Robust and Stochastic Optimization With a Hybrid Coherent Risk Measure With an Application to Supervised Learning

Shutian Liu, *Graduate Student Member, IEEE*, and Quanyan Zhu, *Member, IEEE*

***Abstract*—This letter considers a hybrid risk measure for decision-making under uncertainties that trade-offs between the solutions obtained from the robust optimization and the stochastic optimization techniques. In the proposed framework, the risk measure is shown to satisfy the properties of coherent risk measures. We can control the level of guaranteed robustness using a parameter. We formulate the stochastic and robust optimization problem under the proposed risk measure and show its equivalent formulation and sensitivity result. We introduce the sample approximation of our technique by combining scenario program and sample average approximation, making our framework amenable for practical usage. We present a supervised learning problem as a case study to corroborate our results and show the implications of the proposed method in machine learning.**

***Index Terms*—Optimization, stochastic systems, uncertain systems, machine learning.**

I. Introduction

DECISION-MAKING under uncertainties has been an active research area for decades [1]. It deals with optimal decision-making when systems are exposed to uncertainties. One important step to this problem is to quantify randomness and incorporate it into the decision-making. To this end, many methods have been proposed.

One major field of research is robust optimization (RO), which deals with problems where the decision-maker (DM) takes actions under the worst-case uncertainties. In this formulation, the DM considers the events in the uncertainty set as equally probable and cannot tolerate any uncertainty that lies outside the set. Solutions to RO problems have been extensively studied, such as in [2], [3], and [4]. One key feature of RO is that it protects the system away from the worst risk. RO may overcome the pessimism if a proper uncertainty set is chosen. Another feature of RO is that the optimizer yields a safe or secure decision no matter how unlikely the risk is as long as the corresponding events are included in the uncertainty set. However, considering all feasible events would lead to significant conservatism in solutions and unimplementable decisions.

Stochastic programming (SP) is another decision-making paradigm that captures uncertainties using random variables [1], [5]. In this formulation, not all events are treated equally. Decisions are often made by minimizing the expected loss that considers the average outcome of the random parameter. One advantage of SP is that we can obtain the solutions using a data-driven approach where samples of the uncertain scenarios are used to approximate the expected loss. However, decisions made using SP do not provide a strict guarantee of performance when an event is realized. The performance is assured through the average sense.

SP and RO approaches yield different decisions that are suitable for distinct applications or criteria. One research direction that boasts both features is distributionally robust optimization (DRO) which optimizes the expected loss under a set of possible distributions. DRO makes assumptions on the set of uncertain distributions and guarantees distributional robustness of the solution by finding the worst-case distributions. For its recent advances, readers can refer to [6].

In this letter, we introduce a new approach that couples RO and SP. We consider an unconstrained uncertainty set and partition it into two subsets. In one subset, we aim to consider the worst-case risks to provide a strict guarantee for all the uncertainties in the subset, while in the other subset, we aim to consider the average risk and content with the mean performance over the subset. The partition of the subsets is determined by a parameter that controls the degree of robustness. This approach enables a mixture between two approaches and tradeoffs between the security guarantee of the solution and the actionability of the decision. The method is especially important when handling unbounded uncertainties and unstructured objective functions where the worst-case solution can result in either trivial or meaningless scenarios.

Manuscript received March 17, 2020; revised June 2, 2020; accepted June 22, 2020. Date of publication July 7, 2020; date of current version July 21, 2020. This work was supported in part by the U.S. Department of Homeland Security under Award 2015-ST-061-CIRC01, in part by the National Science of Foundation under Award ECCS-1847056, Award CNS-1544782, Award CNS-2027884, and Award SES-1541164, and in part by Army Research Office under Grant W911NF-19-1-0041. Recommended by Senior Editor G. Cherubini. *(Corresponding author: Shutian Liu.)*

The authors are with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: sl6803@nyu.edu; qz494@nyu.edu).

Digital Object Identifier 10.1109/LCSYS.2020.3007645

We show that the problem formulation is equivalent to a SP problem with a hybrid coherent risk measure (CRM) [7]. The proposed measure can be viewed as a mixture of the potential risk and the inner loss. The proposed approach can be interpreted as a DRO problem with the uncertainty set consisting of density functions. We further investigate the sensitivity of the optimal value with respect to the set partition parameter, which measures the marginal effect of the robustness on the decision-making. We also prove the differentiability of our CRM to ensure its applicability in real-world applications. Furthermore, we explore data-driven methods to enable sample-based computation and use a supervised learning problem as a case study.

This letter is organized as follows. Section II presents the decision-making framework under the hybrid CRM. Section III shows the coherency of the risk measure, the structural and differentiability results and the sensitivity analysis. Section IV provides a sample-based method and a case study is investigated in Section V. Section VI concludes this letter.

