Proceedings of Machine Learning Research vol 144:1–15, 2021

Approximate Distributionally Robust Nonlinear Optimization with Application to Model Predictive Control: A Functional Approach

Yassine Nemmour YNEMMOUR@TUEBINGEN.MPG.DE
Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Jia-Jie Zhu JZHU@TUEBINGEN.MPG.DE
Empirical Inference Department
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

We provide a functional view of distributional robustness motivated by robust statistics and functional analysis. This results in two practical computational approaches for approximate distributionally robust nonlinear optimization based on gradient norms and reproducing kernel Hilbert spaces. Our method can be applied to the settings of statistical learning with small sample size and test distribution shift. As a case study, we robustify scenario-based stochastic model predictive control with general nonlinear constraints. In particular, we demonstrate constraint satisfaction with only a small number of scenarios under distribution shift.

Keywords: Model Predictive Control - Constraint Tightening - Data-Driven Control - Robust Optimization - Distributionally Robust Optimization - Robust Statistics

1. Introduction

This paper studies techniques for handling constraints in the form of uncertain nonlinear inequality

$$f(x, \xi) \leq 0, \tag{1}$$

where x is the decision variable and ξ is an uncertain variable. One principled approach to handle (1) is *robust optimization* (RO) (Soyster, 1973; Ben-Tal et al., 2009; Bertsimas et al., 2011). In RO, uncertain inequalities (1) can be written as their *robust counterparts* (RC)

$$\text{(RC)} \quad f(x, \xi) \leq 0, \forall \xi \in \mathcal{X}, \tag{2}$$

where $\mathcal{X}$ is an uncertainty set where the uncertain ξ lives. The methodology then proceeds to reformulate (2) as solvable deterministic programs for useful choices of uncertainty sets $\mathcal{X}$ such as polyhedra; cf. (Ben-Tal et al., 2013). However, the downside of canonical RO methodology is its potential conservativeness — robustifying against every element in the modeled uncertainty set equally is hardly necessary in practice.

Distributionally robust optimization (DRO) (Delage and Ye, 2010; Goh and Sim, 2010; Scarf, 1958) extends RO techniques to probability measures while mitigating the conservativeness. It is a promising tool to treat many robustness challenges in machine learning and data-driven control. DRO with various constraints have been developed, including finite-order moment bounds (Delage and Ye, 2010; Scarf, 1958; Zymler et al., 2013; Milz and Ulbrich, 2020), f-divergence constraints (Ben-Tal et al., 2013; Iyengar, 2005; Nilim and El Ghaoui, 2005; Wang et al., 2016;

Duchi et al., 2018; Bayraksan and Love, 2015), and, more recently, probability metrics such as the popular Wasserstein distances (Mohajerin Esfahani and Kuhn, 2018; Zhao and Guan, 2018; Gao and Kleywegt, 2016; Blanchet et al., 2019; Xie, 2019) and kernel distances (Zhu et al., 2020b; Xu et al., 2009; Staib and Jegelka, 2019). We refer to (Rahimian and Mehrotra, 2019) for a survey. A significant amount of recent effort has been focused on the Wasserstein DRO due to many of its attractive properties. However, common approaches assume either a restricted function class or the knowledge of Lipschitz constant which is often inaccessible in modern applications as discussed in (Virmaux and Scaman, 2018; Bietti et al., 2019). In particular, this paper is interested in handling general nonlinear inequality (1), which falls outside the scope of commonly-used reformulation techniques.

The RO methodology has been widely adopted in robust constrained control and model predictive control (MPC) (Bemporad and Morari, 1999; Langson et al., 2004; Diehl and Bjornberg, 2004; Mayne et al., 2005) A central component of those approaches is to guarantee constraint satisfaction of the optimal control problem under uncertainty, which may be modeled by (RC) (2). This is also intricately related to recent constraint tightening techniques in stochastic MPC (Köhler et al., 2019; Bonzanini et al., 2019; Santos et al., 2019; Hewing and Zeilinger, 2020; Mark and Liu, 2020); see (Mesbah, 2016; Farina et al., 2016) for more details of the prior art. Using tube-based stochastic MPC as an example, this paper shows how disributionally robust nonlinear optimization techniques can be applied to constraint tightening for general nonlinear constraints. Before going further, we make an important distinction between two different settings in learning and data-driven control. With a possible abuse of terminology, we refer to them as *statistical learning with small datasets* and *distribution shift*.

(i) In the statistical learning with small datasets setting, the training and test data are assumed to be sampled from the same generating distribution P_{true}. Note this is a setting commonly studied in stochastic MPC constraint tightening approaches such as (Hewing and Zeilinger, 2020; Mark and Liu, 2020). In this setting, the empirical distribution $\hat{P}_N$ deviates significantly from P_{true} due to the small sample size. Methods such as scenario approach cannot guarantee high-level of robustness.

(ii) In the distribution shift setting, the test data is generated from a shifted distribution, denoted as P_{shift}, different from P_{true}. This is often the case in practice since the true system is inevitably different from the mathematical model. Then, even with large datasets, common stochastic optimization and control design still performs poorly during test time. This was the topic of (Zhu et al., 2020c) and is also related to the issue of adversarial robustness in machine learning and simulation-to-reality transfer in robotics.

Contribution.

(i) We provide a unified functional view of distributional robustness motivated by robust statistics and robust optimization. Our perspective leads to two simple and practical computational approaches based on gradient norm and Kernel DRO. Our methods are applicable to general nonlinear constraints.

