70] Roberto Rossi. On service level measures in stochastic inventory control. *IFAC Proceedings Volumes*, 46(9):1991–1996, 2013.
- [71] Roberto Rossi, S. Armagan Tarim, Brahim Hnich, and Steven Prestwich. A global chance-constraint for stochastic inventory systems under service level constraints. *Constraints*, 13(4):490–517, 2008.
- [72] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3 edition, 1976.
- [73] Yves Rychener, Adrián Esteban-Pérez, Juan M. Morales, and Daniel Kuhn. Wasserstein distributionally robust optimization with heterogeneous data sources, 2024.
- [74] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2 edition, 2014.
- [75] Haoming Shen and Ruiwei Jiang. Convex chance-constrained programs with Wasserstein ambiguity. *Operations Research*, 73(4):2264–2280, 2025.
- [76] Oleg V. Shylo, Oleg A. Prokopyev, and Andrew J. Schaefer. Stochastic operating room scheduling for high-volume specialties under block booking. *INFORMS Journal on Computing*, 25(4):682–692, 2013.
- [77] Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.
- [78] Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.
- [79] S. Armagan Tarim and Brian G. Kingsman. The stochastic dynamic production/inventory lot-sizing problem with service-level constraints. *International Journal of Production Economics*, 88(1):105–119, 2004.
- [80] Niels van der Laan, Ruud H. Teunter, Ward Romeijnnders, and Onur A. Kilic. The data-driven newsvendor problem: Achieving on-target service-levels using distributionally robust chance-constrained optimization. *International Journal of Production Economics*, 249:108509, 2022.
- [81] L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large automorphisms of measures. *Problems of Information Transmission*, 5(3):47–52, 1969.
- [82] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 2008.
- [83] Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *Operations Research*, 0(0), 2025. Articles in advance; volume and page numbers to be updated after issue assignment.
- [84] Shanshan Wang, Jinlin Li, and Sanjay Mehrotra. Chance-constrained multiple bin packing problem with an application to operating room planning. *INFORMS Journal on Computing*, 33(4):1661–1677, 2021.

- [85] Jinxiang Wei, Zhaolin Hu, Jun Luo, and Shushang Zhu. Enhanced branch-and-bound algorithm for chance constrained programs with gaussian mixture models. *Annals of Operations Research*, 338(2):1283–1315, 2024.
- [86] Weijun Xie. On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming*, 186(1–2):115–155, 2021.
- [87] Chen Xu, Jonghyeok Lee, Xiuyuan Cheng, and Yao Xie. Flow-based distributionally robust optimization. *IEEE Journal on Selected Areas in Information Theory*, 5:62–77, 2024.
- [88] Tianyang Yi, Shibshankar Dey, D. Adrian Maldonado, Sanjay Mehrotra, and Anirudh Subramanyam. Discrete-continuous Gaussian mixture models for wind power generation. In *2024 IEEE Power & Energy Society General Meeting*, pages 1–5. IEEE, 2024.
- [89] YoungChul Yoon, Grani A. Hanasusanto, and Yijie Wang. Data-driven contextual optimization with gaussian mixtures: Flow-based generalization, robust models, and multistage extensions, 2025.
- [90] Luhao Zhang, Jincheng Yang, and Rui Gao. A short and general duality proof for Wasserstein distributionally robust optimization. *Operations Research*, 73(4):2146–2155, 2025. Published online in 2024; assigned to 2025 volume/issue.
- [91] Yuanshi Zhang, Tongxin Xu, Tian Chen, et al. A high-resolution electric vehicle charging transaction dataset with multidimensional features in china. *Scientific Data*, 12:643, 2025. 441,077 charging transactions from 13 stations in Jiaying, China, over two years.
- [92] Linglingzhi Zhu and Yao Xie. Distributionally robust optimization via iterative algorithms in continuous probability spaces, 2024.

Appendix A: Known Results and Tools

A.1 Integration interchangeability.

Theorem 7 (Tonelli-Fubini Theorem for product measure [27]). *Let $(\mathbb{X}, \mathcal{B}(\mathbb{X}), \mu)$ and $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}), \nu)$ be probability measure spaces with probability measures μ and ν respectively. If $f : \mathbb{X} \times \mathbb{Y} \rightarrow [0, \infty]$ be a nonnegative measurable function (Tonelli Condition) or $f : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{R}$ be integrable, i.e. $\int_{\mathbb{X} \times \mathbb{Y}} |f| d(\mu \times \nu) < \infty$ (Fubini Condition), Then:*

i) *In the Tonelli case $f \geq 0$, the maps*

$$\chi \mapsto \int_{\mathbb{Y}} f(\chi, y) d\nu(y), \quad y \mapsto \int_{\mathbb{X}} f(\chi, y) d\mu(\chi)$$

are well-defined as $[0, \infty]$ -valued measurable functions. In the Fubini case $f \in L^1(\mu \otimes \nu)$, these maps are finite almost everywhere and integrable with respect to μ and ν , respectively.

ii) *the iterated integrals and the product-space integral satisfy*

$$\int_{\mathbb{X} \times \mathbb{Y}} f(\chi, y) (\mu \otimes \nu)(d\chi, dy) = \int_{\mathbb{X}} \left(\int_{\mathbb{Y}} f(\chi, y) d\nu(y) \right) d\mu(\chi) = \int_{\mathbb{Y}} \left(\int_{\mathbb{X}} f(\chi, y) d\mu(\chi) \right) d\nu(y).$$

Theorem 8 (Tonelli–Fubini Theorem for Markov Kernels). ([10, Section 10.7]; [16, Section 1.6]; [45, Section 8.3]) *Let $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ be measurable spaces, and let ν_χ be a probability measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. Let $\mathcal{K} : \mathbb{X} \times \mathcal{B}(\mathbb{Y}) \rightarrow [0, 1]$ be a transition kernel, that is:*

- a) *For every $\chi \in \mathbb{X}$, the map $C \mapsto \mathcal{K}(\chi, C)$ is a probability measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$, and*
- b) *For every $C \in \mathcal{B}(\mathbb{Y})$, the map $\chi \mapsto \mathcal{K}(\chi, C)$ is $\mathcal{B}(\mathbb{X})$ -measurable, hence ν_χ -measurable for any probability measure ν_χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$.*

Define the induced probability measure ν_y on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ by

$$\nu_y(C) := \int_{\mathbb{X}} \mathcal{K}(\chi, C) d\nu_\chi(\chi), \quad C \in \mathcal{B}(\mathbb{Y}).$$

Then, if $\Psi : \mathbb{Y} \rightarrow \mathbb{R}$ is measurable, the following two integral rules hold.

(Tonelli case). *If $\Psi \geq 0$, then*

$$\int_{\mathbb{Y}} \Psi(y) d\nu_y(y) = \int_{\mathbb{X}} \left(\int_{\mathbb{Y}} \Psi(y) \mathcal{K}(\chi, dy) \right) d\nu_\chi(\chi). \quad (52)$$

Both sides are equal in $[0, \infty]$.

(Fubini case). If Ψ is integrable with respect to ν_y , i.e., $\int_{\mathbb{Y}} |\Psi(y)| d\nu_y(y) < \infty$, then:

- i) for ν_χ -almost every $\chi \in \mathbb{X}$, the map $y \mapsto \Psi(y)$ is $\mathcal{K}(\chi, \cdot)$ -integrable, i.e., $\int_{\mathbb{Y}} |\Psi(y)| \mathcal{K}(\chi, dy) < \infty$,
- ii) the map $\chi \mapsto \int_{\mathbb{Y}} \Psi(y) \mathcal{K}(\chi, dy)$ is integrable with respect to ν_χ , (since
$$\int_{\mathbb{X}} \left| \int_{\mathbb{Y}} \Psi(y) \mathcal{K}(\chi, dy) \right| d\nu_\chi(\chi) \leq \int_{\mathbb{X}} \int_{\mathbb{Y}} |\Psi(y)| \mathcal{K}(\chi, dy) d\nu_\chi(\chi) = \int_{\mathbb{Y}} |\Psi(y)| d\nu_y(y) < \infty$$
).
- iii) and identity (52) holds with finite values.

Remark 5 (Reading (52)). For $\Psi = \mathbb{1}_C$, the definition of ν_y gives

$$\int_{\mathbb{Y}} \mathbb{1}_C(y) d\nu_y(y) = \int_{\mathbb{X}} \left(\int_{\mathbb{Y}} \mathbb{1}_C(y) \mathcal{K}(\chi, dy) \right) d\nu_\chi(\chi),$$

i.e., identity (52) reduces exactly to the defining relation

$$\nu_y(C) = \int_{\mathbb{X}} \mathcal{K}(\chi, C) d\nu_\chi(\chi).$$

Thus (52) is the function-level extension of the setwise definition of ν_y ; the extension from indicators to nonnegative measurable functions follow from the standard simple-function/monotone convergence argument, and the integrable case follows from the decomposition $\Psi = \Psi^+ - \Psi^-$ [27].

Remark 6 (Interchanging the kernel integration order). The kernel $\mathcal{K}(\chi, dy)$ defines the joint measure

$$\eta(d\chi, dy) = \nu_\chi(d\chi) \mathcal{K}(\chi, dy).$$

If a reverse conditional kernel $\mathcal{L}(y, d\chi)$ exists, i.e., $\eta(d\chi, dy) = \nu_y(dy) \mathcal{L}(y, d\chi)$, then for every nonnegative measurable $F : \mathbb{X} \times \mathbb{Y} \rightarrow [0, \infty]$,

$$\int_{\mathbb{X} \times \mathbb{Y}} F d\eta = \int_{\mathbb{X}} \int_{\mathbb{Y}} F(\chi, y) \mathcal{K}(\chi, dy) d\nu_\chi(\chi) = \int_{\mathbb{Y}} \int_{\mathbb{X}} F(\chi, y) \mathcal{L}(y, d\chi) d\nu_y(y).$$

Thus, the reverse-order expression requires the reverse kernel $\mathcal{L}$, not $\mathcal{K}(d\chi, y)$.

A.2 Gaussian instantiation of transition kernel. We start with a standard result on preservation of measurability under integration, which will be invoked throughout the sequel.

Proposition 11. Let $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $(\mathfrak{Z}, \mathcal{B}(\mathfrak{Z}))$ be measurable spaces, let γ be a measure on $(\mathfrak{Z}, \mathcal{B}(\mathfrak{Z}))$, and let $g : \mathbb{X} \times \mathfrak{Z} \rightarrow [0, \infty]$ be $\mathcal{B}(\mathbb{X}) \otimes \mathcal{B}(\mathfrak{Z})$ -measurable. If

$$\mathcal{G}(\chi) := \int_{\mathfrak{Z}} g(\chi, z) d\gamma(z),$$

then $\mathcal{G} : \mathbb{X} \rightarrow [0, \infty]$ is $\mathcal{B}(\mathbb{X})$ -measurable.

Now, leveraging the above Proposition 11, we show that Gaussian is a special instantiation of the Markov kernel. Let

$$\mathbb{X} := \mathbb{R}^n \times \mathbb{S}_+^n, \quad \mathfrak{Z} := \mathbb{R}^n, \quad \mathbb{Y} := \mathbb{R}^n,$$

equipped with their Borel σ -algebras. Let $\mathbf{A}(\mathbf{Q}) := \mathbf{Q}^{1/2}$, where $\mathbf{Q}^{1/2}$ is the principal symmetric positive semidefinite square root of $\mathbf{Q}$. The map $\mathbf{Q} \mapsto \mathbf{Q}^{1/2}$ is Borel measurable on $\mathbb{S}_+^n$, and

$$\mathbf{A}(\mathbf{Q})\mathbf{A}(\mathbf{Q})^\top = \mathbf{Q} \quad \text{for all } \mathbf{Q} \in \mathbb{S}_+^n.$$

Fix $\gamma := \mathcal{N}(\mathbf{0}, \mathbf{I})$ with identity matrix $\mathbf{I}$ on $(\mathfrak{Z}, \mathcal{B}(\mathfrak{Z}))$. For $\chi = (\mathbf{m}, \mathbf{Q}) \in \mathbb{X}$ and $C \in \mathcal{B}(\mathbb{Y})$, define the map $\mathcal{K} : \mathbb{X} \times \mathcal{B}(\mathbb{Y}) \rightarrow [0, 1]$ by

$$\mathcal{K}(\chi, C) := \int_{\mathfrak{Z}} g_C(\chi, z) d\gamma(z),$$

where the integrand g_C is

$$g_C(\chi, \mathbf{z}) := \mathbb{1}_C(\mathbf{m} + \mathbf{A}(\mathbf{Q})\mathbf{z}).$$

We now verify the two Markov Kernel conditions (a)-(b) stated in Theorem 8 for the Gaussian instantiation:

1. **Condition (a) of Theorem 8:** For each fixed χ , $C \mapsto \mathcal{K}(\chi, C)$ is a probability measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$.

Define $\mathcal{N}(\mathbf{m}, \mathbf{Q})$ as a pushforward of $\gamma = \mathcal{N}(0, \mathbf{I})$ as follows

$$\mathcal{N}(\mathbf{m}, \mathbf{Q}) := (T_\chi)_\# \gamma, \quad T_\chi(\mathbf{z}) := \mathbf{m} + \mathbf{A}(\mathbf{Q})\mathbf{z}.$$

Then for every $C \in \mathcal{B}(\mathbb{Y})$,

$$\mathcal{N}(\mathbf{m}, \mathbf{Q})(C) = (T_\chi)_\# \gamma(C) = \gamma(T_\chi^{-1}(C)) = \int_{\mathfrak{Z}} \mathbb{1}_C(T_\chi(\mathbf{z})) d\gamma(\mathbf{z}) = \int_{\mathfrak{Z}} \mathbb{1}_C(\mathbf{m} + \mathbf{A}(\mathbf{Q})\mathbf{z}) d\gamma(\mathbf{z}) = \mathcal{K}(\chi, C).$$

Clearly, $\mathcal{N}(\mathbf{m}, \mathbf{Q})$ is a probability measure and $\mathcal{N}(\mathbf{m}, \mathbf{Q})(C)$ is the evaluation of that measure on the set C . So $C \mapsto \mathcal{K}(\chi, C) (= \mathcal{N}(\mathbf{m}, \mathbf{Q})(C))$ is the pushforward measure $(T_\chi)_\# \gamma$. Pushforwards of probability measures are probability measures. Concretely:

- $\mathcal{K}(\chi, \emptyset) = 0$, $\mathcal{K}(\chi, \mathbb{Y}) = 1$,
- countable additivity holds because $C \mapsto \gamma(T_\chi^{-1}(C))$ is a measure.

So condition (a) in Theorem 8 is satisfied.

Remark 7. We emphasize that no Lebesgue density was used, so a singular $\mathbf{Q}$ is allowed. Also, $\mathcal{N}(\mathbf{m}, \mathbf{Q})$ is a probability measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$ while $\mathcal{K}(\chi, C)$ is the evaluation of that measure on the set C . Hence,

$$\mathcal{K}(\chi, C) = \mathcal{N}(\mathbf{m}, \mathbf{Q})(C), \quad \chi = (\mathbf{m}, \mathbf{Q}) \in \mathbb{X}, C \in \mathcal{B}(\mathbb{Y}).$$

2. **Condition (b) of Theorem 8:** For each fixed C , $\chi \mapsto \mathcal{K}(\chi, C)$ is ν_χ -measurable.

For each fixed $C \in \mathcal{B}(\mathbb{Y})$, we want to prove ν_χ measurability of

$$\chi \mapsto \mathcal{K}(\chi, C) = \int_{\mathfrak{Z}} g_C(\chi, \mathbf{z}) d\gamma(\mathbf{z}).$$

By Proposition 11, it is enough to show that

$$g_C : \mathbb{X} \times \mathfrak{Z} \rightarrow [0, 1], \quad g_C(\chi, \mathbf{z}) = \mathbb{1}_C(\mathbf{m} + \mathbf{A}(\mathbf{Q})\mathbf{z})$$

is $\mathcal{B}(\mathbb{X}) \otimes \mathcal{B}(\mathfrak{Z})$ -measurable.

But the map

$$h : \mathbb{X} \times \mathfrak{Z} \rightarrow \mathbb{Y}, \quad h((\mathbf{m}, \mathbf{Q}), \mathbf{z}) = \mathbf{m} + \mathbf{A}(\mathbf{Q})\mathbf{z}$$

is measurable because it is built from measurable operations (addition, matrix–vector multiplication) and $\mathbf{A}(\cdot)$ is measurable by construction. Then

$$g_C = \mathbb{1}_C \circ h$$

is measurable since $\mathbb{1}_C$ is Borel measurable on $\mathbb{Y}$. Therefore, by the lemma, $\chi \mapsto \int g_C(\chi, \mathbf{z}) d\gamma(\mathbf{z})$ is measurable. This proves condition (b).

A.3 Notation correspondence of Markov transition kernel for the proof $W_2^{\text{prm}} = W_2^{\text{mix}}$. Let Ψ be a measurable map. Then the following allows us to apply the Markov-kernel Tonelli–Fubini theorem, Theorem 8:

Kernel-based Theorem 8 Notation	Mapped Notation given $\pi \in \Gamma$
$\mathbb{X}$	$\mathcal{S} \times \mathcal{S}$
$\mathbb{Y}$	$\mathbb{R}^{2n}$
$\mathcal{K}(\chi, \cdot)$	$\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\cdot)$
ν_χ	$\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$
$\nu_y(C)$	$\nu_\pi(C) = \int_{\mathcal{S} \times \mathcal{S}} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(C) d\pi(\mathbf{s}_1, \mathbf{s}_2)$
$\Psi(y)$	$\Psi(\mathbf{y}_1, \mathbf{y}_2) = \mathbb{1}_A(\mathbf{y}_1)$

Similarly, for the other directions:

Kernel-based Theorem 8 Notation	Mapped Notation given $\nu \in \Pi^{\text{mix}}(\mu_1, \mu_2)$
$\mathbb{X}$	$\mathcal{Z} = \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{S}_+^{2n}$
$\mathbb{Y}$	$\mathbb{R}^{2n}$
$\mathcal{K}(\chi, \cdot)$	$\mathcal{N}(\mathbf{z})(\cdot)$
ν_χ	$\lambda \in \mathcal{P}(\mathcal{Z})$
$\nu_y(C)$	$\nu(C) = \int_{\mathcal{Z}} \mathcal{N}(\mathbf{z})(C) d\lambda(\mathbf{z})$
$\Psi(y)$	$\Psi(\mathbf{y}_1, \mathbf{y}_2) = \mathbb{1}_B(\mathbf{y}_2)$

For objective identities, the same mapping is used with $\Psi(\mathbf{y}_1, \mathbf{y}_2) = \|\mathbf{y}_1 - \mathbf{y}_2\|^2$, which is nonnegative, so the Tonelli case applies. For the first set of mappings, intuition to define ν_π comes from the Gelbrich OT map Theorem 1, while for the second set of maps, it is a Gaussian mixture measure defined in (5).

