Strong Duality Relations in Nonconvex Risk-Constrained Learning

Dionysis Kalogerias
Department of Electrical Engineering
Yale University
New Haven, USA
dionysis.kalogerias@yale.edu

Spyridon Pougkakiotis
School of Science and Engineering
University of Dundee
Dundee, UK
spougkakiotis001@dundee.ac.uk

2024 58th Annual Conference on Information Sciences and Systems (CISS) | 979-8-3503-6929-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CISS59072.2024.10480186

***Abstract*—We establish strong duality relations for functional two-step compositional risk-constrained learning problems with multiple nonconvex loss functions and/or learning constraints, regardless of nonconvexity and under a minimal set of technical assumptions. Our results in particular imply zero duality gaps within the class of problems under study, both extending and improving on the state of the art in (risk-neutral) constrained learning. More specifically, we consider risk objectives/constraints which involve real-valued convex and positively homogeneous risk measures admitting dual representations with bounded risk envelopes, generalizing expectations and including popular examples, such as the conditional value-at-risk (CVaR), the mean-absolute deviation (MAD), and more generally all real-valued coherent risk measures on integrable losses as special cases. Our results are based on recent advances in risk-constrained nonconvex programming in infinite dimensions, which rely on a remarkable new application of J. J. Uhl's convexity theorem, which is an extension of A. A. Lyapunov's convexity theorem for general, infinite dimensional Banach spaces. By specializing to the risk-neutral setting, we demonstrate, for the first time, that constrained classification and regression can be treated under a unifying lens, while dispensing certain restrictive assumptions enforced in the current literature, yielding a new state-of-the-art strong duality framework for nonconvex constrained learning.**

***Index Terms*—Lagrangian Duality, Strong Duality, Zero Duality Gap, Risk-Constrained Learning, Constrained Regression, Constrained Classification, Constrained Learning with Nonconvex Losses.**

I. Introduction

Classification and regression tasks constitute two core problem classes widely appearing in numerous flavors in signal processing, machine and statistical learning. Recent advances in machine learning and artificial intelligence have rejuvenated this area of research, enabling highly efficient solution methods for wide classes of regression and classification tasks of paramount importance in practical applications, including, among many others, healthcare [1], engineering [2], [3] and computer science [4]. Nonetheless, the literature on classification and regression has, so far, mostly focused on the unconstrained setting, often leading to standard fitting problems. While this class of problems is highly relevant, it fails to ensure that the associated learnt policy will explicitly satisfy key properties of interest, such as safety or lack of bias; see, e.g., [5], [6]. In such cases, highly specialized priors have been utilized, requiring extensive tuning and often unrealistic assumptions, thus leading to fragile methodologies [7]. In light of the ubiquitous nature of systems utilizing regression or classification tasks, there has been a rise in the demand of appropriate ways to incorporate constraints for handling issues like fairness [8], [9], robustness [10], [11], or safety [12], [13], among others. This is especially important in the context of nonconvex (constrained) learning, which typically occurs when utilizing (deep) neural networks (NNs).

Standard machine learning literature, which focuses on the unconstrained case, relies on the use of appropriate penalty functions to satisfy the various specifications that a given learning policy needs to abide by (e.g., see [14]). However, designing appropriate penalties and finding the corresponding parameters for weighting the different objectives is a notorious problem, leading to practically questionable results or fragile "optimal" policies. Such difficulties have recently lead to the study of constrained learning problems [15]–[18].

In this setting, one key question concerns the relation of the constrained learning problem with its associated *Lagrangian dual*. This is particularly important to study in the context of nonconvex constrained learning, in which duality is an especially complicated concept. Specifically, *strong duality* results —which indicate that the primal problem can be substituted by the dual using certain Lagrange multipliers— in constrained learning with general nonconvex losses in both the objective and/or the constraints are particularly important. Indeed, typical constrained learning problems and associated solution methods rely on some form of NN parametrization, which usually leads to nonconvex optimization problems (thus nonconvex constrained learning problems are ubiquitous). At the same time, parametrization in the primal domain allows one to tackle the problem in the dual domain, which is typically finite-dimensional (despite the infinite-dimensional nature of the original —non-parametrized— primal problem).

It is thus not a surprise that constrained learning problems are typically tackled in the Lagrangian dual domain; see, e.g., [13], [16], [17], [19], [20]. Consequently, it is of key importance to find minimal conditions under which the problem under consideration exhibits strong duality, and especially in cases where the associated loss functions are nonconvex. Of

The authors would like to kindly acknowledge support by a Microsoft gift, as well as by the US National Science Foundation (NSF) under Grant CCF 2242215. Author names are listed in alphabetical order.

course, in the context of convex learning this is relatively straightforward via utilizing standard constraint qualifications (CQ), such as Slater's CQ. In the nonconvex setting, however, this question is substantially more challenging. At the same time, strong duality in this context allows for the general study of generalization properties of the associated sample-average approximation (SAA) —or empirical risk minimization (ERM)— problems (as in [8], [9], [15], [18]), generalizing previous established results in the context of traditional statistical learning theory [21]. Except for the above, strong duality is also important for establishing minimal conditions under which saddle points exist for the primal-dual problem, as well as for deriving necessary conditions for optimality.

To the best of our knowledge, current state-of-the-art results concerning strong-duality in nonconvex constrained learning were recently developed in [18]. The theory in [18] distinguishes between two cases: constrained classification and constrained regression. In each case, by devising appropriate assumptions, the authors are able to establish strong duality relations in the general nonconvex setting, via utilizing the celebrated convexity theorem of A. A. Lyapunov (see [22, Corollary IX.1.6]). While the results of [18] are fairly general, there are several issues which we believe can and should be addressed, strengthening the theory of [18] significantly.