(ii) We clearly point out two distinct settings where our distributional robust counterpart (DRC) framework can be useful: statistical learning with small sample size and test distribution shift. Previous works mainly focus on the former setting.

(iii) In a case study, we robustify scenario-based stochastic tube-based MPC with general nonlinear constraint, whereas the previous approaches only consider linear constraint tightening.

Most significantly, we address a common shortcoming of the scenario approach, by achieving constraint satisfaction with only a small number of scenarios.

(iv) While the refined rate of kernel distance in Lemma 1 has already been discovered, it could not be applied to general DRO since the function of interest does not live in a known RKHS. This paper unlocks this potential.

Notation. Depending on the context, $\|\cdot\|$ denotes the norm of a vector, matrix, or operator in the corresponding normed spaces. ∇f denotes the gradient operator. Depending on the context, $\nabla f(x)$ denotes a gradient vector in Euclidean spaces or a Gâteaux derivative in more general Hilbert spaces. The Minkowski set addition is given by $A \oplus B = \{a + b : a \in A, b \in B\}$. The Pontryagin set difference is defined as $A \ominus B = \{a : a + B \subseteq A\}$. We refer to the decision variable as $x \in \mathbb{R}^{n_x}$ and $\xi \in \mathbb{R}^{n_x}$ generally refers to the uncertain variable.

2. Preliminaries

2.1. Reproducing kernel Hilbert space and integral probability metrics

Recall that a kernel is similarity measure modeled by a symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. If there exists a Hilbert space $\mathcal{H}$ and feature map $\phi\colon \mathcal{X} \to \mathcal{H}$ such that $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$ is an inner product on $\mathcal{H}$, a space of real-valued functions on $\mathcal{X}$. Then $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) associated with k. A commonly used kernel is the Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$ where $\sigma > 0$ is the bandwidth parameter.

This paper concerns DRO constrained by sets described by a class of probability metrics, the *integral probability metric* (IPM) (Müller, 1997; Gretton et al., 2012; Sriperumbudur et al., 2012). Given two probability measures P, Q, an IPM generated by some function class $\mathcal{F}$ can be written as $D_{\mathcal{F}}(P, Q) := \sup_{f\in\mathcal{F}}\{|\int f d(P - Q)|\}$, where the optimizer f^* is a witness function. The family of IPMs include, among others, the Kantorovich metric (the dual representation of Wasserstein-1 metric), the total variation distance, and kernel distance (also known as the maximum mean discrepancy (MMD)). In this paper, we focus on the Wasserstein-1 metric and the MMD via a functional approach, i.e., by manipulating the function f in the IPM definition. The Wasserstein-1 metric is obtained by choosing the function class $\mathcal{F} = \{f : \operatorname{lip}(f) \leq 1\}$ where $\operatorname{lip}(f)$ denotes the Lipschitz constant of f. On the other hand, by choosing $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ where $\|\cdot\|_{\mathcal{H}}$ is the norm in an RKHS, we recover the MMD associated with the $\mathcal{H}$. Given N i.i.d. samples $X_i \sim P, Y_i \sim Q$, a sample-based estimator for MMD can be computed by $\widehat{MMD}^2(P, Q) = \frac{1}{N(N-1)}\sum_{i\neq j}[k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)]$.

The statistical guarantees for DRO constrained by probability metrics typically concern the size of the ambiguity set, e.g., the minimum radius ϵ of some metric-ball $B_\epsilon = \{m : D(m, \hat{P}_N) \leq \epsilon\}$ centered at the empirical distribution $\hat{P}_N$ such that it contains the unknown true generating distribution P_{true}. The convergence rate for Wasserstein DRO have been established originally in (Mohajerin Esfahani and Kuhn, 2018, Theorem 3.4) to be $\mathcal{O}((1/n)^{\frac{1}{d}})$ where d is the dimension of data. That rate suffers from the curse of dimensionality and contains unknown constants. However, this is still an active area of research. We refer to recent works that report more refined rates, e.g. (Si et al.; Blanchet et al., 2019). In contrast, the refined rate for kernel distances (MMD) does not involve unknown constants or the dimensions of $\mathcal{X}$. Below is due to (Tolstikhin et al., 2017, Proposition A.1).

Lemma 1 (Tolstikhin et al. (2017)) *Suppose $\hat{P}_N$ is the empirical distribution $\hat{P}_N = \sum_{i=1}^{N} \frac{1}{N}\delta_{\xi_i}$, where ξ_i are i.i.d samples of the true distribution P_{true}. Let C be a constant such that $\sup_{x\in\mathcal{X}} k(x,x) \leq C < \infty$. Then, with probability at least $1-\alpha$,*

$$\mathrm{MMD}(\mathcal{H}, \hat{P}_N, P_{true}) \leq \sqrt{\frac{C}{N}} + \sqrt{\frac{2C\log(1/\alpha)}{N}}. \tag{3}$$

For the commonly used Gaussian RKHS, C is easily computable by $\sup_{x\in\mathcal{X}} k(x,x) = 1$, resulting in a computable rate. It is also possible to choose ambiguity set sizes empirically based on training data, such as bootstrap; see, e.g., (Gretton et al., 2012). The true power of Lemma 1 only manifests when the corresponding kernel k is known, as is the case enabled by this paper's method.