II. Problem Formulations

This section formulates a class of decision-making problems under uncertainties. Let $x \in \mathbb{R}^n$ denote the decision variable; Y is a random variable with image space $S \subseteq \mathbb{R}$; and $(\Omega, \mathcal{F})$ is the underlying measurable space with probability measure P. Denote by $f : \mathbb{R}^n \times S \to \mathbb{R}$ a real-valued loss function. The loss incurred under the decision x and the uncertainties captured by the random variable Y is given by

$$Z = f(x, Y). \tag{1}$$

It is apparent that, for a given x, Z is a random variable defined over the same probability space. Let Z be essentially bounded. Denote the density function of Y by $f_Y(y)$, and assume it to be continuous and have a single peak. Since Z is also a random variable, we need to select a risk measure to evaluate the risk associated with Z.

We partition S into S^+ and S^- such that $S^+ \cup S^- = S$ and $S^+ \cap S^- = \emptyset$. We assume that S^+ is a convex closed set and contains uncertainties on which we aim to impose robustness guarantees while S^- contains uncertainties on which the average outcome of the associated losses is defined.

Define the hybrid risk measure as follows:

$$R_{\bar{\beta}}(f(x, Y)) = \bar{\beta} \cdot \operatorname*{ess\,sup}_{\bar{y} \in S^+} f(x, \bar{y}) + \int_{S^-} f(x, y) f_Y(y) dy, \tag{2}$$

where $\bar{\beta} = \int_{S^+} f_Y(y) dy \in [0, 1]$ is a level parameter that measures the size of S^+. Risk measure (2) uses different criteria to measure the risk associated with two mutually exclusive partitions. Note that in (2), we use $\bar{y}$ to denote the uncertainty parameter on S^+, to distinguish it from its counterpart y on S^-.

The optimization problem that takes into account the risk measure is given as:

$$\min_{x \in X} R_{\bar{\beta}}(f(x, Y)). \tag{3}$$

Problem (3) combines RO on S^+ and SP on S^-. On one hand, as $\bar{\beta}$ gets close to 1, S^+ tends to S, and hence (3) tends to a pure RO; on the other hand, when the level parameter goes to 0, (3) tends to a pure SP.

The choices of S^+ can be based on experience, such as advice from a sophisticated banker in the case of portfolio management. We introduce one special partition based on how frequently outcomes appear. Consider the level sets of f_Y. For a given real number $\beta \in [0, \sup f_Y] \subset \mathbb{R}$, define the super-level set of f_Y:

$$S_\beta^+ = \{y : f_Y(y) \geq \beta\}, \tag{4}$$

which is convex and closed under our assumptions, and

$$S_\beta^- = \{y : f_Y(y) < \beta\} \tag{5}$$

is the sub-level set of f_Y. There are practical reasons to adopt RO on (4) and SP on (5). Firstly, the events that correspond to (4) have high probabilities. In cases where these events correspond to disturbances of the system, it is clear that robustness to these disturbances would yield desirable system performances. In a portfolio management problem, the high probability disturbances can be considered as common fluctuations of the stock prices. A desirable portfolio would secure payoffs under these common fluctuations. Secondly, for the events associated with (5), they may lead to high losses despite low probabilities of their occurances. Decisions based on the worst-case criterion would be either trivial or meaningless. Instead, average loss is a reasonable choice to quantify the risk on (5), since it balances the low likelihood and the high cost of these events. In the portfolio management problem, an example of an event on (5) could be a price shock, which has significant impact on stock prices but rarely happens. A good portfolio design takes into account the low-probability shocks but in a way that weighs its high-impact consequences with the low probability. Thirdly, by combining the worst-case scenario and the average performance, our metric enables a decision that allows the portfolio to achieve a best-effort performance under common disturbances while surviving from the consequences of low probability but high-impact events.

With the partition of (4) and (5), we obtain

$$R_{\bar{\beta}}(f(x, Y)) = \bar{\beta} \cdot \operatorname*{ess\,sup}_{\bar{y} \in S_\beta^+} f(x, \bar{y}) + \int_{S_\beta^-} f(x, y) f_Y(y) dy. \tag{6}$$

The uncertainties captured by Y can be thought of being controlled by an adversary. The set in (4) captures the fact that the DM aims to prepare for the worst-case scenario, while (5) takes into account possible losses from the adversarial events. This problem formulation is applicable to decision problems in cybersecurity, adversarial machine learning, and risk management.

III. Structural Properties and Analysis

A. Coherent Risk Measures

CRMs are a set of risk measures that satisfy four axiomatic properties originally proposed for mathematical finance [7]. The axioms are natural for a certain class of problems.

Definition 1 (CRM, [7]): A risk function $\rho(Z)$ that maps Z to the extended real line $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ is a CRM, if $\rho(\cdot)$ satisfies (a) translation invariance, (b) subadditivity, (c) positive homogeneity, and (d) monotonicity; namely,

(a) $\rho(Z+a)=\rho(Z)+a, \forall a\in\mathbb{R}, \forall Z$.

(b) $\rho(Z_1+Z_2)\le\rho(Z_1)+\rho(Z_2), \forall Z_1, Z_2$.

