2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Distributionally Robust Decision Making Leveraging Conditional Distributions

Yuxiao Chen, Jip Kim, and James Anderson

Abstract— **Distributionally robust optimization (DRO) is a powerful tool for decision making under uncertainty. It is particularly appealing because of its ability to leverage existing data. However, many practical problems call for decision-making with some auxiliary information, and DRO in the context of conditional distributions is not straightforward. We propose a conditional kernel distributionally robust optimization (CKDRO) method that enables robust decision making under conditional distributions through kernel DRO and the conditional mean operator in the reproducing kernel Hilbert space (RKHS). In particular, we consider problems where there is a correlation between the unknown variable y and an auxiliary observable variable x. Given past data of the two variables and a queried auxiliary variable, CKDRO represents the conditional distribution $\mathbb{P}(y|x)$ as the conditional mean operator in the RKHS space and quantifies the ambiguity set in the RKHS as well, which depends on the size of the dataset as well as the query point. To justify the use of RKHS, we demonstrate that the ambiguity set defined in RKHS can be viewed as a ball under a metric that is similar to the Wasserstein metric. The DRO is then dualized and solved via a finite dimensional convex program. The proposed CKDRO approach is applied to a generation scheduling problem and shows that the result of CKDRO is superior to common benchmarks in terms of quality and robustness.**

I. Introduction

Distributionally robust optimization (DRO) has attracted great attention recently as a framework for decision making under uncertainty. Under a classic stochastic decision making setup, the goal is to minimize the expected cost given a distribution of the uncertainty. In practice, this might be overly restrictive as the uncertainty distribution is often not fully known. This concern motivates DRO, which seeks to minimize the worst case expected cost (maximize reward) under a set of distributions referred to as the ambiguity set. If the ambiguity set is large enough to contain the true distribution, the resulting decision provides an upper bound on achievable real-world performance. The simplest (and most conservative) ambiguity set contains any distribution on the support of the uncertainty; in which case DRO reduces to a robust optimization [1]. Several popular constructions of the ambiguity set are based on metrics that measure the difference between probability distributions, such as the Wasserstein metric [2], [3], ϕ-divergence [4], and moments [5]. See the review paper [6] for a comprehensive summary. DRO is also closely related to risk-averse optimization and planning [7] as all coherent risk measures have a dual form that is a DRO problem over the corresponding ambiguity set.

Depending on the probability space, a DRO problem can be discrete (the random variable has discrete support), or continuous (the support is continuous). Computationally, the continuous problem is much more complex, and many methods rely on sampling to gain computation tractability [2], [3], [8]. Most existing data-driven methods assume i.i.d. sampling from the unknown distribution and establish the ambiguity set based on the empirical distribution of the samples. An i.i.d. assumption then provides a probabilistic guarantee that the true distribution will be contained inside an ambiguity set with high probability via concentration inequalities [9], [10].

Unfortunately, the i.i.d. condition does not hold for decision making tasks with conditional distributions. To motivate this formulation, consider the following two examples:

- Generation scheduling [11], [12]: to provide electricity at a minimal cost, generation scheduling needs to be robustified against uncertainties in power demand, which is dominated by ambient factors such as weather conditions. A power system operator can exploit historical data and account for the correlation between the electricity demand and weather forecast to make more reliable and cost-efficient generation scheduling.
- Autonomous driving [13], [14], [15]: to plan safe motions for the autonomous vehicle, we need to reason about the surrounding human drivers' actions, which strongly depend on the scenario. Given past data of how drivers react to different scenarios, for a particular scenario, the motion planning of the autonomous vehicle needs to leverage the distribution of possible actions of surrounding human drivers conditioned on the scenario.

Such problems can be formulated as DRO under ambiguity sets defined on some conditional distributions, yet it is unclear how to quantify the ambiguity set about the conditional distribution and how to leverage past data as the i.i.d. condition does not hold. In such problems, both the conditional mean and uncertainty characterization is critical, motivating methods like quantile regression [16]. A two step process can be applied where the first step uses existing regression tools to predict the distribution of the uncertainty given the auxiliary variables, followed by a subsequent stochastic

Yuxiao Chen is with Nvidia Corporation, Santa Clara, CA, 95051, Email: yuxiaoc@nvidia.com. Jip Kim and James Anderson are with Columbia University, New York, NY, 10027, Emails: {jk4564,ja3451}@columbia.edu. JA is partially supported by grants from the NSF and DOE: EPCN-2144634 and DE-SC0022234 respectively.

Jip Kim is partially supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technology Office Award Number DE-EE0008769. The views expressed herein do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

optimization step given the prediction [17]. The difficulties of DRO with conditional distributions are: (i) The conditional distribution is hard to characterize from data, especially in continuous domains, (ii) It is difficult to determine the size of the ambiguity set for conditional distributions.

The proposed CKDRO method is inspired by two existing tools. First, the kernel embedding method in a reproducing kernel Hilbert space (RKHS) provides a convenient characterization of the conditional distribution with conditional mean operators, and its empirical version can be computed directly from data samples. For reference, see the papers by Song *et al.* [18], Fukumizu *et al.* [19], and the review paper [20]. On a different front, DRO with an RKHS was recently studied in [21] where the authors propose to represent ambiguity sets with mean operator norms defined on an RKHS. Based on these ideas, we propose conditional kernel distributionally robust optimization (CKDRO) with the following contributions: (i) We extend kernel DRO to problems with conditional distributions for data-driven robust conditional decision making, (ii) We establish connections between ambiguity set in RKHS and the Wasserstein metric providing a better understanding the kernel DRO setup, (iii) We propose criteria to determine the ambiguity set for conditional DRO, (iv) We apply the CKDRO method on an optimal power flow (OPF) problem and showcase its benefit over benchmarks.

Section II reviews some key concepts and tools, section III presents the CKDRO method and establishes the connection to DRO through the Wasserstein metric, and section IV presents the application of CKDRO on an OPF problem.

Nomenclature: $\mathbb{R}_+$ denotes the non-negative real line, $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors ($\langle x, y \rangle = x^\intercal y$) or two functions ($\langle f, g \rangle = \int f(x)g(x)dx$). For a random variable X, we use the calligraphic $\mathcal{X}$ denote its domain, $\mathbb{E}_Q[f(X)]$ denotes the expectation of a function $f : \mathcal{X} \to \mathbb{R}^n$ under the probability distribution Q. When X has a fixed distribution, e.g., it's sampled from a given distribution, we simply write $\mathbb{E}_X$ for the expectation. $\mathcal{M}(\mathcal{X})$ denotes the space of all probability distributions Q supported on $\mathcal{X}$ with $\mathbb{E}_Q[||X||] < \infty$. All vector norms are ℓ_2-norms unless otherwise specified.

