ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230287

Credal Learning: Weakly Supervised Learning from Credal Sets

Andrea Campagner [a;*]

[a]IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
ORCiD ID: Andrea Campagner https://orcid.org/0000-0002-0027-5157

Abstract. In this article we study the problem of credal learning, a general form of weakly supervised learning in which instances are associated with credal sets (i.e., closed, convex sets of probabilities), which are assumed to represent the partial knowledge of an annotating agent about the true conditional label distribution. A variety of algorithms have been proposed in this setting, chiefly among them the generalized risk minimization method, a class of algorithms that extend empirical risk minimization. Despite its popularity and promising empirical results, however, the theoretical properties of this algorithm (as well as of credal learning more in general) have not been previously studied. In this article we address this gap by studying the problem of credal learning from the learning-theoretic and complexity-theoretic perspectives. We provide, in particular, three main contributions: 1) we show that, under weak assumptions about the accuracy of the annotating agent, credal learning is learnable in the convex learning setting, providing effective risk bounds; 2) we study the properties of generalized risk minimization and, in particular, identify the optimal instance of this approach, that we call trade-off risk minimization; 3) we study the computational complexity of generalized risk minimization, showing effective algorithms based on gradient descent and providing sufficient and necessary conditions for them being computationally efficient.

1 Introduction

Credal sets [1, 20], that are convex and closed set of probabilities, are a general and widely applied model for uncertainty representation and management. They have attracted interest in both the theoretical and application-oriented literature due to their flexibility as well as for the rich connections with convex analysis [12], optimization [2, 35] and statistics [37]. Also in the context of machine learning (ML), credal sets and related models have recently attracted interest as a way to model weak supervision information in a variety of learning settings, including self-supervised learning [22, 21], learning from noisy data [23, 24], and learning from imprecise data [11, 15, 22], a general family of settings that encompasses, among others, semi-supervised learning, superset learning [17, 26] and fuzzy label learning [10, 32]. In all of these settings the idea is to model the weak supervision by means of credal sets, that are assumed to represent the partial or noisy information available to the annotating agent that produced the data: this general framework for studying weakly supervised learning is called *credal learning*.

Several algorithms have been proposed to tackle the credal learning problem, chiefly among them the *generalized risk minimization* (GRM) paradigm [15, 23]. The intuitive idea underlying GRM is to lift the popular empirical risk minimization (ERM) approach [29] (i.e., identifying the model within a given set of candidate ones that minimizes the value of a specified loss function) to the setting of credal set-valued labels by (implicitly) computing the value of any loss function w.r.t. all probability distributions in any given credal sets and then applying some aggregation operator to obtain a single value: typically, the minimum and maximum operators are used to this purpose, giving rise to the so-called *optimistic* and *pessimistic* risk minimization algorithms [16]. Due to its conceptual simplicity, and to the wide popularity of ERM in supervised learning, GRM has been successfully employed in several applications [6, 22, 24], with promising results. Nonetheless, the theoretical properties of this algorithmic approach, as well as of the credal learning problem more in general, have not been previously investigated. This gap regards not only the learning-theoretic properties of the above mentioned approaches, for which the generalization capacity has not been previously characterized except in some limited settings [4, 25], but also the complexity-theoretic one, in that the actual computational complexity of GRM is not yet generally well-understood [22].

In this work[1], we address this gap by studying the problem of credal learning from the perspective of statistical learning theory and theoretical computer science, focusing on the generalized risk minimization paradigm. To this aim, we show that: 1) the credal learning problem is learnable, i.e., we bound the generalization error for the problem and show that, under weak assumptions, these bounds can be estimated from finite samples; 2) we identify the worst-case optimal instance of generalized risk minimization (i.e., the one that achieves lowest generalization error), under general adversarial assumptions, that we call *trade–of risk minimization*; 3) we show that the above mentioned learnability guarantees can be met in computationally efficient manner, by studying conditions under which trade–off risk minimization can be efficiently computed, and showing learning algorithms that can efficiently solve the credal learning problem even when these assumptions do not hold.

