MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxxx 20xx, pp. xxx–xxx
ISSN 0364-765X | EISSN 1526-5471 |xx|0000|0xxx

DOI 10.1287/moor.xxxx.xxxx

A Distributional Interpretation of Robust Optimization

Huan Xu
Department of Electrical and Computer Engineering, The University of Texas at Austin, USA
email: huan.xu@mail.utexas.edu

Constantine Caramanis
Department of Electrical and Computer Engineering, The University of Texas at Austin, USA
email: caramanis@mail.utexas.edu

Shie Mannor
Department of Electrical Engineering, Technion, Israel
email: shie.mannor@ee.technion.ac.il

Motivated by data-driven decision making and sampling problems, we investigate probabilistic interpretations of Robust Optimization (RO). We establish a connection between RO and Distributionally Robust Stochastic Programming (DRSP), showing that the solution to any RO problem is also a solution to a DRSP problem. Specifically, we consider the case where *multiple* uncertain parameters belong to the same fixed dimensional space, and find the set of distributions of the equivalent DRSP. The equivalence we derive enables us to construct RO formulations for sampled problems (as in stochastic programming and machine learning) that are statistically consistent, even when the original sampled problem is not. In the process, this provides a systematic approach for tuning the uncertainty set. The equivalence further provides a probabilistic explanation for the common shrinkage heuristic, where the uncertainty set used in a RO problem is a shrunken version of the original uncertainty set.

Key words: robust optimization, distributionally robust stochastic program, consistency, machine learning, kernel density estimator

MSC2000 Subject Classification: P90C99

OR/MS subject classification: Programming: Stochastic; Statistics: Nonparametric

History: Received: Xxxx xx, xxxx; Revised: Yyyyyy yy, yyyy and Zzzzzz zz, zzzz.

1. Introduction Robust Optimization (RO) considers deterministic (set-based) uncertainty models in optimization, where a (potentially malicious) adversary has a bounded capability to change the parameters of the function the decision-maker seeks to optimize. Thus, the standard optimization problem[1]

$$\max_{v}: \ f(v), \tag{1}$$

becomes

$$\max_{v}: \ \min_{\mathbf{x}\in\mathcal{Z}}: \ f(v,\mathbf{x}), \tag{2}$$

where the vector $\mathbf{x}$ denotes some uncertain parameters of the objective function f, and may take any value in the set $\mathcal{Z}$. This approach to uncertainty has a long history in control; in optimization it traces back several decades to the early work in Soyster [32]. Particularly in the last decade since the work of Ben-Tal and Nemirovski [3, 4], Bertsimas and Sim [7], and El Ghaoui and Lebret [16], it has become a common approach in operations research, computer science, engineering, and many other related fields (e.g., Shivaswamy et al. [31], Lanckriet et al. [20], El Ghaoui and Lebret [16], Ben-Tal et al. [5, 6], and Boyd et al. [10]); see the recent monograph by Ben-Tal and co-authors [2] for a detailed survey. A key reason for its success has been its computational tractability and the fact that robustified versions of

[1] We give the unconstrained version here without loss of generality, since one can let the objective function be $-\infty$ for infeasible solutions.

many common optimization classes (linear programming, second order cone programming, among others) remain relatively easy to solve.

A much-researched alternative to RO's set-based uncertainty, is to represent the uncertain parameter in a probabilistic way, i.e., assume that the parameter $\mathbf{x}$ is a random variable with distribution μ^*. If we assume the generating distribution, μ^*, is known, the result is the standard stochastic programming paradigm (e.g., Birge and Louveaux [8], Prékopa [23], and Shapiro [28]). If μ^* is not precisely known, and instead μ^* is only known to lie in some set of distributions, $\mathcal{D}$, the resulting optimization formulation is the so-called Distributionally Robust Stochastic Program (DRSP), initially proposed in Scarf [25], almost two decades before the first appearance of RO. In DRSP, the decision maker solves the following problem

$$\max_{v} : \ \min_{\mu\in\mathcal{D}} : \ \mathbb{E}_{\mathbf{x}\sim\mu} f(v,\mathbf{x}). \tag{3}$$

Since its introduction, DRSP has attracted extensive research (e.g., Kall [17], Dupacová [15], Popescu [22], Shapiro [29], Goh and Sim [19], Delage and Ye [12]).

The main focus of this paper is the relationship of these two paradigms. In particular, we show in Section 2 that RO can be reformulated as a DRSP with respect to a particular class of distributions. For the special case where each uncertain parameter belongs to a different space, such a re-interpretation is a well known folk theorem (see Delage and Ye [12]). Yet as we discuss below, in data-driven (or sample-based) optimization problems such as those appearing in stochastic optimization and machine learning, the uncertain parameters belong *to the same space*. To develop a framework for these problems, we generalize the equivalence of RO and DRSP to the setting where multiple uncertain parameters $\mathbf{x}_1,\ldots,\mathbf{x}_n$ belong to the same space $\mathbb{R}^m$. Instead of formulating the RO problem for such a problem as a DRSP with respect to a class of distributions *supported on the product space* $\mathbb{R}^{m\times n}$, as techniques from the standard literature would require, we seek to find a DRSP interpretation with respect to a class of distributions *supported on* $\mathbb{R}^m$. We now elaborate on this, explaining in particular the significance of this generalized equivalence.

Relationship to Optimization from Samples The setup we consider is motivated by sampling problems. In solving

$$\min_{v} : \ \mathbb{E}_{\mathbf{x}\sim\mu^*} f(v,\mathbf{x}),$$

a sampled distribution $(1/n)\sum_{i=1}^n \delta_{\mathbf{x}_i}$ is often used instead of the true (potentially continuous) distribution μ^*. This is often done in machine learning because the true distribution is unknown, and the decision-maker has only access to a finite set of samples generated from that distribution. In stochastic programming this is widely applied, either when the distribution is unknown, or when it is known but too complicated to manipulate directly within an optimization problem, and hence the empirical distribution is used instead.

It is well known that such a sampling approach may not always be consistent — that is, even as the number of samples goes to infinity, the solution recovered may remain a bounded distance away from the true optimal solution. A DRSP interpretation of RO with respect to a class of distributions supported on the same fixed dimensional space would enable us to examine how well or poorly elements of this class of distributions approximate the true (potentially unknown) distribution as the sample size, n, increases. In Section 3, we explore how such an approach can be used to prove that a robust optimization formulation is asymptotically statistically consistent. Moreover, we show how, given a sampled optimization problem, one can design a robustified formulation that is guaranteed to be consistent, even if the original sampled problem fails to be consistent. As an important byproduct, we obtain a systematic way to tune the uncertainty set to guarantee consistency.

In Section 4, we investigate other implications of the equivalence between RO and DRSP. We start from discussing optimization in a stochastic programming setup, and particularly in machine learning in Section 4.1. In Section 4.2, we provide an explanation for the shrinkage heuristic in RO where instead of using the original uncertainty set one uses a shrunken version to obtain better performance in practice. In Section 5 we further illustrate some possible extensions of the equivalence relationship.

Notation: Throughout the paper, without loss of generality, we assume the unknown parameters belong to $\mathbb{R}^m$. We use $\mathcal{P}$ to denote the set of probability distributions on $\mathbb{R}^m$ (with respect to the Borel set). We use $[1:n]$ to denote the set $\{1,2,\cdots,n\}$. A Euclidean ball centered at $\mathbf{x}$ with a radius r is denoted by $\mathcal{B}(\mathbf{x},r)$.