II. Preliminaries

A. Hilbert space embedding

Consider a random variable $X \in \mathcal{X}$. X is endowed with some σ-algebra $\mathcal{A}$, and $\mathcal{P}$ denotes the space of probability distributions over X. A *reproducing kernel Hilbert space* (RKHS) $\mathcal{F}$ on $\mathcal{X}$ with a symmetric kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ whose inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ satisfies the reproducing property:

$$\langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{F}} = f(x),$$
$$\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{F}} = k(x, x').$$

The feature map $k(x, \cdot)$ is also denoted as $\varphi(x)$. A kernel k is *positive definite* if $\forall x_1, ...x_N \in \mathcal{X}, \forall c_1, ...c_N \in \mathbb{R}$, $\sum_{i,j=1}^N c_i c_j k(x_i, x_j) \geq 0$, or in matrix form, $K \succeq 0$, where $K_{ij} = k(x_i, x_j)$. Common choice of kernels include the polynomial kernel $k(x, x') = (x^\intercal x')^d$, the Gaussian RBF kernel $k(x, x') = \exp(-\lambda ||x - x'||^2)$, and the Laplacian kernel $k(x, x') = \exp(-\lambda ||x - x'||_1)$. For every positive definite kernel k, there exists a unique RKHS with k as its kernel up to isometry, and conversely, for every RKHS, its kernel is unique and positive definite [20]. A kernel is called *translation invariant* if $\forall x_1, x_2, x_1', x_2' \in \mathcal{X}$ with $x_1 - x_2 = x_1' - x_2'$, $k(x_1, x_2) = k(x_1', x_2')$. Both the Gaussian kernel and the Laplacian kernel are translation invariant, while the polynomial kernel is not.

The distance between points in an RKHS is also defined via the kernel: $d(x, x') = \sqrt{k(x, x) - 2k(x, x') + k(x', x')}$.

So far, an RKHS is not related to any probability distribution. The central concept of kernel embedding is the expectation operator; $\mu_Q := \mathbb{E}_Q[\varphi(X)]$, which defines the expectation of the feature map under distribution Q. When the context is clear, we also use μ_X to denote the expectation operator when X follows a fixed distribution. Assuming $\mathbb{E}_X[k(X, X)] < \infty$, μ_X is itself an element of $\mathcal{F}$. Given the reproducing property, we have $\forall f \in \mathcal{F}$, $\langle \mu_X, f \rangle_{\mathcal{F}} = \mathbb{E}_X[\langle \varphi(X), f(X) \rangle_{\mathcal{F}}] = \mathbb{E}_X[f(X)]$. Its empirical estimate given a set of samples is simply

$$\hat{\mu}_X := \frac{1}{N} \sum_{i=1}^N \varphi(x_i),$$

where $\{x_i\}_{i=1}^N$ is the sample set, assumed to have been drawn i.i.d. from P. The norm on $\mathcal{F}$ is defined as:

$$||f||_{\mathcal{F}}^2 = \langle f, f \rangle_{\mathcal{F}}.$$

If f has a discrete representation: $f = \sum_i^N \alpha_i \varphi(x_i)$, then by the reproducing property,

$$||f||_{\mathcal{F}}^2 = \sum_{i,j=1}^N \alpha_i \alpha_j k(x_i, x_j) = \alpha^\intercal K \alpha.$$

Now consider a second random variable Y with domain $\mathcal{Y}$, and let $\mathcal{B}$ and $\mathcal{Q}$ denote the σ-algebra and the space of all probability distributions over Y respectively. An RKHS $\mathcal{G}$ is defined on $\mathcal{Y}$, induced by the positive definite kernel $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Let $\phi(y) = l(y, \cdot)$ be its feature map, the (uncentered) covariance operator is defined as

$$C_{YX} := \mathbb{E}_{YX}[\phi(Y) \otimes \varphi(X)],$$

where $\otimes$ is the tensor product. C_{YX} can be viewed as an operator from $\mathcal{F}$ to $\mathcal{G}$ in the sense that for any $f \in \mathcal{F}, \forall g \in \mathcal{G}$, $\langle g, C_{YX} f \rangle_{\mathcal{G}} = Cov[f(X), g(Y)]$. Given samples $\{(x_i, y_i)\}_{i=1}^N$, the empirical evaluation of the covariance operator is

$$\hat{C}_{YX} = \frac{1}{N} \Upsilon \Phi^\intercal$$

where $\Upsilon = [\varphi(x_1), \varphi(x_2)..., \varphi(x_N)]$ is the feature matrix for X, $\Phi = [\phi(y_1), \phi(y_2)..., \phi(y_N)]$ is the feature matrix for Y.

Lemma 1 ([19]). *If $\mathbb{E}_{YX}[g(Y)|X = \cdot] \in \mathcal{G}$ for any $g \in \mathcal{G}$, then*

$$C_{XX} \mathbb{E}_{YX}[g(Y)|X = \cdot] = C_{YX} g.$$

Consequently, [18] shows that if C_{XX} is invertible, then we can define an operator $\mathcal{U}_{Y|X} : \mathcal{F} \to \mathcal{G}$ where $\forall x \in \mathcal{X}$, $\mu_{Y|x} := \mathcal{U}_{Y|X}\varphi(x)$, and $\mathcal{U}_{Y|X}$ is defined as

$$\mathcal{U}_{Y|X} = C_{XX}^{-1} C_{YX}.$$

Similar to the unconditioned expectation operator, $\mu_{Y|x}$ can be evaluated empirically with samples.

Lemma 2 ([18], Theorem 5)**.** *Let $k_x := \Upsilon^\intercal \varphi(x)$, then $\hat{\mu}_{Y|x}$ is the empirical conditioned expectation operator, and is estimated as $\hat{\mu}_{Y|x} = \Phi(K + \lambda N I_N)^{-1} k_x$, where $\lambda > 0$ is the regularization parameter that prevents overfitting.*

B. Distributionally robust optimization

Distributionally robust optimization considers the following problem:

$$\min_{\eta} \max_{Q \in \mathscr{A}} \mathbb{E}_Q[q(\eta, X)], \tag{1}$$

where $\eta \in H$ is the decision variable, X is the random variable, $\mathscr{A} \subseteq \mathcal{P}$ is the ambiguity set that contains all the distribution Q over X under consideration. As mentioned in the introduction, the ambiguity set is typically defined with some metric that measures the difference between distributions, such as the Wasserstein metric or ϕ-divergence. In particular, the Wasserstein metric is defined on $\mathcal{M}(\mathcal{X})$ as follows.

Definition 1 (Wasserstein metric)**.** The Wasserstein metric $d_W : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ is defined as

$$\begin{aligned} d_W(Q_1, Q_2) := \inf_{\Pi} & \int_{\mathcal{X}^2} ||x_1 - x_2|| \Pi(x_1, x_2) dx_1 dx_2 \\ \text{s.t.} & \int_{\mathcal{X}} \Pi(x_1, x_2) dx_2 = Q_1(x_1), \\ & \int_{\mathcal{X}} \Pi(x_1, x_2) dx_1 = Q_2(x_2). \end{aligned}$$

The Wasserstein metric is essentially defined by the optimal transport problem, where Π is a joint distribution supported on $\mathcal{X}^2$ with the two marginal distributions being Q_1 and Q_2.