2 Methods

We first provide an introduction to the theory of credal sets. Credal sets represent a generalization of probability theory, whereas in-

[*] Email: andrea.campagner@unimib.it

[1] For complete proofs, as well as additional material, we refer the reader to the technical appendix available at https://zenodo.org/record/8191602.

stead of quantifying the belief about a set of events (or propositions) in terms of a probability distribution, one can instead consider sets of such distributions. Formally, given a set A, we denote with $\Delta(A)$ the collection of probability densities on A, i.e. $\Delta(A) = \{p : A \to [0,1] : \int_A p = 1\}$. A *credal set* [1] is a convex, closed subset of $\Delta(A)$: that is, $C \subset \Delta(A)$ is a credal set if it is convex and, furthermore, for each convergent sequence $\{p_i\}_i$ of densities in C it holds that $\lim_{i\to\infty} p_i \in C$. In particular, a credal set is a *lower coherent prevision*[2] if there is a function $l : 2^A \to [0,1]$ s.t. $\forall S \subseteq A, l(S) = \inf_{p\in C} p(S)$ and $C = \{p \in \Delta(A) : \forall S \subseteq A, p(S) \geq l(S)\}$. Credal sets are a very general framework for representing uncertainty. While this provides a large amount of flexibility, for practical applications it may lead to intractable problems. For this reason, we also explicitly define some particularly relevant classes of credal sets (specifically, of lower coherent previsions), that are particularly common in applications (and will appear in the theoretical development in Section 2.3). Let $A = \{1,\dots,k\}$ be a finite set. We say that $C \subset \Delta(A)$ is a *probability interval* [8] if, for each $i \in A$ there exist two constants l_i, u_i, and $C = \{p \in \Delta(A) | l_i \leq p_i \leq u_i\}$. We say that C is a *possibility distribution* if, for each $i \in A$, there exists a constant π_i and $C = \{p \in \Delta(A) | p_i \leq \pi_i \wedge 1 - \max_{j\neq i} \pi_j \leq p_i\}$. We say that C is a *comparative probability assessment* [37] if there exists a directed acyclic graph $G = (A, E_G)$ s.t. $C = \{p \in \Delta(A) | \forall i,j \in A, p_i \leq p_j \text{ iff } (i,j) \in E_G\}$. Finally, we say that C is a *linear credal set* if $\exists f_1,\dots,f_k$ and $\exists c_1,\dots,c_k \in \mathbb{R}^k$, s.t. $\forall r \in \{1,\dots,k\}, f_r : \Delta(A) \to \mathbb{R}$, each f_r is linear and $C = \{p \in \Delta(A) | \forall r \in \{1,\dots,k\}, f_r(p) \leq c_r\}$. Intuitively, if we assume the credal sets represent the belief of an agent about some set of events, then, the classes of credal sets defined above correspond to constraints on the belief of the agent: the smaller the class of credal sets is, the more constrained the belief of the agent. Furthermore, as we show in Section 2.3, smaller classes of credal sets correspond to easier learning problems [27, 28, 30].