2.2. Scenario-based approaches to stochastic optimization

In the scenario optimization literature, such as the work in (Campi and Garatti, 2011), a chance constraint $\mathbb{P}(f(x,\xi) \leq 0) \geq 1-\alpha$ is replaced by a sample-based one. Then, the uncertain inequality constraint is required to hold for every sample. Hewing and Zeilinger (2020) combine the data-driven nature of scenario methods with the offline computation of tube-based MPC. As an application of our methodology, we will consider the MPC formulation (Hewing and Zeilinger, 2020, eq. 11(a-g)) for the remainder of the paper. Moreover, they can make use of the bound derived in (Campi and Garatti, 2011, Theorem 2.1). It provides a lower bound on the number of required scenarios such that we find a solution to the chance constrained problem with at least probability $1-\beta$. A simplified version of the scenario bound is found in (Hewing and Zeilinger, 2020)

$$N_s \geq \frac{2}{1-\alpha}((d-1)\ln(2) - \ln(\beta)), \tag{4}$$

where N_s is the number of scenarios and d is the dimensionality of the decision variable. For high satisfaction probabilities and complex problems, i.e., large decision variables, that still suffers from the curse of dimensionality. Recent works investigate regularization in scenario-based approaches, addressing conservativeness and better generalization. As previously stated, penalizing every scenario equally, does increase conservativeness. Campi and Carè (2013) use $\ell 1$ regularization, in form of an $\ell 1$ constraint for the min-max formulation of the scenario-based approach, to encourage sparsity of the decision variable. In (Formentin et al., 2017) constraint relaxation is approached by reducing the number of active scenario constraints. They achieve this with $\ell 2$ regularization. The work in (Zhu et al., 2020a) reduces conservativeness by removing scenarios as long as the kernel mean embedding (KME) of the reduced scenario set does not differ significantly from the KME of all scenarios. That can be formulated as a convex optimization problem and solved efficiently.

3. Theory and Methods

3.1. Distributionally robust counterparts

The starting point of our theory is the distributionally robust counterpart (DRC) of the uncertain nonlinear inequality in (1) and (2).

$$\text{(DRC)} \qquad \sup_{P\in K} \mathbb{E}_P f(x,\xi) \leq 0, \forall P \in K, \tag{5}$$

where K is a set of probability distributions, known as the ambiguity set. (DRC) describes the worst-case expected value of the constraint function depending on an uncertain variable ξ. Throughout the paper, we assume the following non-restrictive conditions for the (DRC).

Assumption 1 *$f(x, \cdot)$ is upper semicontinuous and $f(x, \xi) > -\infty$ for at least some $\xi \in \mathcal{X}$. K is closed convex and non-empty.*

Many existing formulations can be written as special cases of (DRC), such as the canonical worst-case RO, sample average approximation. Zhu et al. (2020b) give an exhaustive list of the special cases. One special case of (DRC) is the polytopic constraints, e.g., by letting $f(x, \xi) = H(x+\xi)-b$, which is considered in (Hewing and Zeilinger, 2020; Mark and Liu, 2020). Those works then used polytopic probabilistic reachable sets (PRS) and Pontryagin difference for constraint tightening. In this paper, we are interested in the more general classes of (DRC). We also note that such functions f may fall outside the scopes of typical Wasserstein DRO reformulation techniques such as (Mohajerin Esfahani and Kuhn, 2018). Hence, this paper's investigation also constitutes a generalization in that regard. We now turn to a probabilistic robustness method — the scenario approach to stochastic optimization.

Example 1 (Scenario approach as DRC) *Let $K = \operatorname{conv}(\delta_{\xi_1}, ..., \delta_{\xi_N})$. Then (DRC) is given by*

$$\sup_{P \in K} \mathbb{E}_P f(x, \xi) = \max_i f(x, \xi_i) \leq 0,$$

which is the scenario approach. It can be seen as a special case of (DRC) where the ambiguity set K is chosen to be a data-driven polytope $\operatorname{conv}(\delta_{\xi_1}, ..., \delta_{\xi_N})$.

Let us consider the setting where N is small and the bound in (Campi and Garatti, 2011) does not provide robustness anymore. We propose to use the following *Minkowski sum* to robustify the scenario approach in this setting. We illustrate our idea in figure 1(*a*)subfigure.

Example 2 (Constraint tightening via Minkowski sums of ambiguity sets) *Let K be the ambiguity set specified by the polytope $K = \operatorname{conv}(\delta_{\xi_1}, ..., \delta_{\xi_N})$ in the scenario approach. Let B be some convex set in the topological vector space of signed measures. Then, we have the following*

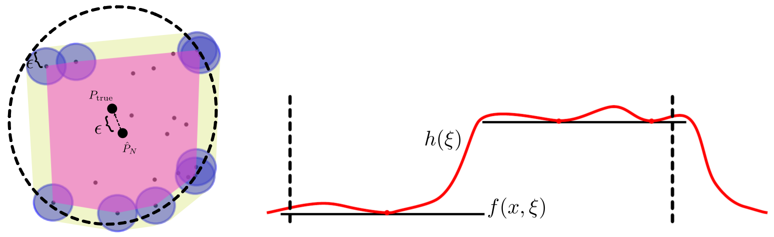
$$\sup_{P \in (K+B) \cap \mathcal{P}} \mathbb{E}_P f(x, \xi) \leq \sup_{P \in K} \mathbb{E}_P f(x, \xi) + \sup_{P \in B} \mathbb{E}_P f(x, \xi) = \max_i f(x, \xi_i) + \delta^*_B(f(x, \cdot)) \leq 0, \tag{6}$$

where $\delta^*_B(f)$ denotes the support function given by $\delta^*_B(f) := \sup_{\mu \in B} \mathbb{E}_\mu f(x, \xi)$.