(c) $\rho(aZ)=a\rho(Z), \forall a\in\mathbb{R}, \forall Z$.

(d) $\rho(Z_1)\le\rho(Z_2), \forall Z_1, Z_2$, such that $Z_1\le Z_2$.

One popular CRM is Conditional Value at Risk (CVaR), a widely used substitute to Value at Risk (VaR) that does not satisfy the subadditivity axiom [7], [8], [9].

Theorem 1: The hybrid risk measure (2) is a CRM.

Proof: For any given x, we compress the notation of $f(x,Y)$ to $f(Y)$ and check the four axioms of CRMs.

(a) Translation invariance.

$$\begin{aligned}&R_{\bar\beta}(f(Y)+a)\\&=\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}(f(\bar y)+a)+\int_{S^-}(f(y)+a)f_Y(y)dy\\&=\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}f(\bar y)+\int_{S^-}f(y)f_Y(y)dy+a\\&=R_{\bar\beta}(f(Y))+a.\end{aligned}$$

(b) Subadditivity.

$$\begin{aligned}&R_{\bar\beta}(f_1(Y)+f_2(Y))\\&=\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}(f_1(\bar y)+f_2(\bar y))+\int_{S^-}(f_1(y)+f_2(y))f_Y(y)dy\\&\le\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}f_1(\bar y)+\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}f_2(\bar y)\\&\quad+\int_{S^-}f_1(y)f_Y(y)dy+\int_{S^-}f_2(y)f_Y(y)dy\\&\le R_{\bar\beta}(f_1(Y))+R_{\bar\beta}(f_2(Y)).\end{aligned}$$

(c) Positive homogeneity.

$$\begin{aligned}R_{\bar\beta}(af(Y))&=\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}(af(\bar y))+\int_{S^-}(af(y))f_Y(y)dy\\&=aR_{\bar\beta}(f(Y)).\end{aligned}$$

(d) Monotonicity. For $f_1(Y)\le f_2(Y)$ almost surely:

$$\begin{aligned}R_{\bar\beta}(f_1(Y))&=\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}f_1(\bar y)+\int_{S^-}f_1(y)f_Y(y)dy\\&\le\bar\beta\,\operatorname*{ess\,sup}_{\bar y\in S^+}f_2(\bar y)+\int_{S^-}f_2(y)f_Y(y)dy\le R_{\bar\beta}(f_2(Y)).\end{aligned}$$

Hence, we arrive at the result. ■

The CRM in (2) differs from CVaR in nature. CVaR_α is a quantile-based measure that averages the upper $1-\alpha$ portion of the loss. In contrast, (2) is not quantile-based. Both S^+ and S^- contribute to the CRM. Furthermore, in (2), we multiply the worst-case loss by $\bar\beta$. It is distinct from the formulation of CVaR. However, our measure coincides with CVaR when both of them reduce to either expectation or essential supremum.

In financial applications, there are inefficiencies associated with CVaR, despite its popularity as a risk measure. For example, [10] presents a case where CVaR fails to characterize the severity of the tail risk. This failure is caused by the fact that conditional expectation cannot handle concerns on the extreme loss. With the RO part of the hybrid measure (2), it can capture the extreme loss and the severity of losses.

B. Equivalent Formulation

In this subsection, we assume that the loss function is nonnegative and discuss the equivalent formulation of (3). First, we introduce the dual problem associated with our formulation [11]. Consider the linear space $\mathcal{Z}:=\mathcal{L}_p(\Omega,\mathcal{F},P)$ of $\mathcal{F}$-measurable functions $\psi:\Omega\to\mathbb{R}$ with $\int_\Omega||\psi(\omega)||^p dP(\omega)<+\infty$, and the norm defined by $||\psi||_p=(\int_\Omega|\psi(\omega)|^p dP(\omega))^{1/p}$. Denote the dual space of $\mathcal{Z}$ by $\mathcal{Z}^*=\mathcal{L}_q(\Omega,\mathcal{F},P)$, where $\frac{1}{p}+\frac{1}{q}=1$, and the corresponding scalar product by:

$$\langle\zeta,Z\rangle=\int_\Omega\zeta(\omega)Z(\omega)dP(\omega),$$

where $Z\in\mathcal{L}_p(\Omega,\mathcal{F},P)$ and $\zeta\in\mathcal{L}_q(\Omega,\mathcal{F},P)$.