An important equivalent definition was proposed in [22], which is used later in this paper.

Lemma 3 (Kantorovich-Rubinstein [22])**.**

$$d_W(Q_1, Q_2) = \sup_{L(f) \le 1} \left\{ \int_{\mathcal{X}} f(x) Q_1(dx) - \int_{\mathcal{X}} f(x) Q_2(dx) \right\},$$

where $L(f)$ denotes the Lipschitz constant of the function f.

Ambiguity sets can then be equivalently defined as $\mathscr{A} = \{Q \mid d_W(Q_0, Q) < \epsilon\}$, where Q_0 is the nominal distribution, which can be given, or estimated from data. Work in [2] presented a convex program that solves the DRO with the above ambiguity set where Q_0 is the empirical distribution of the samples $Q_0 = \frac{1}{N}\delta(x_i)$, where $\delta(\cdot)$ is the Dirac measure.

Kernel DRO was recently proposed in [21] as a convenient data-driven DRO framework that leverages the RKHS. The distributions are represented with kernel embeddings in an RKHS, and the ambiguity set is typically chosen as an RKHS ball: $\mathcal{C} := \{\mu \mid ||\mu - \mu_0||_{\mathcal{F}} \le \epsilon\}$. For clarity, we use $\mathcal{C}$ to denote an ambiguity set defined with respect to the expectation operator, and $\mathscr{A}$ to denote an ambiguity set defined by a probability distribution. They are linked by: $\mathscr{A} = \{P \in \mathcal{P} \mid \int_{\mathcal{X}} \varphi(x) P(dx) \in \mathcal{C}\}$.

There are two common strategies for solving the DRO problem (1); the cutting plane technique [23] and the dual formulation, similar to the ones used in robust optimization [1]. When the inner maximization problem is a concave problem, the dual formulation offers a convex program for the whole DRO.

For clarity, we shall leave the solution of the kernel DRO to the next section, and present it together with the conditional kernel DRO.

III. Conditional kernel DRO

In this section, we present the formulation of the conditional kernel DRO.

A. Problem formulation

Suppose we are given a training data set $\{x_i, y_i\}_{i=1}^N$ consisting of X, Y pairs, where X is an auxiliary variable and the distribution of X and Y are correlated. The conditional DRO problem then tries to minimize the worst case expectation of the cost given an realization of X. In kernel DRO, we are interested in an optimization problem where the auxiliary variable $X = x$ is given. The ambiguity set is defined by the expectation operator $\mu_{Y|x}$, and the DRO is defined as

$$\min_{\eta} \max_{\mu \in \mathcal{C}[x]} \langle \mu, q(\eta, \cdot) \rangle,$$

where $\mathcal{C}[x] \subseteq \mathcal{G}$ is an ambiguity set on the conditional expectation operator $\mu_{Y|X}$ at $X = x$, and $q : H \times \mathcal{Y} \to \mathbb{R}$ is the cost function.

B. Construction of the ambiguity set

We first discuss the construction of the ambiguity set $\mathcal{C}[x]$. Two sources of uncertainty about $\mu_{Y|x}$ are considered. The first is due to the estimation of $\mu_{Y|x}$ from data, the second is due to unseen data.

Lemma 4 (Theorem 6 in [18])**.** *Assuming $\varphi(x)$ is in the range of $\mathcal{C}_{XX}$, $||\hat{\mu}_{Y|x} - \mu_{Y|x}||_{\mathcal{G}}$ converges to zero with rate $O_p(\lambda^{\frac{1}{2}} + (N\lambda)^{-\frac{1}{2}})$ where λ is the regularization term for RKHS norm.* [1]

Faster convergence rates are available under additional assumptions about the underlying distribution in [24], but for simplicity, we stick to the convergence rate provided in Lemma 4. If we choose $\lambda \sim N^{-\frac{1}{4}}$, the convergence rate is $O_p(N^{-\frac{1}{4}})$.

Note that the above convergence result is on $\mu_{Y|X}$, not on any particular $\mu_{Y|x}$, where the former is an average of the later weighted by P, the distribution of X. A naive choice of the ambiguity set would be $\mathscr{A} = \{\mu \mid ||\mu - \hat{\mu}_{Y|x}||_{\mathcal{G}} \le \gamma N^{-\frac{1}{4}}\}$, where $\hat{\mu}_{Y|x}$ is the empirical conditional expectation

[1] O_p stands for probabilistic convergence rate.

operator evaluated at $X = x$, N is the number of data points in the training set. However, note that the size of the ambiguity set does not change with x, indicating that the uncertainty on $\mu_{Y|x}$ is the same whether the queried x is close to the samples or far away from the samples.

To account for the difference in queried x, we propose to use a fictitious sample at $X = x$ to help measure the uncertainty. Now suppose that a fictitious sample of Y drawn from $\mathbb{P}(Y|X = x)$ is added to the samples, denoted as $[x, y_x]$. The empirical evaluation of the conditional mean operator $\mu_{Y|x}$ with the augmented sample set is then

$$\hat{\mu}_{Y|x} = k_x^{A\intercal}(K^A + \lambda(N+1)I_{N+1})^{-1}\Phi^A,$$

where A stands for "augmented",

$$k_x^A = [k(x_1, x), \ldots, k(x_N, x), k(x, x)]^\intercal$$
$$\Phi^A = [\phi(y_1), \phi(y_2), \ldots, \phi(y_N), \phi(y_x)]^\intercal$$

and $K^A = \begin{bmatrix} K & k_x \\ k_x^\intercal & k(x,x) \end{bmatrix}$. Let

$$\beta := k_x^{A\intercal}(K^A + \lambda(N+1)I_{N+1})^{-1}, \tag{2}$$

then $\hat{\mu}_{Y|x} = \sum_{i=1}^N \beta_i\phi(y_i) + \beta_{N+1}\phi(y_x)$.

Proposition 1. *Suppose for all $y, y' \in \mathcal{Y}, l(y, y') \leq R$, the ambiguity set of $\hat{\mu}_{Y|x}$ can be written as $\mathcal{C}[x] = \{\mu \mid ||\mu - \sum_{i=1}^N \beta_i\phi(y_i)||_\mathcal{G} \leq |\beta_{N+1}|R + \gamma(N+1)^{-\frac{1}{4}}\}$ with β defined in* (2)*.*

Proof. By the triangular inequality, $||\mu_{Y|x} - \sum_{i=1}^N \beta_i\phi(y_i)||_\mathcal{G} \leq ||\mu_{Y|x} - \hat{\mu}_{Y|x}||_\mathcal{G} + ||\beta_{N+1}\phi(y_x)||_\mathcal{G}$, where the first term is bounded by $|\beta_{N+1}|R$, and the second term is bounded by $\gamma(N+1)^{-\frac{1}{4}}$ according to Lemma 4. □

The motivation for this formulation is to use the fictitious sample as a probe to measure how a newly added sample at $X = x$ would change the conditional expectation. For a queried x that is close to the samples (as measured by the kernel), $|\beta_{N+1}|$ is small and the uncertainty is small; for an x far away from the samples, the conditional expectation depends heavily on the fictitious sample, which is unknown, thus enlarging the ambiguity set.