Compared with the scenario approach, there is an extra tightening term $\delta^*_B(f)$. In the next section, we provide approaches to compute the tightening by reformulating (DRC).

3.2. A functional view of distributional robustness

We now take a functional approach commonly adopted in robust statistics (Fernholz, 2012; Hampel et al., 2011; v. Mises, 1947) and also in stochastic programming, see, e.g., (Shapiro et al., 2014,



(*a*) (*b*)

Figure 1: The dashed line in (a) is the support of the probability distribution. The ambiguity set of the scenario approach is colored in pink. Intuitively, our way to robustify is a polytope with round corners in the space of probability measures. We can choose the size of B, in blue, according to statistical learning bounds such as in Lemma 1. This way, the size of B is large when N is small and shrinks as N increases, while the ambiguity set, yellow area, converges to the dashed line. Figure (b), from (Zhu et al., 2020b), visualizes the intuition of the majorization in Proposition 3.

Chapter 6.6.3). This view offers a clear picture of the role that function spaces play in distributional robustness. We view the risk measure as a function of the probability measure. Let $P, \hat{P}$ be probability measures. With a slight abuse of terminology, we define the *bias* as

$$\text{Bias}(P; f, \hat{P}) := \int f \, d(P - \hat{P}). \tag{7}$$

Bias describes the risk deviation under a new distribution P from the nominal one $\hat{P}$. Using this quantity, we can write the (DRC) as

$$\sup_{P \in K} \mathbb{E}_P f(x,\xi) = \mathbb{E}_{\hat{P}} f(x,\xi) + \sup_{P \in K} \text{Bias}\,(P; f, \hat{P}) \leq 0, \forall P \in K. \tag{8}$$

In robust statistics, $\sup_{P \in K} \text{Bias}\,(P; f, \hat{P})$ is also referred to as the *maxbias* or *supremum bias*. It is also referred to as the regularizer. Given fixed $\hat{P}_N$, the decomposition in (8) tells us that the key quantity to controlling the distributional robustness is the supremum bias. Let us see this through concrete examples from the DRO literature. Using the functional view in (7), the examples below are straightforward consequences of the definitions for the corresponding probability metrics.

Example 3 (Wasserstein-1 metric) *Suppose the function f is L_f-Lipschitz continuous on the domain $\mathcal{X}$. Let W_1 be the Wasserstein-1 metric. Then,*

$$\sup_{\mathrm{W}_1(P,\hat{P}) \leq \epsilon} \text{Bias}\,(P; f, \hat{P}) = \epsilon \cdot L_f. \tag{9}$$

As the Wasserstein-1 metric belongs to a larger family of the *integral probability metrics* (IPM), we give another instance of this family.

Example 4 (Maximum mean discrepancy (MMD) associated with an RKHS) *Suppose the function $f \in \mathcal{H}_f$ belongs to the RKHS $\mathcal{H}_f$. Let MMD be the MMD associated with $\mathcal{H}_f$. Then,*

$$\sup_{\mathrm{MMD}(\mathcal{H}_f, P, \hat{P}) \leq \epsilon} \text{Bias}\,(P; f, \hat{P}) = \epsilon \cdot \|f\|_{\mathcal{H}_f}. \tag{10}$$

Unfortunately, a limitation of the above approaches is that the Lipschitz constant L_f or the associated RKHS norm $\|\cdot\|_{\mathcal{H}_f}$ is often not accessible in general nonlinear programming and learning tasks. It is henceforth valuable to derive a generally applicable approach.

3.3. A simplified approach via gradient norm regularization

Given a real-valued continuously differentiable function f with Lipschitz constant L_f, we have the straightforward relationship $L_f \geq \sup_{x\in\mathcal{X}} \|\nabla f(x)\|$, where equality is attained with further conditions on $\mathcal{X}$. This motivates the following practical quantity to control the distributional robustness without the access to a Lipschitz constant via the gradient-based lower bound to (9).

$$\sup_{\mathrm{W}_1(P,\hat{P})\leq\epsilon} \operatorname{Bias}(P; f, \hat{P}) = \epsilon \cdot L_f \approx \epsilon \cdot \max_i \|\nabla f(x_i)\|.$$

We note that this optimistic lower bound is exact for constraint tightening in MPC with polyhedral constraints, such as in (Hewing and Zeilinger, 2020). We now propose a regularized scenario approach using the discussion above.

Example 5 (Gradient norm regularized distributionally robust scenario approach) *In the same setting as Example 2, we write the following gradient-norm approximate distributionally robust counterpart reformulation for constraint tightening.*

$$\max_i f(x, \xi_i) + \epsilon \max_i \|\nabla f(\xi_i)\| \leq 0 \tag{11}$$

We are certainly not the first to use gradient norm regularization. There is a large body of literature in machine learning that uses similar methods to improve robustness; see (Bietti et al., 2019) and references therein. However, our perspective here is that the use of gradient norm regularization dates back to even earlier works in robust optimization and optimal control, for example, in (Diehl et al., 2006; Nagy and Braatz, 2004). Compared to those (RC) approximation techniques this paper is interested in the distributional robustness measure instead of worst-case robustness.