Theorem 2: The objective in (2) can be represented using the following dual form:

$$R_{\bar\beta}(Z)=\sup_{\tilde\xi\in\tilde{\mathcal{A}}}\langle\tilde\xi,Z\rangle,\tag{7}$$

where set $\tilde{\mathcal{A}}:=\{\tilde\xi:\tilde\xi=\xi\cdot\mathbb{1}_{\Omega^+}(\omega)+\mathbb{1}_{\Omega^-}(\omega), ||\xi\cdot\mathbb{1}_{\Omega^+}(\omega)||_1=\bar\beta, \xi(w)\ge0\}$, with $\mathbb{1}_\Theta(\cdot)$ denoting the indicator function defined on the subset Θ. Denoting the Dirac delta function by $\delta(\cdot)$, we arrive at the optimal solution:

$$\tilde\xi^*(\omega)=\bar\beta\delta(\omega-\omega^*)+\mathbb{1}_{\Omega^-}(\omega),\tag{8}$$

where ω^* denotes $\arg\operatorname{ess\,sup}_{\omega\in\Omega^+}Z(\omega)$.

Proof: With Z nonnegative, we have the equality

$$||Z||_p=\sup_{||\xi||_q\le1,\xi\ge0}\langle\xi,Z\rangle.$$

Let $p=+\infty$, and the first term in (2) can be written as:

$$\begin{aligned}\bar\beta\cdot\operatorname*{ess\,sup}_{\Omega^+}Z&=\bar\beta\cdot\sup_{||\xi||_1\le1,\xi\ge0}\langle\xi,Z\cdot\mathbb{1}_{\Omega^+}(\omega)\rangle\\&=\bar\beta\cdot\sup_{||\xi||_1\le1,\xi\ge0}\langle\xi\cdot\mathbb{1}_{\Omega^+}(\omega),Z\rangle\\&=\bar\beta\cdot\sup_{\bar\xi\in\bar{\mathcal{A}}}\langle\bar\xi,Z\rangle,\end{aligned}$$

where

$$\bar{\mathcal{A}}=\{\bar\xi:\bar\xi=\xi\cdot\mathbb{1}_{\Omega^+}(\omega), ||\xi||_1\le1, \xi\ge0\}.$$

Then, equivalently,

$$\bar\beta\cdot\operatorname*{ess\,sup}_{\Omega^+}Z=\sup_{\hat\xi\in\hat{\mathcal{A}}}\langle\hat\xi,Z\rangle,$$

where

$$\hat{\mathcal{A}}=\{\hat\xi:\hat\xi=\xi\cdot\mathbb{1}_{\Omega^+}(\omega), ||\xi||_1\le\bar\beta, \xi\ge0\}.$$

Adding the other term in (2), we arrive at

$$\begin{aligned}R_{\bar\beta}(Z)&=\sup_{\hat\xi\in\hat{\mathcal{A}}}\langle\hat\xi,Z\rangle+\mathbb{E}[Z\cdot\mathbb{1}_{\Omega^-}(\omega)]\\&=\sup_{\hat\xi\in\hat{\mathcal{A}}}\langle\hat\xi,Z\rangle+\langle\mathbb{1}_{\Omega^-}(\omega),Z\rangle\\&=\sup_{\tilde\xi\in\tilde{\mathcal{A}}}\langle\tilde\xi,Z\rangle,\end{aligned}\tag{9}$$

where the third equality comes from the fact that adding the expectation term restricted on Ω^- does not affect our choice of the decision variable $\hat\xi$. Finally, since the segment on Ω^- does not contribute when both terms we maximize are positive, hence choosing a function ξ with $||\xi||_1 \leq \bar{\beta}$ and using the segment restricted on Ω^+ to maximize the corresponding scalar product with a positive function Z yields the same result as in the problem where we require $||\xi \cdot \mathbb{1}_{\Omega^+}(\omega)||_1 = \bar{\beta}$. Therefore, we arrive at the third equality.

Furthermore, by inspecting the optimization problem stated in equation (9), we note that on Ω^+, the strategy for obtaining the maximum is to pick a Dirac delta function $\delta(\omega - \omega^*)$. Hence, we arrive at (8). ■

The dual form (7) allows us to interpret (3) as finding the worst-case distribution from a closed convex set. This interpretation reveals that (3) is equivalent to a class of DRO problems [6], which can be further viewed as a game between an adversary who controls the uncertainties and a decision-maker who aims to minimize the loss.

The measure (2) and its equivalent formulation can be applied and extended to control and dynamic decision problems. We can consider multistage scenarios of the risk optimization problem as introduced in [11]. With sigma algebras characterizing information available at different stages, we can use (2) to quantify future risks given the accumulated knowledge. The multistage problem under this metric will yield solutions that make the system sufficiently robust while maintaining a good performance.

C. Sensitivity

From (2), it is clear that the level parameter $\bar{\beta}$, or the threshold on the density β, is a design parameter that indicates the robustness level of the decisions. It is important to understand the role of this design parameter in (3). In this section, we present results that characterize the sensitivity of the optimal value with respect to the design parameter. First, we state the following lemma.

Lemma 1 [12]: Let $x \in \mathbb{R}^n$, $d, p \in \mathbb{R}$, $d \neq 0$, and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a measurable function, such that $\int_0^\infty t^{n+p-1}\phi(t^d)dt < +\infty$. Let g, h be nonnegative positive homogeneous functions of degree d, p, respectively. Then,

$$\int_{\mathbb{R}^n} \phi(g(x))h(x)dx = \frac{d\int_0^{+\infty} t^{n+p-1}\phi(t^d)dt}{\Gamma((n+p)/d)} \int_{\mathbb{R}^n} h \exp(-g)dx,$$

if the integrals are all finite.