C. Connection to Wasserstein ambiguity set

To better motivate the use of the RKHS norm as a metric for the ambiguity set definition, we establish connections between the ambiguity set defined with the RKHS norm and the Wasserstein ball. Specifically, Lemma 3 shows that the Wasserstein metric between two distributions can be characterized by some test function f with Lipschitz constant bounded by 1.

Lemma 5. *Given a translation invariant kernel k, suppose there exists a constant $\mathtt{L}$ such that $\forall x_1, x_2 \in \mathcal{X}$, $d(x_1, x_2) \leq \mathtt{L}||x_1 - x_2||$, then for any $f \in \mathcal{F}$ with a bounded RKHS norm, f is Lipschitz modulo $\mathtt{L}||f||_\mathcal{F}$.*

Proof.

$$\begin{aligned} f(x_1) - f(x_2) &= \langle f, k(\cdot, x_1)\rangle_\mathcal{F} - \langle f, k(\cdot, x_2)\rangle_\mathcal{F} \\ &= \langle f, k(\cdot, x_1) - k(\cdot, x_2)\rangle_\mathcal{F} \\ &\leq ||f||_\mathcal{F} d(x_1, x_2) \\ &\leq \mathtt{L}||f||_\mathcal{F}||x_1 - x_2||. \end{aligned}$$

where the first inequality is by the Cauchy-Schwartz. □

The constant $\mathtt{L}$ varies with the kernel k, for a Gaussian kernel $k(x_1, x_2) = \exp(-\frac{||x_1 - x_2||^2}{2\sigma^2})$, $\mathtt{L}$ is simply $\frac{1}{\sigma}$.

Proposition 2. *Consider all distributions Q satisfying $\mathbb{E}_Q[k(x, x)] < \infty$ for all $x \in \mathcal{X}$, and the following two ambiguity sets:*

$$\mathscr{A}_1 = \{Q \mid d_W(Q, Q_0)|| \leq \epsilon\}$$
$$\mathscr{A}_2 = \{Q \mid ||\mu_Q - \mu_{Q_0}||_\mathcal{F} \leq \epsilon\mathtt{L}\},$$

$\mathscr{A}_2 \subseteq \mathscr{A}_1$.

Proof. By Lemma 3,

$$\begin{aligned} ||\mu_Q - \mu_{Q_0}||_\mathcal{F} &= \mathtt{L} \sup_{f\in\mathcal{F}, ||f||_\mathcal{F} \leq \frac{1}{\mathtt{L}}} \langle \mu_Q - \mu_{Q_0}, f\rangle \\ &= \mathtt{L} \sup_{f\in\mathcal{F}, ||f||_\mathcal{F} \leq \frac{1}{\mathtt{L}}} \{\int_\mathcal{X} f(x)Q_1(dx) - \int_\mathcal{X} f(x)Q_2(dx)\} \\ &\leq \mathtt{L} \sup_{L(f)\leq 1} \{\int_\mathcal{X} f(x)Q_1(dx) - \int_\mathcal{X} f(x)Q_2(dx)\} \\ &= \mathtt{L}d_W(Q, Q_0). \end{aligned}$$

For any $Q \in \mathscr{A}_1$, we have $||\mu_Q - \mu_{Q_0}||_\mathcal{F} \leq \mathtt{L}d_W(Q, Q_0) \leq \mathtt{L}\epsilon$. □

The ambiguity set defined with the RKHS norm can be viewed as a subset of the ambiguity set defined via the Wasserstein metric. In fact, $\mathscr{A}_2$ can be viewed as the Wasserstein ball under a different distance measure.

Proposition 3. *Suppose the Wasserstein metric is defined as*

$$\begin{aligned} \bar{d}_W(Q_1, Q_2) := \inf_\Pi &\int_{\mathcal{X}^2} d(x_1, x_2)\Pi(x_1, x_2)dx_1dx_2 \\ \text{s.t.} &\int_\mathcal{X} \Pi(x_1, x_2)dx_2 = Q_1(x_1), \\ &\int_\mathcal{X} \Pi(x_1, x_2)dx_1 = Q_2(x_2), \end{aligned}$$

where $d(x_1, x_2)$ is the distance metric in an RKHS $\mathcal{F}$, then the dual form of the Wasserstein metric is

$$\bar{d}_W(Q_1, Q_2) = \sup_{||f||_\mathcal{F}\leq 1} \{\int_\mathcal{X} f(x)Q_1(dx) - \int_\mathcal{X} f(x)Q_2(dx)\}.$$

Proof. The proof follows exactly the Kantorovich Rubinstein duality theorem [25] after replacing the Euclidean distance with the RKHS distance metric d, and is omitted here. □

By Proposition 3, the ambiguity set defined as an RKHS norm ball is simply a Wasserstein ball with d as the distance metric. For a Gaussian RBF kernel, the relationship between d and the Euclidean distance is plotted in fig. 1. Comparing to the original Wasserstein ball in 1, the RKHS distance "saturates" at $\sqrt{2}$, which discounts the distance when two points are far away.

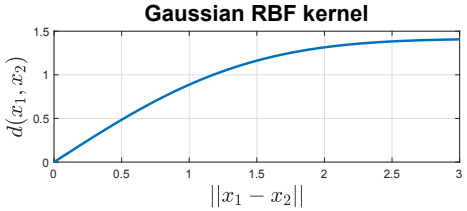

Fig. 1: RKHS distance v.s. Euclidean distance with Gaussian RBF kernel

D. Solution to the conditional kernel DRO

For brevity of notation, given the samples $\{x_i, y_i\}_{i=1}^N$ and a particular $x \in \mathcal{X}$, let $\hat{\mu}[x] := \sum_{i=1}^N \beta_i \phi(y_i)$ with β defined in (2), and $\mathcal{C}[x] = \{\mu \in \mathcal{G} \mid ||\mu - \hat{\mu}[x]||_{\mathcal{G}} \leq \epsilon_x\}$ be the ambiguity set defined in Proposition 1. The conditional kernel DRO is the following optimization:

$$\min_{\eta} \max_{Q \in co(\mathcal{Q}), \mu \in \mathcal{C}[x]} \int_{\mathcal{Y}} q(\eta, y) Q(dy) \quad \text{s.t.} \int_{\mathcal{X}} 1 Q(dy) = 1, \quad \int_{\mathcal{X}} \phi(y) Q(dy) = \mu, \tag{3}$$

where $co(\cdot)$ is the conic hull, $co(\mathcal{Q})$ is the cone of positive measures. The inner maximization looks for a distribution Q whose corresponding expectation operator μ lies inside $\mathcal{C}[x]$ and maximizes the expected cost.