The theoretical development in the next sections is based on the *convex learning* paradigm [34], commonly adopted in statistical learning theory. To this aim, let X be a feature space and Y be the target space: we assume that Y is finite and discrete, while X is a convex set, i.e., $\forall x_1, x_2 \in X, \lambda \in [0,1]$ it holds that $\lambda x_1 + (1-\lambda)x_2 \in X$). In the general setting of *agnostic learning* we assume that instances $(x,y) \in X \times Y$ are generated by drawing from a data-generating distribution $\mathcal{D} \in \Delta(X \times Y)$: the value $\mathcal{D}(x,y)$ represents the probability (or density) of empirically observing the instance (x,y). Note that $\mathcal{D}$ is not necessarily deterministic: in particular, we denote with $\mathcal{D}(x)$ the conditional probability distribution $\mathcal{D}(x)(y) = \mathcal{D}(y|x)$. Let $\mathcal{H}$ be a class of functions $h : X \to \Delta(Y)$, that we call *hypotheses*: we assume $\mathcal{H}$ is a convex set. Intuitively, an hypothesis h associates with each instance x a probabilistic assessment $h(x) \in \Delta(Y)$, which could be understood as an approximation to the conditional distribution $\mathcal{D}(x)$. The quality of such an approximation is evaluated by means of a loss function $l : X \times \Delta(Y) \times \mathcal{H} \to \mathbb{R}$ defined point-wise by $(x,p,h) \mapsto g(p, h(x))$, where $g : \Delta(Y) \times \Delta(Y) \to \mathbb{R}$ is called the *base function* for l. We assume that l is B-bounded for some value B (i.e., $\forall x \in X, p \in \Delta(Y), h \in \mathcal{H}$ it holds that $l(x,p,h) \leq B$), is jointly convex in its second and third arguments (i.e., it is convex in the pair $(p,h) \in \Delta(Y) \times \mathcal{H}$), g is L-Lipschitz in its first and second argument w.r.t. to a given metric d (i.e. $|g(p_1,q) - g(p_2,q)| \leq Ld(p_1,p_2)$ and $|g(p,q_1) - g(p,q_2)| \leq Ld(q_1,q_2)$). We call d the *Lipschitz metric* for loss l, and we assume it is convex in both its arguments. These assumptions are needed to guarantee that the loss function is well-behaved: indeed, Lipschitz-ness implies that the loss function does not change too abruptly, boundedness implies that the function has a finite range of values, while convexity in the pair of arguments (p,h) is required to ensure that optimizing h w.r.t. the loss function is a well-defined and computationally feasible problem. Though we won't focus on any specific loss function in the following, we need to prove that such a loss function exists: the next two results show that two of the most commonly used loss functions (i.e., the Kullback-Leibler divergence and the l_2 loss) satisfy the above assumptions.

Proposition 2.1. *Let KL_ϵ be the Kullback-Leibler divergence defined on the space $\Delta(Y)_\epsilon \times \Delta(Y)_\epsilon$ of distributions s.t. $\forall y \in Y, p(y) \geq \epsilon$. Let l be the loss function defined by $l(x,p,h) = KL_\epsilon(p, h(x))$. Then l is convex in its second and third argument and KL_ϵ is $\log\left(\frac{1}{\epsilon}\right)$-Lipschitz w.r.t. to the l_1 metric on probability distributions. Furthermore, KL_ϵ is $\log\left(\frac{1}{\epsilon}\right)$-smooth and $\log\left(\frac{1}{\epsilon}\right)$-bounded.*

Proof. The KL divergence is convex in the pair (p,q), and when q_θ is a parametric distribution (with parameter θ) in the exponential family it is also convex w.r.t. the pair (p,θ). For Lipschitzness, assume as in the statement that $p_1, p_2, q \in \Delta(Y)_\epsilon$. Then, it holds that $KL_\epsilon(p_1,q) - KL_\epsilon(p_2,q) = p_1 \log(\frac{p_1}{q}) - p_2 \log(\frac{p_2}{q}) = (p1 - p2)\log(\frac{p_1}{q}) + p_2 \log(\frac{p_1}{p_2}) \leq |p_1 - p_2| \log\left(\frac{1}{\epsilon}\right)$. Similarly, KL_ϵ can be shown to be $\log\left(\frac{1}{\epsilon}\right)$-Lipschitz also in its second argument. It is also easy to see that KL_ϵ is $\log\left(\frac{1}{\epsilon}\right)$-smooth by applying the above reasoning to the derivative of KL_ϵ. Finally, it is easy to show that KL_ϵ is $\log\left(\frac{1}{\epsilon}\right)$-bounded. □