A.4 Optimal-transport discrepancy and ambiguity set [8, Theorem 1]. Let $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ be a Polish space with its Borel σ -algebra, and let $\mathcal{P}(\mathcal{S})$ be the set of Borel probability measures on $\mathcal{S}$. Assume $c : \mathcal{S} \times \mathcal{S} \rightarrow [0, \infty]$ is a measurable transportation cost function. Define the optimal transport cost between $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{S})$ by

$$W_c(\mu_1, \mu_2) := \inf_{\pi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{S} \times \mathcal{S}} c(\mathbf{s}_1, \mathbf{s}_2) \pi(d\mathbf{s}_1, d\mathbf{s}_2),$$

where $\Pi(\mu_1, \mu_2)$ is the set of all couplings of μ_1 and μ_2 .

Fix a nominal distribution $\hat{\mu} \in \mathcal{P}(\mathcal{S})$ and a radius $\rho \geq 0$. The corresponding OT ambiguity set is

$$\mathcal{D}(\hat{\mu}, \rho) := \{\mu \in \mathcal{P}(\mathcal{S}) : W_c(\hat{\mu}, \mu) \leq \rho\}.$$

Let $G : \mathcal{S} \rightarrow \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ be a measurable payoff (loss) function. Consider the worst-case expectation

$$\sup_{\mu \in \mathcal{D}(\hat{\mu}, \rho)} \int_{\mathcal{S}} G(\mathbf{s}) \mu(d\mathbf{s}).$$

(A1) Regularity of transport cost. c is lower semicontinuous on $\mathcal{S} \times \mathcal{S}$ and satisfies $c(\mathbf{s}, \mathbf{s}) = 0$ for every $\mathbf{s} \in \mathcal{S}$.

(A2) Upper semicontinuity of the payoff. G is upper semicontinuous on $\mathcal{S}$.

(A3) Integrability condition. There exists some $\beta \geq 0$ such that the following is finite:

$$\int_{\mathcal{S}} \sup_{\mathbf{s} \in \mathcal{S}} \{G(\mathbf{s}) - \beta c(\hat{\mathbf{s}}, \mathbf{s})\} \hat{\mu}(d\hat{\mathbf{s}}) < \infty.$$

Then strong duality holds, i.e.,

$$\sup_{\mu \in \mathcal{D}(\hat{\mu}, \rho)} \int_{\mathcal{S}} G(\mathbf{s}) \mu(d\mathbf{s}) = \inf_{\beta \geq 0} \left\{ \beta \rho + \int_{\mathcal{S}} \sup_{\mathbf{s} \in \mathcal{S}} (G(\mathbf{s}) - \beta c(\hat{\mathbf{s}}, \mathbf{s})) \hat{\mu}(d\hat{\mathbf{s}}) \right\}.$$

Moreover, whenever the integral is well-defined, the right-hand side is a convex optimization problem in the scalar dual variable $\beta \geq 0$, and the inner term

$$\sup_{\mathbf{s} \in \mathcal{S}} (G(\mathbf{s}) - \beta c(\hat{\mathbf{s}}, \mathbf{s}))$$

is the c -transform of G (up to sign conventions).

Empirical nominal distribution $\hat{P}$ with point-mass support

If $\hat{\mu} = \sum_{k=1}^{\hat{K}} \hat{w}_k \delta_{\hat{\mathbf{s}}_k}$, where $\delta_{\hat{\mathbf{s}}_k}$ denotes the Dirac measure, i.e., a point mass at support point $\hat{\mathbf{s}}_k$, then the dual becomes

$$\sup_{\mu \in \mathcal{D}(\hat{\mu}, \rho)} \int_{\mathcal{S}} G(\mathbf{s}) \mu(d\mathbf{s}) = \inf_{\lambda \geq 0} \left\{ \lambda \rho + \sum_{k=1}^{\hat{K}} \hat{w}_k \sup_{\mathbf{s} \in \mathcal{S}} (G(\mathbf{s}) - \lambda c(\hat{\mathbf{s}}_k, \mathbf{s})) \right\}.$$

Special case: If $c(\hat{\mathbf{s}}, \mathbf{s}) = d(\hat{\mathbf{s}}, \mathbf{s})^p$ for a metric d on $\mathcal{S}$, then $W_c(\hat{\mu}, \mu) = W_p(\hat{\mu}, \mu)^p$. The dual above remains the same given the radius ρ remains consistent with the convention adopted.

Appendix B: Omitted Proofs

B.1. Proof of Proposition 2

Let $\mathbf{m} := (\mathbf{m}_1, \mathbf{m}_2) \in \mathbb{R}^{2n}$ and define $X := Z - \mathbf{m}$. Then X is an $\mathbb{R}^{2n}$ -valued random vector with

$$\mathbb{E}[X] = 0 \quad \text{and} \quad \mathbb{E}[XX^\top] = \mathbf{Q}.$$

Using $\|\mathbf{u}\|^2 = \mathbf{u}^\top \mathbf{u}$ for $\mathbf{u} \in \mathbb{R}^{2n}$, we expand

$$\begin{aligned} \mathbb{E}[\|\mathbf{Z}\|^2] &= \mathbb{E}[\|\mathbf{m} + X\|^2] = \mathbb{E}[(\mathbf{m} + X)^\top (\mathbf{m} + X)] \\ &= \mathbb{E}[\mathbf{m}^\top \mathbf{m} + 2\mathbf{m}^\top X + X^\top X] \\ &= \mathbf{m}^\top \mathbf{m} + 2\mathbf{m}^\top \mathbb{E}[X] + \mathbb{E}[X^\top X] \\ &= \|\mathbf{m}\|^2 + \mathbb{E}[X^\top X]. \end{aligned} \tag{53}$$

Next, note that $X^\top X = \text{tr}(X^\top X) = \text{tr}(XX^\top)$, hence by linearity of $\text{tr}(\cdot)$ and $\mathbb{E}[\cdot]$,

$$\mathbb{E}[X^\top X] = \mathbb{E}[\text{tr}(XX^\top)] = \text{tr}(\mathbb{E}[XX^\top]) = \text{tr}(\mathbf{Q}).$$

Substituting this into (53) yields (6). □

B.2. Attainability of the Optimal Solution

We use $\mathcal{X} := \mathbb{R}^n \times \mathbb{R}^n$ with $\mathcal{B}(\mathcal{X})$ its Borel σ -algebra throughout this section toward establishing the attainability claim in Theorem 2. On $\mathcal{X}$, the Gaussian-mixture class we define is:

$$\mathbb{GM}_{2n} := \left\{ \nu \mid \exists \lambda \in \mathcal{P}(\mathcal{Z}) \text{ s.t. } \nu(A) = \int_{\mathcal{Z}} \mathcal{N}(\mathbf{z})(A) \lambda(d\mathbf{z}) \quad \forall A \in \mathcal{B}(\mathcal{X}), \text{ and } \int_{\mathcal{Z}} (\|\mathbf{m}(\mathbf{z})\|^2 + \text{tr}(\mathbf{Q}(\mathbf{z}))) \lambda(d\mathbf{z}) < \infty \right\}. \quad (54)$$

Lemma 10 (Moment identity for Gaussian mixtures). *Let $\lambda \in \mathcal{P}(\mathcal{Z})$ and define $\nu \in \mathcal{P}(\mathcal{X})$ by*

$$\nu(A) = \int_{\mathcal{Z}} \mathcal{N}(\mathbf{z})(A) \lambda(d\mathbf{z}) \quad \forall A \in \mathcal{B}(\mathcal{X}).$$

If $\int_{\mathcal{Z}} (\|\mathbf{m}(\mathbf{z})\|^2 + \text{tr}(\mathbf{Q}(\mathbf{z}))) \lambda(d\mathbf{z}) < \infty$, then

$$\int_{\mathcal{X}} \|\mathbf{u}\|^2 \nu(d\mathbf{u}) = \int_{\mathcal{Z}} (\|\mathbf{m}(\mathbf{z})\|^2 + \text{tr}(\mathbf{Q}(\mathbf{z}))) \lambda(d\mathbf{z}) < \infty. \quad (55)$$

Proof. Let $f_M(\mathbf{u}) := \min\{\|\mathbf{u}\|^2, M\}$ for $M \in \mathbb{N}$. Then $0 \leq f_M \uparrow \|\mathbf{u}\|^2$ pointwise. Using probability kernel Tonelli's theorem (nonnegative integrand),

$$\int_{\mathcal{X}} f_M(\mathbf{u}) \nu(d\mathbf{u}) = \int_{\mathcal{X}} f_M(\mathbf{u}) \left(\int_{\mathcal{Z}} \mathcal{N}(\mathbf{z})(d\mathbf{u}) \lambda(d\mathbf{z}) \right) = \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} f_M(\mathbf{u}) \mathcal{N}(\mathbf{z})(d\mathbf{u}) \right) \lambda(d\mathbf{z}).$$

Letting $M \rightarrow \infty$ and applying monotone convergence theorem [27, Theorem 2.14] to the inner integral (in $\mathbf{u}$) and then the outer integral (in $\mathbf{z}$) gives

$$\int_{\mathcal{X}} \|\mathbf{u}\|^2 \nu(d\mathbf{u}) = \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} \|\mathbf{u}\|^2 \mathcal{N}(\mathbf{z})(d\mathbf{u}) \right) \lambda(d\mathbf{z}).$$

By Proposition 1, $\int_{\mathcal{X}} \|\mathbf{u}\|^2 \mathcal{N}(\mathbf{z})(d\mathbf{u}) = \|\mathbf{m}(\mathbf{z})\|^2 + \text{tr}(\mathbf{Q}(\mathbf{z}))$, yielding (55). $\square$

Lemma 11 (Weak compactness of $\Pi(\mu_1, \mu_2)$). *Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^n)$. Then $\Pi(\mu_1, \mu_2)$ is weakly compact in $\mathcal{P}(\mathcal{X})$.*

Proof. (i) Tightness. Fix $\varepsilon > 0$. Choose compact sets $K_1, K_2 \subset \mathbb{R}^n$ such that $\mu_1(K_1) \geq 1 - \varepsilon/2$ and $\mu_2(K_2) \geq 1 - \varepsilon/2$ (see [10, Vol. II, Theorem 7.4.3] for the existence of such compact high-probability sets).

For any $\nu \in \Pi(\mu_1, \mu_2)$,

$$\nu((K_1 \times K_2)^c) \leq \nu(K_1^c \times \mathbb{R}^n) + \nu(\mathbb{R}^n \times K_2^c) = \mu_1(K_1^c) + \mu_2(K_2^c) \leq \varepsilon,$$

so $\Pi(\mu_1, \mu_2)$ is tight.

(ii) Closedness. Let $\nu_k \xrightarrow{\mathcal{D}} \nu$ weakly in $\mathcal{P}(\mathcal{X})$ with $\nu_k \in \Pi(\mu_1, \mu_2)$. For any bounded continuous $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}$, the map $(\mathbf{u}_1, \mathbf{u}_2) \mapsto \varphi(\mathbf{u}_1)$ is bounded continuous on $\mathcal{X}$, hence

$$\int_{\mathcal{X}} \varphi(\mathbf{u}_1) \nu(d\mathbf{u}_1, d\mathbf{u}_2) = \lim_{k \rightarrow \infty} \int_{\mathcal{X}} \varphi(\mathbf{u}_1) \nu_k(d\mathbf{u}_1, d\mathbf{u}_2) = \int_{\mathbb{R}^n} \varphi(\mathbf{u}_1) \mu_1(d\mathbf{u}_1),$$

so $\nu^{(1)} = \mu_1$. Similarly $\nu^{(2)} = \mu_2$. Thus $\Pi(\mu_1, \mu_2)$ is weakly closed.

Since $\mathcal{X}$ is Polish, Prokhorov's theorem [45, Theorem 13.29] implies that tightness yields relative weak compactness, and closedness implies its weak compactness. $\square$

Lemma 12 (Weak continuity of Gaussian laws in parameters). *Let $d \in \mathbb{N}$. If $\mathbf{m}_k \rightarrow \mathbf{m}$ in $\mathbb{R}^d$ and $\mathbf{Q}_k \rightarrow \mathbf{Q}$ in $\mathbb{S}_+^d$ (entry-wise, equivalently in any matrix norm), then*

$$\mathcal{N}(\mathbf{m}_k, \mathbf{Q}_k) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{m}, \mathbf{Q}) \quad \text{in } \mathcal{P}(\mathbb{R}^d).$$

Consequently, for every bounded continuous $f \in C_b(\mathbb{R}^d)$, the map

$$(\mathbf{m}, \mathbf{Q}) \mapsto \int_{\mathbb{R}^d} f(\mathbf{u}) \mathcal{N}(\mathbf{m}, \mathbf{Q})(d\mathbf{u})$$

is continuous on $\mathbb{R}^d \times \mathbb{S}_+^d$.

Proof. Fix $t \in \mathbb{R}^d$. The characteristic function of $\mathcal{N}(\mathbf{m}_k, \mathbf{Q}_k)$ equals

$$\varphi_k(t) = \exp\left(i t^\top \mathbf{m}_k - \frac{1}{2} t^\top \mathbf{Q}_k t\right).$$

and $\varphi(t) := \exp\left(i t^\top \mathbf{m} - \frac{1}{2} t^\top \mathbf{Q} t\right)$ is the characteristic function of $\mathcal{N}(\mathbf{m}, \mathbf{Q})$, possibly degenerate Gaussian law since $\mathbf{Q} \succeq 0$. Since $\mathbf{m}_k \rightarrow \mathbf{m}$ and $\mathbf{Q}_k \rightarrow \mathbf{Q}$, we have $\varphi_k(t) \rightarrow \varphi(t)$ for every t . By Lévy's continuity theorem [45, Theorem 15.23], $\mathcal{N}(\mathbf{m}_k, \mathbf{Q}_k) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{m}, \mathbf{Q})$. The stated continuity of $(\mathbf{m}, \mathbf{Q}) \mapsto \int f d\mathcal{N}(\mathbf{m}, \mathbf{Q})$ follows because weak convergence implies convergence of integrals against all $f \in C_b$. $\square$

Proposition 12 (Weak compactness of $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}$). *The set $\{\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}\}$ is weakly compact in $\mathcal{P}(\mathcal{X})$.*

Proof. Let $\mathcal{F} := \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}$. If $\mathcal{F} = \emptyset$, then $\mathcal{F}$ is weakly compact and there is nothing to prove. Hence, assume throughout the rest of the proof that $\mathcal{F} \neq \emptyset$.

By Lemma 11, $\Pi(\mu_1, \mu_2)$ is weakly compact. Therefore, it suffices to show that $\mathcal{F}$ is weakly closed. Now, since $\mathcal{F} \neq \emptyset$, choose $\bar{\nu} \in \mathcal{F}$. Since $\bar{\nu} \in \mathbb{GM}_{2n}$, Lemma 10 implies

$$\int_{\mathcal{X}} \|\mathbf{u}\|^2 \bar{\nu}(d\mathbf{u}) < \infty.$$

Because $\bar{\nu} \in \Pi(\mu_1, \mu_2)$, we have

$$\int_{\mathcal{X}} \|\mathbf{u}\|^2 \bar{\nu}(d\mathbf{u}) = \int_{\mathbb{R}^n} \|\mathbf{u}_1\|^2 \mu_1(d\mathbf{u}_1) + \int_{\mathbb{R}^n} \|\mathbf{u}_2\|^2 \mu_2(d\mathbf{u}_2) < \infty.$$

Therefore, $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^n)$. Consequently, every $\nu \in \Pi(\mu_1, \mu_2)$ satisfies

$$\int_{\mathcal{X}} \|\mathbf{u}\|^2 \nu(d\mathbf{u}) = \int_{\mathbb{R}^n} \|\mathbf{u}_1\|^2 \mu_1(d\mathbf{u}_1) + \int_{\mathbb{R}^n} \|\mathbf{u}_2\|^2 \mu_2(d\mathbf{u}_2) =: M < \infty.$$

Since $\mathcal{X}$ is Polish, the weak topology on $\mathcal{P}(\mathcal{X})$ is metrizable. Therefore, to prove weak closedness of $\mathcal{F}$, it suffices to prove sequential weak closedness by [10, Vol II, Chapter 8]. Let $\nu_k \in \mathcal{F}$ and suppose $\nu_k \xrightarrow{\mathcal{D}} \nu$ weakly in $\mathcal{P}(\mathcal{X})$. Because $\nu_k \in \mathcal{F} \subseteq \Pi(\mu_1, \mu_2)$ for every k , and because $\Pi(\mu_1, \mu_2)$ is weakly closed by Lemma 11, the weak limit satisfies $\nu \in \Pi(\mu_1, \mu_2)$. Therefore, it remains to prove that $\nu \in \mathbb{GM}_{2n}$.