Theorem 1. *Assuming q is convex in η, $,\forall \eta \in H, q(\eta, \cdot) \in \mathcal{G}$, the dual form of* (3) *is*

$$\min_{\eta, f \in \mathcal{G}, f_0 \in \mathbb{R}} f_0 + \sum_{i=1}^N \beta_i f(y_i) + \epsilon_x ||f||_{\mathcal{G}} \quad \text{s.t. } \forall y \in \mathcal{Y}, q(\eta, y) \leq f(y) + f_0. \tag{4}$$

Proof. The dual form was derived in [21], we extend it to the conditional distribution case and present the derivation for completeness. First focus on the inner maximization, the Lagrangian of the of which is

$$\begin{aligned}\mathcal{L}(\mu, P, f, f_0) &= \int_{\mathcal{Y}} q(\eta, y) Q(dy) - \delta_{\mathcal{C}[x]}(\mu) \\ &+ \langle f, \mu - \int_{\mathcal{Y}} \phi(y) Q(dy) \rangle + f_0 (1 - \int_{\mathcal{Y}} 1 Q(dy)) \\ &= f_0 + \langle f, \mu \rangle - \delta_{\mathcal{C}[x]}(\mu) + \int_{\mathcal{Y}} q(\eta, y) - f(y) - f_0 Q(dy),\end{aligned}$$

where $f_0 \in \mathbb{R}$ is the Lagrange multiplier of the first constraint in (3) and $f \in \mathcal{G}$ is the Lagrange multiplier of the second constraint. $\delta_{\mathcal{C}[x]}$ is the indicator function of $\mathcal{C}[x]$, which is 0 if $\mu \in \mathcal{C}[x]$, ∞ otherwise. The only way to upper bound $\mathcal{L}$ is to make $q(\eta, y) - f(y) - f_0 \leq 0$ for all $y \in \mathcal{Y}$. With f_0 and f fixed that satisfy $q(\eta, y) - f(y) - f_0$, the maximization is only over $\mu \in \mathcal{C}[x]$ of the function $\langle f, \mu \rangle - \delta_{\mathcal{C}[x]}(\mu)$. Since $\delta_{\mathcal{C}[x]}$ is convex, by Fenchel duality,

$$\max_{\mu \in \mathcal{G}} \langle f, \mu \rangle - \delta_{\mathcal{C}[x]}(\mu) = \delta^*_{\mathcal{C}[x]}(f) = \max_{\mu \in \mathcal{C}[x]} \langle f, \mu \rangle,$$

where $\delta^*_{\mathcal{C}[x]}$ is the Fenchel dual of $\delta_{\mathcal{C}[x]}$. By the definition of $\mathcal{C}[x]$ in Proposition 1,

$$\begin{aligned}\delta^*_{\mathcal{C}[x]}(f) &= \max_{\mu \in \mathcal{C}[x]} \langle f, \mu \rangle = \max_{\mu \in \mathcal{C}[x]} \langle f, \hat{\mu}[x] \rangle + \langle f, \mu - \hat{\mu}[x] \rangle \\ &= \langle f, \hat{\mu}[x] \rangle + \epsilon_x ||f||_{\mathcal{G}}.\end{aligned}$$

The last equality comes from the fact that we can choose μ such that $\mu - \hat{\mu}[x] = \epsilon_x \frac{f}{||f||_{\mathcal{G}}}$. The dual optimization is then

$$\min_{f \in \mathcal{G}, f_0 \in \mathbb{R}} f_0 + \langle f, \hat{\mu}[x] \rangle + \epsilon_x ||f||_{\mathcal{G}} \quad \text{s.t.} \forall y \in \mathcal{Y}, q(\eta, y) \leq f_0 + f(y).$$

Combining with the outer minimization, the dual form of (3) is the optimization in (4). Since the inner optimization is strictly feasible, strong duality holds, and there is no duality gap. □

Given an $x \in \mathcal{X}$, $\hat{\mu}[x] = \sum_{i=1}^N \beta_i \phi(y_i)$, by the reproducing property, $\langle f, \hat{\mu}[x] \rangle = \sum_{i=1}^N \beta_i f(y_i)$. However, the inequality constraint in (4) cannot be strictly enforced. Instead, the condition is enforced on a finite certification set of $\{y_j\}_{j=1}^M$. f is a function in RKHS, which is also finitely parameterized by the certification set as $f = \sum_{j=1}^M \alpha_j \phi(y_j)$. The finite optimization we end up solving is then

$$\min_{\eta, \alpha \in \mathbb{R}^M, f_0 \in \mathbb{R}} f_0 + \sum_{i=1}^N \beta_i f(y_i) + \epsilon_x ||f||_{\mathcal{G}} \quad \text{s.t. } q(\eta, y_j) \leq f(y_j) + f_0, \ j = 1, ..., M, \tag{5}$$

where $f(y) = \sum_{j=1}^M \alpha_j l(y_j, y)$, $||f||_{\mathcal{G}} = \sqrt{\alpha^\intercal L \alpha}$. Here $L \in \mathbb{R}^{M \times M}$ is computed as $L_{ij} = l(y_i, y_j)$.

IV. Application to Generation Scheduling

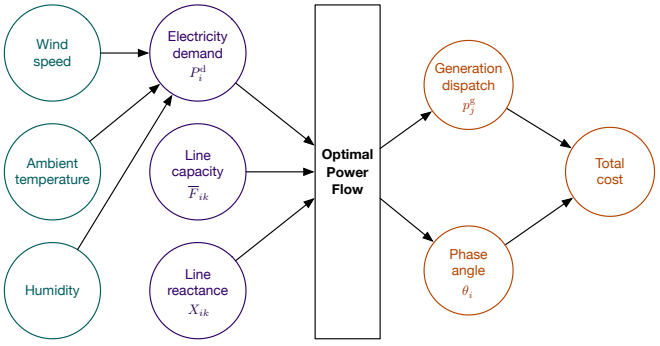

Fig. 2: Causality map of the OPF problem

As an application of the conditional kernel DRO, we consider the problem of scheduling power generation for power grid. We formulate this via a DC optimal power flow (DC-OPF) problem; a constrained optimization problem that determines generation dispatch to serve electricity demand in the transmission network at a minimal cost subject to operational constraints.