This lemma reformulates an integral over a level set as an integral on $\mathbb{R}^n$ under the appropriate assumptions.

In (2), we assume that f is a non-negative bounded function, and $f; f_Y$ are positive homogeneous with degree p and d, respectively. An example of a positive homogeneous density function is Pareto distribution with $f_Y(y) = \frac{\alpha y_m^\alpha}{y^{\alpha+1}}$, $y \in (y_m, +\infty)$, and $\alpha > 1$. Assume that the expectations exist under the probability density function f_Y. Further, we assume that the density function has a single peak which results in nested convex super-level sets. Consider the reformulation of (6) with partition of S as in (4) and (5):

$$R_{\bar{\beta}}(f(x,Y)) = \left[\int_S \mathbb{1}_{[\beta, f_Y^*]}(f_Y) \cdot f_Y(y)dy\right] \cdot P^*(\beta) + \int_S \mathbb{1}_{[0,\beta)}(f_Y) \cdot f(x,y) f_Y(y)dy, \tag{10}$$

where f_Y^* denotes $\max_y f_Y(y)$; (4) and (5) are reformulated as $\mathbb{1}_{[\beta, f_Y^*]}(f_Y)$ and $\mathbb{1}_{[0,\beta)}(f_Y)$ respectively; and $P^*(\beta)$ is the optimal solution of the following optimization problem:

$$\begin{aligned} \min_{\bar{y}} \quad & -f(x, \bar{y}) \\ \text{s.t.} \quad & \beta - f_Y(\bar{y}) \leq 0. \end{aligned} \tag{11}$$

The $P^*(\beta)$ term results in an unperturbed optimization problem. To apply Lemma 1 to the two terms on the right hand side of (10), we pick functions $\phi = \mathbb{1}_{[\beta, f_Y^*]}$, $g = f_Y$, and $h = f_Y$ for the first term and $\phi = \mathbb{1}_{[0,\beta)}$, $g = f_Y$, and $h = f(x,y)f_Y$ for the second term. Then, we can transform (10) to the following:

$$\begin{aligned} R_{\bar{\beta}}(f(x,Y)) &= \left[\frac{d\int_S t^d \mathbb{1}_{[\beta, f_Y^*]}(t^d)dt}{\Gamma((1+d)/d)} \int_S f_Y \exp(-f_Y)dy\right] \cdot P^*(\beta) \\ &\quad + \frac{d\int_S t^{p+d}\mathbb{1}_{[0,\beta)}(t^d)dt}{\Gamma((1+p+d)/d)} \int_S f(x,y) f_Y \exp(-f_Y)dy \quad (12a) \\ &= \left[\frac{f_Y^{*\frac{1+d}{d}} - \beta^{\frac{1+d}{d}}}{\Gamma(1+\frac{1+d}{d})} \int_S f_Y \exp(-f_Y)dy\right] \cdot P^*(\beta) \\ &\quad + \frac{\beta^{\frac{1+p+d}{d}}}{\Gamma(1+\frac{1+p+d}{d})} \int_S f(x,y) f_Y \exp(-f_Y)dy, \quad (12b) \end{aligned}$$

where (12b) directly follows from the properties of the integration and the gamma function. Denote by $\gamma_1 = \Gamma(1 + \frac{1+d}{d})$, $\gamma_2 = \Gamma(1 + \frac{1+p+d}{d})$, $C_1 = \int_S f_Y \exp(-f_Y)dy$, $C_2 = \int_S f(x,y) f_Y \exp(-f_Y)dy$, and $\theta = f_Y^{*\frac{1+d}{d}}$. Our result is summarized as follows.

Theorem 3: Under the assumptions that f is a non-negative bounded function; f, f_Y are positive homogeneous with degree p and d, respectively; and f_Y has a single peak, the sensitivity can be obtained from the partial derivative:

$$\frac{\partial R_{\bar{\beta}}(f(x,Y))}{\partial \beta} = -\frac{C_1}{\gamma_1}\frac{1+d}{d}\beta^{\frac{1}{d}}P^*(\beta) + \frac{C_1}{\gamma_1}(\theta - \beta^{\frac{1+d}{d}})\frac{\partial P^*(\beta)}{\partial \beta} + \frac{C_2}{\gamma_2}\frac{1+p+d}{d}\beta^{\frac{1+p}{d}}. \tag{13}$$

Proof: By directly taking the partial derivative of (12b) with respect to β, we obtain (13). ■

Note that the term $\frac{\partial P^*(\beta)}{\partial \beta}$ indicates the sensitivity of the optimization problem (11) when we perturb β. $\frac{\partial P^*(\beta)}{\partial \beta}$ characterizes the difference of the optimal value of the optimization problem per unit change in the value of the constraint of (11). It is also referred to as the shadow price in the literature of economics. The result of (13) characterizes how the parameters influence the optimal value.