Remark 2 (Tightening for constraints learned from data) *It is sometimes not practical to explicitly encode all the real-world constraints in control problems. Instead, they can be learned from data. Such learned constraints are typically nonlinear which cannot be used with many existing DRO reformulation methods. For example, support vector machines (SVM) have been applied in the context of data-driven stochastic MPC, e.g., in (Shang and You, 2018). Suppose C denotes the (nonlinear) smooth decision function of a learned constraint, i.e., the constraint set is given by $\{\xi : C(x,\xi) \leq 0\}$, where x is a decision variable and ξ corresponds to the data samples.*

Moreover, if $C(x,\cdot)$ is an RKHS function associated with a known kernel, as is the case of SVM, (10) implies the (DRC) reformulation $\max_i C(x,\xi_i) + \epsilon\|C\|_{\mathcal{H}} \leq 0$, where ϵ can be chosen as the MMD radius in (10). This tightening has an elegant interpretation of RKHS feature space perturbation discussed in (Xu et al., 2009), but only applies to kernelized models such as SVMs.

3.4. Kernel DRC of uncertain (nonlinear) inequalities

This section provides an approach that enjoys wide-applicability to general constraints and the elegance of RKHS norm based tightening using the *Kernel DRO* framework of (Zhu et al., 2020b). It introduces a dual form of general DRO problems. The dual problem optimizes a RKHS function h, preserving convexity of the objective function f, while guaranteeing that h majorizes f wherever required, see 1(b)subfigure for an intuition. Using the functional view in Section 3.2, their result implies

$$\sup_{\mathrm{MMD}(\mathcal{H},P,\hat{P})\leq\epsilon} \mathrm{Bias}\,(P; f, \hat{P}) = \inf_{h\geq f} \|h\|_{\mathcal{H}} \cdot \mathrm{MMD}(\mathcal{H}, P, \hat{P}).$$

Compared to (10), we can simply choose $\mathcal{H}$ to be commonly used RKHSs, e.g., Gaussian RKHS, and f needs not be an RKHS function. We make our statement precise in the following proposition, which adapts (Zhu et al., 2020b, Theorem 3.1) to (DRC).

Proposition 3 (Kernel DRC) [1] *Suppose Assumption 1 is satisfied. Let $\mathcal{H}$ be any RKHS and the set $\mathcal{C} \subseteq \mathcal{H}$ be defined by $\mathcal{C} = \{\mu : \int \phi dP = \mu, P \in K\}$. Then, x satisfies (DRC) (5) if there exists $h \in \mathcal{H}$ such that*

$$\textit{(Kernel DRC)} \quad \delta^*_{\mathcal{C}}(h) \leq 0,\; f(x,\xi) \leq h(\xi) \quad \forall\xi \in \mathcal{X}. \tag{12}$$

where $\delta^*_{\mathcal{C}}(h) := \sup_{\mu\in\mathcal{C}}\langle h, \mu\rangle_{\mathcal{H}}$ is the support function.

The reformulation in Proposition 3 is convexity preserving: if the inequality in (DRC) is convex in the decision variable x, so is (12). Proposition 3 provides a flexible tool to handle (DRC) for general nonlinear functions f. For different choices of K, $\delta^*_{\mathcal{C}}(h)$ can be computed in closed form; see (Zhu et al., 2020b, Table 1,3) and (Ben-Tal et al., 2014) for a list of expressions. As a concrete example, we now use Proposition 3 in the setting of scenario approach constraint tightening to account for the distribution shift.

Corollary 4 (RKHS norm regularized distributionally robust scenario approach) [1] *Suppose the ambiguity set in (DRC) (5) is the Minkowski sum $K = \mathrm{conv}(\delta_{\xi_1}, ..., \delta_{\xi_N}) \oplus B_\epsilon$, where $B_\epsilon = \{P : \sup_{\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_P\, f \leq \epsilon\}$. Then, under the same assumption as Proposition 3, x satisfies (DRC) (5)* if *there exists $h \in \mathcal{H}$ such that*

$$\max_i h(\xi_i) + \epsilon\|h\|_{\mathcal{H}} \leq 0, \quad f(x,\xi) \leq h(\xi), \forall\xi \in \mathcal{X}.$$

4. Experiments

In this section we report numerical studies of earlier derived robust reformulations on a double integrator system. We compare our derived MPC methods, gradient regularization and kernel distributionally robust MPC (KDR-MPC), in settings of distribution shift and learning with small datasets. The uncertain inequality constraint is referred to as $C(x, \xi) \leq 0$. We refer to (Hewing and Zeilinger, 2020, (11)(a-g)) for the detailed MPC formulation. Note that x here corresponds to the nominal state z and ξ to the error state e in that MPC formulation. For details about the implementation we refer to the appendix in extended online version of this paper.

Double Integrator We consider the double integrator with quadratic cost that regulates the system to the origin. The dynamics are given by

$$x(k+1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} u + w, \tag{13}$$

where $x(k) \in \mathbb{R}^2$ is the state, $u \in \mathbb{R}$ is the control variable and $w \sim \mathcal{N}(0, 0.1 \cdot I)$ is the additive noise. We consider the simple regulator problem with multiple different constraints and cost

[1] Proofs are provided in the appendix of the extended online version of this paper.