A. Background of DC-OPF

The DC-OPF problem is formulated as:

$$\min_{p^{\mathrm{g}},\theta} \quad \sum_{j\in\mathcal{G}} \left[a_j (p_j^{\mathrm{g}})^2 + b_j p_j^{\mathrm{g}} + c_j\right] \tag{6a}$$

$$\text{s.t.} \quad \sum_{j\in\mathcal{G}_i} p_j^{\mathrm{g}} - P_i^{\mathrm{d}} = \sum_{k\in\mathcal{N}} \frac{\theta_i - \theta_k}{X_{ik}}, \quad \forall i \in \mathcal{N}, \tag{6b}$$

$$\underline{P}_j^{\mathrm{g}} \le p_j^{\mathrm{g}} \le \overline{P}_j^{\mathrm{g}}, \quad \forall j \in \mathcal{G}, \tag{6c}$$

$$\underline{F}_{ik} \le \frac{1}{X_{ik}}(\theta_i - \theta_k) \le \overline{F}_{ik}, \quad \forall i,k \in \mathcal{N}. \tag{6d}$$

where the active power output of generator $j \in \mathcal{G}$ is denoted as p_j^{g} and the nodal phase angle of bus $i \in \mathcal{N}$ is denoted as θ_i. $\mathcal{G}_i$ is the set of generators connected to bus i. The objective function in (6a) minimizes the total quadratic generation cost with parameters a_j, b_j and c_j. Equation (6b), i.e. a linear approximation of the power flow equation, enforces the active nodal power balance where X_{ik} is the reactance of transmission line (i,k). Active power output of generators is constrained in (6c) using their lower and upper bound limits ($\underline{P}_j^{\mathrm{g}}$, $\overline{P}_j^{\mathrm{g}}$). Lastly, the power flow limits of each transmission line are constrained with lower and upper bounds ($\underline{F}_{ik}$, $\overline{F}_{ik}$) in (6d).

Remark 1. In practice, we add slack variables to relax the equality constraint (6b) to improve feasibility, but we found that the slack variable is always close to zero.

B. Conditional DC-OPF problem setup

Unfortunately, the power demand P_i^{d} cannot be predicted perfectly at the time of OPF. To robustify the OPF solution against the prediction error, we consider a two-stage optimization where the decision variables are divided into two groups, the here-and-now (HAN) variables and the wait-and-see (WAS) variables [26], where the HAN variables are determined in the first stage, the WAS variables can be determined after an observation of the uncertain parameters is made. Similar formulations have been applied to generation scheduling problems (c.f. [27], [28]), here we consider a simplified two-stage DC-OPF problem. The first stage determines the nominal generation dispatch $p^{\mathrm{g,n}}$, then the second stage determine the actual generation by solving the following optimization:

$$\begin{aligned}
\mathrm{OPF}(P^d, p^{\mathrm{g,n}}) = \min_{\theta, p^{\mathrm{g}}} \; & \mathcal{J}_n(p^{\mathrm{g}}) + \mathcal{J}_a(p^{\mathrm{g,n}}, p^{\mathrm{g}}) \\
\text{s.t. } & p^{\mathrm{g}} \in \mathcal{F}(p^{\mathrm{g,n}}) \\
& \sum_{j\in\mathcal{G}_i} p_j^{\mathrm{g}} - P_i^{\mathrm{d}} = \sum_{k\in\mathcal{N}} \frac{\theta_i - \theta_k}{X_{ik}}, \forall i \in \mathcal{N}, \\
& \underline{P}_j^{\mathrm{g}} \le p_j^{\mathrm{g}} \le \overline{P}_j^{\mathrm{g}}, \; \forall j \in \mathcal{G}, \\
& \underline{F}_{ik} \le \frac{1}{X_{ik}}(\theta_i - \theta_k) \le \overline{F}_{ik}, \; \forall i,k \in \mathcal{N},
\end{aligned} \tag{7}$$

where p^{g} is the actual power generation. $\mathcal{J}_n(p^{\mathrm{g}}) = \sum_{j\in\mathcal{G}} a_j (p_j^{\mathrm{g}})^2 + b_j p_j^{\mathrm{g}} + c_j$ is the generation cost, $\mathcal{J}_a$ is the adjusting cost, which penalizes the difference between $p^{\mathrm{g,n}}$ and p^{g}, and $\mathcal{F}(p^{\mathrm{g,n}})$ is the feasible set of the actual generation given the nominal generation dispatch. The HAN variable is the nominal generation dispatch $p^{\mathrm{g,n}}$, the WAS variables are the actual generation dispatch p^{g} and phase angle θ. The feasible set $\mathcal{F}$ is defined as $\mathcal{F}(p^{\mathrm{g,n}}) = \{p \in \mathbb{R}^{|\mathcal{G}|} | \forall j \in \mathcal{G}\backslash\mathcal{G}_{\mathrm{s}}, p_j = p_j^{\mathrm{g,n}}\}$, where $\mathcal{G}_{\mathrm{s}}$ is the set of flexibility resources, i.e. the generation dispatch can be adjusted after the nominal dispatch. Additional constraints can be added such as ramping constraints on the change of power generation. The adjusting cost then penalizes the adjustment of flexible generators $j \in \mathcal{G}_{\mathrm{s}}$. In particular, we consider the following adjusting cost function:

$$\mathcal{J}_a(p^{\mathrm{g,n}}, p^{\mathrm{g}}) = \sum_{j\in\mathcal{G}\backslash\mathcal{G}_s} \zeta_j^+ [p_j^{\mathrm{g,n}} - p_j^{\mathrm{g}}]_+ + \zeta_j^- [p_j^{\mathrm{g}} - p_j^{\mathrm{g,n}}]_+,$$

which is an asymmetric linear cost that penalizes generation surplus and generation deficiency differently. $[\cdot]_+ = \max\{\cdot, 0\}$, and ζ^+ is chosen to be smaller than ζ^- so that generation deficiency is more heavily penalized. Under the piecewise linear $\mathcal{J}_a$, the second stage is a convex optimization problem. $\mathrm{OPF}(P^d, p^{\mathrm{g,n}})$ denotes the optimal cost function of (7) whenever (7) is strictly feasible, and $\mathrm{OPF}(P^d, p^{\mathrm{g,n}}) = \infty$ if (7) is infeasible. Clearly, $\mathrm{OPF}(P^d, p^{\mathrm{g,n}})$ depends heavily on P^d, which is unknown at the first stage.

While the knowledge on P^d is not perfect when deciding $p^{\mathrm{g,n}}$ in the first stage, there is a causal relationship between the weather data and the electricity demand as shown in fig. 2. Therefore, given historical data $\{[\text{weather,time}], \text{demand}\}_{i=1}^N$, we build a conditional Kernel DRO for the DC-OPF problem to better capture the conditional distribution of P^d. The weather and time signal serves as the random variable being conditioned on (X), P^d is the random variable whose conditional distribution we are interested in (Y). The first stage decision making is then solved with the following DRO:

$$\min_{p^{\mathrm{g,n}}} \max_{Q\in\mathscr{A}_\xi} \mathbb{E}_Q[\mathrm{OPF}(P^d, p^{\mathrm{g,n}})], \tag{8}$$

where ξ is the auxiliary variable (weather, time, etc.), and $\mathscr{A}_\xi$ is the ambiguity set over the conditional distribution of P^d given ξ.