2. Distributional interpretation for robust optimization In this section, we turn our attention to the relationship between RO and DRSP. We consider a general case when *multiple* uncertain parameters lie in the same space, as opposed to the Cartesian product of the space of each parameter. As discussed above, this setting arises naturally in data-driven problems. We prove a statement that is slightly stronger than the equivalence of RO and DRSP: fixing a candidate solution, the worst case reward of RO is equivalent to the minimal expected reward of DRSP. That is, the inner minimization of RO is equivalent to the inner minimization of DRSP. Hence, in Theorem 2.1 and the proof, we suppress the decision variable v, in order to reduce unnecessary notation.

THEOREM 2.1 *Given a measurable function $f : \mathbb{R}^m \to \mathbb{R} \bigcup \{-\infty\}$, $c_1, \ldots, c_n > 0$ such that $\sum_{i=1}^n c_i = 1$, and non-empty Borel sets $\mathcal{Z}_1, \ldots, \mathcal{Z}_n \subseteq \mathbb{R}^m$, let*

$$\mathcal{P}_n \triangleq \{\mu \in \mathcal{P} \,|\, \forall S \subseteq [1:n] : \mu(\bigcup_{i \in S} \mathcal{Z}_i) \geq \sum_{i \in S} c_i\}.$$

Then the following holds

$$\sum_{i=1}^n c_i \inf_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i) = \inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}). \tag{4}$$

Note that the uncertainty sets $\mathcal{Z}_1, \cdots, \mathcal{Z}_n$ can have nonempty intersection, or even be identical, as is the case when the points are sampled from the same space.

PROOF. We can assume without loss of generality that $\inf_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i) > -\infty$ for $i = 1, 2, \ldots, n$, since otherwise both sides equal $-\infty$ and the theorem holds trivially.

Let $\hat{\mathbf{x}}_i$ be an ϵ-optimal solution to $\inf_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i)$. We first show that the left-hand-side of Equation (4) is larger or equal to the right-hand-side.

Consider the following distribution

$$\hat{\mu}(\{\hat{\mathbf{x}}_i\}) = c_i + \sum_{j \neq i : \hat{\mathbf{x}}_j = \hat{\mathbf{x}}_i} c_j.$$

Observe that $\hat{\mu}$ belongs to $\mathcal{P}_n$, and

$$\sum_{i=1}^n c_i f(\hat{\mathbf{x}}_i) = \int_{\mathbb{R}^m} f(\mathbf{x}) d\hat{\mu}(\mathbf{x}).$$

Thus we have

$$\sum_{i=1}^n c_i \inf_{\mathbf{x} \in \mathcal{Z}_i} f(\mathbf{x}_i) + \epsilon \geq \inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}).$$

Since ϵ can be arbitrarily close to zero, we have

$$\sum_{i=1}^n \inf_{\mathbf{x}_i \in \mathcal{Z}_i} c_i f(\mathbf{x}_i) \geq \inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}). \tag{5}$$

We next prove the reverse inequality to complete the proof. By re-indexing if necessary, assume

$$f(\hat{\mathbf{x}}_1) \geq f(\hat{\mathbf{x}}_2) \geq \cdots \geq f(\hat{\mathbf{x}}_n). \tag{6}$$

Now construct the following function

$$\hat{f}(\mathbf{x}) \triangleq \begin{cases} \max_{i | \mathbf{x} \in \mathcal{Z}_i} f(\hat{\mathbf{x}}_i) & \text{if } \mathbf{x} \in \bigcup_{j=1}^n \mathcal{Z}_j; \\ f(\mathbf{x}) & \text{otherwise.} \end{cases}$$

Observe that $f(\mathbf{x}) \geq \hat{f}(\mathbf{x}) - \epsilon$ for all $\mathbf{x}$.

Furthermore, fixing a $\mu \in \mathcal{P}_n$, we have

$$\int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}) + \epsilon \geq \int_{\mathbb{R}^m} \hat{f}(\mathbf{x}) d\mu(\mathbf{x}) = \sum_{k=1}^n f(\hat{\mathbf{x}}_k) \Big[\mu(\bigcup_{i=1}^k \mathcal{Z}_i) - \mu(\bigcup_{i=1}^{k-1} \mathcal{Z}_i) \Big].$$

Here the inequality holds because $f(\mathbf{x}) \geq \hat{f}(\mathbf{x}) - \epsilon$, and the equality holds because $f(\hat{\mathbf{x}}_1) \geq f(\hat{\mathbf{x}}_2) \geq \cdots \geq f(\hat{\mathbf{x}}_n)$.

Define $\alpha_k \triangleq \left[\mu(\bigcup_{i=1}^{k} \mathcal{Z}_i) - \mu(\bigcup_{i=1}^{k-1} \mathcal{Z}_i)\right]$. Then, by telescoping and the fact that $\mu \in \mathcal{P}_n$, we have

$$\sum_{i=k}^{t} \alpha_k = \mu(\bigcup_{k=1}^{t} \mathcal{Z}_k) \geq \sum_{k=1}^{t} c_k; \quad \sum_{k=1}^{n} \alpha_k = \mu(\bigcup_{k=1}^{n} \mathcal{Z}_k) = 1.$$

Hence, since $f(\hat{\mathbf{x}}_1) \geq f(\hat{\mathbf{x}}_2) \geq \cdots \geq f(\hat{\mathbf{x}}_n)$, we have

$$\sum_{k=1}^{n} \alpha_k f(\hat{\mathbf{x}}_k) \geq \sum_{k=1}^{n} c_k f(\hat{\mathbf{x}}_k).$$

Thus, for any $\mu \in \mathcal{P}_n$, we have

$$\int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}) + \epsilon \geq \sum_{k=1}^{n} c_k f(\hat{\mathbf{x}}_k) \geq \sum_{k=1}^{n} c_k \inf_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_k).$$

Since ϵ and μ are arbitrary, this leads to

$$\inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}) \geq \sum_{k=1}^{n} c_k \inf_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_k).$$

Combining this with Equation (5), we establish the theorem. □

We observe that if $\mathcal{Z}_i$ are disjoint, then the equivalent distributional set $\mathcal{P}_n$ has the following form:

$$\mathcal{P}_n = \{\mu \in \mathcal{P} | \mu(\mathcal{Z}_i) = c_i, \ i = 1, \cdots, n\}.$$

Taking $n = 1$, this reduces to the following theorem, well known in the literature (e.g., Delage and Ye [12]), stating that the solution to a robust optimization problem is the solution to a special DRSP problem, where the distributional set is the one that contains all distributions whose support is contained in the uncertainty set, *in the Cartesian product of the space of each parameter.*

COROLLARY 2.1 *Given a measurable function $f : \mathbb{R}^m \to \mathbb{R} \bigcup \{-\infty\}$, and a non-empty Borel set $\mathcal{Z} \subseteq \mathbb{R}^m$, the following holds:*

$$\inf_{\mathbf{x}' \in \mathcal{Z}} f(\mathbf{x}') = \inf_{\mu \in \mathcal{P} | \mu(\mathcal{Z}) = 1} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}). \tag{7}$$

3. Consistency of robust optimization The theoretical equivalence established in Theorem 2.1 has algorithmic consequences as well. In this section, we consider sampled stochastic optimization problems. As discussed above, it is well-known that these problems may fail to be asymptotically consistent, i.e., even as the number of samples goes to infinite, we may not recover the optimal solution. We use the equivalence relationship stated in Theorem 2.1 to construct a sequence of robust optimization problems that are asymptotically consistent in a statistical sense, even if the original sampled problem fails consistency. En route, this construction provides a systematic approach to choosing the appropriate size of the uncertainty set.