D. Differentiability of the Risk Measure

Many numerical algorithms rely on the gradient of the objective with respect to the decision variable. To characterize the gradient of our risk measure, we rewrite (2) as follows:

$$R_{\bar{\beta}}(f(x,Y)) = \bar{\beta} \cdot \operatorname*{ess\,sup}_{\bar{y} \in S^+} f(x, \bar{y}) + \mathbb{E}[f(x,Y) \cdot \mathbb{1}_{S^-}(y)]. \tag{14}$$

The following result states that with mild assumptions, the objective function in Problem (3) is differentiable.

Theorem 4: Let $f(\cdot, Y)$ be a differentiable and convex function. Assume that S^+ is a compact set and there exist a unique solution of the problem $\operatorname{ess\,sup}_{\bar{y}\in S^+} f(x,\bar{y})$, denoted by $\bar{y}^*$. And further assume that $f(x,y)$ is integrable for all x, and differentiable for all y, and $\mathbb{E}[f(x,Y)] < +\infty$. Then, (2) is differentiable with respect to x, and the gradient is given by:

$$\frac{\partial}{\partial x} R_{\bar{\beta}}(f(x,Y)) = \bar{\beta}\cdot\frac{\partial}{\partial x} f(x,\bar{y}^*) + \mathbb{E}[\frac{\partial}{\partial x} f(x,Y)\cdot \mathbb{1}_{S^-}(y)].$$

Proof: The result follows directly from Danskin's theorem in [13] and [14, Sec. 3] by chosing the bounding random variable corresponding to $f(x,Y)$ since multiplying with the indicator function only suppresses the absolute value. ■

IV. Sample-Based Method

Computation is key to bridge theory and applications for mathematical programming problems. In this section, we provide a data-driven method to solve (3). Assume that there are $N = q + r$ partitioned i.i.d. samples of Y, and denote q of N samples from $S^+_{\bar{\beta}}$ by $\{\bar{y}_1, \ldots, \bar{y}_q\}$, and r of N samples from $S^-_{\bar{\beta}}$ by $\{\underline{y}_1, \ldots, \underline{y}_r\}$. We leverage techniques from scenario program (ScP) [15], [16] and sample average approximation (SAA) [1] to develop a sample-based algorithm.

In the literature of ScP [15], a given x is feasible if the constraint $x \in X_\delta$ is satisfied for all uncertainties $\delta \in \Delta$. The set Δ is of infinite cardinality, and the sets X_δ are assumed to be closed and convex. Instead of solving for infinitely many realizations, we directly create scenarios of the constraint with samples $\delta^{(i)}, i = 1, \ldots, K$ of δ, and obtain:

$$\begin{aligned} &\min_{x\in X} \ c^T x \\ &s.t. \ \ x \in \cap \ X_{\delta^{(i)}}, i = 1, \ldots, K. \end{aligned}$$

The violation probability $V(x)$ of a given x is defined as $\mathbb{P}\{\delta \in \Delta : x \notin X_\delta\}$. Define a solution x^* to a ScP to be ϵ-robust if $V(x^*) \le \epsilon$. Besides, $\{x \in X : f(x) \le f^* + \epsilon\}$ is defined as the ϵ-optimal solution set to the problem $\min_{x\in X} f(x)$ with f^* denoting the optimal value.

Denote by d the number of support constraints as defined in [15], and by λ, D and L the parameters as defined in assumptions (M5) and (M6) in [1, Corollary 5.19].

Corollary 1: With N samples of Y, (3) can be approximated as:

$$\begin{aligned} &\min_{x\in X, M\in\mathbb{R}} \ M \\ &s.t. \ \ \bar{\beta}\cdot f(x,\bar{y}_j) + \frac{1}{N}\sum_{i=1}^{r} f(x,\underline{y}_i) \le M, \quad j = 1, 2, \ldots, q. \end{aligned}$$

Suppose that there is a unique solution x^*_s to the problem. The probability of x^*_s being in the ϵ_1-optimal solution set to (3) is at least $1-\alpha$, and x^*_s is an ϵ_2-robust solution with probability at least $1-\theta$, if the sample size N satisfies:

$$N \ge \max\left\{\frac{q}{\bar{\beta}}, \frac{r}{1-\bar{\beta}}\right\}, \tag{15}$$

where $q = \frac{e}{e-1}\frac{1}{\epsilon_2}(ln\frac{1}{\theta} + d)$, and $r = \frac{O(1)\lambda^2 D^2}{(\epsilon_1)^2}[\ln(\frac{O(1)LD}{\epsilon_1}) + \ln(\frac{1}{\alpha})]$.