Following the conditional kernel DRO framework, the ambiguity set on conditional distribution is turned into an ambiguity set in the conditional mean operator in RKHS. Given samples $\{\xi_i, P_i^d\}_{i=1}^N$, we use Gaussian RBF kernel for both ξ and P^d. In particular, each dimension of ξ is equipped with its own kernel and the distances are added to form the total distance. For cyclic dimensions such as hour in the day, the distance is wrapped around, i.e., the distance between 23:00 and 1:00 is 2 hours, not 22 hours. We denote the two kernel functions as k_ξ and k_{P^d}.

Given a particular auxiliary variable evaluation ξ, the conditional mean operator is computed as $\hat{\mu}_{P^d|\xi} = \sum_{i=1}^N \beta_i \phi(P_i^d) + \beta_{N+1}\phi(P_\xi^d)$ with β computed following (2), where the last term is associated with the unknown P^d under the query point ξ. The ambiguity set is defined

following Proposition 1, and the conditional kernel DRO is solved in the dual form (5).

$$\min_{p^{\mathrm{g,n}},\alpha\in\mathbb{R}^M,f_0\in\mathbb{R}} \quad f_0+\sum_{i=1}^{N}\beta_i f(P_i^d)+\epsilon_x||f||_{\mathcal{G}}$$
$$\text{s.t. } \mathrm{OPF}(p^{\mathrm{g,n}},P_j^d)\le f(P_j^d)+f_0,\ \ j=1,...,M, \tag{9}$$

where $f(P^d)=\sum_{j=1}^{M}\alpha_j l(P_j^d,P^d)$. Note that the objective depends on $f(P_i^d)$, which enumerates over the observation dataset $\xi_i,P_i^d{}_{i=1}^{N}$, whereas the constraints enumerates over $\{P_j^d\}_{j=1}^{M}$, which is the certification set. The certification set's role is to approximate the support of P^d, and we construct it by random sampling P^d from the observation dataset with added Gaussian noise.

In practice, to reduce the computation complexity, we only consider the closest m points in the observation dataset measured by the kernel distance. This is because points with a large distance to the query point in the RKHS will not influence the result much, thus can be ignored. The choice of m depends on the kernel, i.e., how fast does $k(x,x')$ decrease to near zero as x' moves away from x.

C. Numerical experiments

The case study uses the 11-bus ERCOT transmission system from [29] with the historical hourly demand profile from ERCOT market information data hub [30]. The corresponding historical weather data for the ERCOT zones was obtained from the National Solar Radiation Database [31]. All models in the case study are implemented using the CVXPY package [32] and the code and data are available in https://github.com/chenyx09/CKDRO.

We consider three benchmarks.

- The first one is the optimal cost, i.e., assuming that P^d is perfectly known in the first stage, which is not realistic, yet it serves as the lower bound of the best cost attainable.
- The second benchmark is to solve the first stage as an OPF problem with the average P^d in the dataset $\overline{P^d}$, i.e., replacing the expectation in (8) with $\overline{P^d}$. This setup ignores the auxiliary variable completely.
- The third benchmark is to perform a kernel interpolation with the same kernels used in DRO and perform OPF with the interpolated P^d. To be specific, given the auxiliary variable x, first compute β following (2), then the interpolated P_x^d is computed as $P_x^d=\sum_{i=1}^{N}\frac{\beta_i}{\overline{\beta}}P_i^d$ with $\overline{\beta}=\sum_{i=1}^{N}\beta_i$, then $p^{\mathrm{g,n}}$ is obtained by solving the OPF with P_x^d.

The auxiliary variable has 18 dimensions, consisting of hours in the year (cyclic with period 8760, indicating the time in the year), hours in the day (cyclic with period 24, indicating time in the day), and temperature and precipitation in 8 regions.

We sample 60 points from the history dataset and for each data point $[\xi_i,P_i^d]$, the generation cost with the CKDRO solution is compared with the three benchmarks. To be specific, the CKDRO is solved with ξ_i, the two stage cost is then evaluated as $\mathrm{OPF}(p^{\mathrm{g,n}},P_i^d)$, where $p^{\mathrm{g,n}}$ is the solution of the first stage CKDRO. For the three benchmarks, the first stage is solved as an OPF problem with P_i^d, $\overline{P^d}$ and $P_{\xi_i}^d$, respectively, then the two stage cost is evaluated with the solution from the first stage. Since the DRO setup aims at optimizing over the worst performance over the ambiguity set, in addition to the nominal two stage cost, we also plot the worst case cost, where Gaussian noise $w\sim\mathcal{N}(0,\sigma_w)$ is added to P_i^d and the worst cost among m samples of $\mathrm{OPF}(p^{\mathrm{g,n}},P_i^d+w)$ is recorded.

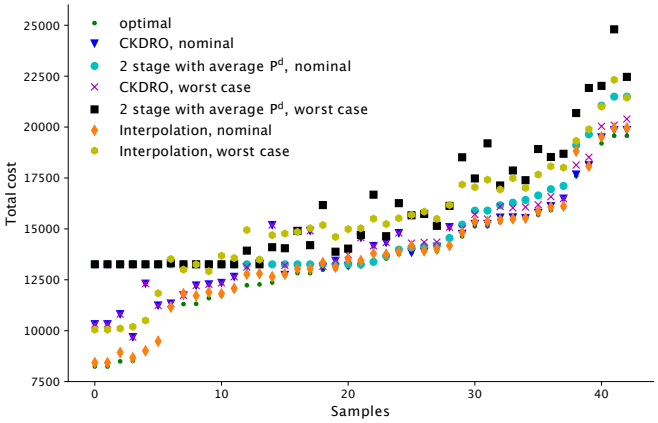

Fig. 3: Comparison between CKDRO result and benchmarks for sample data set (50 points sorted by the objective value). The nominal cost of the CKDRO result is close to the optimal value whereas its worst case cost is very close to the nominal cost, showing greater robustness than the benchmarks.

Figure 3 shows the cost comparison between the CKDRO result and the benchmarks. The results are sorted with the optimal cost in ascending order. The kernel interpolation cost is almost identical to the optimal cost, showing that the kernel interpolation is able to predict P_d^i fairly accurately. However, since it takes no uncertainty into account, its worst case cost is worse than the CKDRO result. The cost with the mean load profile $\overline{P^d}$ is significantly higher than the CKDRO result, and there is a big gap between its nominal cost and worst case cost, indicating that it is not robust to uncertainty. CKDRO is able to achieve close-to-optimal nominal cost while being much more robust to uncertainty, which is demonstrated by the little gap between its nominal cost and worst case cost.