The main theorem of the section states that as long as the utility function $f(\cdot, \cdot)$ is bounded, and satisfies a mild continuity condition, then an explicitly stated robust optimization formulation for the sampled stochastic optimization, is asymptotically consistent. That is, it recovers the optimal solution to

$$\max: \ \mathbb{E}_{\mathbf{x} \sim \mu}[f(v, \mathbf{x})],$$

where μ is given only through a sequence of i.i.d. samples. We note that these conditions are weaker than those required for consistency of sampled stochastic programs, e.g., as in King and Wets [18]. Thus, Theorem 3.1 provides a computationally tractable avenue for developing algorithms for solution of stochastic programming with stronger consistency guarantees (that is, they require weaker assumptions on the problem). We provide several examples of this in the next section.

THEOREM 3.1 *Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots$ are i.i.d. samples of a distribution h^* on $\mathbb{R}^m$, and the utility function $f(\cdot, \cdot)$ satisfies*

(*i*) *Boundedness:* $\max_{v,\mathbf{x}} |f(v,x)| \leq C$.

(*ii*) *Equicontinuity:* $d(\epsilon) \downarrow 0$ *where*

$$d(\epsilon) \triangleq \max_{v,\mathbf{x},\|\boldsymbol{\delta}\|_\infty \leq \epsilon} |f(v,\mathbf{x}) - f(v,\mathbf{x}+\boldsymbol{\delta})|.$$

If $\{\epsilon(n)\}$ *satisfies*

$$\epsilon(n) \downarrow 0; \quad n\epsilon(n)^m \uparrow \infty,$$

then the sequence of optimal solutions to the RO formulation

$$v(n) \triangleq \arg\max_v \ \frac{1}{n}\sum_{i=1}^n \inf_{\|\boldsymbol{\delta}_i\|_\infty \leq \epsilon(n)} f(v, \mathbf{x}_i + \boldsymbol{\delta}_i)$$

satisfies

$$\lim_n \int_{\mathbb{R}^m} f(v(n),\mathbf{x})h^*(\mathbf{x})d\mathbf{x} = \inf_v \int_{\mathbb{R}^m} f(v,\mathbf{x})h^*(\mathbf{x})d\mathbf{x}.$$

That is, the RO formulation is consistent.

Proof. The key to the proof rests on the equivalence established in Theorem 2.1. This equivalence then allows us to show that the RO formulation given in the theorem statement, is equivalent to a DSRP, whose distribution set contains a Kernel Density Estimator (KDE). From here, the proof is essentially immediate: exploiting the fact that a KDE converges to the generating distribution in the ℓ_1 sense, one can easily conclude that the sequence of solutions to the RO problems given, are consistent. For completeness, we include a brief introduction to Kernel Density Estimators in the Appendix.

Consider a fixed n. Define the sets

$$\mathcal{Z}_i \triangleq \left\{\mathbf{x}_i + \boldsymbol{\delta} \,\middle|\, \|\boldsymbol{\delta}_i\|_\infty \leq \epsilon(n)\right\}.$$

Thus, the RO formulation is to maximize $\sum_{i=1}^n \frac{1}{n} \inf_{\mathbf{x}'_i \in \mathcal{Z}_i} f(v,\mathbf{x}'_i)$. By Theorem 2.1, we know that for any v, the following holds:

$$\begin{aligned} &\sum_{i=1}^n \frac{1}{n} \inf_{\mathbf{x}'_i \in \mathcal{Z}_i} f(v,\mathbf{x}'_i) = \inf_{\tilde{\mu} \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(v,\mathbf{x}) d\tilde{\mu}(\mathbf{x}); \\ &\text{where:} \quad \mathcal{P}_n = \{\mu \in \mathcal{P} | \forall S \subseteq [1:n] : \mu(\bigcup_{i\in S} \mathcal{Z}_i) \geq |S|/n\}. \end{aligned} \tag{8}$$

Next we show that the set of distributions, $\mathcal{P}_n$, contains a kernel density estimator. Consider a distribution h_n defined as

$$\begin{aligned} h_n(\mathbf{x}) &= (n\epsilon(n)^m)^{-1} \sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\epsilon}\right); \\ \text{where:} \quad K(\mathbf{z}) &= \frac{\mathbf{1}(\|\mathbf{z}\|_\infty \leq 1)}{2^m}. \end{aligned}$$

Indeed, observe that h_n is a kernel density estimator. Now, for any $S \subseteq [1:n]$, we have

$$\begin{aligned} &\int_{\mathbb{R}^m} \mathbf{1}(\mathbf{x} \in \bigcup_{j\in S} \mathcal{Z}_j) h_n(\mathbf{x}) d\mathbf{x} \\ =& \int_{\mathbb{R}^m} \mathbf{1}(\mathbf{x} \in \bigcup_{j\in S} \mathcal{Z}_j)(n\epsilon(n)^m)^{-1} \sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\epsilon(n)}\right) d\mathbf{x} \\ \geq& \int_{\mathbb{R}^m} \mathbf{1}(\mathbf{x} \in \bigcup_{j\in S} \mathcal{Z}_j)(n\epsilon(n)^m)^{-1} \sum_{i\in S} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\epsilon(n)}\right) d\mathbf{x} \\ =& \sum_{i\in S} \int_{\mathbb{R}^m} \mathbf{1}(\mathbf{x} \in \bigcup_{j\in S} \mathcal{Z}_j)(n\epsilon(n)^m)^{-1} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\epsilon(n)}\right) d\mathbf{x} \\ \geq& \sum_{i\in S} \int_{\mathbb{R}^m} \mathbf{1}(\mathbf{x} \in \mathcal{Z}_i)(n\epsilon(n)^m)^{-1} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\epsilon(n)}\right) d\mathbf{x} \\ \overset{(a)}{=}& \sum_{i\in S} \int_{\mathbb{R}^m} (n\epsilon(n)^m)^{-1} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\epsilon(n)}\right) d\mathbf{x} = |S|/n. \end{aligned}$$

Here, the second-to-last equality, (a), holds because, due to the definition of K and $\mathcal{Z}_i$, $K((\mathbf{x}-\mathbf{x}_i)/\epsilon(n))$ is non-zero only when $\mathbf{x} \in \mathcal{Z}_i$. Hence, $h_n \in \mathcal{P}_n$, [2] which by Equation (8) implies

$$\sum_{i=1}^{n} \frac{1}{n} \inf_{\mathbf{x}_i' \in \mathcal{Z}_i} f(v, \mathbf{x}_i') \leq \int_{\mathbb{R}^m} f(v, \mathbf{x}) h_n(\mathbf{x}) d\mathbf{x}.$$

Since h_n is a kernel density estimator, it is well-known (e.g., see Devroye and Györfi [13]) that under the condition that $\epsilon(n) \to 0$ and $n\epsilon(n)^m \to \infty$ the following holds,

$$\int_{\mathbb{R}^m} \left| h_n(\mathbf{x}) - h^*(\mathbf{x}) \right| d\mathbf{x} \overset{n}{\to} 0.$$

Therefore, since $|f(v, \mathbf{x})| \leq C$, there exists $\{M_n\} \to 0$ such that the following holds for all v,

$$\int_{\mathbb{R}^m} f(v, \mathbf{x}) h_n(\mathbf{x}) d\mathbf{x} \leq \int_{\mathbb{R}^m} f(v, \mathbf{x}) h^*(\mathbf{x}) d\mathbf{x} + M_n,$$

which leads to for all v,

$$\sum_{i=1}^{n} \frac{1}{n} \inf_{\mathbf{x}_i' \in \mathcal{Z}_i} f(v, \mathbf{x}_i') - M_n \leq \int_{\mathbb{R}^m} f(v, \mathbf{x}) h^*(\mathbf{x}) d\mathbf{x}.$$