First, treating the classification and the regression tasks separately introduces some counter-intuitive and hard-to-evaluate assumptions (e.g., see [18, Assumption 6]). Instead, our claim herein is that the classification and regression tasks can be naturally studied in a unified manner. Second, [18] assumes that the underlying functional space of feasible policies is closed, decomposable, and convex. We claim that only decomposability is required (as far as strong duality is concerned), and this is quite important, since convexity is often a property that decomposable spaces fail to satisfy (except in trivial cases). Third, we address some technical issues in [18, Proof of Proposition III.2], and propose a slightly different approach to resolve them. Apart from the above, our analysis also relaxes some additional assumptions utilized in [18].

In a nutshell, our contributions in this work enable the use of *nonlinear risk functionals* in place of linear operators in the underlying model, i.e., our technical approach allows for general *risk measures* [23] in place of (linear) expectations. In fact, risk-aware (constrained) learning is an increasingly relevant subject that has recently gained significant traction (see, e.g., [24]–[26]). There are several reasons for this. For instance, risk-aware learning can be incorporated to boost the tail behaviour of a classifier [27]. Additionally, while on-average behaviour of existing systems might be optimal, they face failure when faced with atypical data [28]. Further, by utilizing risk measures such as the CVaR, one can substantially robustify indicator-type (chance) constraints that are typically employed in the constrained learning literature (as in [18, Section V.B]), thus resulting in more favorable problem formulations. Overall, incorporating risk-measures in constrained learning has a plethora of useful applications.

Nonetheless, the focus of this work is mostly theoretical; indeed, we lay the theoretical foundations on how one can incorporate such risk measures in the objective and/or the constraints of a learning problem, and establish strong duality relations of such problems under minimal conditions, in fact under the exact same conditions required for strong duality in the risk-neutral setting. More specifically, we propose a general two-step compositional risk-constrained learning framework, where the risk incurred by each corresponding random loss (for each objective/constraint) is evaluated *hierarchically*, *first* relative to the response (e.g., label) posterior given the features, and *then* relative to the feature prior. Essentially, we advocate for a *decomposition of risk* into *likelihood risk* and *prior risk*, where each component may be evaluated by a risk measure of different type.

In light of the above, our results pave the way for a more widespread use of risk measures in the context of constrained learning, and also for the development of dual-domain algorithms for actually solving such risk-aware constrained learning problems. At the same time, by specializing our results to the risk-neutral case, we resolve some technical issues present in the available literature (as discussed above), while also providing state-of-the-art results using more general and realistic assumptions, indeed verifiable in a practical setting.

Notation: Bold capital letters (such as $\mathbf{A}$), or calligraphic letters (such as $\mathcal{A}$) will denote finite-dimensional sets/spaces, such as Euclidean spaces. Math script letters (such as $\mathscr{A}$) will denote σ-algebras. Boldsymbol letters (such as $\boldsymbol{A}$ or $\boldsymbol{a}$) will denote (random) vectors. The space of p-integrable functions from a measurable space $(\Omega, \mathscr{F})$ equipped with a (σ-)finite measure $\mu\colon \mathscr{F} \to \mathbb{R}_+$ to a Banach space $\mathbb{A}$, with standard notation $\mathcal{L}_p(\Omega, \mathscr{F}, \mu; \mathbb{A})$, is abbreviated as $\mathcal{L}_p(\mu, \mathbb{A})$. The rest of the notation is standard.

II. Strong Duality in Risk-Constrained Functional Programming

Let us fix an arbitrary complete probability space $(\Omega, \mathscr{F}, \mu)$, and consider a random element $\boldsymbol{H} : \Omega \to \mathcal{H} \triangleq \mathbb{R}^{N_H}$ with induced Borel measure $\mathrm{P} \equiv \mathrm{P}_{\boldsymbol{H}} : \mathscr{B}(\mathcal{H}) \to [0,1]$, modeling some observable random phenomenon. Following the recent developments in [29], we are interested in the class of risk-constrained nonconvex functional programs formulated as

$$\boxed{\begin{array}{ll} \underset{\boldsymbol{x}, \boldsymbol{p}(\cdot)}{\text{maximize}} & g^o(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \leqslant -\boldsymbol{\rho}(-\boldsymbol{f}(\boldsymbol{p}(\boldsymbol{H}), \boldsymbol{H})) \\ & \boldsymbol{g}(\boldsymbol{x}) \geqslant \mathbf{0} \\ & (\boldsymbol{x}, \boldsymbol{p}) \in \mathcal{X} \times \Pi \end{array}}\,, \tag{RCP}$$

where $\mathbf{C} \triangleq \mathbb{R}^N$, $g^o : \mathbb{R}^N \to \mathbb{R}$ and $\boldsymbol{g} : \mathbb{R}^N \to \mathbb{R}^{N_g}$ are given concave functions, $\boldsymbol{p} : \mathcal{H} \to \mathbf{R} \triangleq \mathbb{R}^{N_p}$ is an allocation policy on observables $\boldsymbol{H}$, $\boldsymbol{f} : \mathbf{R} \times \mathcal{H} \to \mathbf{C}$ is a (generally nonconvex) function measuring the quality of a policy $\boldsymbol{p}$ at each realization $\boldsymbol{H}$ in $\mathcal{H}$ and such that $\boldsymbol{f}(\boldsymbol{p}(\cdot), \cdot) \in \mathcal{L}_1(\mathrm{P}, \mathbf{C})$ on Π, and where $\boldsymbol{\rho} : \mathcal{L}_1(\mathrm{P}, \mathbf{C}) \to \mathbf{C}$ is a finite-valued vector risk measure, assumed to be *convex, lower semicontinuous and positively*

homogeneous in every dimension (i.e., component-wise), such that for each $i \in \mathbb{N}_N^+$,

$$\rho_i(\boldsymbol{Z}) = \rho_i(Z_i), \quad \text{for all } \boldsymbol{Z} \in \mathcal{L}_1(\mathrm{P}, \mathbf{C}).$$

As indicated in (RCP), the finite-dimensional variables $\boldsymbol{x}$ are further restricted to the set $\mathcal{X} \subseteq \mathbb{R}^N$ and the policies denoted by $\boldsymbol{p}$ are restricted to some infinite-dimensional set Π.