As shown in (9), ϵ determines the level of robustness for the CKDRO, and changes with the number of data points and the location of the auxiliary variable. To showcase the effect of ϵ, fig. 4 shows the comparison of the CKDRO result with the adaptive ϵ leveraging Proposition 1 and two benchmarks, one fixing $\epsilon=0.15$, and the other fixing $\epsilon=3$. A smaller ϵ causes the DRO to consider distributions that are more concentrated around the conditional mean, resulting in better nominal cost, yet may struggle when P^d deviates from the conditional mean, which is shown in the worst case cost. A large ϵ, on the other hand, leads to a smaller gap between the nominal and worst case cost, sometimes even 0 gap, yet significantly hurt the nominal performance. The adaptive ϵ

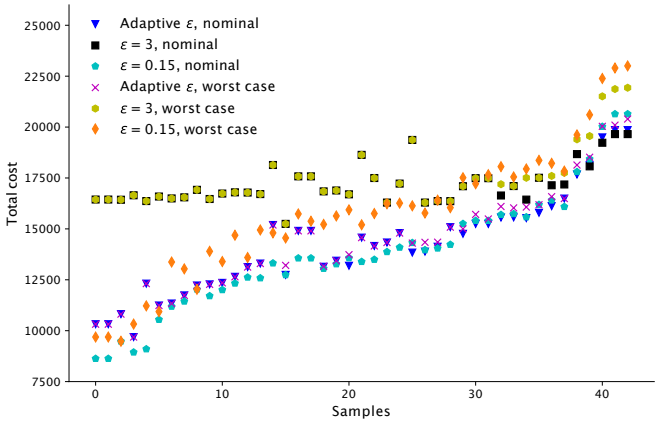

Fig. 4: Influence of ϵ on the CKDRO result. A small ϵ results in a smaller nominal cost, yet the worst case cost can be huge; a large ϵ leads to a smaller gap between the nominal and worst case cost, yet can significantly hurt the nominal cost. The adaptive ϵ achieves the smallest worst case cost.

achieves the best worst case performance, as is the goal of the DRO.

V. Conclusion

We proposed a conditional kernel DRO framework capable of robust decision making that leverages the conditional distribution of the unknown variable. The main application of CKDRO is when an auxiliary variable is observable and correlated with the unknown variable, and one needs to make a decision after observing the auxiliary variable. The key idea is to use the conditional mean operator in RKHS space to define the ambiguity set, whose size depends on the data volume as well as the query point position. We draw connections between the ambiguity set used in CKDRO and one defined with the Wasserstein distance and show that they are related via the change of test function. The CKDRO is solved in its dual form and approximated by a finite-dimensional convex program. We show its application on a optimal power flow problem and the results show that CKDRO is able to outperform several common benchmarks by leveraging the auxiliary variable and the uncertainty in conditional distribution.

References

[1] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*. Princeton university press, 2009.

[2] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.

[3] I. Yang, "Wasserstein distributionally robust stochastic control: A data-driven approach," *IEEE Transactions on Automatic Control*, 2020.

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

[5] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.

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

[7] A. Ruszczyński and A. Shapiro, "Optimization of convex risk functions," *Mathematics of operations research*, vol. 31, no. 3, pp. 433–452, 2006.

[8] P. Coppens and P. Patrinos, "Data-driven distributionally robust mpc for constrained stochastic systems," *arXiv preprint arXiv:2103.03006*, 2021.

[9] N. Fournier and A. Guillin, "On the rate of convergence in wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.

[10] J. A. Tropp, "An introduction to matrix concentration inequalities," *arXiv preprint arXiv:1501.01571*, 2015.

[11] E. Dall'Anese, K. Baker, and T. Summers, "Optimal power flow for distribution systems under uncertain forecasts," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 7502–7507.

[12] H. Zhang and P. Li, "Chance constrained programming for optimal power flow under uncertainty," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2417–2424, 2011.

[13] Y. Chen, N. Sohani, and H. Peng, "Modelling of uncertain reactive human driving behavior: a classification approach," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 3615–3621.

[14] Y. Chen, S. Dathathri, T. Phan-Minh, and R. M. Murray, "Counter-example guided learning of bounds on environment behavior," in *Conference on Robot Learning*. PMLR, 2020, pp. 898–909.

[15] Y. Chen, U. Rosolia, C. Fan, A. D. Ames, and R. Murray, "Reactive motion planning with probabilistics safety guarantees," in *Conference on Robot Learning*. PMLR, 2021.

[16] N. Meinshausen and G. Ridgeway, "Quantile regression forests." *Journal of Machine Learning Research*, vol. 7, no. 6, 2006.

[17] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *Management Science*, vol. 66, no. 3, pp. 1025–1044, 2020.

[18] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 961–968.

[19] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces," *Journal of Machine Learning Research*, vol. 5, no. Jan, pp. 73–99, 2004.

[20] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *arXiv preprint arXiv:1605.09522*, 2016.

[21] J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf, "Kernel distributionally robust optimization," *arXiv preprint arXiv:2006.06981*, 2020.

[22] L. V. Kantorovich and S. Rubinshtein, "On a space of totally additive functions," *Vestnik of the St. Petersburg University: Mathematics*, vol. 13, no. 7, pp. 52–59, 1958.

[23] G. C. Calafiore, "Ambiguous risk measures and optimal robust portfolios," *SIAM Journal on Optimization*, vol. 18, no. 3, pp. 853–877, 2007.

[24] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil, "Conditional mean embeddings as regressors-supplementary," *arXiv preprint arXiv:1205.4656*, 2012.

[25] D. A. Edwards, "On the kantorovich–rubinstein theorem," *Expositiones Mathematicae*, vol. 29, no. 4, pp. 387–398, 2011.

[26] R. J. Wets, "Stochastic programming models: wait-and-see versus here-and-now," in *Decision Making Under Uncertainty*. Springer, 2002, pp. 1–15.

[27] Y. Zhang, J. Le, F. Zheng, Y. Zhang, and K. Liu, "Two-stage distributionally robust coordinated scheduling for gas-electricity integrated energy system considering wind power uncertainty and reserve capacity configuration," *Renewable energy*, vol. 135, pp. 122–135, 2019.

[28] P. Zhao, C. Gu, D. Huo, Y. Shen, and I. Hernando-Gil, "Two-stage distributionally robust optimization for energy hub systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3460–3469, 2019.

[29] S. Battula, L. Tesfatsion, and T. E. McDermott, "An ercot test system for market design studies," *Applied Energy*, vol. 275, p. 115182, 2020.

[30] ERCOT, "ERCOT Market Information List," 2022. [Online]. Available: https://www.ercot.com/mp/data-products

[31] NREL, "NSRDB: National Solar Radiation Database," 2022. [Online]. Available: https://nsrdb.nrel.gov/

[32] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.