Proof: By substituting the expectation term with the sample mean and introducing the variable M, we obtain the formulation of (15) through a similar argument as in [15]. Since $\bar{y}$ is independent of the portion of Y restricted on S^-, we can rewrite the minimization of (14) in a SP formulation:

$$\min_{x\in X} \mathbb{E}[\bar{\beta}\cdot \operatorname*{ess\,sup}_{\bar{y}\in S^+} f(x,\bar{y}) + f(x,Y)\cdot \mathbb{1}_{S^-}(y)].$$

Hence, according to [1, Corollary 5.19], with r satisfying the conditions in the corollary, we obtain the probabilistic optimality guarantee of x^*_s. To show the robustness guarantee of x^*_s, with the independence of the expectation term and $\bar{y}$, we obtain an equivalent minimization problem of (14):

$$\min_{x\in X}\{\operatorname*{ess\,sup}_{\bar{y}\in S^+}[\bar{\beta}\cdot f(x,\bar{y}) + \mathbb{E}[f(x,Y)\cdot \mathbb{1}_{S^-}(y)]]\}.$$

Observing that ess sup operates on the whole expression, we use an epigraph form to obtain a problem suitable for ScPs. Invoking [15] and [17], we obtain the results. ■

With the reformulation, we obtain a deterministic problem that becomes easily solvable using existing algorithms.

V. Case Study: Supervised Learning

In this section, we test the performance of our hybrid measure in handling randomness. We compare different risk measures in obtaining an efficient hinge loss for support-vector machines (SVMs) when the data points are affected by noise.

We use the wireless indoor localization dataset from [18] and [19]. Each data point $x_i, i = 1, \ldots, N$, represents signal strengths of seven routers. We add random noise e to x_i, $i \in I_{+1}$ corresponding to one class to make them difficult to be distinguished from x_i, $i \in I_{-1}$ corresponding to the other class. Noise e can be considered as either consequences of endogenous failures of the routers or results of exogenous disturbances from an attacker aiming to mislead the classifier.

We use (2) to measure the risk associated with the random distances between the margin and the data points. By introducing variables $\zeta_i, i = 1, \ldots, N$, we formulate the following problem to obtain an SVM with ω denoting the normal vector of the margin and b denoting a shift from the origin:

$$\begin{aligned} \min_{w,b,\zeta_i,i=1,\ldots,N} \ & \frac{1}{N}\sum_{i=1}^{N}\zeta_i + ||w||^2 \\ s.t. \ & \zeta_i \ge 0, i = 1, \ldots, N, \\ & \zeta_i \ge 1 - (-1)\cdot(w^T x_i - b), i \in I_{-1}, \\ & \zeta_i \ge R_{\bar{\beta}}[1 - (1)\cdot(w^T(x_i + e) - b)], i \in I_{+1}. \end{aligned} \tag{16}$$

We compare (16) with SVMs whose hinge loss functions are obtained using expectation, worst-case losses, and CVaR. The formulations of these counterparts take the same form as (16) except that the right hand side of the last constraint is replaced by $\mathbb{E}[1-(1)\cdot(w^T(x_i+e)-b)]$, $\max_{e\in E}[1-(1)\cdot(w^T(x_i+e)-b)]$, and $\mathrm{CVaR}_\alpha[1-(1)\cdot(w^T(x_i+e)-b)]$, respectively.

Let $e_1 \sim \mathcal{N}(-30, 49)$. We pick $e = (e_1, 0, 0, 0, 0, 0, 0)^T$ to reduce the problem dimension. We have $K = 500$ i.i.d. samples of e_1. We use the histogram of sampled e_1 to partition (4) and (5). Given a choice of $\bar{\beta}$, denoted by $\tilde{\beta}$, we can determine the number of samples that belong to (4) using $N^+ = N\tilde{\beta}$.

TABLE I
EXPERIMENT RESULTS FOR HYBRID RISK MEASURE

$\bar{\beta}$	0.164	0.286	0.390	0.490	0.586	0.680
Accuracy Rate	94.0%	94.0%	94.0%	93.5%	92.5%	92.5%
$\bar{\beta}$	0.768	0.830	0.882	0.918	0.950	
Accuracy Rate	93.0%	94.5%	94.5%	95.0%	94.5%	

Then, we group highly populated bins until the number of included samples is at least N^+. This method yields two partitioned sets over the uncertainties created by e_1. The parameter $\bar{\beta}$ follows directly from the partition.

The size of training set is $N = 800$. Parameters $\bar{\beta}$ in (2) and α in CVaR take values ranging from 0.05 to 0.95 with the lengths of the intervals set to 0.05. We use a histogram of 20 bins. Uncertainty set E is the interval between the two extreme samples. There are 200 testing data points.