By symmetry, we also have

$$\int_{\mathbb{R}^m} f(v, \mathbf{x}) h^*(\mathbf{x}) d\mathbf{x} \leq \sum_{i=1}^{n} \frac{1}{n} \sup_{\mathbf{x}_i' \in \mathcal{Z}_i} f(v, \mathbf{x}_i') + M_n.$$

Further note that

$$\sup_{\mathbf{x} \in \mathcal{Z}_i} f(v, \mathbf{x}) - \inf_{\mathbf{x} \in \mathcal{Z}_i} f(v, \mathbf{x}) \leq d(2\epsilon(n)).$$

Thus we have for all v

$$\sum_{i=1}^{n} \frac{1}{n} \inf_{\mathbf{x}_i' \in \mathcal{Z}_i} f(v, \mathbf{x}_i') - M_n \leq \int_{\mathbb{R}^m} f(v, \mathbf{x}) h^*(\mathbf{x}) d\mathbf{x} \leq \sum_{i=1}^{n} \frac{1}{n} \inf_{\mathbf{x}_i' \in \mathcal{Z}_i} f(v, \mathbf{x}_i') + M_n + d(2\epsilon(n)).$$

Since both M_n and $d(2\epsilon(n))$ go to zero, the theorem follows easily. □

REMARK 3.1 Observe from the proof that if we relax the requirement of equicontinuity of $\{f(v, \cdot)\}$, then RO is essentially maximizing an asymptotic lower bound of the true expected reward. Furthermore, if we instead only require the equicontinuity of $\{f(v(n), \cdot) | n = 1, 2, \cdots\}$, then the consistency result still holds. In fact, as $v(n)$ is the optimal solution of a robust optimization, this condition is much easier to satisfy than the equicontinuity of $\{f(v, \cdot)\}$.

REMARK 3.2 Theorem 3.1 suggests a methodology for choosing an appropriate size $\epsilon(n)$ of the uncertainty set. Previous work on RO (e.g., Bertsimas and Sim [7]) considers the setting where the observed parameter is the result of corruption of the true parameter (via additive noise). Consequently, the decision maker tunes the size of the uncertainty set used in the RO formulation, to satisfy some probabilistic bounds or risk measure constraints, given *a priori* information of the noise and in particular the noise magnitude. Theorem 3.1 provides a different paradigm for the design of the uncertainty set: when the uncertainty is due to inherent randomness of the parameters, then to approximately solve the stochastic program, the decision maker can use RO, with the size of uncertainty set slowly *decreasing* in the number of samples. To be more specific, the theorem shows that if the size scales like $o(1)$ and $\omega(n^{-1/m})$, then we achieve asymptotic consistency. Finally, also note that it is straightforward to modify the uncertainty set from a ℓ_∞ ball to other parameterized uncertainty sets. Indeed, the only requirement is that the family of distributions obtained using Theorem 2.1, contains a kernel density estimator.

4. Implications of Theorem 3.1 In this section, we describe the applications of Theorem 3.1 to consistency of sampled optimization problems in machine learning (Support Vector Machines, and also Lasso, or ℓ_1-regularized regression) and then to the popular but as-of-yet not well-understood technique known as shrinkage, used, e.g., in Chapter 2 of [2], and [40].

[2] More precisely, h_n is the density function of a probability measure that belongs to $\mathcal{P}_n$.

4.1 Sampled Optimization. Many decision problems have the form

$$\max_v : \quad \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\{f(v,\mathbf{x})\}, \tag{9}$$

where the expectation is difficult or impossible to optimize, or even evaluate exactly. In Machine Learning problems, the typical set up is that the distribution $\mathbb{P}$ is unknown, and the decision-maker has knowledge of $\mathbb{P}$ only through a set of samples (cf. textbooks such as Anthony and Bartlett [1]). As mentioned above, one often sees the same setting in SP, where even if the distribution is known, it may be too complicated to evaluate (cf textbooks such as Birge and Louveaux [8]). A sampling approach is often used instead (Shapiro and de Mello [30]), i.e., to make a decision simply by solving the sampled optimization problem:

$$v_n^* \triangleq \arg\max_v \frac{1}{n}\sum_{i=1}^n f(v,\mathbf{x}_i).$$

One would hope that the solution to such a problem is a good approximation of the solution of Problem (9). However, due to the fact that the decision obtained is dependent on the samples, the empirical utility for v_n^* is a biased estimate of its expected utility. Thus, the sampling technique often yields overly optimistic solutions. Even worse, it may be the case that as $n \uparrow \infty$, the sequence $\{v_n^*\}$ does not converge to the optimal decision. This is often termed "over-fitting" in the machine learning literature (Vapnik and Chervonenkis [36]), and has attracted extensive research (see textbooks such as Anthony and Bartlett [1], Devroye et al. [14], and many others). There are various approaches, often custom-tailored to the problem at hand, to try to control the problem of overfitting, with regularization being one of the foremost such tools.

Theorem 3.1 provides a unified approach to mitigate this unjustified optimism: one may assume some set-based uncertainty in the samples observed, and subsequently solve the corresponding robust counterpart. If the uncertainty sets are chosen properly (i.e., as in Theorem 3.1), then the *corresponding class* $\mathcal{P}_n$ to this robust optimization problem "approximately" contains the true distribution $\mathbb{P}$, i.e., it contains a sequence of distributions that converges to $\mathbb{P}$ uniformly with respect to v. The theorem then guarantees that the sequence of min-max decisions converges to the optimal solution.

It turns out that this property is in fact exploited implicitly (i.e., unwittingly) by many widely used learning algorithms, and hence one can show (post hoc) that this serves as an explanation for their success. We now give two examples which we adapt from our work in [37] and [38]. We show that the classical and much used technique of regularization, is in fact a special case of the more general approach outlined in Theorem 3.1.

Example 4.1 (Support Vector Machines) Classification is a fundamental problem in machine learning. Here, a decision maker observes a set of training samples $\{\mathbf{x}_i\}_{i=1}^m$ (each sample is assumed in $\mathbb{R}^k$) and their labels $\{y_i\}_{i=1}^m$ (each label is in $\{-1,1\}$). The goal is to learn the labeling rule, so as to be able to label future points $\mathbf{x}\in\mathbb{R}^m$. The Support Vector Machine (SVM) approach searches for a linear classification rule, of the form $(\mathbf{w}^\top\mathbf{x}+b)$, to separate the space into points labeled $+1$ and -1. The linear rule given by $(\mathbf{w},b)$ is selected by attempting to solve the expected classification loss on future samples, i.e., the testing loss:

$$\min_{\mathbf{w},b} : \mathbb{E}_{y,\mathbf{x}\sim\mu^*}[1-y\big(\langle\mathbf{w},\,\mathbf{x}\rangle+b\big)\big]^+. \tag{10}$$

Since the distribution, μ^* is unknown, instead the decision-maker minimizes the loss on the observed samples, i.e., minimizes the training error,

$$\min_{\mathbf{w},b} : \quad \sum_{i=1}^m \big[1-y_i\big(\langle\mathbf{w},\,\mathbf{x}_i\rangle+b\big)\big]^+. \tag{11}$$