Table 1: Closed-loop MPC constraint satsifaction evaluation

Method	Empirical Probability $\lambda = 0.075$	Empirical Probability $\lambda = 0.1$
Stochastic Tube MPC	0.907	0.812
Gradient regularization	0.999	0.994
Kernel DRC	0.999	0.996

$l(x, u) = x^T Qx + u^T Ru$, with $Q = I$ and $R = 1$. Not only do we treat linear constraints but also other nonlinear constraints, such as an exponential constraint and a learned constraint represented by a SVM, see Remark 2, i.e.,

$$C_{lin}(x, \xi) = |x_2 + \xi_2| - 3 \leq 0, \qquad C_{exp}(x, \xi) = -5 + e^{0.1(x_1+\xi_1)} - x_2 - \xi_2 \leq 0. \tag{14}$$

Experimental setup We present results on the double integrator in the case of model-mismatch. By adding a constant offset to the system matrix $\tilde{A} = A + \lambda\mathbb{I}$, we can vary model-mismatch. Further, we use (4) to compute the desired number of scenarios in our setting. For the empirical results reported in Table 1, the chance constraint satisfaction level is chosen to be $\alpha = 0.85$ and $\beta = 0.15$, resulting in 34 scenarios.

Small sample sizes In Figure 2(*c*)subfigure, we also show results when using Lemma 1 to robustify in small sample-size settings. See appendix for more details.

Gradient regularization As mentioned earlier in Section 3.3, gradient regularization is motivated by the lower bound on the Lipschitz constant and can be used to robustify our MPC problem against shifts in distributions. Setting ϵ to 0 in (11), we recover the regular scenario approach.

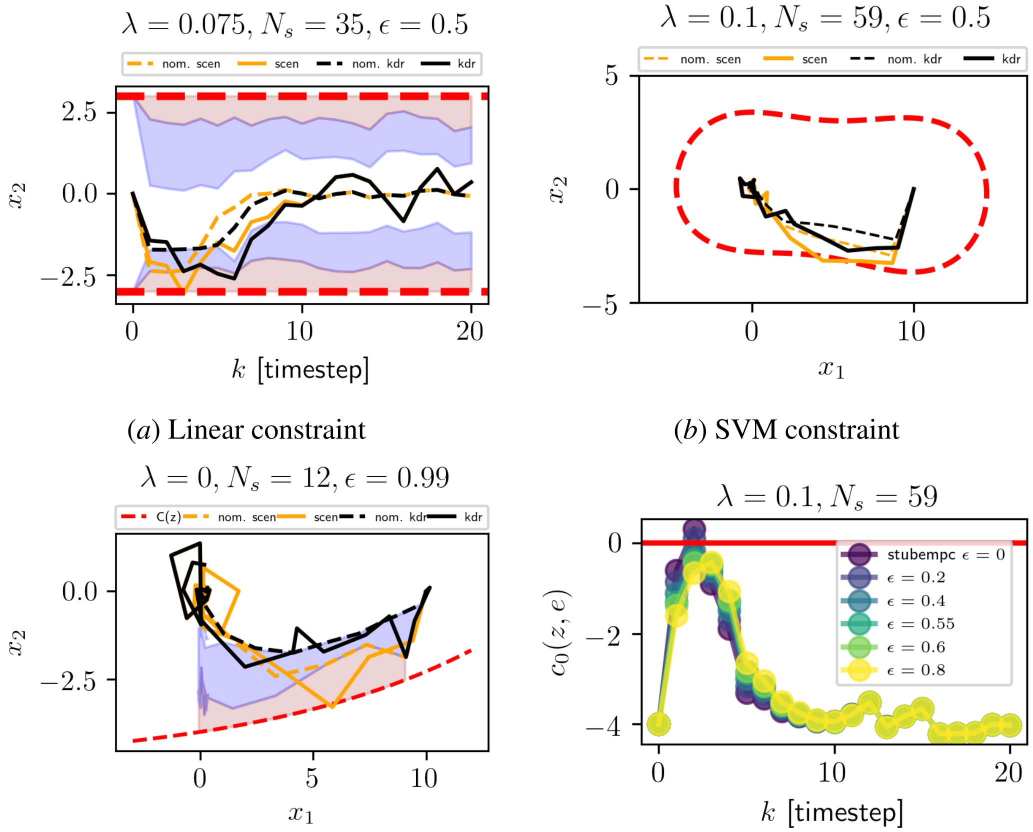
KDR-MPC Following Proposition 3, we can reformulate constraints as the distributionally robust counterpart. We enforce the function $f(\cdot)$ to majorize $C(x, \cdot)$ everywhere. The RKHS function $f(\xi)$ takes the from of $f(\xi) = f_0 + \sum_i \alpha_i k(sv_i, \xi)$, where α_i and f_0 are decision variables and sv_i are the scenarios. Figure 2(*d*)subfigure shows KDR-MPC for various robustness levels with the exponential function constraint from above. We can estimate the radius ϵ of the MMD-ball, using the bound in (1) and an estimate of the MMD between the train set $X_{\text{train}} \sim P_{\text{true}}$ and the test set $X_{\text{test}} \sim P_{\text{shift}}$. Then, we get

$$\epsilon_{verify} = \sqrt{\frac{1}{N}} + \sqrt{\frac{2\log(1/\alpha)}{N}} + \widehat{\text{MMD}}(P_{\text{true}}, P_{\text{shift}}). \tag{15}$$

Learned SVM constraint One can frame the left-hand-side of the constraints as the *decision function* of a classification problem, i.e.,

$$y = +1, \text{ if } C(x) \leq 0, \quad y = -1, \text{ otherwise.} \tag{16}$$

For example, in SVM classification, the kernelized decision function is given by $C(x) = \alpha^T \phi(x) + b$, where α is a vector, $\phi(x)$ are the features associated with the kernel and b is a constant. We follow the (DRC) reformulation of the SVM constraint as detailed in Remark 2. The tightening term involving the RKHS norm $\|C\|_{\mathcal{H}}$ can be computed offline. In that experiment, we create an artifical dataset to learn the SVM. An illustration of an SVM constraint is found in figure 2(*b*)subfigure.