For each k , pick $\lambda_k \in \mathcal{P}(\mathcal{Z})$ witnessing $\nu_k \in \mathbb{GM}_{2n}$, i.e.,

$$\nu_k(A) = \int_{\mathcal{Z}} \mathcal{N}(\mathbf{z})(A) \lambda_k(d\mathbf{z}) \quad \forall A \in \mathcal{B}(\mathcal{X}), \quad (56)$$

and

$$\int_{\mathcal{Z}} (\|\mathbf{m}(\mathbf{z})\|^2 + \text{tr}(\mathbf{Q}(\mathbf{z}))) \lambda_k(d\mathbf{z}) < \infty. \quad (57)$$

Moreover, since $\nu_k \in \Pi(\mu_1, \mu_2)$ and $\|(\mathbf{u}_1, \mathbf{u}_2)\|^2 = \|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2$, by the preceding derivation of $M < \infty$,

$$\int_{\mathcal{X}} \|\mathbf{u}\|^2 \nu_k(d\mathbf{u}) = \int_{\mathbb{R}^n} \|\mathbf{u}_1\|^2 \mu_1(d\mathbf{u}_1) + \int_{\mathbb{R}^n} \|\mathbf{u}_2\|^2 \mu_2(d\mathbf{u}_2) =: M < \infty. \quad (58)$$

By Lemma 10 applied to (57), we obtain

$$\int_{\mathcal{Z}} (\|\mathbf{m}(\mathbf{z})\|^2 + \text{tr}(\mathbf{Q}(\mathbf{z}))) \lambda_k(d\mathbf{z}) = \int_{\mathcal{X}} \|\mathbf{u}\|^2 \nu_k(d\mathbf{u}) \leq M \quad \forall k. \quad (59)$$

Step 1: Tightness of $\{\lambda_k\}$. Let $g(\mathbf{z}) := \|\mathbf{m}(\mathbf{z})\|^2 + \text{tr}(\mathbf{Q}(\mathbf{z}))$. For $R > 0$, the sublevel set $K_R := \{\mathbf{z} \in \mathcal{Z} : g(\mathbf{z}) \leq R\}$ is compact in $\mathcal{Z}$. K_R is closed because $\mathcal{Z} = \mathbb{R}^{2n} \times \mathbb{S}_+^{2n}$ is closed in the finite-dimensional Euclidean space $\mathbb{R}^{2n} \times \mathbb{S}^{2n}$, and g is continuous on $\mathcal{Z}$. Moreover, if $\mathbf{z} = (\mathbf{m}, \mathbf{Q}) \in K_R$, then $\|\mathbf{m}\|^2 \leq R$ and $\text{tr}(\mathbf{Q}) \leq R$. Since $\mathbf{Q} \succeq 0$, its eigenvalues are nonnegative, and hence

$$\|\mathbf{Q}\|_F = \left(\sum_i \lambda_i(\mathbf{Q})^2 \right)^{1/2} \leq \sum_i \lambda_i(\mathbf{Q}) = \text{tr}(\mathbf{Q}) \leq R.$$

Thus K_R is bounded. Hence, K_R is closed and bounded in a finite-dimensional Euclidean space, and therefore compact by the Heine–Borel theorem [72]. By Markov's inequality and (59),

$$\lambda_k(K_R^c) = \lambda_k(\{g > R\}) \leq \frac{1}{R} \int_{\mathcal{Z}} g(\mathbf{z}) \lambda_k(d\mathbf{z}) \leq \frac{M}{R},$$

uniformly in k . Given $\varepsilon > 0$, choosing $R > M/\varepsilon$ yields $\lambda_k(K_R) \geq 1 - \varepsilon$ for all k . Since K_R is compact, $\{\lambda_k\}$ is tight on $\mathcal{Z}$. By Prokhorov's theorem [45, Theorem 13.29], there exists a subsequence (not relabeled) and $\lambda \in \mathcal{P}(\mathcal{Z})$ such that

$$\lambda_k \xrightarrow{\mathcal{D}} \lambda \quad \text{weakly in } \mathcal{P}(\mathcal{Z}).$$

Step 2: Identify the weak limit ν as a Gaussian mixture. Fix any bounded continuous $f \in C_b(\mathcal{X})$. Let

$$\Psi_f : \mathcal{Z} \rightarrow \mathbb{R}, \quad \Psi_f(\mathbf{z}) := \int_{\mathcal{X}} f(\mathbf{u}) \mathcal{N}(\mathbf{z})(d\mathbf{u}).$$

By Lemma 12 applied with $d = 2n$, if $\mathbf{z}_k \rightarrow \mathbf{z}$ in $\mathcal{Z}$, then

$$\mathcal{N}(\mathbf{z}_k) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{z}).$$

Therefore,

$$\Psi_f(\mathbf{z}_k) = \int_{\mathcal{X}} f(\mathbf{u}) \mathcal{N}(\mathbf{z}_k)(d\mathbf{u}) \longrightarrow \int_{\mathcal{X}} f(\mathbf{u}) \mathcal{N}(\mathbf{z})(d\mathbf{u}) = \Psi_f(\mathbf{z}).$$

Hence, Ψ_f is continuous on $\mathcal{Z}$. Moreover, since

$$|\Psi_f(\mathbf{z})| \leq \|f\|_{\infty} \quad \forall \mathbf{z} \in \mathcal{Z},$$

we have $\Psi_f \in C_b(\mathcal{Z})$.

Hence, using (56) and the kernel Fubini theorem for bounded measurable functions,

$$\int_{\mathcal{X}} f(\mathbf{u}) \nu_k(d\mathbf{u}) = \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} f(\mathbf{u}) \mathcal{N}(\mathbf{z})(d\mathbf{u}) \right) \lambda_k(d\mathbf{z}) = \int_{\mathcal{Z}} \Psi_f(\mathbf{z}) \lambda_k(d\mathbf{z}).$$

Since $\lambda_k \xrightarrow{\mathcal{D}} \lambda$ and $\Psi_f \in C_b(\mathcal{Z})$, we have

$$\int_{\mathcal{Z}} \Psi_f(\mathbf{z}) \lambda_k(d\mathbf{z}) \longrightarrow \int_{\mathcal{Z}} \Psi_f(\mathbf{z}) \lambda(d\mathbf{z}).$$

On the other hand, since $\nu_k \xrightarrow{\mathcal{D}} \nu$ and $f \in C_b(\mathcal{X})$, by [10, Corollary 8.2.10]

$$\int_{\mathcal{X}} f(\mathbf{u}) \nu_k(d\mathbf{u}) \longrightarrow \int_{\mathcal{X}} f(\mathbf{u}) \nu(d\mathbf{u}).$$

Therefore,

$$\int_{\mathcal{X}} f(\mathbf{u}) \nu(d\mathbf{u}) = \int_{\mathcal{Z}} \Psi_f(\mathbf{z}) \lambda(d\mathbf{z}) \quad \forall f \in C_b(\mathcal{X}). \quad (60)$$

Now define

$$\tilde{\nu}(A) := \int_{\mathcal{Z}} \mathcal{N}(\mathbf{z})(A) \lambda(d\mathbf{z}), \quad A \in \mathcal{B}(\mathcal{X}).$$

Because $z \mapsto \mathcal{N}(z)(\cdot)$ is a probability kernel and $\lambda \in \mathcal{P}(\mathcal{Z})$, the set function $\tilde{\nu}$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, i.e., $\tilde{\nu} \in \mathcal{P}(\mathcal{X})$. Moreover, by the kernel Fubini theorem again, for every $f \in C_b(\mathcal{X})$,

$$\int_{\mathcal{X}} f(\mathbf{u}) \tilde{\nu}(d\mathbf{u}) = \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} f(\mathbf{u}) \mathcal{N}(z)(d\mathbf{u}) \right) \lambda(dz) = \int_{\mathcal{Z}} \Psi_f(z) \lambda(dz).$$

Combining this identity with (60) yields

$$\int_{\mathcal{X}} f(\mathbf{u}) \nu(d\mathbf{u}) = \int_{\mathcal{X}} f(\mathbf{u}) \tilde{\nu}(d\mathbf{u}) \quad \forall f \in C_b(\mathcal{X}). \quad (61)$$

Let

$$\bar{C}_b(\mathcal{X}) := \{f : \mathcal{X} \rightarrow \mathbb{R} : f \text{ is bounded and uniformly continuous}\}.$$

Since $\bar{C}_b(\mathcal{X}) \subseteq C_b(\mathcal{X})$, (61) implies

$$\int_{\mathcal{X}} f(\mathbf{u}) \nu(d\mathbf{u}) = \int_{\mathcal{X}} f(\mathbf{u}) \tilde{\nu}(d\mathbf{u}) \quad \forall f \in \bar{C}_b(\mathcal{X}).$$

Because $\mathcal{X} = \mathbb{R}^n \times \mathbb{R}^n$ is a metric space, bounded uniformly continuous functions determine probability measures on $\mathcal{X}$ [10, Vol II, Chapter 8]. Therefore, $\nu = \tilde{\nu}$. Equivalently,

$$\nu(A) = \int_{\mathcal{Z}} \mathcal{N}(z)(A) \lambda(dz) \quad \forall A \in \mathcal{B}(\mathcal{X}). \quad (62)$$

Step 3: verify the moment condition for λ . Let $g(z) := \|\mathbf{m}(z)\|^2 + \text{tr}(\mathbf{Q}(z))$. Since g is lower semicontinuous on $\mathcal{Z}$ and $\lambda_k \xrightarrow{\mathcal{D}} \lambda$, Portmanteau theorem [45, Theorem 13.16] yields

$$\int_{\mathcal{Z}} g(z) \lambda(dz) \leq \liminf_{k \rightarrow \infty} \int_{\mathcal{Z}} g(z) \lambda_k(dz) \leq M < \infty.$$

Together with (62), this shows $\nu \in \mathbb{GM}_{2n}$. Hence, $\nu \in \mathcal{F}$, so $\mathcal{F}$ is weakly closed.

Therefore, $\mathcal{F}$ is a weakly closed subset of the weakly compact set $\Pi(\mu_1, \mu_2)$, and is thus weakly compact. $\square$

Corollary 2 (Attainment of solution in Theorem 2). *Let $c : \mathcal{X} \rightarrow (-\infty, \infty]$ be lower semicontinuous and bounded below. Then the optimization problem*

$$\inf_{\nu \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}} \int_{\mathcal{X}} c(\mathbf{u}) \nu(d\mathbf{u})$$

admits a minimizer in $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}$. In particular, this holds for $c(\mathbf{u}_1, \mathbf{u}_2) = \|\mathbf{u}_1 - \mathbf{u}_2\|^2$.

Proof. By Proposition 12, the feasible set $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}$ is weakly compact. The feasible set is nonempty because the product coupling $\nu := \mu_1 \otimes \mu_2 \in \Pi(\mu_1, \mu_2)$ belongs to $\mathbb{GM}_{2n}$ through the point-mass Gaussian representation allowed by $\mathbb{S}_+^{2n}$. Additionally, since c is lower semicontinuous and bounded below, the map $\nu \mapsto \int_{\mathcal{X}} c d\nu$ is weakly lower semicontinuous (cf. the direct-method argument in [82, Theorem 4.1]). Hence, the infimum is attained over the nonempty weakly compact feasible set. $\square$

B.3. Second Moment Preservation

Lemma 13 (Geometric-mean domination and a Frobenius bound). *Let $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathbb{S}_+^n$ and let Σ be defined by (9). Then*

$$\Sigma \preceq \frac{\mathbf{Q}_1 + \mathbf{Q}_2}{2}. \quad (63)$$

Consequently,

$$\|\Sigma\|_F \leq \frac{\|\mathbf{Q}_1\|_F + \|\mathbf{Q}_2\|_F}{2}, \quad \|\bar{\mathbf{Q}}\|_F^2 \leq 2(\|\mathbf{Q}_1\|_F^2 + \|\mathbf{Q}_2\|_F^2), \quad (64)$$

where $\bar{\mathbf{Q}} = \begin{pmatrix} \mathbf{Q}_1 & \Sigma \\ \Sigma & \mathbf{Q}_2 \end{pmatrix}$.

Proof. Inequality (63) is the standard operator inequality for the (Kubo–Ando) [2, 49] matrix geometric mean:

$$\mathbf{Q}_1 \odot \mathbf{Q}_2 \preceq \frac{\mathbf{Q}_1 + \mathbf{Q}_2}{2}.$$

Since (9) is precisely the geometric mean $\mathbf{Q}_1 \odot \mathbf{Q}_2$, we obtain (63).

Next, for PSD matrices, any unitarily invariant norm is monotone with respect to the Loewner order: if $0 \preceq A \preceq B$, then $\|A\| \leq \|B\|$ (see, e.g., [7, 35]). Applying this with the Frobenius norm and (63) gives

$$\|\Sigma\|_F \leq \left\| \frac{\mathbf{Q}_1 + \mathbf{Q}_2}{2} \right\|_F \leq \frac{\|\mathbf{Q}_1\|_F + \|\mathbf{Q}_2\|_F}{2},$$

using the triangle inequality.

Finally, since $\bar{\mathbf{Q}}$ is block-symmetric,

$$\|\bar{\mathbf{Q}}\|_F^2 = \|\mathbf{Q}_1\|_F^2 + \|\mathbf{Q}_2\|_F^2 + 2\|\Sigma\|_F^2.$$

With $a := \|\mathbf{Q}_1\|_F$ and $b := \|\mathbf{Q}_2\|_F$, the bound above implies

$$2\|\Sigma\|_F^2 \leq 2 \left(\frac{a+b}{2} \right)^2 = \frac{(a+b)^2}{2} \leq a^2 + b^2,$$

where we used $(a+b)^2 \leq 2(a^2 + b^2)$. Hence, $\|\bar{\mathbf{Q}}\|_F^2 \leq 2(\|\mathbf{Q}_1\|_F^2 + \|\mathbf{Q}_2\|_F^2)$, proving (64). $\square$

B.3.1. Proof of Proposition 2. Let

$$f(\mathbf{z}) := \|\mathbf{m}_1(\mathbf{z})\|^2 + \|\mathbf{m}_2(\mathbf{z})\|^2 + \|\bar{\mathbf{Q}}(\mathbf{z})\|_F^2.$$

Similar to (12), by the pushforward (change-of-variables) identity for nonnegative measurable map f ,

$$\int_{\mathcal{Z}} f(\mathbf{z}) d\lambda(\mathbf{z}) = \int_{\mathcal{Z}} f(\mathbf{z}) d(\mathcal{T}_{\#}\pi)(\mathbf{z}) = \int_{\mathcal{S} \times \mathcal{S}} f(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2), \quad (65)$$

where the first term is the one we have in the claim (14).

Now, for $(\mathbf{s}_1, \mathbf{s}_2) = ((\mathbf{m}_1, \mathbf{Q}_1), (\mathbf{m}_2, \mathbf{Q}_2))$ we have $f(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) = \|\mathbf{m}_1(\mathbf{s}_1)\|^2 + \|\mathbf{m}_2(\mathbf{s}_2)\|^2 + \|\bar{\mathbf{Q}}(\mathbf{s}_1, \mathbf{s}_2)\|_F^2$. By Lemma 13 in Appendix B (which uses the identity (9) and the Kubo–Ando matrix geometric mean [2, 49]), $\|\bar{\mathbf{Q}}(\mathbf{s}_1, \mathbf{s}_2)\|_F^2 \leq 2(\|\mathbf{Q}_1(\mathbf{s}_1)\|_F^2 + \|\mathbf{Q}_2(\mathbf{s}_2)\|_F^2)$. Therefore,

$$f(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) \leq \|\mathbf{m}_1(\mathbf{s}_1)\|^2 + \|\mathbf{m}_2(\mathbf{s}_2)\|^2 + 2\|\mathbf{Q}_1(\mathbf{s}_1)\|_F^2 + 2\|\mathbf{Q}_2(\mathbf{s}_2)\|_F^2.$$

Integrating both sides with respect to π yields

$$\int_{\mathcal{S} \times \mathcal{S}} f(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2) \leq \int_{\mathcal{S} \times \mathcal{S}} \left(\|\mathbf{m}_1(\mathbf{s}_1)\|^2 + \|\mathbf{m}_2(\mathbf{s}_2)\|^2 + 2\|\mathbf{Q}_1(\mathbf{s}_1)\|_F^2 + 2\|\mathbf{Q}_2(\mathbf{s}_2)\|_F^2 \right) d\pi(\mathbf{s}_1, \mathbf{s}_2).$$

Combining this with (65) gives us:

$$\int_{\mathcal{Z}} f(\mathbf{z}) d\lambda(\mathbf{z}) \leq \int_{\mathcal{S} \times \mathcal{S}} \left(\|\mathbf{m}_1(\mathbf{s}_1)\|^2 + \|\mathbf{m}_2(\mathbf{s}_2)\|^2 + 2\|\mathbf{Q}_1(\mathbf{s}_1)\|_F^2 + 2\|\mathbf{Q}_2(\mathbf{s}_2)\|_F^2 \right) d\pi(\mathbf{s}_1, \mathbf{s}_2).$$

The right-hand side is finite by assumption (13). Hence (14) holds. $\square$

B.3.2. Proof of Proposition 3. Let $f(z) := \|\mathbf{m}_1(z)\|^2 + \|\mathbf{m}_2(z)\|^2 + \|\mathbf{Q}(z)\|_F^2$, $g(s_1, s_2) := \|\mathbf{m}_1(s_1)\|^2 + \|\mathbf{m}_2(s_2)\|^2 + \|\mathbf{Q}_{11}(s_1)\|_F^2 + \|\mathbf{Q}_{22}(s_2)\|_F^2$. Since the measurable map g is nonnegative, similar to the pushforward identity (17),

$$\int_{S \times S} g(s_1, s_2) d\pi(s_1, s_2) = \int_{S \times S} g(s_1, s_2) d(\mathcal{M}_\# \lambda)(s_1, s_2) = \int_{\mathcal{Z}} g(\mathcal{M}(z)) d\lambda(z), \quad (66)$$

where the left-most term is exactly the one desired in our claim (19). Now fix $z = ((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q}) \in \mathcal{Z}$ and write $\mathbf{Q}$ in blocks as defined in (15). Then,

$$g(\mathcal{M}(z)) = \|\mathbf{m}_1(z)\|^2 + \|\mathbf{m}_2(z)\|^2 + \|\mathbf{Q}_{11}(z)\|_F^2 + \|\mathbf{Q}_{22}(z)\|_F^2.$$

Additionally, the following inequality holds pointwise since by Frobenius norm definition $\|\mathbf{Q}\|_F^2 = \|\mathbf{Q}_{11}\|_F^2 + \|\mathbf{Q}_{22}\|_F^2 + \|\mathbf{Q}_{12}\|_F^2 + \|\mathbf{Q}_{21}\|_F^2$,

$$\begin{aligned} & \|\mathbf{Q}_{11}(z)\|_F^2 + \|\mathbf{Q}_{22}(z)\|_F^2 \leq \|\mathbf{Q}(z)\|_F^2 \\ \implies & \|\mathbf{m}_1(z)\|^2 + \|\mathbf{m}_2(z)\|^2 + \|\mathbf{Q}_{11}(z)\|_F^2 + \|\mathbf{Q}_{22}(z)\|_F^2 \leq \|\mathbf{Q}(z)\|_F^2 + \|\mathbf{m}_1(z)\|^2 + \|\mathbf{m}_2(z)\|^2. \end{aligned}$$

Hence, for all $z \in \mathcal{Z}$, $g(\mathcal{M}(z)) \leq f(z)$. Integrating both sides with respect to λ ,

$$\int_{\mathcal{Z}} g(\mathcal{M}(z)) d\lambda(z) \leq \int_{\mathcal{Z}} f(z) d\lambda(z).$$

Then, combining the above inequality with (66) we obtain

$$\int_{S \times S} g(s_1, s_2) d\pi(s_1, s_2) \leq \int_{\mathcal{Z}} f(z) d\lambda(z).$$

The right-hand side is finite by assumption (18), thus, (19) holds. $\square$

B.4. Proof of Wasserstein-Metric Equivalence with Restricted Parameter Support Set

Throughout this subsection, we assume $\mathcal{X} := \mathbb{R}^n \times \mathbb{R}^n$ and a compact support set $\mathcal{S} := \Theta \times \Xi \subset \mathbb{R}^n \times \mathbb{S}_{++}^n$.