It was recently proven in [29] that (RCP) exhibits strong duality under rather standard assumptions, strictly generalizing state-of-the-art results on the risk-neutral case (i.e., where all associated risk measures are expectations), e.g., in [30]–[32].

Assumption 1. The following conditions are in effect:

1) The utilities g^o and $\boldsymbol{g}$ are concave.
2) The service feasible set $\mathcal{X}$ is convex.
3) The policy feasible set Π is decomposable.
4) The Borel measure P is nonatomic[a].
5) Problem (RCP) satisfies Slater's CQ (i.e., it is strictly feasible).

[a]Recall that P is nonatomic if for any event E with $\mathrm{P}(E) > 0$, an event $E' \subseteq E$ exists such that $\mathrm{P}(E) > \mathrm{P}(E') > 0$.

In particular, the following fact was established in [29].

Theorem 1 (Kalogerias, Pougkakiotis)**.** *Let Assumption 1 be in effect. Then problem* (RCP) *has zero (Lagrangian) duality gap. In fact,* (RCP) *exhibits strong duality, i.e., optimal dual variables exist.*

In fact, Theorem 1 is applicable in seemingly more general formulations of (RCP). In particular, by letting ρ_0 be some real-valued convex, lower semicontinuous and positively homogeneous risk measure, $f_0 \colon \boldsymbol{R} \times \mathcal{H} \to \mathbb{R}$ be an arbitrary function such that $f_0(\boldsymbol{p}(\cdot), \cdot) \in \mathcal{L}_1(P, \mathbb{R})$, and $\boldsymbol{r} \colon \mathcal{X} \to \mathbb{R}^N$ be some component-wise convex function, it readily follows that the problem

$$\begin{array}{ll} \underset{\boldsymbol{x}, \boldsymbol{p}(\cdot)}{\text{maximize}} & g^o(\boldsymbol{x}) - \rho_0(-f_0(\boldsymbol{p}(\boldsymbol{H}), \boldsymbol{H})) \\ \text{subject to} & \boldsymbol{r}(\boldsymbol{x}) \preccurlyeq -\boldsymbol{\rho}(-\boldsymbol{f}(\boldsymbol{p}(\boldsymbol{H}), \boldsymbol{H})) \\ & \boldsymbol{g}(\boldsymbol{x}) \succcurlyeq \mathbf{0} \\ & (\boldsymbol{x}, \boldsymbol{p}) \in \mathcal{X} \times \Pi \end{array},$$

also exhibits strong duality. For a detailed derivation of this fact, we refer the reader to [29, Section 6.1].

III. Nonconvex Risk-Constrained Learning

In this section, we will provide precise conditions to show that a wide range of risk-constrained learning problems, involving nonconvex losses, can be cast in the form of (RCP), thus exhibiting strong duality under Assumption 1.

We consider a general formulation of a supervised risk-constrained learning problem, where the associated loss functions are allowed to be nonconvex. On our usual probability space $(\Omega, \mathscr{F}, \mu)$, we consider random *example vectors* $(\boldsymbol{X}_i, Y_i) \colon \Omega \to \mathbb{R}^d \times \mathbb{R}$, $i \in \mathbb{N}_m$ together with their induced Borel probability distributions $\mathcal{D}_i : \mathscr{B}(\mathbb{R}^d \times \mathbb{R}) \to [0, 1]$, instantiated over data pairs $(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathbb{R}^d$ represents a realized feature or system input and $y \in \mathbb{R}$ represents a realized label or measurement (response). We denote by $\mathcal{D}_{\boldsymbol{X}_i}$ the marginal probability distribution of $\boldsymbol{X}_i$, and by $\mathcal{D}_{Y_i|\boldsymbol{X}_i}$ the conditional probability distribution of Y_i given a realization of $\boldsymbol{X}_i$, for $i \in \mathbb{N}_m$.

Assumption 2. The marginal probability distributions $\mathcal{D}_{\boldsymbol{X}_i}$, $i \in \mathbb{N}_m^+$, are absolutely continuous with respect to the "common denominator" $\mathcal{D}_{\boldsymbol{X}_0}$ (without loss of generality), which in turn is assumed to be nonatomic.

Remark 1. We note that Assumption 2 implies, by virtue of the Radon-Nikodym theorem, that there must exist integrable functions $w_i \colon \mathbb{R}^d \to \mathbb{R}^+$, such that $w_i \triangleq d\mathcal{D}_{\boldsymbol{X}_i}/d\mathcal{D}_{\boldsymbol{X}_0}$, for all $i \in \mathbb{N}_m^+$. Our assumption of absolute continuity of the marginal distributions $\mathcal{D}_{\boldsymbol{X}_i}$ with respect to $\mathcal{D}_{\boldsymbol{X}_0}$ is standard in the literature. Indeed, it holds if each distribution $\mathcal{D}_{\boldsymbol{X}_i}$ is assumed to have a Lebesgue density (see, e.g., [18] and [29]). On the other hand, nonatomicity of the common denominator $\mathcal{D}_{\boldsymbol{X}_0}$ is also standard (see, e.g., [18, Theorem 1]) and very intuitive, since the feature or input space is naturally expected to be continuous in many applications of interest yielding (constrained) classification or constrained regression tasks.