(a) Linear constraint

(b) SVM constraint

(c) No distribution shift, small sample size

(d) Kernel DRC

Figure 2: Comparison of scenario MPC and KDR-MPC on the double integrator with different constraint functions (14), linear constraint in (a), SVM constraint in (b) and exponential constraint in (c) and (d). The orange lines show the scenario approach and black refer to our KDR-MPC. The dashed lines visualize the nominal solutions against the noisy realization shown as a solid line. In (a) and (c) the brown shaded are corresponds to the tightening from the scenario approach and the blue area shows the additional tightening due to our regularization. The additive noise $w \sim \mathcal{N}(0, 0.2 \cdot I)$ has slightly increased variance in (c). In (d) we verify the regularization with $\epsilon_{verify} = 0.55$, according to (15).

5. Discussion

Future directions of this paper include the application of our methods to nonlinear stochastic MPC as well as learning-based MPC. Notably, coupling our (Kernel DRC) approach with the convergence result in Lemma 1 is a promising direction for new statistical guarantees. Another direction is to design tailored numerical methods for DRO in MPC, instead of the off-the-shelf solvers used in our current experiments.

References

Güzin Bayraksan and David K. Love. Data-Driven Stochastic Programming Using Phi-Divergences. In Dionne Aleman, Aurélie Thiele, J. Cole Smith, and Harvey J. Greenberg, editors, *The Operations Research Revolution*, pages 1–19. INFORMS, September 2015. ISBN 978-0-9843378-8-0. doi: 10.1287/educ.2015.0134.

Alberto Bemporad and Manfred Morari. Robust model predictive control: A survey. In *Robustness in Identification and Control*, pages 207–226. Springer, 1999.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, February 2013. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.1120.1641.

Aharon Ben-Tal, Dick den Hertog, and Jean Philippe Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1-2):265–299, 2014. ISSN 14364646. doi: 10.1007/s10107-014-0750-8.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and Applications of Robust Optimization. *SIAM Review*, 53(3):464–501, January 2011. ISSN 0036-1445, 1095-7200. doi: 10.1137/080734510.

Alberto Bietti, Grégoire Mialon, Dexiong Chen, and Julien Mairal. A Kernel Perspective for Regularizing Deep Neural Networks. *arXiv:1810.00363 [cs, stat]*, May 2019.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein Profile Inference and Applications to Machine Learning. *Journal of Applied Probability*, 56(03):830–857, September 2019. ISSN 0021-9002, 1475-6072. doi: 10.1017/jpr.2019.49.

Angelo D. Bonzanini, Tito L.M. Santos, and Ali Mesbah. Tube-based stochastic nonlinear model predictive control: A comparative study on constraint tightening. *IFAC-PapersOnLine*, 52(1): 598–603, 2019. ISSN 24058963. doi: 10.1016/j.ifacol.2019.06.128.

M. C. Campi and S. Garatti. A Sampling-and-Discarding Approach to Chance-Constrained Optimization: Feasibility and Optimality. *Journal of Optimization Theory and Applications*, 148(2): 257–280, 2011. ISSN 00223239. doi: 10.1007/s10957-010-9754-6.

Marco C. Campi and Algo Carè. Random convex programs with L1-regularization: Sparsity and generalization. *SIAM Journal on Control and Optimization*, 51(5):3532–3557, sep 2013. ISSN 03630129. doi: 10.1137/110856204.

Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58(3):595–612, June 2010. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.1090.0741.

Moritz Diehl and Jakob Bjornberg. Robust dynamic programming for min-max model predictive control of constrained uncertain systems. *IEEE Transactions on Automatic Control*, 49(12):2253–2257, 2004.

Moritz Diehl, Hans Georg Bock, and Ekaterina Kostina. An approximation technique for robust nonlinear optimization. *Mathematical Programming*, 107(1-2):213–230, June 2006. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-005-0685-1.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *arXiv:1610.03425 [stat]*, June 2018.

Marcello Farina, Luca Giulioni, and Riccardo Scattolini. Stochastic linear Model Predictive Control with chance constraints - A review. *Journal of Process Control*, 44:53–67, aug 2016. ISSN 09591524. doi: 10.1016/j.jprocont.2016.03.005.

Luisa Turrin Fernholz. *Von Mises Calculus for Statistical Functionals*, volume 19. Springer Science & Business Media, 2012.

Simone Formentin, Simone Garatti, Marco C. Campi, and Sergio M. Savaresi. Tuning regularization via scenario optimization. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 635–640, Melbourne, Australia, December 2017. IEEE. ISBN 978-1-5090-2873-3. doi: 10.1109/CDC.2017.8263732.