For a joint covariance matrix

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix} \in \mathbb{S}_+^{2n}, \quad (67)$$

define the restricted joint-parameter set

$$\mathcal{Z}_{\mathcal{S}} := \{z = ((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{S}_+^{2n} : (\mathbf{m}_1, \mathbf{Q}_{11}) \in \mathcal{S}, (\mathbf{m}_2, \mathbf{Q}_{22}) \in \mathcal{S}\}.$$

We define the restricted $2n$ -variate Gaussian-mixture class by

$$\mathbb{GM}_{2n}^{\mathcal{S}} := \left\{ \nu \in \mathcal{P}(\mathbb{R}^{2n}) : \nu(A) = \int_{\mathcal{Z}_{\mathcal{S}}} \mathcal{N}(z)(A) \lambda(dz) \forall A \in \mathcal{B}(\mathbb{R}^{2n}) \text{ for some } \lambda \in \mathcal{P}(\mathcal{Z}_{\mathcal{S}}) \right\}.$$

Remark 8. We note that $\mathcal{Z}_S$ is not the full set of all joint PSD covariances; it is the set of valid joint Gaussian parameters whose two marginal Gaussian parameters lie in $\mathcal{S}$. Thus, for $\mathbf{z} = ((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q}) \in \mathcal{Z}_S$, the full joint covariance $\mathbf{Q}$ is required only to satisfy $\mathbf{Q} \succeq 0$, while its diagonal blocks satisfy

$$\mathbf{Q}_{11}, \mathbf{Q}_{22} \in \Xi \subseteq \mathbb{S}_\delta^n.$$

The joint covariance may therefore be singular even though the marginal covariances are uniformly positive definite. For example, if $(\mathbf{m}, \mathbf{Q}_0) \in \mathcal{S}$, then

$$\begin{pmatrix} \mathbf{Q}_0 & \mathbf{Q}_0 \\ \mathbf{Q}_0 & \mathbf{Q}_0 \end{pmatrix} \succeq 0$$

is singular, while both diagonal blocks equal $\mathbf{Q}_0 \in \Xi$. Hence

$$((\mathbf{m}, \mathbf{m}), \begin{pmatrix} \mathbf{Q}_0 & \mathbf{Q}_0 \\ \mathbf{Q}_0 & \mathbf{Q}_0 \end{pmatrix}) \in \mathcal{Z}_S.$$

On the other hand, $\mathcal{Z}_S$ also contains the independent Gaussian couplings. Indeed, if $\mathbf{s}_i = (\mathbf{m}_i, \mathbf{Q}_i) \in \mathcal{S}$, $i = 1, 2$, then

$$\mathbf{Q}^{\text{ind}} := \begin{pmatrix} \mathbf{Q}_1 & 0 \\ 0 & \mathbf{Q}_2 \end{pmatrix}$$

satisfies $\mathbf{Q}^{\text{ind}} \succeq 0$, and

$$((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q}^{\text{ind}}) \in \mathcal{Z}_S.$$

Thus $\mathcal{Z}_S$ contains at least the independent Gaussian couplings of pairs of Gaussian components in $\mathcal{S}$.

Lemma 14 (Compactness of the restricted joint-parameter set). *Under Assumption 1, the set $\mathcal{Z}_S$ is compact.*

Proof. Since $\mathcal{S}$ is compact, the sets of admissible marginal means and diagonal covariance blocks are bounded. Thus $\mathbf{m}_1, \mathbf{m}_2, \mathbf{Q}_{11}$, and $\mathbf{Q}_{22}$ are bounded over $\mathcal{Z}_S$. Moreover, if $\mathbf{Q} \succeq 0$, then every entry of the off-diagonal block satisfies

$$|(\mathbf{Q}_{12})_{ij}|^2 \leq (\mathbf{Q}_{11})_{ii}(\mathbf{Q}_{22})_{jj}.$$

Hence, $\mathbf{Q}_{12}$ and $\mathbf{Q}_{21}$ are also bounded. Therefore, $\mathcal{Z}_S$ is bounded.

It remains to note that $\mathcal{Z}_S$ is closed. Indeed, $\mathbb{S}_+^{2n}$ is closed, and the conditions $(\mathbf{m}_1, \mathbf{Q}_{11}) \in \mathcal{S}$ and $(\mathbf{m}_2, \mathbf{Q}_{22}) \in \mathcal{S}$ are closed because $\mathcal{S}$ is compact, hence closed. Thus $\mathcal{Z}_S$ is closed and bounded in a finite-dimensional Euclidean space. By Heine–Borel [72], $\mathcal{Z}_S$ is compact. $\square$

Lemma 15 (Nonemptiness under restricted Gaussian-mixture support). *Suppose that $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^n)$ admit mixing laws over $\mathcal{S}$; that is, there exist $\eta_1, \eta_2 \in \mathcal{P}(\mathcal{S})$ such that, for every $C \in \mathcal{B}(\mathbb{R}^n)$,*

$$\mu_1(C) = \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(C) \eta_1(d\mathbf{s}), \quad \mu_2(C) = \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(C) \eta_2(d\mathbf{s}).$$

Then $\Pi(\mu_1, \mu_2) \cap \text{GM}_{2n}^{\mathcal{S}} \neq \emptyset$.

Proof. Let $\pi := \eta_1 \otimes \eta_2 \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$.

For $\mathbf{s}_i = (\mathbf{m}_i, \mathbf{Q}_i) \in \mathcal{S}$, $i = 1, 2$, let $\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)$ denote the Gelbrich coupling between $\mathcal{N}(\mathbf{s}_1)$ and $\mathcal{N}(\mathbf{s}_2)$.

Thus

$$\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2) = \mathcal{N}((\mathbf{m}_1, \mathbf{m}_2), \overline{\mathbf{Q}}(\mathbf{s}_1, \mathbf{s}_2)),$$

$$\overline{\mathbf{Q}}(\mathbf{s}_1, \mathbf{s}_2) = \begin{pmatrix} \mathbf{Q}_1 & \Sigma(\mathbf{Q}_1, \mathbf{Q}_2) \\ \Sigma(\mathbf{Q}_1, \mathbf{Q}_2)^\top & \mathbf{Q}_2 \end{pmatrix} \in \mathbb{S}_+^{2n}.$$

where

Since the diagonal covariance blocks of $\bar{\mathbf{Q}}(\mathbf{s}_1, \mathbf{s}_2)$ are $\mathbf{Q}_1$ and $\mathbf{Q}_2$, the map $\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2) := ((\mathbf{m}_1, \mathbf{m}_2), \bar{\mathbf{Q}}(\mathbf{s}_1, \mathbf{s}_2))$ satisfies $\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{Z}_S \quad \forall (\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{S} \times \mathcal{S}$. Hence, $\lambda := \mathcal{T}_\# \pi \in \mathcal{P}(\mathcal{Z}_S)$.

Define $\nu \in \mathcal{P}(\mathbb{R}^{2n})$ by

$$\nu(A) := \int_{\mathcal{Z}_S} \mathcal{N}(\mathbf{z})(A) \lambda(d\mathbf{z}), \quad A \in \mathcal{B}(\mathbb{R}^{2n}).$$

Equivalently,

$$\nu(A) = \int_{\mathcal{S} \times \mathcal{S}} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(A) \pi(ds_1, ds_2).$$

Therefore, $\nu \in \mathbb{GM}_{2n}^S$.

It remains to show that $\nu \in \Pi(\mu_1, \mu_2)$. Let $B \in \mathcal{B}(\mathbb{R}^n)$. Since $\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)$ is a coupling of $\mathcal{N}(\mathbf{s}_1)$ and $\mathcal{N}(\mathbf{s}_2)$,

$$\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(B \times \mathbb{R}^n) = \mathcal{N}(\mathbf{s}_1)(B).$$

Thus $\nu(B \times \mathbb{R}^n) = \int_{\mathcal{S} \times \mathcal{S}} \mathcal{N}(\mathbf{s}_1)(B) (\eta_1 \otimes \eta_2)(ds_1, ds_2) = \int_{\mathcal{S}} \mathcal{N}(\mathbf{s}_1)(B) \eta_1(ds_1) = \mu_1(B)$.

Hence, $\nu^{(1)} = \mu_1$. The same argument gives

$$\nu(\mathbb{R}^n \times B) = \int_{\mathcal{S}} \mathcal{N}(\mathbf{s}_2)(B) \eta_2(ds_2) = \mu_2(B),$$

so $\nu^{(2)} = \mu_2$. Therefore, $\nu \in \Pi(\mu_1, \mu_2)$.

Combining $\nu \in \Pi(\mu_1, \mu_2)$ and $\nu \in \mathbb{GM}_{2n}^S$, we obtain

$$\nu \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S.$$

Hence, the intersection is nonempty. $\square$

Corollary 3 (Weak compactness under restricted support). *Under Assumption 1, $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S$ is weakly compact.*

Proof. The set $\mathcal{Z}_S$ is compact. Indeed, $\mathcal{S}$ is compact, so the diagonal blocks $\mathbf{Q}_{11}$ and $\mathbf{Q}_{22}$, as well as the means $\mathbf{m}_1, \mathbf{m}_2$, are bounded. Moreover, for any $\mathbf{Q} \succeq 0$,

$$|(\mathbf{Q}_{12})_{ij}|^2 \leq (\mathbf{Q}_{11})_{ii} (\mathbf{Q}_{22})_{jj},$$

so the off-diagonal block $\mathbf{Q}_{12}$ is also bounded. The defining conditions of $\mathcal{Z}_S$ are closed, because $\mathcal{S}$ and $\mathbb{S}_+^{2n}$ are closed. Hence, $\mathcal{Z}_S$ is closed and bounded in a finite-dimensional Euclidean space and therefore compact.

Now the proof of Proposition 12 applies with $\mathcal{Z}$ replaced by $\mathcal{Z}_S$. In fact, the tightness of the mixing laws is immediate because every λ_k is supported on the compact set $\mathcal{Z}_S$. The weak limit of a subsequence of λ_k 's remains in $\mathcal{P}(\mathcal{Z}_S)$, and the weak continuity of Gaussian laws again identifies the weak limit as a Gaussian mixture supported on $\mathcal{Z}_S$. Hence, $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S$ is weakly closed inside the weakly compact set $\Pi(\mu_1, \mu_2)$, and is therefore weakly compact. $\square$

Corollary 4 (Attainment of solution in Theorem 3). *Assume the setting of Lemma 15. Let $c : \mathcal{X} \rightarrow (-\infty, \infty]$ be lower semicontinuous and bounded below. Then*

$$\inf_{\nu \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S} \int_{\mathcal{X}} c(\mathbf{u}) \nu(d\mathbf{u})$$

admits a minimizer in $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S$. In particular, this holds for $c(\mathbf{u}_1, \mathbf{u}_2) = \|\mathbf{u}_1 - \mathbf{u}_2\|^2$.

Proof. By Corollary 3, the feasible set $\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S$ is weakly compact. By Lemma 15, it is nonempty. Since c is lower semicontinuous and bounded below, the map $\nu \mapsto \int_{\mathcal{X}} c(\mathbf{u}) \nu(d\mathbf{u})$ is weakly lower semicontinuous. Therefore, the infimum is attained over the nonempty weakly compact feasible set. $\square$

B.4.1. Moment preservation under restricted parameter support. In the restricted-support setting, the two parameter maps $\mathcal{T}$ and $\mathcal{M}$ from the full-support case will be used with restricted domains. First, for $\mathbf{s}_i = (\mathbf{m}_i, \mathbf{Q}_i) \in \mathcal{S}$, the Gelbrich joint covariance has diagonal blocks $\mathbf{Q}_1$ and $\mathbf{Q}_2$. Therefore,

$$\mathcal{T}(\mathcal{S} \times \mathcal{S}) \subseteq \mathcal{Z}_{\mathcal{S}},$$

and the forward map is $\mathcal{T} : \mathcal{S} \times \mathcal{S} \rightarrow \mathcal{Z}_{\mathcal{S}}$. Second, for $\mathbf{z} = ((\mathbf{m}_1, \mathbf{m}_2), \mathbf{Q}) \in \mathcal{Z}_{\mathcal{S}}$ with $\mathbf{Q}$ as denoted in (67), define $\mathcal{M}(\mathbf{z}) := ((\mathbf{m}_1, \mathbf{Q}_{11}), (\mathbf{m}_2, \mathbf{Q}_{22}))$. By the definition of $\mathcal{Z}_{\mathcal{S}}$, this is well-defined as $\mathcal{M} : \mathcal{Z}_{\mathcal{S}} \rightarrow \mathcal{S} \times \mathcal{S}$ and

$$\mathcal{M}(\mathcal{Z}_{\mathcal{S}}) \subseteq \mathcal{S} \times \mathcal{S}.$$

Because $\mathcal{S}$ is compact in the finite-dimensional ambient space $\mathbb{R}^n \times \mathbb{S}^n$, the product set $\mathcal{S} \times \mathcal{S}$ is compact and hence bounded. Therefore, under the ambient Euclidean–Frobenius gauge, every $\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$ satisfies

$$\int_{\mathcal{S} \times \mathcal{S}} (\|\mathbf{m}_1\|^2 + \|\mathbf{Q}_1\|_F^2 + \|\mathbf{m}_2\|^2 + \|\mathbf{Q}_2\|_F^2) d\pi(s_1, s_2) < \infty.$$

Similarly, by Lemma 14, $\mathcal{Z}_{\mathcal{S}}$ is compact. Hence, every $\lambda \in \mathcal{P}(\mathcal{Z}_{\mathcal{S}})$ has a finite second moment with respect to the ambient gauge on $\mathcal{Z}_{\mathcal{S}}$. We formally state the claim below.

Corollary 5 (Restricted moment preservation). *Under Assumption 1, the moment-preservation statements of Propositions 2 and 3 remain valid after replacing $\mathcal{Z}$ by $\mathcal{Z}_{\mathcal{S}}$. In particular,*

$$\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S}) \implies \mathcal{T}_{\#}\pi \in \mathcal{P}(\mathcal{Z}_{\mathcal{S}}),$$

and

$$\lambda \in \mathcal{P}(\mathcal{Z}_{\mathcal{S}}) \implies \mathcal{M}_{\#}\lambda \in \mathcal{P}(\mathcal{S} \times \mathcal{S}).$$

Moreover, all parameter second moments appearing in Propositions 2 and 3 are finite.

Proof. The two push-forward inclusions follow from the restricted domains of $\mathcal{T}$ and $\mathcal{M}$:

$$\mathcal{T}(\mathcal{S} \times \mathcal{S}) \subseteq \mathcal{Z}_{\mathcal{S}}, \quad \mathcal{M}(\mathcal{Z}_{\mathcal{S}}) \subseteq \mathcal{S} \times \mathcal{S}.$$

By Assumption 1, $\mathcal{S}$ is compact. By Lemma 14, $\mathcal{Z}_{\mathcal{S}}$ is compact. Therefore, the parameter gauges used in Propositions 2 and 3 are bounded on the relevant supports. Hence, the corresponding moment integrals are finite for every probability measure supported on these compact sets. $\square$

B.4.2. Containment under restricted support

For clarity, let

$$\mathcal{S}_{\text{full}} := \mathbb{R}^n \times \mathbb{S}_+^n, \quad \mathcal{Z}_{\text{full}} := \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{S}_+^{2n}.$$

Then $\mathcal{S} \subseteq \mathcal{S}_{\text{full}}$ and $\mathcal{Z}_{\mathcal{S}} \subseteq \mathcal{Z}_{\text{full}}$.

Define

$$\Gamma_{\mathcal{S}} := \left\{ \pi \in \mathcal{P}_2(\mathcal{S} \times \mathcal{S}) \mid \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(A) d\pi_1(\mathbf{s}) = \mu_1(A), \quad \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(B) d\pi_2(\mathbf{s}) = \mu_2(B), \quad \forall A, B \in \mathcal{B}(\mathbb{R}^n) \right\}.$$

We keep the notation $\pi \in \mathcal{P}_2(\mathcal{S} \times \mathcal{S})$ only to remain consistent with (23).

Corollary 6. For every $\pi \in \Gamma_{\mathcal{S}}$, define

$$\nu_{\pi}(A) := \int_{\mathcal{S} \times \mathcal{S}} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(A) d\pi(\mathbf{s}_1, \mathbf{s}_2), \quad A \in \mathcal{B}(\mathbb{R}^{2n}).$$

Then $\nu_{\pi} \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^{\mathcal{S}}$.