Results for different values of $\bar{\beta}$ for our measure is presented in Table I. The accuracy rates are 93.5% and 93.0% for the SVMs with expectation and worst-case scenario. The same highest accuracy rate is 94.5% for the SVMs with CVAR_α with $\alpha = 0.70, 0.75, 0.80, 0.85, 0.90, 0.95$. We can note that the overall best accuracy rate 95.0% is achieved by the hybrid measure with $\bar{\beta} = 0.918$. Besides, for most of the choices of $\bar{\beta}$, the SVMs with the hybrid measure outperforms the RO and SP counterparts. Moreover, the results for our measure and CVaR share a similar trend. The accuracy rates of the SVMs with CVaR are around 93.5% when $\alpha \leq 0.25$. They achieve 92.5% when $0.30 \leq \alpha \leq 0.50$. They gradually reach the peak for larger values of α. This observation shows the effectiveness of (2) in quantifying the risk of uncertain events.

VI. CONCLUSION

In this letter, we have proposed a hybrid coherent risk measure that arises from a mixture of two perspectives toward uncertainties. One is the robust optimization approach that views each uncertain event equally probable and makes the worst-case decision. The other is the stochastic programming that views uncertainties as a random variable and makes decisions under average-performance criteria. We have formulated a decision-making problem under the hybrid measure and obtained an equivalent problem of finding the worst-case density function. The sensitivity analysis provided a way of capturing how the optimal value changes with respect to perturbations of the level parameter. Differentiability of the proposed measure is also proven to ensure the availability in using gradient based algorithms to optimize. Furthermore, we have proposed an approximation method combining scenario programming and sample average approximation to compute the optimal solution numerically. A supervised learning problem is used as a case study to corroborate the analytical results and computational algorithms.

The proposed methodology provides a fundamental tradeoff between the risk neutralism from the stochastic programming approach and the pessimism from the robust optimization approach. As future work, we would like to study a game setting with an opponent controlling the randomness. The equilibrium behaviors of the players would provide insights on the worst-case uncertainty and guide the efficient selection of a level parameter. Besides, we would like to extend our work to multistage optimization problems, and leverage the ideas of the proposed measure to provide robust yet resilient performances for dynamic systems.

REFERENCES

[1] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, PA, USA: SIAM, 2014.
[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, vol. 28. Princeton, NJ, USA: Princeton Univ. Press, 2009.
[3] D. Bertsimas and M. Sim, "The price of robustness," *Oper. Res.*, vol. 52, no. 1, pp. 35–53, 2004.
[4] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM Rev.*, vol. 53, no. 3, pp. 464–501, 2011.
[5] A. Shapiro and A. Philpott. (2007). *A Tutorial on Stochastic Programming*. [Online]. Available: www2.isye.gatech.edu/ashapiro/publications.html
[6] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Oper. Res.*, vol. 62, no. 6, pp. 1358–1376, 2014.
[7] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathe. Finance*, vol. 9, no. 3, pp. 203–228, 1999.
[8] S. Uryasev, S. Sarykalin, G. Serraino, and K. Kalinchenko, "VaR vs CVaR in risk management and optimization," in *Proc. CARISMA Conf.*, 2010, pp. 1–75.
[9] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *J. Risk*, vol. 2, pp. 21–42, Apr. 2000.
[10] Y. Yamai and T. Yoshiba, "Comparative analyses of expected shortfall and value-at-risk (3): Their validity under market stress," IMES Discussion Paper No 2002-E-2, Bank Japan, Tokyo, Japan, 2002.
[11] A. Ruszczyński and A. Shapiro, "Optimization of risk measures," in *Probabilistic and Randomized Methods for Design Under Uncertainty*, G. Calafiore and F. Dabbene, Eds. London, U.K.: Springer, 2006, pp. 119–157.
[12] J. B. Lasserre, "Level sets and non Gaussian integrals of positively homogeneous functions," *Int. Game Theory Rev.*, vol. 17, no. 1, 2015, Art. no. 1540001.
[13] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Sci., 1999.
[14] J. Norris, "Probability and measure," *Notes for Students*. London, U.K.: Museums Assoc., 2010.
[15] M. C. Campi and S. Garatti, "The exact feasibility of randomized solutions of uncertain convex programs," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1211–1230, 2008.
[16] G. C. Calafiore, "Random convex programs," *SIAM J. Optim.*, vol. 20, no. 6, pp. 3427–3464, 2010.
[17] T. Alamo, R. Tempo, and A. Luque, "On the sample complexity of randomized approaches to the analysis and design under uncertainty," in *Proc. IEEE Amer. Control Conf.*, 2010, pp. 4671–4676.
[18] R. Bhatt, "Fuzzy-rough approaches for pattern classification: Hybrid measures, mathematical analysis, feature selection algorithms, decision tree algorithms, neural learning, and applications," in *Decision Tree Algorithms, Neural Learning, and Applications*, Amazon Books, 2017.
[19] J. G. Rohra, B. Perumal, S. J. Narayanan, P. Thakur, and R. B. Bhatt, "User localization in an indoor environment using fuzzy hybrid of particle swarm optimization & gravitational search algorithm with neural networks," in *Proc. 6th Int. Conf. Soft Comput. Prob. Solving*, 2017, pp. 286–295.