Before proceeding with our proposed formulation of a general risk-constrained learning problem, we introduce the notion of a *conditional risk mapping*. In particular, for any $i \in \mathbb{N}_m$, we consider the vector spaces $\mathcal{Z}_1^i \triangleq \mathcal{L}_p(\Omega, \sigma(\boldsymbol{X}_i), \mu; \mathbb{R})$ and $\mathcal{Z}_2^i \triangleq \mathcal{L}_{p'}(\mathbb{R}^d \times \mathbb{R}, \mathscr{B}(\mathbb{R}^d \times \mathbb{R}), \mathcal{D}_i) \equiv \mathcal{L}_{p'}(\mathcal{D}_i, \mathbb{R})$, where $p, p' \in [1, \infty]$. We define a conditional risk mapping as a functional $\widehat{\rho}\left(\cdot \middle| \boldsymbol{X}_i\right) : \mathcal{Z}_2^i \to \mathcal{Z}_1^i$, with the property that for every qualifying Borel function $Z \colon \mathbb{R}^{d+1} \to \mathbb{R}$, $\widehat{\rho}(\cdot|\boldsymbol{X}_i)$ obeys the relation

$$\widehat{\rho}\left(Z(\boldsymbol{X}_i, Y_i)|\boldsymbol{X}_i\right)(\omega) = \widehat{\rho}\left(Z(\boldsymbol{x}, Y_i)|\boldsymbol{X}_i\right)(\omega)\big|_{\boldsymbol{x}=\boldsymbol{X}_i(\omega)},$$

for every elementary event $\omega \in \Omega$, where the instantiation $[\widehat{\rho}\left(Z(\boldsymbol{x}, \cdot)|\boldsymbol{X}_i\right)](\omega) \colon \mathcal{Z}_3^i \to \mathbb{R}$ is a properly chosen risk functional, with $\mathcal{Z}_3^i \triangleq \mathcal{L}_{p'}(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathcal{D}_{Y_i|\boldsymbol{X}_i}) \equiv \mathcal{L}_p(\mathcal{D}_{Y_i|\boldsymbol{X}_i}, \mathbb{R})$. A typical example of such a conditional risk mapping is, of course, the conditional expectation. We note that the definition of a conditional risk mapping above is compatible with our particular construction in Section II, relying on a Borel space setting. Conditional risk mappings are usually defined more generally, and are often required to satisfy certain conditions of interest, and thus associated definitions and assumptions vary (see, e.g., [33, Definition 2.1] and [34, Definition 1]). Concrete examples of conditional risk mappings obeying our definition will be given later on.

Letting $\ell_i \colon \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$ be given (possibly nonconvex or discontinuous) "loss" functions, for $i \in \mathbb{N}_m$, our proposed risk-constrained learning framework relies on the constrained nonconvex functional program

$$\boxed{\begin{array}{ll} \underset{\boldsymbol{f}(\cdot) \in \mathcal{F}}{\text{minimize}} & \rho_0\big(\widehat{\rho}_0\big(\ell_0\big(\boldsymbol{f}(\boldsymbol{X}_0), Y_0\big)\big|\boldsymbol{X}_0\big)\big) \\ \text{subject to} & \rho_i\big(\widehat{\rho}_i\big(\ell_i\big(\boldsymbol{f}(\boldsymbol{X}_i), Y_i\big)\big|\boldsymbol{X}_i\big)\big) \leqslant c_i, i \in \mathbb{N}_m^+ \end{array}}, \quad \text{(RCL)}$$

where $\boldsymbol{f} \colon \mathbb{R}^d \to \mathbb{R}^k$ belongs to an appropriate policy space $\mathcal{F}$, and for $i \in \mathbb{N}_m$, $c_i \in \mathbb{R}$, ρ_i is a real-valued convex, lower

semicontinuous, and positively homogeneous risk measure, while $\widehat{\rho}_i(\cdot|\boldsymbol{X}_i)$ is some conditional risk mapping.

Assumption 3. For each $i \in \mathbb{N}_m$, we have $\ell_i\left(\boldsymbol{f}(\bullet),\cdot\right) \in \mathcal{Z}_2^i$, for any $\boldsymbol{f} \in \mathcal{F}$, and the space of policies $\mathcal{F}$ is decomposable.

Remark 2. Let us emphasize that Assumption 3 is very general. The decomposability of the policy space $\mathcal{F}$ is standard, and is already utilized in (RCP). Furthermore, the assumption that $\ell_i\left(\boldsymbol{f}(\bullet),\cdot\right) \in \mathcal{Z}_2^i$ is also standard and very mild. In particular, it also appears in Assumption 1, and includes various nonconvex or even vastly discontinuous functions ℓ_i.

Fix $i \in \mathbb{N}_m$. Under Assumption 3, and using our definition of a conditional risk mapping $\widehat{\rho}(\cdot|\boldsymbol{X}_i)$, it readily follows that

$$\begin{aligned}\widehat{\rho}_i(\ell_i(\boldsymbol{f}(\boldsymbol{X}_i),Y_i)|\boldsymbol{X}_i) &\equiv \widehat{\rho}_i(\ell_i(\boldsymbol{z},Y_i)|\boldsymbol{X}_i)\big|_{\boldsymbol{z}=\boldsymbol{f}(\boldsymbol{X}_i)}\\ &\equiv F_i\left(\boldsymbol{f}(\boldsymbol{X}),\boldsymbol{X}\right),\end{aligned}$$

such that $F_i\left(\boldsymbol{f}(\cdot),\cdot\right) \in \mathcal{L}_p\big(\mathcal{D}_{\boldsymbol{X}_i},\mathbb{R}\big)$, for any $\boldsymbol{f} \in \mathcal{F}$. We now showcase that our definition of conditional risk mappings is general and includes various useful examples as special cases, as follows:

- As already mentioned, the most typical example of a conditional risk mapping is that of conditional expectation. In the context of (RCL), we set $\widehat{\rho}(\cdot|\boldsymbol{X}_i) = \mathbb{E}\left\{\cdot|\boldsymbol{X}_i\right\}$. Since we assume that the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ exists, it then readily follows that

$$\mathbb{E}\left\{\ell_i\left(\boldsymbol{f}(\boldsymbol{X}_i),Y_i\right)|\boldsymbol{X}_i\right\} = \mathbb{E}\left\{\ell_i\left(\boldsymbol{z},Y_i\right)\big|\boldsymbol{X}_i\right\}\big|_{\boldsymbol{z}=\boldsymbol{f}(\boldsymbol{X}_i)}.$$

  The latter equality is well-known as the *substitution rule*.
- Let us now consider two popular cases in the class of *coherent* conditional risk mappings (see [23, Chapter 6]). The first is the conditional version of the conditional value at risk (CVaR). For any $i \in \mathbb{N}_m$, and any $\alpha \in (0,1)$, we can define $\mathrm{CVaR}_\alpha^i\left(\cdot|\boldsymbol{X}_i\right) : \mathcal{Z}_2^i \to \mathcal{Z}_1^i$, as

$$\mathrm{CVaR}_\alpha^i\left(Z|\boldsymbol{X}_i\right) = \inf_{W\in\mathcal{Z}_1^i}\left\{W + \alpha^{-1}\mathbb{E}\left\{(Z-W)_+\,|\boldsymbol{X}_i\right\}\right\},$$

  for $Z \in \mathcal{Z}_2^i$, where $(\cdot)_+ \equiv \max\{\cdot,0\}$. As before, the instantiation $\left[\mathrm{CVaR}_\alpha^i\left(Z|\boldsymbol{X}_i\right)\right](\boldsymbol{X}(\omega))$ can be considered to take values from $\mathcal{Z}_3^i$. As in the case of conditional expectation, it is not hard to see that by letting $Z = \ell_i(\boldsymbol{f}(\boldsymbol{X}_i),Y_i)$, we obtain

$$\begin{aligned}&\mathrm{CVaR}_\alpha^i\left(\ell_i(\boldsymbol{f}(\boldsymbol{X}_i),Y_i)|\boldsymbol{X}_i\right)\\ &\quad= \mathrm{CVaR}_\alpha^i\left(\ell_i(\boldsymbol{z},Y_i)|\boldsymbol{X}_i\right)\big|_{\boldsymbol{z}=\boldsymbol{f}(\boldsymbol{X}_i)}.\end{aligned}$$

  The second popular case we consider is the conditional mean-upper-semideviation. In particular, for $Z \in \mathcal{Z}_2^i$, and some $c \in [0,1]$, this conditional risk mapping takes the form

$$\widehat{\rho}(Z|\boldsymbol{X}_i) = \mathbb{E}\left\{Z|\boldsymbol{X}_i\right\} + c\big(\mathbb{E}\big\{\left(Z-\mathbb{E}\{Z|\boldsymbol{X}_i\}\right)_+^{p'}\big|\boldsymbol{X}_i\big\}\big)^{1/p'}.$$

  Again, one can readily verify that this conditional risk mapping is well-defined in the sense of our definition, and thus satisfies the substitution rule.
- Lastly, in order to stress the generality of the conditional risk mappings considered in this work, let us define conditional extensions of the generalized mean semideviations introduced in [35]. In fact, we may consider a larger class of generalized mean semideviations than those discussed in [35]. To that end, let $R\colon \mathbb{R} \to \mathbb{R}$ be any nonnegative (possibly nonconvex) function. For any $Z \in \mathcal{Z}_2^i$, and any $c \in [0,+\infty)$, we consider conditional generalized mean semideviation risk measures, defined as

$$\begin{aligned}&\widehat{\rho}(Z|\boldsymbol{X}_i)\\ &= \mathbb{E}\left\{Z|\boldsymbol{X}_i\right\} + c\big(\mathbb{E}\big\{\left(R\left(Z-\mathbb{E}\{Z|\boldsymbol{X}_i\}\right)\right)^{p'}\big|\boldsymbol{X}_i\big\}\big)^{1/p'},\end{aligned}$$

  provided that $R(Z-\mathbb{E}\{Z|\boldsymbol{X}_i\}) \in \mathcal{Z}_1^i$. Apparently, this conditional risk mapping is also well-defined according to our definition and satisfies the substitution rule. At the same time, we observe that, depending on the choice of R, such conditional risk mappings might fail to satisfy several properties such as convexity, monotonicity, positive homogeneity, etc. The model in (RCL) allows such general constructions, showcasing the wide range of conditional risk mappings enabled in our framework.

Our first main result in this section may now be formulated, as follows.

Theorem 2. *Let problem (RCL) satisfy Assumptions 2 and 3. Then, there exist convex, proper, lower-semicontinuous and positively homogeneous risk measures $\widetilde{\rho}_i\colon \mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_0},\mathbb{R}) \to \mathbb{R}$, as well as functions $G_i(\bullet,\cdot)$ satisfying $G_i(\boldsymbol{f}(\cdot),\cdot) \in \mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_0},\mathbb{R})$, for all $i \in \mathbb{N}_m^+$, such that problem (RCP) can be equivalently written as*

$$\begin{array}{ll}\underset{\boldsymbol{f}(\cdot)\in\mathcal{F}}{\text{minimize}} & \rho_0\left(F_0\big(\boldsymbol{f}(\boldsymbol{X}_0),\boldsymbol{X}_0\big)\right)\\ \text{subject to} & \widetilde{\rho}_i\left(G_i\big(\boldsymbol{f}(\boldsymbol{X}_0),\boldsymbol{X}_0\big)\right) \leqslant c_i, i \in \mathbb{N}_m^+\end{array}. \quad \text{(RCL0)}$$