Proof. Fix $\pi \in \Gamma_{\mathcal{S}}$. Since $\pi \in \Gamma_{\mathcal{S}}$, its marginals satisfy

$$\int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(A) d\pi_1(\mathbf{s}) = \mu_1(A), \quad \int_{\mathcal{S}} \mathcal{N}(\mathbf{s})(B) d\pi_2(\mathbf{s}) = \mu_2(B), \quad \forall A, B \in \mathcal{B}(\mathbb{R}^n)$$

Since π_1 and π_2 are supported on $\mathcal{S}$, these are the same marginal mixing-law identities used in Lemma 1 viewing π as a probability measure on $\mathcal{S}_{\text{full}} \times \mathcal{S}_{\text{full}}$ supported on $\mathcal{S} \times \mathcal{S}$. Moreover, since $\mathcal{S}$ is compact, the second moment of π is finite. Therefore Lemma 1 applies and gives $\nu_{\pi} \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}$.

It remains to check that the Gaussian-mixture representation of ν_{π} is supported on $\mathcal{Z}_{\mathcal{S}}$. Let $\lambda := \mathcal{T}_{\#}\pi$. Since $\mathcal{T}(\mathcal{S} \times \mathcal{S}) \subseteq \mathcal{Z}_{\mathcal{S}}$, $\lambda \in \mathcal{P}(\mathcal{Z}_{\mathcal{S}})$. Now fix any $A \in \mathcal{B}(\mathbb{R}^{2n})$, and define

$$\Phi_A(\mathbf{z}) := \mathcal{N}(\mathbf{z})(A).$$

The map Φ_A is measurable because $\mathbf{z} \mapsto \mathcal{N}(\mathbf{z})(\cdot)$ is a Gaussian transition kernel. Additionally, since $\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)$ is the joint Gaussian parameter corresponding to the Gelbrich coupling $\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)$, we have

$$\Phi_A(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) = \mathcal{N}(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2))(A) = \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(A).$$

Since $\lambda = \mathcal{T}_{\#}\pi$, the pushforward identity gives

$$\int_{\mathcal{S} \times \mathcal{S}} \Phi_A(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2) = \int_{\mathcal{Z}_{\mathcal{S}}} \Phi_A(\mathbf{z}) d\lambda(\mathbf{z}).$$

Then for $A \in \mathcal{B}(\mathbb{R}^{2n})$, we obtain

$$\begin{aligned} \nu_{\pi}(A) &= \int_{\mathcal{S} \times \mathcal{S}} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(A) d\pi(\mathbf{s}_1, \mathbf{s}_2) \\ &= \int_{\mathcal{S} \times \mathcal{S}} \Phi_A(\mathcal{T}(\mathbf{s}_1, \mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2) \\ &= \int_{\mathcal{Z}_{\mathcal{S}}} \Phi_A(\mathbf{z}) d\lambda(\mathbf{z}) \\ &= \int_{\mathcal{Z}_{\mathcal{S}}} \mathcal{N}(\mathbf{z})(A) d\lambda(\mathbf{z}). \end{aligned}$$

Since this identity holds for every $A \in \mathcal{B}(\mathbb{R}^{2n})$, we have $\nu_{\pi} \in \mathbb{GM}_{2n}^{\mathcal{S}}$. Together with the conclusion from Lemma 1 that $\nu_{\pi} \in \Pi(\mu_1, \mu_2)$, this proves $\nu_{\pi} \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^{\mathcal{S}}$. $\square$

Corollary 7. *Let*

$$\nu \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S,$$

and let $\lambda \in \mathcal{P}(\mathcal{Z}_S)$ be a restricted mixing law for ν , i.e.,

$$\nu(A) = \int_{\mathcal{Z}_S} \mathcal{N}(\mathbf{z})(A) \lambda(d\mathbf{z}), \quad A \in \mathcal{B}(\mathbb{R}^{2n}).$$

If $\pi_\nu := \mathcal{M}_\# \lambda$, then $\pi_\nu \in \Gamma_S$.

Proof. Since $\mathcal{Z}_S \subseteq \mathcal{Z}$, we can regard $\lambda \in \mathcal{P}(\mathcal{Z}_S)$ as a probability measure on the ambient joint-parameter space $\mathcal{Z}$, supported on $\mathcal{Z}_S$. By Corollary 5, parameter second-moment requirements appearing in Proposition 3 are finite under the restricted support. Hence Lemma 2 applies and shows that the first marginal of $\pi_\nu := \mathcal{M}_\# \lambda$ is a mixing law for μ_1 , and the second marginal is a mixing law for μ_2 . Moreover, Corollary 5 gives $\pi_\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$, and compactness of $\mathcal{S}$ gives $\pi_\nu \in \mathcal{P}_2(\mathcal{S} \times \mathcal{S})$. Therefore, by definition, $\pi_\nu \in \Gamma_S$. $\square$

B.4.3. Objective identity under restricted parameter support. We now show the objective identity (35) of Theorem 3.

Proof of Theorem 3. We prove the two inequalities.

First, let $\pi \in \Gamma_S$. By Corollary 6, the measure

$$\nu_\pi(A) := \int_{\mathcal{S} \times \mathcal{S}} \mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(A) d\pi(\mathbf{s}_1, \mathbf{s}_2), \quad A \in \mathcal{B}(\mathbb{R}^{2n}),$$

belongs to

$$\Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S.$$

Therefore, by the definition of $W_{2,S}^{\text{mix}}$,

$$W_{2,S}^{\text{mix}} \leq \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_\pi(\mathbf{y}_1, \mathbf{y}_2).$$

Since $\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \geq 0$, the kernel Tonelli theorem gives

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu_\pi(\mathbf{y}_1, \mathbf{y}_2) = \int_{\mathcal{S} \times \mathcal{S}} \left[\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2) \right] d\pi(\mathbf{s}_1, \mathbf{s}_2).$$

For each fixed $(\mathbf{s}_1, \mathbf{s}_2)$, $\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)$ is the Wasserstein-2 optimal Gaussian coupling between $\mathcal{N}(\mathbf{s}_1)$ and $\mathcal{N}(\mathbf{s}_2)$. Hence,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2)(\mathbf{y}_1, \mathbf{y}_2) = d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)).$$

Consequently,

$$W_{2,S}^{\text{mix}} \leq \int_{\mathcal{S} \times \mathcal{S}} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi(\mathbf{s}_1, \mathbf{s}_2).$$

Taking the infimum over $\pi \in \Gamma_S$ yields

$$W_{2,S}^{\text{mix}} \leq W_{2,S}^{\text{prn}}.$$

Conversely, let

$$\nu \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S,$$

and let $\lambda \in \mathcal{P}(\mathcal{Z}_S)$ be a restricted mixing law for ν , so that

$$\nu(A) = \int_{\mathcal{Z}_S} \mathcal{N}(z)(A) \lambda(dz), \quad A \in \mathcal{B}(\mathbb{R}^{2n}).$$

By Corollary 7,

$$\pi_\nu := \mathcal{M}_\# \lambda \in \Gamma_S.$$

For each $z \in \mathcal{Z}_S$, the Gaussian law $\mathcal{N}(z)$ has marginals

$$\mathcal{N}(\mathcal{M}_1(z)) \quad \text{and} \quad \mathcal{N}(\mathcal{M}_2(z)).$$

Therefore $\mathcal{N}(z)$ is one feasible coupling between these two Gaussian marginals. By optimality of the Gelbrich coupling,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\mathcal{N}(z)(\mathbf{y}_1, \mathbf{y}_2) \geq d_{\text{BW}}^2(\mathcal{N}(\mathcal{M}_1(z)), \mathcal{N}(\mathcal{M}_2(z))).$$

Since $\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \geq 0$, the kernel Tonelli theorem and the restricted mixture representation of ν give

$$\begin{aligned} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2) &= \int_{\mathcal{Z}_S} \left[\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\mathcal{N}(z)(\mathbf{y}_1, \mathbf{y}_2) \right] d\lambda(z) \\ &\geq \int_{\mathcal{Z}_S} d_{\text{BW}}^2(\mathcal{N}(\mathcal{M}_1(z)), \mathcal{N}(\mathcal{M}_2(z))) d\lambda(z). \end{aligned}$$

Now use the pushforward identity for

$$\pi_\nu = \mathcal{M}_\# \lambda.$$

With

$$\Phi(\mathbf{s}_1, \mathbf{s}_2) := d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)),$$

we have

$$\int_{\mathcal{Z}_S} \Phi(\mathcal{M}(z)) d\lambda(z) = \int_{S \times S} \Phi(\mathbf{s}_1, \mathbf{s}_2) d\pi_\nu(\mathbf{s}_1, \mathbf{s}_2).$$

Equivalently,

$$\int_{\mathcal{Z}_S} d_{\text{BW}}^2(\mathcal{N}(\mathcal{M}_1(z)), \mathcal{N}(\mathcal{M}_2(z))) d\lambda(z) = \int_{S \times S} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi_\nu(\mathbf{s}_1, \mathbf{s}_2).$$

Therefore,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2) \geq \int_{S \times S} d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_1), \mathcal{N}(\mathbf{s}_2)) d\pi_\nu(\mathbf{s}_1, \mathbf{s}_2).$$

Since $\pi_\nu \in \Gamma_S$, the right-hand side is bounded below by $W_{2,S}^{\text{prm}}$. Hence

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\mathbf{y}_1 - \mathbf{y}_2\|^2 d\nu(\mathbf{y}_1, \mathbf{y}_2) \geq W_{2,S}^{\text{prm}}.$$

Taking the infimum over $\nu \in \Pi(\mu_1, \mu_2) \cap \mathbb{GM}_{2n}^S$ yields

$$W_{2,S}^{\text{mix}} \geq W_{2,S}^{\text{prm}}.$$

Combining the two inequalities proves identity (35) □

B.5. Properties of Burer-Wasserstein Distance d_{BW}^2 and Duality

B.5.1. Proof of Lemma 4. Fix $(\hat{m}, \hat{Q}) \in \Theta \times \Xi$. For $Q \in \mathbb{S}_+^n$, define

$$F(Q) := \text{tr}(\hat{Q}) + \text{tr}(Q) - 2 \text{tr}\left((Q^{1/2} \hat{Q} Q^{1/2})^{1/2}\right). \quad (68)$$

Hence,

$$d_{\text{BW}}^2(m, Q; (\hat{m}, \hat{Q})) = \|m - \hat{m}\|_2^2 + F(Q).$$

We prove the two claims in the order stated: first, convexity on $\mathbb{R}^n \times \mathbb{S}_+^n$, and then continuity on $\Theta \times \Xi$.

Convexity proof. Since the map $m \mapsto \|m - \hat{m}\|_2^2$ is convex on $\mathbb{R}^n$, it remains to prove that F is convex on $\mathbb{S}_+^n$. We first justify the following trace identity

$$\text{tr}\left((Q^{1/2} \hat{Q} Q^{1/2})^{1/2}\right) = \text{tr}\left((\hat{Q}^{1/2} Q \hat{Q}^{1/2})^{1/2}\right). \quad (69)$$

Let $R := Q^{1/2} \hat{Q}^{1/2}$. Then,

$$RR^\top = Q^{1/2} \hat{Q} Q^{1/2}, \quad R^\top R = \hat{Q}^{1/2} Q \hat{Q}^{1/2} \quad (70)$$

Both $RR^\top$ and $R^\top R$ are symmetric positive semidefinite. We show that they have the same characteristic polynomial. For every $t \neq 0$, Sylvester's determinant identity [35] gives

$$\det(tI - RR^\top) = t^n \det\left(I - \frac{1}{t} RR^\top\right) = t^n \det\left(I - \frac{1}{t} R^\top R\right) = \det(tI - R^\top R).$$

The two sides are polynomials in t . Since they agree for every $t \neq 0$, they agree for every $t \in \mathbb{R}$. Hence,

$$\det(tI - RR^\top) = \det(tI - R^\top R) \quad \forall t \in \mathbb{R}.$$

Therefore, $RR^\top$ and $R^\top R$ have the same characteristic polynomial.

Since $RR^\top$ and $R^\top R$ are symmetric positive semidefinite, the spectral theorem [35] gives orthogonal matrices U, V and scalars $\lambda_1, \dots, \lambda_n \geq 0$ such that

$$RR^\top = U \text{diag}(\lambda_1, \dots, \lambda_n) U^\top, \quad R^\top R = V \text{diag}(\lambda_1, \dots, \lambda_n) V^\top.$$

By the definition of the principal symmetric positive semidefinite square root,

$$(RR^\top)^{1/2} = U \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}) U^\top, \quad \text{and} \quad (R^\top R)^{1/2} = V \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}) V^\top.$$

Using the invariance of trace under orthogonal similarity,

$$\text{tr}((RR^\top)^{1/2}) = \sum_{i=1}^n \sqrt{\lambda_i} = \text{tr}((R^\top R)^{1/2}).$$

Substitution with (70) gives (69).

We next prove that $H(\mathbf{Q}) := \text{tr}\left((\hat{\mathbf{Q}}^{1/2}\mathbf{Q}\hat{\mathbf{Q}}^{1/2})^{1/2}\right)$ is concave on $\mathbb{S}_+^n$. Let $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)} \in \mathbb{S}_+^n$, $\alpha \in [0, 1]$ and

$$\mathbf{Q}^{(\alpha)} := (1 - \alpha)\mathbf{Q}^{(1)} + \alpha\mathbf{Q}^{(2)}. \quad (71)$$

Since the map

$$\mathcal{L} : \mathbb{S}^n \rightarrow \mathbb{S}^n, \quad \mathcal{L}(\mathbf{Q}) := \hat{\mathbf{Q}}^{1/2}\mathbf{Q}\hat{\mathbf{Q}}^{1/2},$$

is linear, for any $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)} \in \mathbb{S}^n$ and any $a, b \in \mathbb{R}$,

$$\begin{aligned} \mathcal{L}(a\mathbf{Q}^{(1)} + b\mathbf{Q}^{(2)}) &= \hat{\mathbf{Q}}^{1/2}(a\mathbf{Q}^{(1)} + b\mathbf{Q}^{(2)})\hat{\mathbf{Q}}^{1/2} \\ &= a\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(1)}\hat{\mathbf{Q}}^{1/2} + b\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(2)}\hat{\mathbf{Q}}^{1/2} \\ &= a\mathcal{L}(\mathbf{Q}^{(1)}) + b\mathcal{L}(\mathbf{Q}^{(2)}). \end{aligned}$$

Therefore, following (71) we obtain

$$\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(\alpha)}\hat{\mathbf{Q}}^{1/2} = (1 - \alpha)\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(1)}\hat{\mathbf{Q}}^{1/2} + \alpha\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(2)}\hat{\mathbf{Q}}^{1/2}.$$

Additionally, since the principal square-root map is operator concave on $\mathbb{S}_+^n$ [7, Theorem 4.2.3],

$$\left(\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(\alpha)}\hat{\mathbf{Q}}^{1/2}\right)^{1/2} \succeq (1 - \alpha)\left(\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(1)}\hat{\mathbf{Q}}^{1/2}\right)^{1/2} + \alpha\left(\hat{\mathbf{Q}}^{1/2}\mathbf{Q}^{(2)}\hat{\mathbf{Q}}^{1/2}\right)^{1/2}.$$

Taking traces preserves the Löwner order [7]: if $\mathbf{A} \succeq \mathbf{B}$, then $\text{tr}(\mathbf{A}) \geq \text{tr}(\mathbf{B})$. Hence,

$$H(\mathbf{Q}^{(\alpha)}) \geq (1 - \alpha)H(\mathbf{Q}^{(1)}) + \alpha H(\mathbf{Q}^{(2)}).$$

Thus H is concave on $\mathbb{S}_+^n$.

Finally, we show convexity of F on $\mathbb{S}_+^n$. Using (69), (68) gives us

$$F(\mathbf{Q}) = \text{tr}(\hat{\mathbf{Q}}) + \text{tr}(\mathbf{Q}) - 2H(\mathbf{Q}),$$

where $\text{tr}(\hat{\mathbf{Q}})$ is constant in $\mathbf{Q}$ and $\text{tr}(\mathbf{Q})$ is linear in $\mathbf{Q}$ since $\text{tr}(\mathbf{Q}^{(\alpha)}) = (1 - \alpha)\text{tr}(\mathbf{Q}^{(1)}) + \alpha\text{tr}(\mathbf{Q}^{(2)})$. Also, concavity of H implies $-2H(\mathbf{Q})$ is convex. Hence,

$$F(\mathbf{Q}^{(\alpha)}) \leq (1 - \alpha)F(\mathbf{Q}^{(1)}) + \alpha F(\mathbf{Q}^{(2)}),$$

i.e., F is convex on $\mathbb{S}_+^n$. Combining this with the convexity of $\mathbf{m} \mapsto \|\mathbf{m} - \hat{\mathbf{m}}\|_2^2$, the map

$$(\mathbf{m}, \mathbf{Q}) \mapsto d_{\text{BW}}^2(\mathbf{m}, \mathbf{Q}; (\hat{\mathbf{m}}, \hat{\mathbf{Q}})) \quad (72)$$

is convex on $\mathbb{R}^n \times \mathbb{S}_+^n$.

Continuity proof. Let $\mathcal{U} := \mathbb{R}^n \times \mathbb{S}_{++}^n$. The set $\mathcal{U}$ is open and convex in the finite-dimensional vector space $\mathbb{R}^n \times \mathbb{S}^n$. Since $\mathbb{S}_\delta^n = \{\mathbf{Q} \in \mathbb{S}_{++}^n : \mathbf{Q} \succeq \delta \mathbf{I}\}$, we have $\Theta \times \Xi \subseteq \mathcal{U}$. From the convexity result above, the map (72) is convex on $\mathbb{R}^n \times \mathbb{S}_+^n$. Hence, its restriction to $\mathcal{U}$ is a finite convex function on the open convex set $\mathcal{U}$. By the standard continuity theorem for finite convex functions on open convex sets [69, Corollary 2.36], this restriction is continuous on $\mathcal{U}$. Since $\Theta \times \Xi \subseteq \mathcal{U}$, the further restriction to $\Theta \times \Xi$ is continuous. Hence, $d_{\text{BW}}^2(\mathbf{m}, \mathbf{Q}; (\hat{\mathbf{m}}, \hat{\mathbf{Q}}))$ is continuous on $\Theta \times \Xi$.

B.5.2. Convexity on restricted domains.