While we do not go into details here (but see Schölkopf and Smola [26]) this problem in fact is often very high-dimensional, possibly infinite-dimensional, because one may use a so-called kernel mapping to non-linearly map the data into a higher dimensional space, and look for a linear classifier in that space. Because of this, it has long been known that the empirical optimization as formulated in (11) will not be consistent in general. To correct for this fact, a long-standing technique in machine learning has been to add an ℓ^2-norm regularizer on $\mathbf{w}$, as a so-called complexity penalty. The resulting problem is the norm-regularized SVM (Boser et al. [9], and Vapnik and Chervonenkis [35]):

$$\min_{\mathbf{w},b} : \quad c\|\mathbf{w}\|_2 + \sum_{i=1}^m \big[1-y_i\big(\langle\mathbf{w},\,\mathbf{x}_i\rangle+b\big)\big]^+. \tag{12}$$

If the training samples $\{\mathbf{x}_i, y_i\}_{i=1}^m$ are non-separable, then we have shown in [37] that this regularized SVM is equivalent (i.e., has the same set of solutions) to the robust optimization problem

$$\min_{\mathbf{w},b} : \max_{(\hat{y}_1,\hat{\mathbf{x}}_1,\ldots,\hat{y}_m,\hat{\mathbf{x}}_m)\in\mathcal{T}_m} \sum_{i=1}^m \big[1-\hat{y}_i\big(\langle \mathbf{w},\, \hat{\mathbf{x}}_i\rangle + b\big)\big]^+,$$

where the uncertainty set is given by

$$\mathcal{T}_m \;\triangleq\; \Big\{(\hat{y}_1,\hat{\mathbf{x}}_1,\ldots,\hat{y}_1,\hat{\mathbf{x}}_m) \,:\, \hat{y}_j \equiv y_j; \sum_{i=1}^m \|\mathbf{x}_i-\hat{\mathbf{x}}_i\|_2 \le c\Big\}.$$

Then, using Theorem 3.1, one can show that regularized SVM is consistent in a much more direct fashion than previous equivalent results (e.g., Steinwart [33]), that rely on concepts such as algorithmic stability or VC-dimension.

EXAMPLE 4.2 (LASSO) The next example considers regression. In this setup we are given m vectors in $\mathbb{R}^k$ denoted by $\{\mathbf{x}_i\}_{i=1}^m$ and m associated real values $\{b_i\}_{i=1}^m$. We are looking for a k dimensional linear regressor $\mathbf{v}$ such that the expected regression error for a new testing sample is minimized, i.e., we want to solve

$$\min_{\mathbf{v}} : \mathbb{E}_{b,\mathbf{x}\sim\mu^*}(b-\mathbf{x}^\top\mathbf{v})^2. \tag{13}$$

As in the previous example, the distribution is unknown except through the given training samples that are i.i.d. realizations of μ^*. There are many ways to solve this regression problem and we consider a specific popular framework known as Lasso (Tibshirani [34]). Let $\mathbf{b}$ denote the vector form of $b_1,\ldots,b_m$ and X denote a $m\times k$ matrix such that its i^{th} row is $\mathbf{x}_i^\top$.

In [38], we show that the l_1 regularized regression problem (also known as Lasso)

$$\min_{\mathbf{v}} : \quad \|\mathbf{b}-X\mathbf{v}\|_2 + c\|\mathbf{v}\|_1,$$

is equivalent to a robust regression

$$\min_{\mathbf{v}} : \quad \max_{\Delta\in\mathcal{U}_m} \|\mathbf{b}-(X+\Delta)\mathbf{v}\|_2,$$

with the uncertainty set

$$\mathcal{U}_m \triangleq \Big\{\big[\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_k\big]\Big|\|\boldsymbol{\delta}_j\|_2\le c,\;\; j=1,\cdots,k\Big\}.$$

Thus, as in the previous example, using Theorem 3.1, we can establish that the robust formulation above, and hence the well-known Lasso procedure, is statistically consistent.

4.2 Uncertainty Set Shrinkage When parameter deviation is not adversarial in nature, the standard RO formulation

$$\max_v \min_{\mathbf{x}_\delta\in\Delta} f(v,\mathbf{x}_0+\mathbf{x}_\delta),$$

where Δ is the set of all possible deviation, often leads to conservative solutions (cf Xu and Mannor [39], and Delage and Mannor [11]). A natural remedy is to shrink the uncertainty set, i.e., fix $\alpha\in(0,1)$ and solve

$$\max_v \min_{\mathbf{x}_\delta\in\alpha\Delta} f(v,\mathbf{x}_0+\mathbf{x}_\delta).$$

This method, though intuitively appealing, easily implemented and hence widely used in application, lacks justification. Specifically, the physical meaning of the set $\alpha\Delta$ is unclear.

Based on the probabilistic interpretation of RO, we provide a justification for such an approach: indeed, the shrinkage approach *approximately* solves (see Theorem 4.1 for the precise statement) the following DRSP

$$\inf_{\mu\in\hat{\mathcal{P}}'} \int_{\mathbb{R}^m} f(v,\mathbf{x})d\mu(\mathbf{x}),$$

where $\hat{\mathcal{P}}' = \{\mu\in\mathcal{P}\,|\,\mu(\{\mathbf{x}_0\})\ge 1-\alpha,\;\mu(\mathbf{x}_0+\Delta)=1\}$. Thus, the uncertainty set shrinkage method indeed describes a system having a two-scenario setup: with probability at least $1-\alpha$ the system is in a

"normal state," where the parameters take the nominal value $\mathbf{x}_0$; otherwise the system is in an "abnormal state," where the parameter has a deviation that belongs to Δ.

Let us first consider the important special case where $f(\cdot,\cdot)$ is linear w.r.t. to the second argument. For example, linear programs with uncertain cost, or MDPs with uncertain reward fall into this setting. In this case, the shrinkage approach *exactly* solves the DRSP. We formalize this statement in the following corollary. The corollary follows immediately from Theorem 4.1, which we provide below.

COROLLARY 4.1 *If for all v, $f(v,\cdot)$ is linear, then*

$$\min_{\mathbf{x}_\delta\in\alpha\Delta} f(v,\mathbf{x}_0+\mathbf{x}_\delta) = \inf_{\mu\in\hat{\mathcal{P}}'}\int_{\mathbb{R}^m} f(v,\mathbf{x})d\mu(\mathbf{x}),$$

where, as above,

$$\hat{\mathcal{P}}' = \{\mu\in\mathcal{P}\,|\,\mu(\{\mathbf{x}_0\})\geq 1-\alpha,\ \mu(\mathbf{x}_0+\Delta)=1\}.$$

In general, when $f(\cdot,\cdot)$ is not linear w.r.t. the second argument, such an equivalence relationship does not hold exactly. For example, let $x_0=0$, $\Delta=[-1:1]$ and take $f:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ to be given by

$$f(v,x) = v\cdot\mathbf{1}(v-x\leq 1).$$

Then the optimal solution to the α-shrinkage problem $\max_v\min_{x_\delta\in\alpha\Delta} f(v,x_0+x_\delta)$ is $1-\alpha$, whereas the solution to the distributionally robust problem remains 1. Nevertheless, under some assumptions on the curvature of $f(\cdot,\cdot)$, the equivalence continues to *approximately hold* in much greater generality. This is the content of the following theorem.

THEOREM 4.1 *Suppose for any v, $f(v,\cdot)$ is twice differentiable with a uniformly bounded Hessian. That is, there exists $h\geq 0$ such that for all v, $\mathbf{x}$*

$$-hI\preceq H_v(\mathbf{x})\preceq hI,$$

where $H_v(\mathbf{x})$ is the Hessian of $f(v,\cdot)$ evaluated at $\mathbf{x}$. Then for all v

$$\inf_{\mu\in\hat{\mathcal{P}}'}\int_{\mathbb{R}^m} f(v,\mathbf{x})d\mu(\mathbf{x})-\alpha D^2h\leq\min_{\mathbf{x}_\delta\in\alpha\Delta} f(v,\mathbf{x}_0+\mathbf{x}_\delta)\leq\inf_{\mu\in\hat{\mathcal{P}}'}\int_{\mathbb{R}^m} f(v,\mathbf{x})d\mu(\mathbf{x})+\alpha D^2h,$$

where $D=\max_{\mathbf{x}\in\Delta}\|\mathbf{x}\|_2$ and $\hat{\mathcal{P}}'=\{\mu\in\mathcal{P}\,|\,\mu(\{\mathbf{x}_0\})\geq 1-\alpha,\ \mu(\mathbf{x}_0+\Delta)=1\}$.