Proof. From Assumption 2, we observe that each $\mathcal{D}_{\boldsymbol{X}_i}$, for $i \in \mathbb{N}_m^+$, is absolutely continuous with respect to $\mathcal{D}_{\boldsymbol{X}_0}$, while, from Assumption 3, we have shown that $F_i(\boldsymbol{f}(\cdot),\cdot) \in \mathcal{L}_p(\mathcal{D}_{\boldsymbol{X}_i},\mathbb{R}) \subset \mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_i},\mathbb{R})$. It then follows that $G_i(\boldsymbol{f}(\cdot),\cdot) \triangleq F_i(\boldsymbol{f}(\cdot),\cdot)w_i \in \mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_0},\mathbb{R})$, $i \in \mathbb{N}_m^+$. Let $\mathbb{A}_i \subseteq \mathcal{L}_\infty(\mathcal{D}_{\boldsymbol{X}_i},\mathbb{R})$, for all $i \in \mathbb{N}_m$, be the uncertainty sets corresponding to $\rho_i(\cdot)$ and consider another (related) convex, lower-semicontinuous, and positively homogeneous risk measure $\widetilde{\rho}_i(\cdot)$ defined on $w_i\mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_i},\mathbb{R}) \subseteq \mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_0},\mathbb{R})$, such that $\widetilde{\rho}_i(Zw_i) = \rho(Z)$, for any $Z \in \mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_i},\mathbb{R})$. In particular, we define

$$\begin{aligned}\widetilde{\rho}_i\left(G_i\big(\boldsymbol{f}(\boldsymbol{X}_0),\boldsymbol{X}_0\big)\right) &\triangleq \sup_{\zeta\in\mathbb{A}_i}\int \zeta(\boldsymbol{x})G_i\big(\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{x}\big)d\mathcal{D}_{\boldsymbol{X}_0}(\boldsymbol{x})\\ &= \sup_{\zeta\in\mathbb{A}_i}\int \zeta(\boldsymbol{x})F_i\big(\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{x}\big)d\mathcal{D}_{\boldsymbol{X}_i}(\boldsymbol{x})\\ &= \rho_i\left(F_i\big(\boldsymbol{f}(\boldsymbol{X}_i),\boldsymbol{X}_i\big)\right)\\ &= \rho_i\big(\widehat{\rho}_i\big(\ell_i\big(\boldsymbol{f}(\boldsymbol{X}_i),Y_i\big)\big|\boldsymbol{X}_i\big)\big),\end{aligned}$$

where we have used the fact that $\rho_i(\cdot)$ is convex, lower-semicontinuous, and positively homogeneous. However, the supremum in the second integral, defining $\rho_i(\cdot)$, would not change by restricting $\mathbb{A}_i$ to an appropriately selected bounded

set $\widetilde{\mathbb{A}}_i \subseteq \mathcal{L}_\infty(\mathcal{D}_{\boldsymbol{X}_0}, \mathbb{R})$ chosen independently of Zw_i and with $\mathbb{A}^i \supseteq \widetilde{\mathbb{A}}^i$, since the integral is taken with respect to $d\mathcal{D}_{\boldsymbol{X}_i}$, and does not change for functions differing on $\mathcal{D}_{\boldsymbol{X}_i}$-measure zero sets. In other words, $\widetilde{\rho}(\cdot)$ admits a representation

$$\widetilde{\rho}_i\left(G_i\big(\boldsymbol{f}(\boldsymbol{X}_0), \boldsymbol{X}_0\big)\right) \equiv \sup_{\zeta \in \widetilde{\mathbb{A}}_i} \int \zeta(\boldsymbol{x}) G_i\big(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{x}\big) d\mathcal{D}_{\boldsymbol{X}_0}(\boldsymbol{x}),$$

for some $\widetilde{\mathbb{A}}_i \subseteq \mathcal{L}_\infty(\mathcal{D}_{\boldsymbol{X}_0}, \mathbb{R})$. Using the introduced notation, we recast (RCL) into the equivalent (given our assumptions) form of (RCL0). □

Leveraging Theorem 2, we can now state our second main result of this section.

Theorem 3. *Let Assumptions 2, 3 hold for problem (RCL). If, additionally, (RCL) satisfies Slater's constraint qualification, it exhibits strong duality.*

Proof. We note that under Assumptions 2, 3, and by utilizing Theorem 2, (RCL0) is an instance of (RCP), and hence Staler's constraint qualification suffices to ensure that it satisfies the conditions given in Assumption 1. It then follows from Theorem 1 that (RCL0) exhibits strong duality, and of course the same holds for (RCL). □

IV. The Risk-Neutral Case

So far, we have stated problem (RCL) (and consequently (RCL0)) in full generality. A highly important instance of (RCL) is in the risk-neutral setting, where $\rho_i(\cdot) = \mathbb{E}_{\mathcal{D}_{\boldsymbol{X}_i}}\{\cdot\}$, and $\widehat{\rho}_i(\cdot|\boldsymbol{X}_i) = \mathbb{E}\{\cdot|\boldsymbol{X}_i\}$, for all $i \in \mathbb{N}_m$. Using the tower property, it is easy to see that, for all $i \in \mathbb{N}_m$,

$$\mathbb{E}_{\mathcal{D}_{\boldsymbol{X}_i}}\left\{\mathbb{E}\Big\{\ell_i\big(\boldsymbol{f}(\boldsymbol{X}_i), Y_i\big)\Big|\boldsymbol{X}_i\Big\}\right\} = \mathbb{E}_{\mathcal{D}_i}\left\{\ell_i\big(\boldsymbol{f}(\boldsymbol{X}_i), Y_i\big)\right\},$$

where $\mathcal{D}_i$ is the Borel probability distribution of $(\boldsymbol{X}_i, Y_i)$. Thus, the risk-neutral constrained learning problem reads

$$\begin{array}{ll} \underset{\boldsymbol{f}(\cdot)\in\mathcal{F}}{\text{minimize}} & \mathbb{E}_{\mathcal{D}_0}\left\{\ell_0\big(\boldsymbol{f}(\boldsymbol{X}_0), Y_0\big)\right\} \\ \text{subject to} & \mathbb{E}_{\mathcal{D}_i}\left\{\ell_i\big(\boldsymbol{f}(\boldsymbol{X}_i), Y_i\big)\right\} \leqslant c_i, \quad i \in \mathbb{N}_m^+ \end{array}. \tag{CL}$$