Corollary 8 (Convexity on restricted domains). *Fix $(\hat{\mathbf{m}}, \hat{\mathbf{Q}}) \in \Theta \times \Xi$, and define*

$$d_{\text{BW}}^2(\mathbf{m}, \mathbf{Q}; (\hat{\mathbf{m}}, \hat{\mathbf{Q}})) = \|\hat{\mathbf{m}} - \mathbf{m}\|_2^2 + \text{tr}\left(\hat{\mathbf{Q}} + \mathbf{Q} - 2(\mathbf{Q}^{1/2} \hat{\mathbf{Q}} \mathbf{Q}^{1/2})^{1/2}\right).$$

Then the following holds.

- i) *The map $(\mathbf{m}, \mathbf{Q}) \mapsto d_{\text{BW}}^2(\mathbf{m}, \mathbf{Q}; (\hat{\mathbf{m}}, \hat{\mathbf{Q}}))$ is convex on $\mathbb{R}_+^n \times \mathbb{S}_{++}^n$.*
- ii) *If $\Theta \times \Xi$ is convex, then the same map is convex on $\Theta \times \Xi$.*

Proof. By Lemma 4, the map is convex on $\mathbb{R}^n \times \mathbb{S}_+^n$.

We first show that $\mathbb{R}_+^n \times \mathbb{S}_{++}^n$ is convex. Let

$$(\mathbf{m}^{(1)}, \mathbf{Q}^{(1)}), (\mathbf{m}^{(2)}, \mathbf{Q}^{(2)}) \in \mathbb{R}_+^n \times \mathbb{S}_{++}^n, \quad \alpha \in [0, 1].$$

Since $\mathbb{R}_+^n$ is convex,

$$(1 - \alpha)\mathbf{m}^{(1)} + \alpha\mathbf{m}^{(2)} \in \mathbb{R}_+^n.$$

Moreover, for every $x \neq 0$,

$$x^\top ((1 - \alpha)\mathbf{Q}^{(1)} + \alpha\mathbf{Q}^{(2)})x = (1 - \alpha)x^\top \mathbf{Q}^{(1)}x + \alpha x^\top \mathbf{Q}^{(2)}x > 0.$$

Thus

$$(1 - \alpha)\mathbf{Q}^{(1)} + \alpha\mathbf{Q}^{(2)} \in \mathbb{S}_{++}^n.$$

Therefore, $\mathbb{R}_+^n \times \mathbb{S}_{++}^n$ is a convex subset of $\mathbb{R}^n \times \mathbb{S}_+^n$. The restriction of a convex function to a convex subset of its domain is convex, so the first claim follows.

For the second claim, suppose $\Theta \times \Xi$ is convex. Then it is a convex subset of $\mathbb{R}^n \times \mathbb{S}_+^n$. Therefore, the restriction of the convex function from Lemma 4 to $\Theta \times \Xi$ is convex. Finally, if Θ and Ξ are convex, then $\Theta \times \Xi$ is convex because convex combinations are taken componentwise. $\square$

Remark 9 (A common convex compact specification of $\Theta \times \Xi$). *One common way to specify the compact mean–covariance support is to take*

$$\Theta := \{\mathbf{m} \in \mathbb{R}^n : |\mathbf{m}_i - \hat{\mathbf{m}}_i| \leq \varsigma |\hat{\mathbf{m}}_i|, \quad i = 1, \dots, n\},$$

for some $\varsigma \geq 0$, and

$$\Xi := \left\{ \mathbf{Q} \in \mathbb{S}^n : \underline{\lambda}_f \hat{\mathbf{Q}} \preceq \mathbf{Q} \preceq \bar{\lambda}_f \hat{\mathbf{Q}} \right\},$$

where $\hat{\mathbf{Q}} \succ 0$ and $0 < \underline{\lambda}_f \leq \bar{\lambda}_f$. Under this specification, Θ is a Cartesian product of closed intervals and is therefore compact and convex. The set Ξ is also compact and convex. Indeed, if $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)} \in \Xi$ and $\alpha \in [0, 1]$, then

$$\underline{\lambda}_f \hat{\mathbf{Q}} \preceq (1 - \alpha)\mathbf{Q}^{(1)} + \alpha\mathbf{Q}^{(2)} \preceq \bar{\lambda}_f \hat{\mathbf{Q}},$$

so $(1 - \alpha)\mathbf{Q}^{(1)} + \alpha\mathbf{Q}^{(2)} \in \Xi$. Therefore, $\Theta \times \Xi$ is compact and convex. Moreover, with $\hat{\delta} := \underline{\lambda}_f \lambda_{\min}(\hat{\mathbf{Q}}) > 0$,

$$\mathbf{Q} \succeq \underline{\lambda}_f \hat{\mathbf{Q}} \succeq \underline{\lambda}_f \lambda_{\min}(\hat{\mathbf{Q}}) \mathbf{I}.$$

Hence, $\Xi \subseteq \mathbb{S}_{\hat{\delta}}^n$.

B.5.3. Proof of Theorem 4. Fix any $\mathbf{x} \in \mathcal{X}$. Since $\mathcal{D} \neq \emptyset$, the robust constraint in (CDR), $\sum_{k=1}^{\hat{K}} \int_{\mathcal{S}} G(\mathbf{s}, \mathbf{x}) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) \geq \theta, \forall \pi \in \mathcal{D}$, is equivalent to $\inf_{\pi \in \mathcal{D}} \sum_{k=1}^{\hat{K}} \int_{\mathcal{S}} G(\mathbf{s}, \mathbf{x}) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) \geq \theta$. By Lemma 4, for any fixed outer decision $\mathbf{x}$, the strong duality result for this inner problem provides the following objective-value identity

$$\begin{aligned} & \inf_{\pi \in \mathcal{D}} \sum_{k=1}^{\hat{K}} \int_{\mathcal{S}} G(\mathbf{s}, \mathbf{x}) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) \\ &= \sup_{\beta \in \mathbb{R}^{\hat{K}}, \beta_{\hat{K}+1} \geq 0} \left\{ \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} : \beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) \leq G(\mathbf{s}, \mathbf{x}), \forall \mathbf{s} \in \mathcal{S}, \forall k \in [\hat{K}] \right\}. \end{aligned} \quad (73)$$

We next show that the supremum on the right-hand side of (73) is attained. This attainment is needed because the semi-infinite formulation uses an actual decision vector β , not a supremum operator. The point $(\beta, \beta_{\hat{K}+1}) = (0, 0)$ is feasible because $G(\mathbf{s}, \mathbf{x}) \geq 0$ on $\mathcal{S}$, so the supremum is at least zero. No component β_k needs to be negative: if $\beta_k < 0$, replacing it by 0 preserves feasibility, since $0 - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) \leq 0 \leq G(\mathbf{s}, \mathbf{x})$, and weakly improves the objective. Hence a maximizing sequence may be chosen with $\beta_k \geq 0$ for every k . Evaluating the dual constraint at $\mathbf{s} = \hat{\mathbf{s}}_k$ gives $\beta_k \leq G(\hat{\mathbf{s}}_k, \mathbf{x}) \leq 1$, since $d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\hat{\mathbf{s}}_k)) = 0$. Thus $0 \leq \beta_k \leq 1$ for all k . Along any maximizing sequence with nonnegative objective value, $0 \leq \sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \leq 1 - \rho \beta_{\hat{K}+1}$, where $\hat{w} \in \Delta_{\hat{K}} := \left\{ \hat{w} \in \mathbb{R}_+^{\hat{K}} : \sum_{k=1}^{\hat{K}} \hat{w}_k = 1 \right\}$ and $\beta_k \leq 1$ were used. Since $\rho > 0$, we obtain $0 \leq \beta_{\hat{K}+1} \leq 1/\rho$. Therefore the supremum can be taken over the compact set $[0, 1]^{\hat{K}} \times [0, 1/\rho]$. The feasible set is closed because the constraint functions are continuous in $\mathbf{s}$ on compact $\mathcal{S}$, and the dual objective is continuous. Hence the supremum is a maximum.

We now prove the feasibility equivalence. Suppose first that $\mathbf{x}$ satisfies the robust constraint in (CDR). Then $\inf_{\pi \in \mathcal{D}} \sum_{k=1}^{\hat{K}} \int_{\mathcal{S}} G(\mathbf{s}, \mathbf{x}) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) \geq \theta$. By the objective-value equality above and the attained maximum on the dual side, there exist $\beta \in \mathbb{R}^{\hat{K}}$ and $\beta_{\hat{K}+1} \geq 0$ such that $\sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \rho \beta_{\hat{K}+1} \geq \theta$ and $\beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) \leq G(\mathbf{s}, \mathbf{x}), \forall \mathbf{s} \in \mathcal{S}, \forall k \in [\hat{K}]$. Hence we obtain $\beta \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$ satisfying the constraints of (S-Inf).

Conversely, suppose there exists $\beta \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$ satisfying the constraints of (S-Inf). Then $(\beta_1, \dots, \beta_{\hat{K}}, \beta_{\hat{K}+1})$ is feasible for the dual problem above, and its objective value is at least θ . Therefore, the dual supremum is at least θ . By the objective-value equality, $\inf_{\pi \in \mathcal{D}} \sum_{k=1}^{\hat{K}} \int_{\mathcal{S}} G(\mathbf{s}, \mathbf{x}) d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) \geq \theta$. This is exactly the robust constraint in (CDR). Hence, for every $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}$ is feasible for (CDR) if and only if there exists $\beta \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$ such that $(\mathbf{x}, \beta)$ is feasible for (S-Inf).

Since both formulations minimize the same objective $\mathbf{c}^\top \mathbf{x}$, the feasibility equivalence gives equality of the optimal objective values. It also gives the stated optimal-solution correspondence. If $\mathbf{x}^*$ is optimal for (CDR), then the feasibility equivalence gives some β^* such that $(\mathbf{x}^*, \beta^*)$ is feasible for (S-Inf). If this pair were not optimal for (S-Inf), then some feasible $(\bar{\mathbf{x}}, \bar{\beta})$ of (S-Inf) would satisfy $\mathbf{c}^\top \bar{\mathbf{x}} < \mathbf{c}^\top \mathbf{x}^*$. The feasibility equivalence would make $\bar{\mathbf{x}}$ feasible for (CDR), contradicting optimality of $\mathbf{x}^*$. Thus $(\mathbf{x}^*, \beta^*)$

is optimal for (S-Inf). Conversely, if $(\mathbf{x}^*, \beta^*)$ is optimal for (S-Inf), then $\mathbf{x}^*$ is feasible for (CDR). If $\mathbf{x}^*$ were not optimal for (CDR), there would exist a feasible $\bar{\mathbf{x}}$ of (CDR) with $\mathbf{c}^\top \bar{\mathbf{x}} < \mathbf{c}^\top \mathbf{x}^*$. The feasibility equivalence would give some $\bar{\beta}$ such that $(\bar{\mathbf{x}}, \bar{\beta})$ is feasible for (S-Inf), contradicting optimality of $(\mathbf{x}^*, \beta^*)$. Therefore the optimal $\mathbf{x}$ -solutions coincide whenever an optimal solution exists. $\square$

Remark 10. The assumption $\rho > 0$ is used only to guarantee attainment of the dual supremum in the semi-infinite reformulation. It is not needed for finiteness of the primal inner value, since $0 \leq G(\mathbf{s}, \mathbf{x}) \leq 1$ and $\mathcal{D} \neq \emptyset$. When $\rho > 0$, the term $-\rho\beta_{\hat{K}+1}$ bounds $\beta_{\hat{K}+1}$ along any maximizing sequence, which yields compactness of the relevant dual level set. If $\rho = 0$, this coercive term disappears; the dual supremum may remain finite but may fail to be attained because $\beta_{\hat{K}+1}$ can diverge while relaxing all positive-distance constraints. Thus, without an additional dual-attainment argument, the value equality still supports an optimal-value statement, but the existential semi-infinite reformulation with a finite β requires $\rho > 0$.

B.6. Reformulated DR-Chance Constrained Program

B.6.1. Proof of Proposition 6. By definition (38), $G(\mathbf{s}, \mathbf{x}) = 0$. As a result, since $\theta \in (0, 1]$, constraint (39) always fails to satisfy, i.e., $\min_{\pi \in \mathcal{P}} \sum_{k=1}^{\hat{K}} \int_{\mathbf{s} \in (\Theta \times \Xi)} 0 d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) = 0 \not\geq \theta$.

Now, for $\mathbf{x} = \mathbf{0}$ with $b \geq 0$, $G(\mathbf{s}, \mathbf{x}) = 1$, which translates the worst-case distribution finding problem (39) to a feasibility problem. In particular, any value of $\pi \in \Pi(\hat{\mathbf{w}}, \mathbf{w})$ with some $\mathbf{w} \in \mathcal{P}_2(\mathcal{S})$ yields $\min_{\pi \in \mathcal{P}} \sum_{k=1}^{\hat{K}} \int_{\mathbf{s} \in (\Theta \times \Xi)} d\pi(\hat{\mathbf{s}}_k, \mathbf{s}) = 1 > \theta$. Therefore, since $\mathbf{w} = \hat{\mathbf{w}} \in \mathcal{P}_2(\mathcal{S})$, $\pi \in \Pi(\hat{\mathbf{w}}, \hat{\mathbf{w}})$ is also a solution to the problem (39) yielding same objective values. $\square$

B.6.2. Proof of Proposition 7. If $\|\beta\| = 0$, i.e., $\beta^k = 0 \forall k \in [\hat{K}]$, $\beta_{\hat{K}+1} \geq 0$ implies that $\sum_{k=1}^{\hat{K}} \hat{w}_k \beta_k - \beta_{\hat{K}+1} \rho \leq 0 \not\geq \theta$; a contradiction since $\theta \in (0, 1]$. Hence, $\|\beta\| > 0$.

We also prove *part-(ii)* by contradiction. When $b < 0$, for $\mathbf{x} = \mathbf{0}$, $G(\mathbf{s}; \mathbf{x}) = \mathbb{1}_{\geq 0}(b) = 0$ by definition (38). Then for $\hat{\beta} \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$ be feasible it must satisfy all the semi-infinite constraints:

$$\beta_k - \beta_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) - G(\mathbf{s}; \mathbf{x}) \leq 0, \quad \forall \mathbf{s} \in \Theta \times \Xi, \forall k \in [\hat{K}].$$

Hence, nominal distribution parameter also belongs to the set $\Theta \times \Xi$, i.e., $\mathbf{s}_k = \hat{\mathbf{s}}_k \in \Theta \times \Xi$ which leads $d_{\text{BW}}^2(\mathcal{N}(\mathbf{s}_k), \mathcal{N}(\hat{\mathbf{s}}_k)) = 0$, $\forall k \in [\hat{K}]$. To satisfy such a constraint, there must be a solution $\hat{\beta}$ such that $\hat{\beta}_k \leq 0$, $\forall k \in [\hat{K}]$. However, since $\beta_{\hat{K}+1} \in \mathbb{R}_+$, we obtain $\sum_{k=1}^{\hat{K}} \hat{w}_k \hat{\beta}_k - \hat{\beta}_{\hat{K}+1} \rho \leq 0 \not\geq \theta \in (0, 1]$, which contradicts the feasibility of $\hat{\beta}$.

For *part-iii*, since $b \geq 0$, given $\hat{\beta} \in \mathbb{R}^{\hat{K}} \times \mathbb{R}_+$ and $\mathbf{x} = \mathbf{0}$, $G(\mathbf{s}; \mathbf{x}) = G(\mathbf{s}; \mathbf{0}) = 1$. Hence, any element $\mathbf{s} \in \Theta \times \Xi$ can be a solution to $\max_{\mathbf{s}} \hat{\beta}_k - \hat{\beta}_{\hat{K}+1} d_{\text{BW}}^2(\mathcal{N}(\hat{\mathbf{s}}_k), \mathcal{N}(\mathbf{s})) - 1$, so does the solution $\mathbf{s}_k = \hat{\mathbf{s}}_k$ for any $k \in [\hat{K}]$. $\square$

B.6.3. Proof of Proposition 8. First, consider the forward direction of the claim starting from $\bar{\Phi}(z; \tilde{z}) \geq \zeta$. Let us take some $z \geq 0$, ζ such that $\bar{\Phi}(z; \tilde{z}) \geq \zeta$. Then, $z = y_3 \geq 0$ with $t_3 = 1$ implies that $t_1, t_2 = 0$, $g_i z + g_i^0 \geq \zeta$, $i \in [R]_0$ and $1 \geq \zeta$. Thus, for a given solution $z \geq 0$ satisfying $\bar{\Phi}(z; \tilde{z}) \geq \zeta$, one can construct a feasible solution to the set $\mathcal{H}^O(\tilde{z})$ using $t_1, t_2 = 0, t_3 = 1, \alpha = \mathbf{0}, y_1, y_2 = 0, z = y_3 \geq 0$.

Next we consider some $z \in [\tilde{z}_{-L}, 0)$. Then, z will be in $[\tilde{z}_{-i}, \tilde{z}_{-i+1}]$ for some $i \in [L]$ which indicates $t_2 = 1, t_1 = t_3 = 0$, and therefore, $y_1 = y_3 = 0$. Then there exists some weights $\hat{\alpha}_i, \hat{\alpha}_{i-1} \geq 0$ such that $\hat{\alpha}_i + \hat{\alpha}_{i-1} = 1 = t_2$. Furthermore, from constraint $y_2 = \sum_{i=0}^L \alpha_i \tilde{z}_{-i}$ we obtain $y_2 = \hat{\alpha}_i \tilde{z}_{-i} + \hat{\alpha}_{i-1} \tilde{z}_{-i+1}$. Hence, $\alpha_i \Phi(\tilde{z}_{-i}) + \alpha_{i-1} \Phi(\tilde{z}_{-i+1}) \geq \zeta$. Thus for any given $z \in [\tilde{z}_{-L}, 0)$, one can find an index $i \in [L]$ such that $\alpha_i = \hat{\alpha}_i, \alpha_{i-1} = \hat{\alpha}_{i-1}$ with $y_2 = \hat{\alpha}_i \tilde{z}_{-i} + \hat{\alpha}_{i-1} \tilde{z}_{-i+1}, y_1 = y_3 = 0$, and $\alpha_j = 0 \forall j \in [L] \setminus \{i-1, i\}$, which satisfy all the constraints of $\mathcal{H}^O(\tilde{z})$. Similarly, if we take some $z < \tilde{z}_{-L}$, then $t_1 = 1$, and $y_1 = z$. This implies $t_2, t_3, y_2, y_3 = 0, \Phi(\tilde{z}_{-L}) \geq \zeta$. Therefore, $\mathcal{H}^O(\tilde{z}) \neq \emptyset$.