Proposition 2.2. *Let l be the loss function defined by $l(x,p,h) = \frac{1}{2}\sum_{y\in Y}(p(y) - h(x)_y)^2$. Then, l is convexin its second and third argument. Furthermore, assume that $\max_{x\in X} ||x|| \leq R$: then, $\frac{1}{2}\sum_{y\in Y}(p(y) - z_y)^2$ is $2R^2$-smooth and $2R^2$-Lipschitz w.r.t. to the l_2 metric. If $\mathcal{H}$ is B-bounded (i.e., $\max_{h\in\mathcal{H}} ||h|| \leq B$), then l is also BR^2-bounded.*

Proof. The l_2 loss is easily shown to be convex in the pair (p,q). Indeed, considering the summands $(p(y) - z_y)^2$, it holds that the Hessian of each summand is semi-definite positive and, hence, the summands are all convex functions: therefore, as the l_2 is a sum of convex function, it is also a convex function. For the parts on Lipschitzness, smoothness and boundedness, see e.g. [33]. □

In the following, we will generally denote a loss function with l and the corresponding base function and Lipschitz metric as, respectively, g and d. Note that we will always assume that the above mentioned assumptions hold true for l, g and d.

2.1 Credal Learning

As described in the Introduction, credal learning arises as a generalization of standard supervised learning (as well as of other weakly supervised learning tasks), in which we assume the data is sampled from a data generating distribution $\mathcal{D} \in \Delta(X, Y, \mathcal{C}(Y))$, where $\mathcal{C}(Y) \subseteq 2^{\Delta Y}$ is the collection of credal sets over Y (i.e., the collection of convex, closed sets of probability distributions over Y). Thus, an instance in credal learning is a triple (x, y, C): y represents the true label associated with x and is typically assumed to be *unobserved*; C instead represents the partial knowledge of the annotating

[2] Lower coherent previsions are of particular interest in the theory of imprecise probabilities as they represent the largest class of credal sets that can be interpreted as expressing the imprecise belief of a rational agent having a linear utility function [36].

agent about y, represented in the form of a credal set. Given $x \in X$, we denote with $\mathcal{D}(x)$ the conditional probability distribution of the true label, i.e., $\mathcal{D}(x) = \int_{C \in \mathcal{C}(Y)} \mathcal{D}(\cdot, C|x) d\mathcal{D} \in \Delta(Y)$. Intuitively, an instance (x, y, C) in a credal learning problem represents the information that the human labeller has an imprecise belief about $\mathcal{D}(x)$ which is represented as a credal set C: ideally, C will contain $\mathcal{D}(x)$ and will be as small (i.e., as precise) as possible. Formally:

Definition 2.1. *Let $\mathcal{D} \in \Delta(X \times Y \times \mathcal{C}(Y))$ be a data-generating distribution. The* degree of ambiguity *α of $\mathcal{D}$ is defined as $\mathbb{E}_{(x,y,\mathcal{C}(x))\sim\mathcal{D}}\left[diam_d(\mathcal{C}(x))\right]$.*

Thus, the degree of ambiguity α is defined as the expected size (measured as the d-diameter) of the credal set-valued annotations produced by the annotating agent: intuitively, the greater α the more the uncertainty of the annotating agent. In the extreme case where $\alpha = 1$ it holds, with probability 1, that $C(x) = \Delta(Y)$: hence, the annotating agent knows nothing about the true label y. In the other extreme case where $\alpha = 0$, it holds that, with probability 1, $|C(x)| = 1$ and hence the annotating agent always reports a single probability measure p_x over labels for each instance x. As mentioned above, while one desires the credal sets C to be as small as possible (i.e., α to be as close to 0 as possible), this is not enough: one also desires that C actually contains $\mathcal{D}(x)$ (that is, the imprecise belief of the human labeller is compatible with the actual truth) or, more generally, that there exists $p \in C$ which is as close as possible to $\mathcal{D}(x)$ (that is, C is a good approximation to the actual truth). However, this assumption may not hold true, and the annotating agent may make errors, in the sense that the credal set C, for a given instance (x, y, C) may not be compatible with the true conditional distribution of labels given x. We formalize this requirement, for a general distribution $\mathcal{D}$, through the following definition:

Definition 2.2. *The data generating distribution $\mathcal{D}$ is $(1-\eta)$-calibrated if $\mathbb{E}_{(x,y,\mathcal{C}(x))\sim\mathcal{D}}\left[d(\mathcal{D}(x), \mathcal{E}(x))\right] \leq \eta$, where $\mathcal{E}(x)$ is the d-projection of $\mathcal{D}(x)$ onto $C(x)$, i.e. $\mathcal{E}(x) = \min_{p \in C(x)} d(\mathcal{D}(x), p)$.*

Intuitively, $\mathcal{D}$ is $(1-\eta)$-calibrated if the annotations provided by the annotator agent are sufficiently close to the true label distribution: if, in particular, it is assumed that what the annotator agent is always compatible with the truth (i.e. $\mathcal{D}$ is 1-calibrated), then we simply say that $\mathcal{D}$ is calibrated. Though the notion of calibration may seem strong, we note that it is not. For example, calibration is trivially satisfied by any vacuous data generating distribution having $\alpha = 1$. More generally, calibration is typically assumed in several weakly supervised learning settings that arise as natural restrictions of credal learning. To this aim, we note that supervised learning, semi-supervised learning, superset learning, as well as noisy label learning, that have been mentioned in the Introduction, can all be formalized as special cases of credal learning:

- Supervised learning: $\forall x, y$ it holds that $D(\{p_y\}|x, y) = 1$, where p_y is the probability measure s.t. $p_y(y') = \mathbb{1}_{y=y'}$;
- Semi-supevised learning: $\forall x, y$ it holds that $D(A|x, y) > 0$ iff $A = \{p_y\}$ or $A = \mathcal{C}(Y)$;
- Superset learning [15, 25]: $\forall x, y$ it holds that $D(A|x, y) > 0$ iff $\exists y_1, \ldots, y_k \in Y$ s.t. $y \in \{y_1, \ldots, y_k\}$ and $A = \{\sum_i w_i p_{y_i} : w_i \in [0, 1], \sum_i w_i = 1\}$;
- Learning from fuzzy labels [4, 5, 15]: $\forall x, y$ it holds that $D(C|x, y) > 0$ iff C is a possibility distribution s.t. $\exists y' \in Y, \pi_{y'} = 1$ and $\pi_y > 0$;
- Noisy label learning: let $\mathcal{D}' \in \Delta(X, Y, Y)$ s.t. for each $x \in X$, $\mathcal{D}'(y \neq y'|x) \leq \eta$. Then, the problem of learning from $\mathcal{D}'$ can be equivalently formulated as the problem of learning from $\mathcal{D} \in \Delta(X, Y, \mathcal{C}(Y))$ with $\mathcal{D}(A|x, y) > 0$ iff $d(A, \mathcal{D}(x)) \leq \eta$.

More generally, different classes of credal sets (see Section 2) can be used to formalize different non-standard learning problems as instances of credal learning, as shown in the following examples.

Example 2.1. Differential diagnosis *refers to the diagnostic approach by which information about a patient is used to exclude or rank different diagnostic hypotheses about the patient. It is easy to see that the problem of learning from differential diagnosis can be formulated as a credal learning problem in which the features x represent the information about the patients and the corresponding credal sets C represent the belief of the doctor about the patients' health status: these credal sets can be expressed as a comparative probability assessment, in which the graph G is expressed over the possible set of diseases and an arc (c_i, c_j) exists among two conditions if c_i is considered to be less likely than c_j.*

Example 2.2. *An online betting site wants to quote the buying prices for a collection of gambles G defined on a set of events E: each gamble g pays 1\$ if a specific event $e_g \in E$ occurs, otherwise it pays 0\$. Each of the situations in which the events in E could occur can be described in terms of a feature vector x: to quote the buying prices, the betting site draws a collection of such situations x and asks an expert bettor to state, for each gamble g, its corresponding highest buying price $L(g)$ (i.e., the highest value $v \in [0, 1]$ s.t. the bettor would consider buying g a rational choice). If the expert bettor is rational, then the function $L : G \to [0, 1]$ that assigns to each gamble g the corresponding value $L(g)$ is a coherent lower prevision [36]. Therefore, the problem of learning from instances in the form $(x, L(g))$ can be formulated as a credal learning problem, in which the credal sets are constrained to be lower coherent previsions*[3].