Variations of this problem have been heavily studied and utilized in the machine learning literature (e.g. [15]–[18]). By considering (CL), let us assume that $\mathcal{D}_{\boldsymbol{X}_0}$ is nonatomic (without loss of generality), and that $\mathcal{D}_{\boldsymbol{X}_0} \gg \mathcal{D}_{\boldsymbol{X}_i}$, that Slater's CQ holds, and that $\ell_i(\boldsymbol{f}(\bullet), \cdot) \in \mathcal{L}_1(\mathcal{D}_i, \mathbb{R})$ for any $\boldsymbol{f} \in \mathcal{F}$, where $\mathcal{F}$ is a decomposable functional space. Under this very general framework, we claim that the problem exhibits strong duality. In fact, we recover the results of [18, Proposition III.2 and Proposition B.1] (unifying the classification and regression regimes), while dispensing several additional assumptions.

Before we compare our result to that given in [18], let us first discuss its proof. Obviously, we could directly apply Theorem 3 (based on Theorem 1, the proof of which can be found in [29, Section 5]). While this would certainly be a possibility, the analysis in [29] utilizes Uhl's weak extension of A. A. Lyapunov convexity theorem (see [36, Theorem 1]), since the latter is applicable to a wide range of nonlinear functionals (and strictly more general than expectations). While this is a very general result, it requires that the underlying nonatomic measure is finite (thus not directly covering σ-finite measures). Nonetheless, using the developments of the previous section, we observe that problem (CL) can be equivalently written as

$$\begin{array}{ll} \underset{\boldsymbol{f}(\cdot)\in\mathcal{F}}{\text{minimize}} & \mathbb{E}_{\mathcal{D}_0}\left\{F_0\big(\boldsymbol{f}(\boldsymbol{X}_0), \boldsymbol{X}_0\big)\right\} \\ \text{subject to} & \mathbb{E}_{\mathcal{D}_0}\left\{G_i\big(\boldsymbol{f}(\boldsymbol{X}_0), \boldsymbol{X}_0\big)\right\} \leqslant c_i, i \in \mathbb{N}_m^+ \end{array}, \tag{CL0}$$

where $F_0(\boldsymbol{f}(\cdot), \cdot)$, $G_i(\boldsymbol{f}(\cdot), \cdot) \in \mathcal{L}_1(\mathcal{D}_{\boldsymbol{X}_0}, \mathbb{R})$, for all $i \in \mathbb{N}_m^+$. In then follows that a simple application of the standard A. A. Lyapunov's convexity theorem (see [22, Corollary IX.1.6]), which also supports σ-finite nonatomic measures (such as the Lebesgue measure), combined with an appropriate application of the supporting hyperplane theorem (e.g., see [37, Proposition 1.5.1]), would immediately yield the desired result, that is (CL0) (and thus (CL)) exhibits strong duality. Indeed, since we have already shown the equivalence between (CL) and (CL0), the strong duality result follows immediately by [32, Theorem 1] (see also prior developments in [30], [31]).

Let us now compare our (risk-neutral) strong duality result with that shown in [18]. The major difference lies in the way that the authors in [18] condition the problem. In particular, [18] relies on the nonatomicity of conditional distributions of the form $\mathcal{D}_{\boldsymbol{X}_i|Y_i}$, in contrast to our work, that requires nonatomicity of the marginal distributions $\mathcal{D}_{\boldsymbol{X}_i}$. We argue that our approach has significant benefits. Firstly, the model studied in [18] assumes that $\mathcal{F}$ is closed, convex, and decomposable. Instead, we have shown that only decomposability is required for showing strong duality of (CL). Secondly, our result is applicable to both classification and regression tasks, under the same set of (minimal) assumptions. In contrast, the aforementioned work separates these two cases. Specifically, the strong duality result for regression problems given in [18, Proposition B.1] is shown here to hold without the additional requirement postulated in [18, Assumption 6] (which, in essence, requires uniform continuity of $\ell_i(\boldsymbol{f}(\bullet), y)w_i(\cdot|y)$, for all $\boldsymbol{f} \in \mathcal{F}$, where $w_i(\cdot|Y_i)$ is the density induced by $\mathcal{D}_{\boldsymbol{X}_i|Y_i}$; note that this condition might be extremely difficult to verify and is quite restrictive). In fact, [18] makes some further implicit assumptions, including boundedness of the range of Y_i, for all $i \in \mathbb{N}_m$, as well as Lipschitz continuity of $\ell_i(\bullet, y)$ for all y in the range of Y_i (although the latter is also required to show some results that are not related to strong duality).

Apart from the benefits of the proposed approach, we should also mention a technical issue appearing in the analysis given in [18], while proposing a way to fix it. In particular, in the proof of [18, Lemma B.2], the last argument utilized by the authors to show that the corresponding "cost-constrained set" (denoted as $\mathcal{C}$ in [18]) is convex, is not accurate (to the best of our knowledge). Indeed, what the authors may show at most is that the *closure* of this set is convex, and thus their proof is incomplete. Nonetheless, this can be fixed by utilizing the developments in [29, Section 5.3], where it is shown that convexity of the closure of the "cost-constrained set" suffices to ensure strong duality of (CL). Of course, in our analysis, this step is completely bypassed.