Now we prove that any point in the set $\mathcal{H}^O(\tilde{z})$ also satisfies $\bar{\Phi}(z; \tilde{z}) \geq \zeta$ based on the definition of $\bar{\Phi}(z; \tilde{z})$ in (45). It is formally shown below using three separate cases:

- i) Let us choose $y_3 = z \geq 0$ with $t_3 = 1$ and $t_1, t_2, y_1, y_2 = 0$ such that $g_i y_3 + g_i^0 \geq \zeta$ for all $i \in [R]_0$. Then $\bar{\Phi}(z; \tilde{z}) = \min_{i \in [R]_0} \{g_i y_3 + g_i^0\} \geq \zeta$.
- ii) Let $\tilde{z}_{-L} \leq y_2 = z < 0$ with $t_2 = 1$ and $t_1, t_3, y_1, y_3 = 0$ such that $y_2 = \alpha_i \tilde{z}_{-i} + \alpha_{i-1} \tilde{z}_{-i+1}, \alpha_i + \alpha_{i-1} = 1 = t_2$, and $\alpha_i \Phi(\tilde{z}_{-i}) + \alpha_{i-1} \Phi(\tilde{z}_{-i+1}) \geq \zeta$ for some $i \in [L]$ and $\alpha_j = 0$ for all $j \in [L] \setminus \{i-1, i\}$. Thus there exists an index $i \in [L]$ that satisfies $\bar{\Phi}(z; \tilde{z}) = h_i y_2 + h_i^0 \geq \zeta$. Hence, $\bar{\Phi}(z; \tilde{z}) = \max_{i \in [L]} h_i y_2 + h_i^0 \geq \zeta$ is satisfied.
- iii) Finally, $y_1 = z < \tilde{z}_{-L}$ with $t_1 = 1, t_2, t_3, y_2, y_3 = 0$ indicates that $\Phi(\tilde{z}_{-L}) = \bar{\Phi}(z; \tilde{z}) \geq \zeta$.

Additionally, since all of these cases allow ζ to be at most 1, using the definition of $\bar{\Phi}(z; \tilde{z})$ in (45), $\bar{\Phi}(z; \tilde{z}) \geq \zeta$ is satisfied. $\square$

B.6.4. Proof of Proposition 9. First, consider the forward direction of the claim starting from $-\Phi(z; \tilde{z}) \geq \zeta'$. Let us take some $z \leq 0$, ζ' such that $-\Phi(z; \tilde{z}) \geq \zeta'$. Then, $z = y_1 \leq 0$ with $t_1 = 1$ implies that $t_2, t_3 = 0$, $-h_i z - h_i^0 \geq \zeta', i \in [L]_0$ and $\zeta' \leq 0$. Thus, for a given solution $z \leq 0$ satisfying $-\Phi(z; \tilde{z}) \geq \zeta'$, one can construct a feasible solution to the set $\mathcal{H}^I(\tilde{z})$ using $t_3, t_2 = 0, t_1 = 1, \alpha = \mathbf{0}, y_2, y_3 = 0, z = y_1 \leq 0$.

Next we consider some $z \in (0, \tilde{z}_R]$. Then, z will be in $[\tilde{z}_{i-1}, \tilde{z}_i]$ for some $i \in [R]$ which indicates $t_2 = 1, t_1 = t_3 = 0$, and therefore, $y_1 = y_3 = 0$. Then there exists some weights $\hat{\alpha}_{i-1}, \hat{\alpha}_i \geq 0$ such that $\hat{\alpha}_{i-1} + \hat{\alpha}_i = 1 = t_2$. Furthermore, from constraint $y_2 = \sum_{i=0}^R \alpha_i \tilde{z}_i$ we obtain $y_2 = \hat{\alpha}_{i-1} \tilde{z}_{i-1} + \hat{\alpha}_i \tilde{z}_i$. Hence, $-\alpha_{i-1} \Phi(\tilde{z}_{i-1}) - \alpha_i \Phi(\tilde{z}_i) \geq \zeta'$. Thus for any given $z \in (0, \tilde{z}_R]$, one can find an index $i \in [R]$ such that $\alpha_{i-1} = \hat{\alpha}_{i-1}, \alpha_i = \hat{\alpha}_i$ with $y_2 = \hat{\alpha}_{i-1} \tilde{z}_{i-1} + \hat{\alpha}_i \tilde{z}_i, y_1 = y_3 = 0$, and $\alpha_j = 0 \forall j \in [R_0] \setminus \{i-1, i\}$, which satisfy all the constraints of $\mathcal{H}^I(\tilde{z})$. Similarly, if we take some $z > \tilde{z}_R$, then $t_3 = 1$, and $y_3 = z$. This implies $t_1, t_2, y_1, y_2 = 0, -\Phi(\tilde{z}_R) \geq \zeta'$. Therefore, $\mathcal{H}^I(\tilde{z}) \neq \emptyset$.

Now we prove that any point in the set $\mathcal{H}^I(\tilde{z})$ also satisfies $-\underline{\Phi}(z; \tilde{z}) \geq \zeta'$ based on the definition of $\underline{\Phi}(z; \tilde{z})$ in (46). It is formally shown below using three separate cases:

i) Let us choose $y_1 = z \leq 0$ with $t_1 = 1$ and $t_2, t_3, y_2, y_3 = 0$ such that $-h_i y_1 - h_i^0 \geq \zeta'$ for all $i \in [L]_0$.

Then $-\underline{\Phi}(z; \tilde{z}) = \min_{i \in [L]_0} \{-h_i y_1 - h_i^0\} \geq \zeta'$.

ii) Let $0 < y_2 = z \leq \tilde{z}_R$ with $t_2 = 1$ and $t_1, t_3, y_1, y_3 = 0$ such that $y_2 = \alpha_{\check{i}-1} \tilde{z}_{\check{i}-1} + \alpha_{\check{i}} \tilde{z}_{\check{i}}$, $\alpha_{\check{i}-1} + \alpha_{\check{i}} = 1 = t_2$, and $-\alpha_{\check{i}-1} \Phi(\tilde{z}_{\check{i}-1}) - \alpha_{\check{i}} \Phi(\tilde{z}_{\check{i}}) \geq \zeta'$ for some $\check{i} \in [R]$ and $\alpha_j = 0$ for all $j \in [R_0] \setminus \{\check{i} - 1, \check{i}\}$.

Thus there exists an index $\check{i} \in [R]$ that satisfies $-\underline{\Phi}(z; \tilde{z}) = -g_{\check{i}} y_2 - g_{\check{i}}^0 \geq \zeta'$. Hence, $-\underline{\Phi}(z; \tilde{z}) = \max_{i \in [R]} -g_i y_2 - g_i^0 \geq \zeta'$ is satisfied.

iii) Finally, $y_3 = z > \tilde{z}_R$ with $t_3 = 1$, $t_1, t_2, y_1, y_2 = 0$ indicates that $-\Phi(\tilde{z}_R) = -\underline{\Phi}(z; \tilde{z}) \geq \zeta'$.

Additionally, since all of these cases allow ζ' to be at most 0, using the definition of $\underline{\Phi}(z; \tilde{z})$ in (46), $-\underline{\Phi}(z; \tilde{z}) \geq \zeta'$ is satisfied. $\square$

B.6.5. Proof of Lemma 8. To prove the claim, following [29, Definition 1.3] we first show that $\mathcal{F}(0) := \{(x, \beta) \mid x \in \mathcal{X}, \beta \in \overline{\mathfrak{B}}, J_k(s_l) \leq 0, s_l \in \mathcal{S}_\ell, k \in [\hat{K}]\}$ is uniformly compact near 0 if the following condition holds: there is a neighborhood $[-\varepsilon, \varepsilon]$ of 0 for some $\varepsilon > 0$ such that the closure of the set $\bigcup_{\tau' \in [-\varepsilon, \varepsilon]} \mathcal{F}(\tau')$ is compact, i.e., closed and bounded. Note that by Assumption 3, $\bigcup_{\tau' \in [-\varepsilon, \varepsilon]} \mathcal{F}(\tau')$ is not empty. Additionally, by Lemma 7, $J_k(s_l)$ is continuous everywhere in $(x, \beta) \in \mathcal{X} \times \overline{\mathfrak{B}}$ for any $s_l \in \mathcal{S}_\ell, k \in [\hat{K}]$. Then, given any $\tau' \in [-\varepsilon, \varepsilon]$, each convergent sequence $\{(x_t, \beta_t)\}$ with $(x_t, \beta_t) \in \mathcal{F}(\tau') \forall t$ implies that $(\bar{x}, \bar{\beta}) := \lim_{t \rightarrow \infty} (x_t, \beta_t) \in \mathcal{F}(\tau')$. Thus, $\mathcal{F}(\tau')$ also includes all its limit points $(\bar{x}, \bar{\beta})$. Therefore, $\bigcup_{\tau' \in [-\varepsilon, \varepsilon]} \mathcal{F}(\tau')$ is closed. Additionally, since $\mathcal{X}$ and $\overline{\mathfrak{B}}$ are bounded (since compact) set, $\mathcal{X} \cap \{(x, \beta) \mid \beta \in \overline{\mathfrak{B}}, J_k(s_l) \leq \tau', s_l \in \mathcal{S}_\ell, k \in [\hat{K}]\} \subseteq \mathcal{X} \times \overline{\mathfrak{B}}$ is also bounded and this is true for every $\tau' \in [-\varepsilon, \varepsilon]$. Therefore, $\bigcup_{\tau' \in [-\varepsilon, \varepsilon]} \mathcal{F}(\tau')$ is compact.

Next, we show that the MFCQ condition holds at the optimal solution $(x_*^{(\ell)}, \beta_*^{(\ell)})$ and thereby continuity of the optimal value function $V_\ell^*(0)$ at 0. Let us define the indicator $\delta_{\mathcal{X}}(x)$ and $\delta_{\overline{\mathfrak{B}}}(\beta)$ respectively for the set $\mathcal{X}$ and $\overline{\mathfrak{B}}$ and the Lagrangian for some $q \in \mathbb{R}^{\hat{K} \times |\mathcal{S}_\ell|}$:

$$\mathcal{L}(x, \beta, u, q) = c^\top x + \delta_{\mathcal{X}}(x) + \delta_{\overline{\mathfrak{B}}}(\beta) + q^\top H(x, \beta).$$

Since $(x_*^{(\ell)}, \beta_*^{(\ell)})$ is a local minimizer, continuous differentiability of $G(x; \cdot)$ implies continuous differentiability of $J((x, \beta); s)$ for all $(x, \beta) \in \mathcal{X} \times \overline{\mathfrak{B}}$ by Lemma 7, $\exists \check{q} \geq 0$, normal-cone multipliers $\check{v} \in \partial_x \delta_{\mathcal{X}}(x_*^{(\ell)}) = \mathcal{N}_{\mathcal{X}}(x_*^{(\ell)})$ and $\check{u} \in \partial_\beta \delta_{\overline{\mathfrak{B}}}(\beta_*^{(\ell)}) = \mathcal{N}_{\mathcal{S}}(\beta_*^{(\ell)})$ such that primal feasibility, dual feasibility, and complementary conditions hold with

$$0 = (c, 0) + (\check{v}, \check{u}) + \nabla_{x, \beta} H^\top(x_*^{(\ell)}, \beta_*^{(\ell)}) \check{q}.$$

Additionally, by Assumption 4, there are no non-zero (q, v) with $v \in \mathcal{N}_{\mathcal{X}}(\check{x})$, $q \geq 0$ satisfying

$$0 = (\check{v}, \check{u}) + \nabla_{x, \beta} H^\top(x_*^{(\ell)}, \beta_*^{(\ell)}) \check{q}. \quad (74)$$

Then Gordan's alternative for cones (Thm. 6.4 in [69]; Thm. 2.4.9 in [4]) gives

$$\exists \mathbf{z} \in \mathcal{T}(\mathbf{x}_*^{(\ell)}, \beta_*^{(\ell)}) \text{ such that } \nabla_{\mathbf{x}, \beta} \mathbf{H}(\mathbf{x}_*^{(\ell)}, \beta_*^{(\ell)})^\top \mathbf{z} < 0,$$

where $\mathcal{T}(\mathbf{x}_*^{(\ell)}, \beta_*^{(\ell)}) := \text{cl}\{\alpha(\mathbf{t} - (\mathbf{x}_*^{(\ell)}, \beta_*^{(\ell)})) \mid \alpha > 0, \mathbf{t} \in (\mathcal{X} \times \beta)\}$ is the Bouligand tangent cone of $\mathcal{X} \times \beta$ at $(\mathbf{x}_*^{(\ell)}, \beta_*^{(\ell)})$. Thus, the Mangasarian–Fromovitz constraint qualification (MFCQ) holds at $(\mathbf{x}_*^{(\ell)}, \beta_*^{(\ell)})$. Then, by Theorem 2.6 of [29], optimal value function $V_\ell^*(0)$ corresponding to master problem is continuous 0.

Finally, Theorem 6 and Proposition 10 ensure that using $O(\frac{1}{\sqrt{\epsilon_j}} \sqrt{\log(\frac{1}{\epsilon_j})})$ breakpoints, a ϵ_j -feasible solution $(\mathbf{x}^{(\ell)}, \beta^{(\ell)})$ can be obtained. Furthermore, continuity of $V_\ell^*(0)$ at 0 implies that for any given $\epsilon_j > 0$, $\exists \tau^+(\epsilon_j)$ such that $|V_\ell^*(0) - V_\ell^*(\tau)| \leq \epsilon_j$ can be achieved for some τ such that $0 < \tau \leq \tau^+(\epsilon_j)$, i.e., for some $\tau \in (0, \epsilon_j]$. Since $V_\ell^*(\epsilon_j) \leq \mathbf{c}^\top \mathbf{x}^{(\ell)}, \mathbf{c}^\top \mathbf{x}_*^{(\ell)} - \mathbf{c}^\top \mathbf{x}^{(\ell)} = V_\ell^*(0) - \mathbf{c}^\top \mathbf{x}^{(\ell)} \leq V_\ell^*(0) - V_\ell^*(\epsilon_j) \leq \epsilon_j$. $\square$

Remark 11. We note that strict interior, hence, MFCQ can also be shown by considering a perturbation to $\check{s}$ (see Bonnans and Shapiro [11] for details).

Appendix C: Piecewise linear approximation of a Gaussian CDF

Here, we present a breakpoint construction approach in Algorithm 2 that we followed in our numerical study. Such a construction scheme guarantees that $O(\frac{1}{\sqrt{\tau}} \sqrt{\log(\frac{1}{\tau})})$ breakpoints suffice to achieve piecewise-linear approximation accuracy τ ; see Theorem 6. We also illustrate the resulting outer and inner piecewise-linear approximations of the standard normal CDF $\Phi(z)$ in Figure 3

Algorithm 2 Breakpoints for approximating $\Phi(z)$

(a) Tangent approximation of $\Phi(z), z \geq 0$	(b) Secant approximation of $\Phi(z), z < 0$
Require: τ Ensure: $\mathcal{A}^R$ 1: Part 1: $0 \leq z \leq 1$ 2: $i \leftarrow 1, \check{z}_i \leftarrow 1, \mathcal{A}^R \leftarrow \mathcal{A}^R \cup \{\check{z}_i\}$ 3: $\check{z}_{i+1} = \check{z}_i - \sqrt{\frac{2\tau}{e^{-0.5} \sqrt{2\pi}}}$ 4: while $\check{z}_{i+1} > 0$ do 5: $\mathcal{A}^R \leftarrow \mathcal{A}^R \cup \{\check{z}_{i+1}\}, i \leftarrow i + 1$ 6: $\check{z}_{i+1} = \check{z}_i - \sqrt{\frac{2\tau}{\phi(\check{z}_i)\check{z}_i}}$ 7: end while 8: $\mathcal{A}^R \leftarrow \mathcal{A}^R \cup \{0\}, i \leftarrow i + 1$ 9: Reverse the order of $\mathcal{A}^R$ 10: $\mathcal{A}^R \leftarrow \mathcal{A}^R \setminus \{1\}$ 11: Part 2: $z \geq 1$ 12: $i \leftarrow i + 1, \check{z}_{i-1} \leftarrow 1, \mathcal{A}^R \leftarrow \mathcal{A}^R \cup \{\check{z}_{i-1}\}$ 13: $\check{z}_i = \check{z}_{i-1} + \sqrt{\frac{2\tau}{\phi(\check{z}_{i-1})\check{z}_{i-1}}}$ 14: while $1 - \Phi(\check{z}_i) < \tau$ do 15: $\mathcal{A}^R \leftarrow \mathcal{A}^R \cup \{\check{z}_i\}, i \leftarrow i + 1$ 16: $\check{z}_i = \check{z}_{i-1} + \sqrt{\frac{2\tau}{\phi(\check{z}_{i-1})\check{z}_{i-1}}}$ 17: end while 18: $\check{z}_R \leftarrow \check{z}_i$ 19: $\mathcal{A}^R \leftarrow \mathcal{A}^R \cup \{\check{z}_R\}$	Require: τ Ensure: $\mathcal{A}^L$ 1: Part 1: $-1 \leq z \leq 0$ 2: $i \leftarrow 1, \check{z}_{-i} \leftarrow -1, \mathcal{A}^L \leftarrow \mathcal{A}^L \cup \{\check{z}_{-i}\}$ 3: $\check{z}_{-i-1} = \check{z}_{-i} - 2\sqrt{\frac{2\tau}{\phi(\check{z}_{-i})\check{z}_{-i}}}$ 4: while $\check{z}_{-i-1} < 0$ do 5: $\mathcal{A}^L \leftarrow \mathcal{A}^L \cup \{\check{z}_{-i-1}\}, i \leftarrow -i - 1$ 6: $\check{z}_{-i-1} = \check{z}_{-i} - 2\sqrt{\frac{2\tau}{\phi(\check{z}_{-i})\check{z}_{-i}}}$ 7: end while 8: Reverse the order of $\mathcal{A}^L$ 9: $\mathcal{A}^L \leftarrow \mathcal{A}^L \setminus -1$ 10: Part 2: $z \leq -1$ 11: $i \leftarrow i + 1, \check{z}_{-i+1} \leftarrow -1, \mathcal{A}^L \leftarrow \mathcal{A}^L \cup \{\check{z}_{-i+1}\}$ 12: $\check{z}_{-i} = \check{z}_{-i+1} + 2\sqrt{\frac{2\tau}{\phi(\check{z}_{-i+1})\check{z}_{-i+1}}}$ 13: while $\Phi(\check{z}_{-i}) < \tau$ do 14: $\mathcal{A}^L \leftarrow \mathcal{A}^L \cup \{\check{z}_{-i}\}, i \leftarrow i + 1$ 15: $\check{z}_{-i} = \check{z}_{-i+1} + 2\sqrt{\frac{2\tau}{\phi(\check{z}_{-i+1})\check{z}_{-i+1}}}$ 16: end while 17: $\check{z}_{-L} \leftarrow \check{z}_{-i}$ 18: $\mathcal{A}^L \leftarrow \mathcal{A}^L \cup \{\check{z}_{-L}\}$