PROOF. Denote the gradient of $f(v,\cdot)$ by $g_v(\cdot)$. Fix v and $\mathbf{x}_1\in\Delta$, and let $\mathbf{x}_1'=\alpha\mathbf{x}_1$, which by definition belongs to $\alpha\Delta$. Since $f(v,\cdot)$ is twice differentiable, then there exists $\beta\in[0,1]$ such that

$$f(v,\mathbf{x}_0+\mathbf{x}_1)=f(v,\mathbf{x}_0)+g_v((1-\beta)\mathbf{x}_0+\beta(\mathbf{x}_0+\mathbf{x}_1))\mathbf{x}_1=f(v,\mathbf{x}_0)+g_v(\mathbf{x}_0+\beta\mathbf{x}_1)\mathbf{x}_1.$$

Similarly, there exists $\beta'\in[0,1]$ such that

$$f(v,\mathbf{x}_0+\mathbf{x}_1')=f(v,\mathbf{x}_0)+g_v((1-\beta')\mathbf{x}_0+\beta'(\mathbf{x}_0+\mathbf{x}_1'))\mathbf{x}_1'=f(v,\mathbf{x}_0)+\alpha g_v(\mathbf{x}_0+\alpha\beta'\mathbf{x}_1)\mathbf{x}_1.$$

Due to the boundedness of Hessian we have

$$\|g_v(\mathbf{x}_0+\beta\mathbf{x}_1)-g_v(\mathbf{x}_0+\alpha\beta'\mathbf{x}_1)\|\leq h\|\beta\mathbf{x}_1-\alpha\beta'\mathbf{x}_1\|\leq h\|\mathbf{x}_1\|\leq hD,$$

which implies that

$$f(v,\mathbf{x}_0+\mathbf{x}_1')\leq f(v,\mathbf{x}_0)+\alpha g_v(\mathbf{x}_0+\beta\mathbf{x}_1)\mathbf{x}_1+\alpha hD\|\mathbf{x}_1\|=(1-\alpha)f(v,\mathbf{x}_0)+\alpha f(v,\mathbf{x}_0+\mathbf{x}_1)+\alpha hD^2,$$

and similarly

$$f(v,\mathbf{x}_0+\mathbf{x}_1')\geq(1-\alpha)f(v,\mathbf{x}_0)+\alpha f(v,\mathbf{x}_0+\mathbf{x}_1)-\alpha hD^2.$$

Thus,

$$(1-\alpha)f(v,\mathbf{x}_0)+\alpha f(v,\mathbf{x}_0+\mathbf{x}_1)-\alpha D^2h\leq f(v,\mathbf{x}_0+\mathbf{x}_1')\leq(1-\alpha)f(v,\mathbf{x}_0)+\alpha f(v,\mathbf{x}_0+\mathbf{x}_1)+\alpha D^2h.$$

Since this holds for all $\mathbf{x}_1\in\Delta$, and recall that $\mathbf{x}_1'=\alpha\mathbf{x}_1$, by definition of $\alpha\Delta$ we have

$$(1-\alpha)f(v,\mathbf{x}_0)+\alpha\min_{\mathbf{x}_\delta\in\Delta} f(v,\mathbf{x}_0+\mathbf{x}_\delta)-\alpha D^2h\leq\min_{\mathbf{x}_\delta'\in\alpha\Delta} f(v,\mathbf{x}_0+\mathbf{x}_\delta')\leq(1-\alpha)f(v,\mathbf{x}_0)+\alpha\min_{\mathbf{x}_\delta\in\Delta} f(v,\mathbf{x}_0+\mathbf{x}_\delta)+\alpha D^2h.$$

The theorem thus holds due to the following equality implied by Corollary 5.2,

$$(1-\alpha)f(v,\mathbf{x}_0)+\alpha\min_{\mathbf{x}_\delta\in\Delta}f(v,\mathbf{x}_0+\mathbf{x}_\delta)=\inf_{\mu\in\hat{\mathcal{P}}'}\int_{\mathbb{R}^m}f(v,\mathbf{x})d\mu(\mathbf{x}).$$

□

The following corollary considers shrinkage in the general case where $f(v,\cdot)$ is not linear and where the uncertainty set is star-shaped. Instead of additive upper and lower bounds as in Theorem 4.1 we provide a probabilistic interpretation for both the upper bound and the lower bound.

COROLLARY 4.2 *Let Δ be star shaped, i.e., if $\mathbf{x}\in\Delta$, then $\gamma\mathbf{x}\in\Delta$ for all $\gamma\in[0,1]$. If for all v, $f(v,\cdot)$ is convex, $f(v,\mathbf{x}_0)-\min_{\mathbf{x}_\delta\in\Delta}f(v,\mathbf{x}_0+\mathbf{x}_\delta)\geq 1$ (this can be achieved by normalization), and is twice differentiable with a bounded Hessian, then*

$$\inf_{\mu\in\hat{\mathcal{P}}''}\int_{\mathbb{R}^m}f(v,\mathbf{x})d\mu(\mathbf{x})\leq\min_{\mathbf{x}_\delta\in\alpha\Delta}f(v,\mathbf{x}_0+\mathbf{x}_\delta)\leq\inf_{\mu\in\hat{\mathcal{P}}'}\int_{\mathbb{R}^m}f(v,\mathbf{x})d\mu(\mathbf{x});$$

where D, h, and $\hat{\mathcal{P}}'$ are the same as in Theorem 4.1, and

$$\hat{\mathcal{P}}''=\{\mu\in\mathcal{P}\,|\,\mu(\{\mathbf{x}_0\})\geq\max(0,1-\alpha-\alpha D^2h),\ \mu(\mathbf{x}_0+\Delta)=1\}.$$

PROOF. The right hand side holds due to the following equation implied by convexity of $f(v,\cdot)$,

$$f(v,\mathbf{x}_0+\alpha\mathbf{x}_1)\leq(1-\alpha)f(v,\mathbf{x}_0)+\alpha f(v,\mathbf{x}_1);\quad\forall\mathbf{x}_1\in\Delta.$$

By Theorem 4.1 we have

$$(1-\alpha)f(v,\mathbf{x}_0)+\alpha\min_{\mathbf{x}_\delta\in\Delta}f(v,\mathbf{x}_0+\mathbf{x}_\delta)-\alpha D^2h\leq\min_{\mathbf{x}'_\delta\in\alpha\Delta}f(v,\mathbf{x}_0+\mathbf{x}'_\delta).$$

Since $f(v,\mathbf{x}_0)-\min_{\mathbf{x}_\delta\in\Delta}f(v,\mathbf{x}_0+\mathbf{x}_\delta)\geq 1$, this implies

$$(1-\alpha-\alpha D^2h)f(v,\mathbf{x}_0)+(\alpha+\alpha D^2h)\min_{\mathbf{x}_\delta\in\Delta}f(v,\mathbf{x}_0+\mathbf{x}_\delta)\leq\min_{\mathbf{x}'_\delta\in\alpha\Delta}f(v,\mathbf{x}_0+\mathbf{x}'_\delta).$$

We thus establish the left hand side by combining with the following inequality implied by Δ is star shaped,

$$\min_{\mathbf{x}_\delta\in\Delta}f(v,\mathbf{x}_0+\mathbf{x}_\delta)\leq\min_{\mathbf{x}'_\delta\in\alpha\Delta}f(v,\mathbf{x}_0+\mathbf{x}'_\delta).$$

□

Corollary 4.2 shows that for convex functions, the shrinkage method indeed solves a problem that is bounded by two DRSPs, both having a two-scenario setup: with a certain probability the system is in a "normal state," where the parameters take the nominal value $\mathbf{x}_0$; otherwise the system is in an "abnormal state," where the parameter has a deviation that belongs to Δ. The difference of the "normal state" probability of these two DRSPs depends on the curvature of the function, and diminishes to zero when the function $f(v,\cdot)$ is linear.