V. Conclusion

In this paper, we established strong duality for a wide class of risk-constrained learning problems. Our results rely on a recent result relying on an application of Uhl's extension of Lyapunov's convexity theorem for general, infinite dimensional Banach spaces in the context of infinite-dimensional optimization, and are applicable to problems involving a wide range of risk measures, strictly generalizing existing results available for the risk-neutral constrained learning setting, and without imposing additional assumptions. We proposed a general risk-constrained functional learning framework involving nonconvex (possibly even discontinuous) losses and two-step compositional risk measures. The outer risk measures (evaluating feature risk) are assumed to be real-valued, convex, lower semicontinuous and positively homogeneous, with support over the space $\mathcal{L}_1$, while the inner (conditional) risk mappings (evaluating posterior risk) are significantly more general, and are even allowed to be nonconvex. In the special case of risk-neutral constrained learning, we unified existing results for constrained regression and constrained classification tasks, while dispensing several assumptions utilized in the current literature. Overall, we have presented new state-of-the-art strong duality relations for a rich risk-constrained learning framework, hopefully paving the way for a more widespread utilization of risk-constraints in this setting.

References

[1] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[2] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: a survey," *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, 2018.

[3] T. Samad and A. M. Annaswamy, "Controls for smart grids: Architectures and applications," *Proceedings of the IEEE*, vol. 105, no. 11, pp. 2244–2261, 2017.

[4] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke, "The limits and potentials of deep learning for robotics," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 405–420, 2018.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[6] M. Kay, C. Matuszek, and S. A. Munson, "Unequal representation and gender stereotypes in image search results for occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3819–3828.

[7] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, vol. 48, pp. 2990–2999.

[8] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Proceedings of the 2017 Conference on Learning Theory*, 2017, vol. 65, pp. 1920–1953.

[9] G. Goh, A. Cotter, M. Gupta, and M. P Friedlander, "Satisfying real-world goals with dataset constraints," in *Advances in Neural Information Processing Systems*, 2016, vol. 29.

[10] Y. Yang, C. Lin, X. Ji, Q. Tian, Q. Li, H. Yang, Z. Wang, and C. Shen, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[11] C. Cianfarani, A. N. Bhagoji, V. Sehwag, B. Zhao, H. Zheng, and P. Mittal, "Understanding robust learning through the lens of representation similarities," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34912–34925, 2022.

[12] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, vol. 70, pp. 22–31.

[13] S. Paternain, M. Calvo-Fullana, L. F. O. Chamon, and A. Ribeiro, "Learning safe policies via primal-dual methods," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 6491–6497.

[14] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. Van den Broeck, "A semantic loss function for deep learning with symbolic knowledge," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, vol. 80, pp. 5502–5511.

[15] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, "A sample average approximation method for chance constrained programming," *Journal of Optimization Theory and Applications*, vol. 142, pp. 399–416, 2009.

[16] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[17] A. Robey, L. F. O. Chamon, G. J. Pappas, H. Hassani, and A. Ribeiro, "Adversarial robustness with semi-infinite constrained learning," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 6198–6215.

[18] L. F. O. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro, "Constrained learning with non-convex losses," *IEEE Transactions on Information Theory*, vol. 69, pp. 1739–1760, 2023.

[19] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, vol. 97, pp. 7472–7482.

[20] I. Hounie, A. Ribeiro, and L. F. O. Chamon, "Resilient constrained learning," *arXiv:2306.02426v3*, 2023.

[21] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer New York, NY, 2000.

[22] J. Diestel and J. R. Uhl, *Vector Measures*, American Mathematical Society, Providence, math. surv edition, 1977.

[23] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.

[24] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *Journal of Machine Learning Research*, vol. 19, no. 13, pp. 1–48, 2018.

[25] R. Chen and I. C. Paschalidis, *Distributionally Robust Learning*, Foundations and Trends in Optimization, 2020.

[26] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, vol. 49, no. 3, pp. 1378–1406, 2021.

[27] R. Zhai, C. Dan, A. Suggala, J. Z. Kolter, and P. Ravikumar, "Boosted CVaR classification," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 21860–21871.

[28] S. Sagawa, P. Koh, T. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *International Conference on Learning Representations*, 2020.

[29] D. S. Kalogerias and S. Pougkakiotis, "Strong duality in risk-constrained nonconvex functional programming," *arXiv:2206.11948v3*, 2023.

[30] Z. Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 57–73, 2008.

[31] A. Ribeiro and G. B. Giannakis, "Separation principles in wireless networking," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4488–4505, 2010.

[32] A. Ribeiro, "Optimal resource allocation in wireless communication and networking," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, 2012.

[33] Andrzej Ruszczyński and Alexander Shapiro, "Conditional risk mappings," *Mathematics of Operations Research*, vol. 31, no. 3, pp. 544–561, 2006.

[34] F. Riedel, "Dynamic coherent risk measures," *Stochastic Processes and their Applications*, vol. 112, no. 2, pp. 185–200, 2004.

[35] D. S. Kalogerias and W. B. Powell, "Recursive optimization of convex risk measures: Mean-semideviation models," *arXiv:2206.11948*, 2018.

[36] J. J. Uhl, "The range of a vector-valued measure," *Proceedings of the American Mathematical Society*, vol. 23, no. 1, pp. 158–163, oct 1969.

[37] D. P. Bertsekas, *Convex optimization theory*, Athena Scientific, Nashua NH, US, 2009.