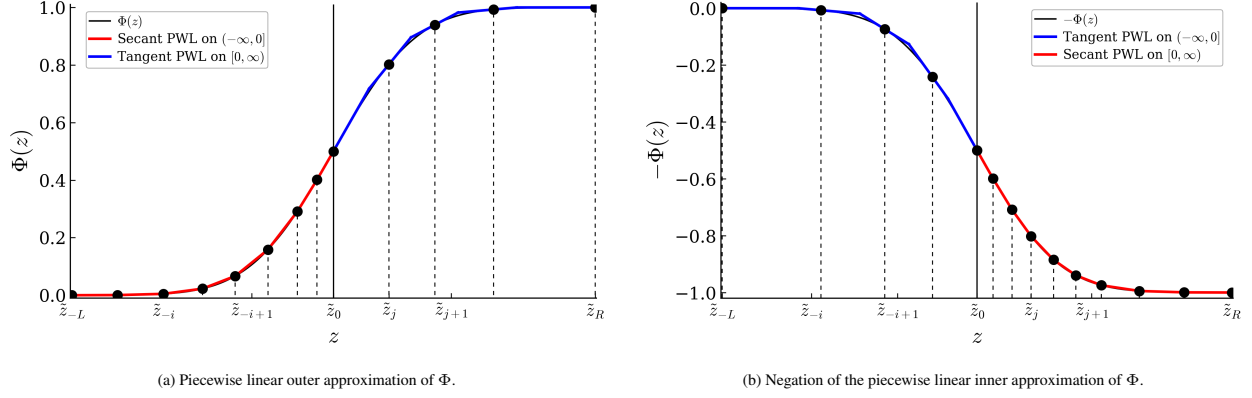


Figure 3 Piecewise linear approximations used in the formulation.

Appendix D: Computational Study Details

D.1. Data Preparation Details

For the EV application, the nominal distribution represents daily hourly charging demand and is estimated from session-level charging records. The dataset, sourced from Figshare [26], contains one record per charging session, including start and end times, delivered energy in kWh, user and charging-post identifiers, location or district name, charging costs in Yuan, termination reason, and weather attributes such as temperature, relative humidity, and precipitation. Its structure is consistent with the previously published high-resolution Jiaxing EV charging transaction dataset [91]. Since aggregating the full two-year sample can obscure temporal heterogeneity and collapse the demand distribution toward a dominant mode, we partition the data by quarter and use Q1 observations, i.e., January–March of 2020 and 2021, for the study.

The session-level records are converted into daily 24-dimensional demand vectors. For a session s with delivered energy E_s , let Δ_{st} denote its overlap duration with hour $t \in [24]$. The energy allocated to hour t is $E_s \Delta_{st} / \sum_{t' \in [24]} \Delta_{st'}$. Summing these allocated contributions over all sessions occurring on the same day yields a per-day hourly-demand vector. The data are processed in Python using `pandas`: start and end timestamps are parsed as date-time variables, the energy field is converted to numeric values, and records with missing or nonpositive delivered energy are removed. After a fixed random shuffle, 60% of the daily vectors are used for estimation, while the remaining days are reserved as holdout data for evaluating out-of-sample satisfaction probability.

The objective $\min \sum_{t=1}^{24} (C^o(x_t) + c_t x_t)$ having a *linear (fixed) term* and a *piecewise-linear (operational) term* are built as follows. The linear term is $\sum_{t=1}^{24} c_t x_t$, where c_t is the time-of-use cost (Yuan/kWh) in hour t ; the 24 values c_t are drawn uniformly from $[300, 350]$. The operational cost in hour t is a convex piecewise-linear function $C^o(x_t)$ of the capacity fraction x_t , built from a pivot $(\tau_0, y_0) = (0.8, 200)$, 16 segments (8 left of the pivot, 8 right), base slope $\eta = 1$, and growth factor $\gamma = 2$: left-segment slopes are $-\eta\gamma^i$ for $i = 0, \dots, 7$ (reversed so the steepest negative slope is near τ_0), right-segment slopes are $\eta\gamma^i$ for $i = 0, \dots, 7$; knot positions are even-spaced on $[0, \tau_0]$ and $[\tau_0, 1]$; intercepts are determined by continuity at

the pivot. Thus $C^o(x_t) = \max_{t=1}^{16} (a_t x_t + b_t)$ with (a_t, b_t) from this construction. Model parameters of the considered EV model are further summarized in Table 4 below.

Table 4 Polytopic constraints and coefficients for the EV chance-constrained model.

Parameter	Value	Description
<i>Variable bounds</i>		
$x_{\min}, x_{\max}$	0, 1	Bounds on each x_t (fraction of capacity in hour t).
<i>Chance constraint and throughput</i>		
$P_{\max}$	560.18	Peak capacity (0.98 quantile of daily total demand).
E_{Total}	1,151,359.45	Total energy delivered (0.98 quantile of daily total delivered energy)
$E_{\min}$	$E_{Total}/P_{\max}$	Minimum daily energy delivery (throughput target);
<i>Objective (two-term structure)</i>		
Linear term	$\sum_{t=1}^{24} c_t x_t$	$c_t \in [300, 350]$ Yuan/kWh (time-of-use); 24 values per instance.
PWL term	$C^o(x) = \max_{a \in \mathcal{A}_- \cup \mathcal{A}_+} f(a)$ where $f(a) = \{a x + (y_0 - a x_0)\}$	Pivot $(\tau_0, y_0) = (0.8, 200)$; $\mathcal{A}_{\pm} = \{\pm 2^i : i = 0, \dots, 7\}$. 8 left + 8 right segments;

D.2. Master Iteration Runtime Scheduling Strategy

Because every admitted cut increases the mixed-integer master problem constraint pool, the master solver uses a highly customized stopping rule that balances tractability and solution quality across iterations. Early termination is never allowed before a feasible incumbent has been found. Once feasibility is achieved, the solver may stop only after a preset soft time budget has elapsed and the incumbent has reached the prescribed mixed-integer optimality gap of 0.1%; otherwise, the solve continues until an iteration-dependent hard cap is reached. This hard cap increases with the iteration count, particularly 2 hours for iterations 0 and 1, then 4, 6, and 8 hours. The attained incumbent gap is recorded at the end of each master solve and carried into the next iteration as a state variable for the stopping logic. This schedule is designed to reflect the systematic growth of the master problem as cuts accumulate: early iterations are kept relatively inexpensive, whereas substantially more time is allowed for the later iterations so that the performance gain attributable to additional cuts can be assessed under a solve budget commensurate with the enlarged constraint set. Because each CDR master solve was stopped at its iteration-wise time budget regardless of whether the desired optimality gap of 0.1% was met, the zeroth CDR solution can provide a solution to the FDR problem. Accordingly, the zeroth iteration problem is also solved separately outside the scheduled CDR iterations. For this standalone solve, we impose the same 0.1% gap target and a 22-hour time limit, equal to the cumulative budget used by CDR across four iterations, to obtain a fair benchmark for comparison with the CDR objective and OSS probability achieved at its fourth-iteration solution.

D.2.1. Iteration-wise cutting surface algorithmic progression. Since CDR models are harder to solve, and we do not solve them to optimality (only four master iterations are performed), we discuss the iteration-wise behavior of the cutting-surface algorithm to investigate the properties of the violated constraints identified in separation steps when solving this model. Figure 4 shows that violation magnitudes are not necessarily monotonic; however, they stabilize in most reported settings.

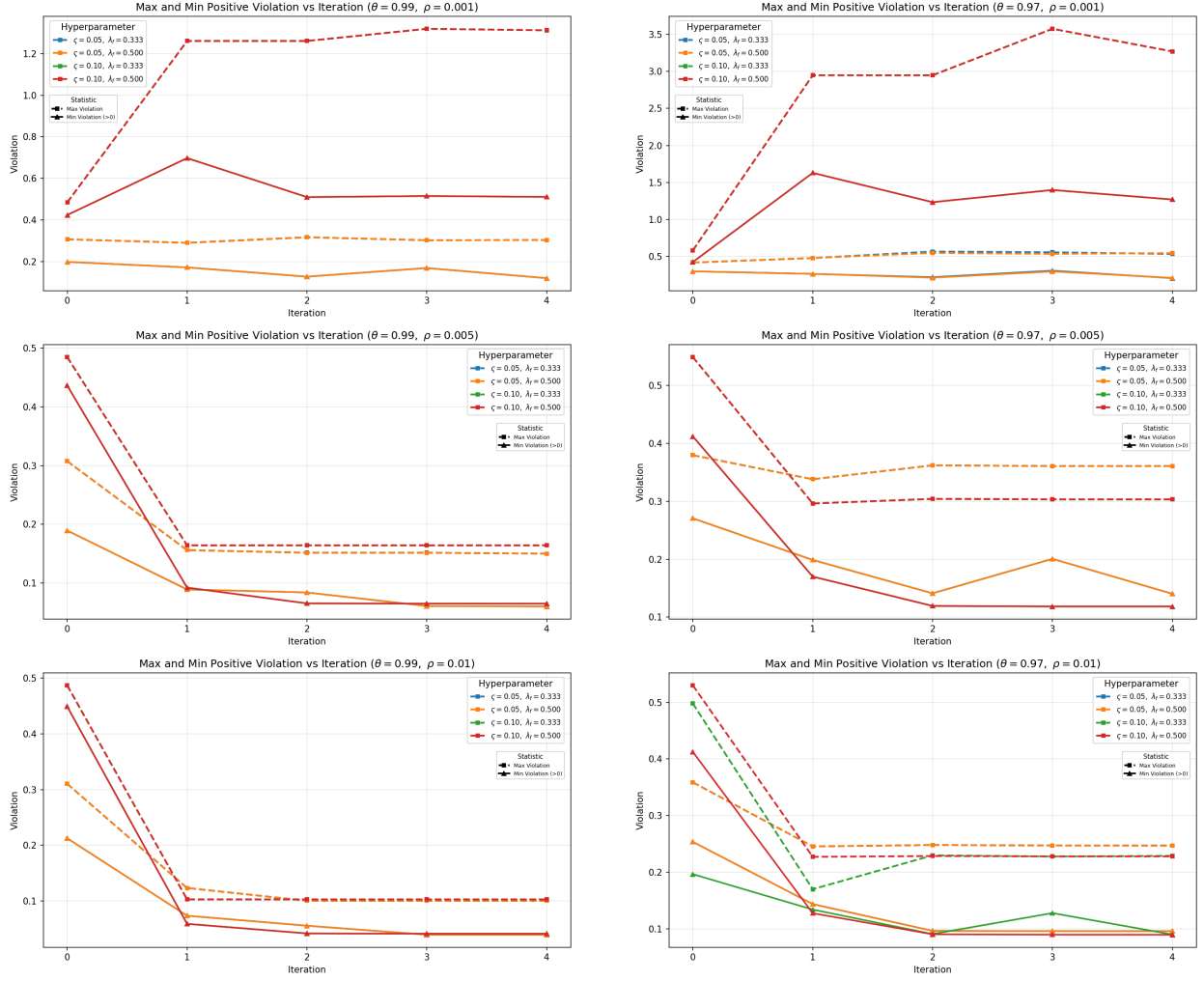


Figure 4 Maximum and minimum violations throughout iterations for all $(\zeta, \lambda_f) \in \{(0.05, 0.333), (0.05, 0.5), (0.1, 0.333), (0.1, 0.5)\}$, reported in the same plot. The first, second, and third rows in this figure correspond to $\rho = 0.001$, $\rho = 0.005$, and $\rho = 0.01$, respectively. Within each row, the left and right panels correspond to $\theta = 0.99$ and $\theta = 0.97$, respectively. Plots for $\theta = 0.95$ are omitted due to their similarity with those from $\theta \in \{0.97, 0.99\}$.

For $\rho \in \{0.005, 0.01\}$, the maximum positive violation reduces rapidly within two iterations and then remains small. This pattern is especially pronounced for the more stringent service levels. At $\theta = 0.99$, for example, the final maximum positive violation is about 0.15 when $\rho = 0.005$ and about 0.10 when $\rho = 0.01$, with minimum positive violations around 0.06 and 0.04, respectively. Thus, in the moderate- and large-radius regimes, the cutting surface algorithm appears to identify the important cut locations (the mean and covariance of the unknown Gaussian in the mixture) early, and later iterations mainly refine them. The only regime in which this stabilization is not as apparent is the small-radius case $\rho = 0.001$ combined with the larger mean-support range $\zeta = 0.10$ and the service level θ is small. This indicates that an enlarged mean support set can result in a harder separation problem, even when the radius of the Wasserstein ball is small, and the target service level is not stringent.

D.3. Plots from Numerical Findings

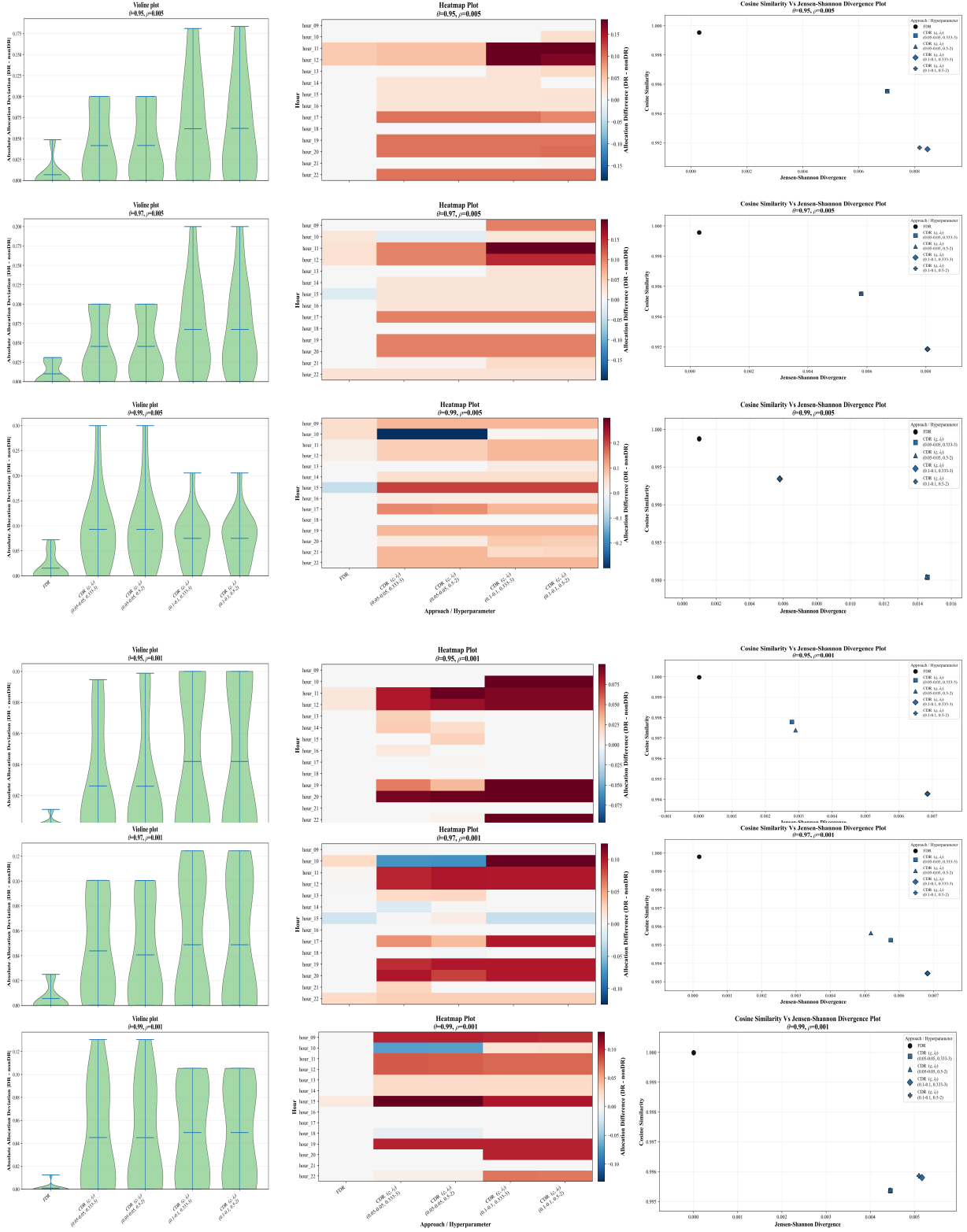


Figure 5

Comparison of finite- and continuous-support solutions for $(\varsigma, \lambda_{\min}) \in \{(0.05, 0.333), (0.05, 0.5), (0.1, 0.333), (0.1, 0.5)\}$. The upper and lower row blocks correspond to $\rho = 0.005$ and $\rho = 0.001$, respectively, with rows in each block ordered from top to bottom by $\theta = 0.95, 0.97, 0.99$. In each row, the left, middle, and right panels show, respectively, the violin plot of absolute deviations from the nominal solution, the heatmap of allocation differences relative to the nominal solution, and the cosine-similarity–Jensen–Shannon-divergence scatter plot.