Rui Gao and Anton J. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv:1604.02199 [math]*, July 2016.

Joel Goh and Melvyn Sim. Distributionally Robust Optimization and Its Tractable Approximations. *Operations Research*, 58(4-part-1):902–917, August 2010. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.1090.0795.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, March 2012. ISSN 1533-7928.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*, volume 196. John Wiley & Sons, 2011.

Lukas Hewing and Melanie N. Zeilinger. Scenario-Based Probabilistic Reachable Sets for Recursively Feasible Stochastic Model Predictive Control. *IEEE Control Systems Letters*, 4(2): 450–455, 2020. ISSN 24751456. doi: 10.1109/LCSYS.2019.2949194.

Garud N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. ISSN 0364-765X.

Johannes Köhler, Raffaele Soloperto, Matthias A. Müller, and Frank Allgöwer. A computationally efficient robust model predictive control framework for uncertain nonlinear systems. 2019. ISSN 0018-9286. doi: 10.1109/tac.2020.2982585.

W. Langson, I. Chryssochoos, S. V. Raković, and D. Q. Mayne. Robust model predictive control using tubes. *Automatica*, 40(1):125–133, jan 2004. ISSN 00051098. doi: 10.1016/j.automatica.2003.08.009.

Christoph Mark and Steven Liu. Stochastic MPC with Distributionally Robust Chance Constraints. *arXiv preprint arXiv:2005.00313 [math]*, 2020.

D. Q. Mayne, M. M. Seron, and S. V. Raković. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica*, 41(2):219–224, feb 2005. ISSN 00051098. doi: 10.1016/j.automatica.2004.08.019.

Ali Mesbah. Stochastic model predictive control: An overview and perspectives for future research. *IEEE Control Systems Magazine*, 36(6):30–44, Dec 2016. ISSN 1941-000X. doi: 10.1109/MCS.2016.2602087.

Johannes Milz and Michael Ulbrich. An Approximation Scheme for Distributionally Robust Nonlinear Optimization. *SIAM Journal on Optimization*, 30(3):1996–2025, January 2020. ISSN 1052-6234, 1095-7189. doi: 10.1137/19M1263121.

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, sep 2018. ISSN 14364646. doi: 10.1007/s10107-017-1172-1.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, September 2018. ISSN 1436-4646. doi: 10.1007/s10107-017-1172-1.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Zoltan K. Nagy and Richard D. Braatz. Open-loop and closed-loop robust optimal control of batch processes using distributional and worst-case analysis. *Journal of Process Control*, 14(4):411–422, June 2004. ISSN 09591524. doi: 10.1016/j.jprocont.2003.07.004.

Arnab Nilim and Laurent El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, October 2005. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.1050.0216.

Hamed Rahimian and Sanjay Mehrotra. Distributionally Robust Optimization: A Review. *arXiv:1908.05659 [cs, math, stat]*, 2019.

Tito L.M. Santos, Angelo D. Bonzanini, Tor Aksel N. Heirung, and Ali Mesbah. A constraint-tightening approach to nonlinear model predictive control with chance constraints for stochastic systems. *Proceedings of the American Control Conference*, 2019-July(July):1641–1647, 2019. ISSN 07431619. doi: 10.23919/acc.2019.8814623.

Herbert Scarf. A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, 1958.

C. Shang and F. You. Chance Constrained Model Predictive Control via Active Uncertainty Set Learning and Calibration. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2605–2610, December 2018. doi: 10.1109/CDC.2018.8619665.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

Nian Si, Jose Blanchet, Soumyadip Ghosh, and Mark Squillante. Quantifying the Empirical Wasserstein Distance to a Set of Measures: Beating the Curse of Dimensionality. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, page 11.

A. L. Soyster. Technical Note—Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming. *Operations Research*, 21(5):1154–1157, October 1973. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.21.5.1154.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. ISSN 1935-7524. doi: 10.1214/12-EJS722.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ilya Tolstikhin, Bharath K. Sriperumbudur, Krikamol Mu, and et. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017. URL http://jmlr.org/papers/v18/17-032.html.

R. v. Mises. On the Asymptotic Distribution of Differentiable Statistical Functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947. ISSN 0003-4851.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3835–3844. Curran Associates, Inc., 2018.

Zizhuo Wang, Peter W. Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, April 2016. ISSN 1619-6988. doi: 10.1007/s10287-015-0240-3.

Weijun Xie. On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming*, November 2019. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-019-01445-5.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and Regularization of Support Vector Machines. *Journal of machine learning research*, 10(7), 2009.

Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, March 2018. ISSN 01676377. doi: 10.1016/j.orl.2018.01.011.

J.-J. Zhu, M. Diehl, and B. Schölkopf. A kernel mean embedding approach to reducing conservativeness in stochastic programming and control. In *2nd Annual Conference on Learning for Dynamics and Control (L4DC)*, volume 120 of *Proceedings of Machine Learning Research*, pages 915–923. PMLR, June 2020a. URL http://proceedings.mlr.press/v120/zhu20a.html.

Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel Distributionally Robust Optimization. *arXiv:2006.06981 [cs, math, stat]*, June 2020b.

Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Worst-case risk quantification under distributional ambiguity using kernel mean embedding in moment problem. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3457–3463, Dec 2020c. doi: 10.1109/CDC42340.2020.9303938.

Steve Zymler, Daniel Kuhn, and Berç Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, February 2013. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-011-0494-7.