5. Extensions: Correlated Uncertainty and Computation In this section we show two extensions of Theorem 2.1 that address situations that arise naturally in practice. First, we show that it is possible to consider uncertainty sets (of different samples) that are coupled, and derive the set of distributions of the equivalent DRSP. This is useful in settings where the samples are not generated in an independent manner, but also when they are generated i.i.d., and one seeks to reduce the conservativeness of the robust formulation by modeling them as satisfying some joint constraints. Note that as stated, Theorem 2.1 cannot capture this, as it implicitly assumes that the realization of each parameter does not depend on that of others. In practice it is often the case that the uncertain parameters need to satisfy some joint constraints (e.g., Bertsimas and Sim [7]).

In particular, we consider the setting where each sample is subject to some perturbations, and the total perturbation is bounded. This would be a sensible constraint to place on a robust optimization formulation, if the perturbations of each sample are generated independently according to some unknown distribution, in which case one would expect the total variation of the perturbations to be small.

Note that uncertain parameters of Example 4.1 (SVM) essentially satisfy such a joint constraint. We omit the proof of Corollary 5.1 since it is a straightforward extension of Theorem 2.1.

Corollary 5.1 (Bounded Total Perturbation) *The RO formulation of the bounded total perturbation is equivalent to DRSP, i.e., the following holds*

$$\max_{\sum_i \|\mathbf{x}_i - \mathbf{x}_i^*\|_2 \le r} \sum_{i=1}^n c_i f(\mathbf{x}_i) = \inf_{\mu \in \overline{\mathcal{P}}} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x})$$

where the set of distributions are given by

$$\overline{\mathcal{P}} = \bigcup_{r_1,\cdots,r_n \ge 0, \sum_i r_i = r.} \left\{ \mu \in \mathcal{P} \,|\, \forall S \subseteq [1:n] :\ \mu\big(\bigcup_{i \in S} \mathcal{B}(\mathbf{x}_i^*, r_i)\big) \ge \sum_{i \in S} c_i \right\}.$$

Corollary 5.1 shows that Theorem 2.1 can be generalized to joint constraints on uncertain parameters. Indeed, it is straightforward to adapt Corollary 5.1 to accommodate other types of joint constraints – the main limitation simply being that the resulting robust optimization problem remains computationally tractable.

The next corollary demonstrates that the equivalence relationship of Theorem 2.1 has computational consequences. These consequences stem from the fact that, generically, it is considerably easier to solve a (deterministic) RO problem, than it is to solve a DRSP. Indeed, we show that it is possible to solve certain DRSPs that arise naturally in practice by solving their equivalent RO formulation.

To illustrate this point, consider the setting where the admissible distributions have a nested-set structure. That is, given sets $\mathcal{Z}_1 \subseteq \mathcal{Z}_2 \subseteq \cdots \subseteq \mathcal{Z}_n$, and $p_1 < p_2 < \cdots < p_n = 1$, suppose each of the admissible distributions satisfies $\mu(\mathcal{Z}_i) \ge p_i$. This nested structure is natural; it comes from the setting where a distribution is estimated from samples. A common technique in such settings is to consider modeling the uncertainty using probabilistic confidence, which naturally decreases in the number of samples, hence yielding a nested structure.

Corollary 5.2 (Nested distribution) *Let $\mathcal{Z}_1 \subseteq \mathcal{Z}_2 \subseteq \cdots \subseteq \mathcal{Z}_n$, and $0 = p_0 < p_1 < p_2 < \cdots < p_n = 1$. Consider the set of nested distributions*

$$\hat{\mathcal{P}} = \{\mu \in \mathcal{P} \,|\, \mu(\mathcal{Z}_i) \ge p_i,\ \forall i = 1, \cdots, n\}.$$

Then the DRSP is equivalent to the RO:

$$\inf_{\mu \in \hat{\mathcal{P}}} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}) = \sum_{i=1}^n (p_n - p_{n-1}) \inf_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i).$$

Proof. Let $c_i = p_i - p_{i-1}$, and define

$$\hat{\mathcal{P}}' = \{\mu \in \mathcal{P} | \forall S \subseteq [1:n] :\ \mu(\bigcup_{i \in S} \mathcal{Z}_i) \ge \sum_{i \in S} c_i\}.$$

Since $\mathcal{Z}_i \subseteq \mathcal{Z}_j$ for $i < j$, we have

$$\bigcup_{i \in S} \mathcal{Z}_i = \mathcal{Z}_{i^*(S)}, \quad \text{where } i^*(S) \triangleq \max_{i \in S} i,$$

and

$$\sum_{i \in S} c_i \le \sum_{i \le i^*(S)} c_i = p_{i^*(S)}.$$

Thus, $\hat{\mathcal{P}} = \hat{\mathcal{P}}'$, and note that by Theorem 2.1 we have

$$\inf_{\mu \in \hat{\mathcal{P}}'} \int_{\mathbb{R}^m} f(\mathbf{x}) d\mu(\mathbf{x}) = \sum_{i=1}^n (p_n - p_{n-1}) \inf_{\mathbf{x}_i \in \mathcal{Z}_i} f(\mathbf{x}_i),$$

which implies the corollary. □

6. Conclusion We show that robust optimization problems can be re-formulated as distributionally robust stochastic problems. That is, robust optimization is equivalent to maximizing the worst-case expected value over a class of distributions. While such an equivalence is well known in the special case where each uncertain parameter belongs to a different space, we generalize it to the case where *multiple* parameters belong to the same fixed dimensional space. This setting arises naturally in stochastic problems that are attacked via sampling (as in machine learning or stochastic programming). Using this reformulation, we show how to construct robust optimization problems that are statistically consistent, even when the original empirical optimization is not. Our approach further provides a probabilistic interpretation to the common practice of shrinking the uncertainty set in robust optimization to avoid over conservativeness.

Acknowledgements We thank Aharon Ben-Tal for useful discussions and for pointing us to the shrinkage heuristic. The research of C.C. was partially supported by NSF grants EFRI-0735905, CNS-0721532, CNS-0831580, and DTRA grant HDTRA1-08-0029. The research of S.M. was partially supported by the Israel Science Foundation (contract 890015).

References

[1] M. Anthony and P. L. Bartlett, *Neural network learning: Theoretical foundations*, Cambridge University Press, 1999.

[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*, Princeton University Press, 2009.

[3] A. Ben-Tal and A. Nemirovski, *Robust solutions of uncertain linear programs*, Operations Research Letters **25** (1999), no. 1, 1–13.

[4] _____, *Robust solutions of linear programming problems contaminated with uncertain data*, Mathematical Programming, Serial A **88** (2000), 411–424.

[5] A. Ben-Tal, B. Golany, A. Nemirovski, and J. P. Vial, *Supplier-retailer flexible commitments contracts: A robust optimization approach-manufacturing and service*, Manufacturing and Service Operations Management **7** (2005), no. 3, 248–271.

[6] A. Ben-Tal, T. Margalit, and A. Nemirovski, *Robust modeling of multi-stage portfolio problems*, High Performance Optimization (H. Frenk, C. Roos, T. Terlaky, and S. Zhang, eds.), 2000, pp. 303–328.

[7] D. Bertsimas and M. Sim, *The price of robustness*, Operations Research **52** (2004), no. 1, 35–53.

[8] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*, Springer-Verlag, New York, 1997.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik, *A training algorithm for optimal margin classifiers*, Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (New York, NY), 1992, pp. 144–152.

[10] S. P. Boyd, S. J. Kim, D. D. Patil, and M. A. Horowitz, *Digital circuit optimization via geometric programming*, Operations Research **53** (2005), no. 6, 899–932.

[11] E. Delage and S. Mannor, *Percentile optimization for Markov decision processes with parameter uncertainty*, Operations Research (2010), no. 1, 203–213.

[12] E. Delage and Y. Ye, *Distributional robust optimization under moment uncertainty with applications to data-driven problems*, To appear in *Operations Research*, 2010.

[13] L. Devroye and L. Györfi, *Nonparametric density estimation: the l_1 view*, John Wiley & Sons, 1985.

[14] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, Springer, New York, 1996.

[15] J. Dupacová, *The minimax approach to stochastic programming and an illustrative application*, Stochastics **20** (1987), 73–88.

[16] L. El Ghaoui and H. Lebret, *Robust solutions to least-squares problems with uncertain data*, SIAM Journal on Matrix Analysis and Applications **18** (1997), 1035–1064.

[17] P. Kall, *Stochastic programming with recourse: Upper bounds and moment problems, a review*, Advances in Mathematical Optimization, Academie-Verlag, Berlin, 1988.

[18] A.J. King and R.J.B. Wets, *Epi-consistency of convex stochastic programs*, Stochastics and Stochastic Reports **34** (1991), no. 1, 1045–1129.

[19] J. Goh and M. Sim, *Distributionally Robust Optimization and its tractable approximations*, Operations Research (in press), 2010.

[20] G. R. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan, *A robust minimax approach to classification*, Journal of Machine Learning Research **3** (2003), 555–582.

[21] E. Parzen, *On the estimation of a probability density function and the mode*, The Annals of Mathematical Statistics **33** (1962), 1065–1076.

[22] I. Popescu, *Robust mean-covariance solutions for stochastic optimization*, Operations Research **55** (2007), no. 1, 98–112.

[23] A. Prékopa, *Stochastic programming*, Kluwer, 1995.

[24] M. Rosenblatt, *Remarks on some nonparametric estimates of a density function*, The Annals of Mathematical Statistics **27** (1956), 832–837.

[25] H. Scarf, *A min-max solution of an inventory problem*, Studies in Mathematical Theory of Inventory and Production, Stanford University Press, 1958, pp. 201–209.

[26] B. Schölkopf and A. J. Smola, *Learning with kernels*, MIT Press, 2002.

[27] D. W. Scott, *Multivariate density estimation: Theory, practice and visualization*, John Wiley & Sons, New York, 1992.

[28] A. Shapiro, *Asymptotic analysis of stochastic programs*, Annals of Operations Research **30** (1991), no. 1-4, 169–186.

[29] _____, *Worst-case distribution analysis of stochastic programs*, Mathematical Programming **107** (2006), no. 1, 91–96.

[30] A. Shapiro and H. de Mello, *On rate of convergence of monte carlo approximations of stochastic programs*, SIAM J. Optimization **11** (2000), 70–86.

[31] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, *Second order cone programming approaches for handling missing and uncertain data*, Journal of Machine Learning Research **7** (2006), 1283–1314.

[32] A. L. Soyster, *Convex programming with set-inclusive constraints and applications to inexact linear programming*, Operations Research **21** (1973), 1154–1157.

[33] I. Steinwart, *Consistency of support vector machines and other regularized kernel classifiers*, IEEE Transactions on Information Theory **51** (2005), no. 1, 128–142.

[34] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society, Series B **58** (1996), no. 1, 267–288.

[35] V. N. Vapnik and A. Chervonenkis, *Theory of pattern recognition*, Nauka, Moscow, 1974.

[36] _____, *The necessary and sufficient conditions for consistency in the empirical risk minimization method*, Pattern Recognition and Image Analysis **1** (1991), no. 3, 260–284.

[37] H. Xu, C. Caramanis, and S. Mannor, *Robustness and regularization of support vector machines*, Journal of Machine Learning Research **10** (2009), no. Jul, 1485–1510.

[38] _____, *Robust regression and Lasso*, *To appear in* IEEE Transactions on Information Theory, 2010.

[39] H. Xu and S. Mannor, *The robustness-performance tradeoff in Markov decision processes*, Advances in Neural Information Processing Systems 19 (B. Schölkopf, J. C. Platt, and T. Hofmann, eds.), MIT Press, 2007, pp. 1537–1544.

[40] A. Ben-Tal and A. Goryashko and E. Guslitzer and A. Nemirovski, *Adjustable robust solutions of uncertain linear programs*, Math. Programming, 2003, pp. 351–376.

Kernel Density Estimation The *kernel density estimator* for a density $\hat{h}$ in $\mathbb{R}^d$, originally proposed in Rosenblatt [24] and Parzen [21], is defined by

$$h_n(\mathbf{x}) = (nc_n^d)^{-1} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \hat{\mathbf{x}}_i}{c_n}\right),$$

where $\{c_n\}$ is a sequence of positive numbers, $\hat{\mathbf{x}}_i$ are i.i.d. samples generated according to $\hat{h}$, and K is a Borel measurable function (kernel) satisfying $K \geq 0$, $\int K = 1$. See Devroye and Györfi [13], Scott [27],

and the reference therein for detailed discussions. Figure 1 illustrates a kernel density estimator using Gaussian kernel for a randomly generated sample-set. A celebrated property of a kernel density estimator is that it converges in ℓ^1 to $\hat{h}$, i.e., $\int_{\mathbb{R}^d} |h_n(\mathbf{x}) - \hat{h}(\mathbf{x})| d\mathbf{x} \to 0$, when $c_n \downarrow 0$ and $nc_n^d \uparrow \infty$ (Devroye and Györfi [13]).

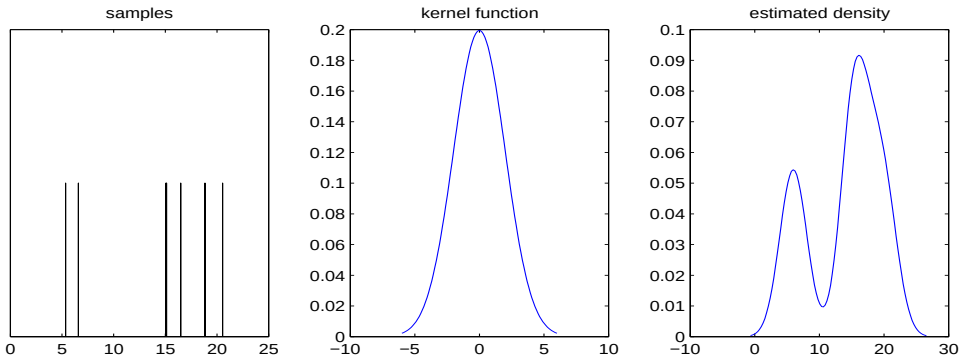

Figure 1: Illustration of Kernel Density Estimation.