From the above definitions, it is easy to see that in supervised, semi-supervised and superset learning the data generating distribution is always assumed to be calibrated: hence, the annotating agent is asumed to not commit any errors (but may have an arbitrary degree of ambiguity α). Indeed, for any instance (x, y, C), the singleton distribution p_y that assigns full probability to the ground truth label y is always guaranteed to belong to the credal set C. While this assumption is usually carried over also in the setting of credal learning, by contrast, in this article we won't make any such assumption, and instead allow arbitrary errors in the credal set labels. In this sense, the relaxed notion of $(1-\eta)$-calibration allows for some degree of error in the annotations and hence relaxes the above mentioned settings. Similarly, it is easy to see that in noisy label learning one generally considers a data generating distribution that is $(1-\eta)$-calibrated (with $\eta > 0$) but with zero ambiguity: this problem can be equivalently formulated as an instance of credal learning in which the data generating distribution is calibrated but has non-zero ambiguity degree $\alpha = \eta$. This transformation has been exploited in the literature to solve noisy label learning problem using the GRM method [9, 23, 24]. Thus, credal learning is a general learning setting that bridges between noisy label learning (in which there may be errors but no ambiguity) and imprecise label learning (where, conversely,

[3] The learning problem described in Example 2.2 is extremely general. Let X be any feature space, and Y be a target space. If we identify, the collection of events E with Y, the betting site with a data-generating distribution $\mathcal{D} \in \Delta(X \times Y)$, and define for each $(x, y) \sim \mathcal{D}$ the lower prevision $L_x : Y \to [0, 1]$ as in Example 2.2, then any credal learning problem in which the credal sets are constrained to be lower coherent previsions can be formulated as an instance of the learning problem in Example 2.2.

there may be some ambiguity but no labeling errors). Furthermore, even though the above mentioned learning problems need not be formulated as credal learning problems, we will show in Section 2.3 that doing so entails several advantages.

The problem of credal learning amounts to the problem of, given a class of hypotheses $\mathcal{H}$, finding a h with small true error, i.e. s.t.:

$$L_{\mathcal{D}}(h) := \int_{X\times Y\times \mathcal{C}(Y)} l(x,y,h)d\mathcal{D}$$

is small, where, as above, l is a loss function over $X \times \Delta(Y) \times \mathcal{H}$. Obviously, since the data generating distribution $\mathcal{D}$ is not known, this usually reduces to the problem of minimizing some approximation of the true error based on a finite sample $\{(x_i, C_i)\}_{i=1}^m$ drawn from the marginal of $\mathcal{D}$ on $X \times \mathcal{C}(Y)$. Among such approaches, we focus, in particular, on the framework of *generalized risk minimization* (GRM). This is a family of approaches that aim to extend empirical risk minimization to the setting of weakly supervised learning: in the context of credal learning, it has been proposed and studied in [22], based on previous work in superset learning [7, 14, 15, 16].

Generalized risk minimization is based around the idea of using an aggregation operator (see below) to extend empirical risk minimization to more general settings. Formally, let $\mathcal{C}$ be the class of compact, convex sets defined on $\mathbb{R}$. A (generalized) *aggregation operator* [13] is a monotonic function $A : \mathcal{C} \to \mathbb{R}$ satisfying:

1. $\forall S \in \mathcal{C}$ it holds that $\inf S \leq A(S) \leq \sup S$;
2. $\forall v \in \mathbb{R}$, $A(\{v\}) = v$.

The aggregation operator A can be used to *lift* any given loss function $l : X \times \Delta(Y) \times \mathcal{H} \to \mathbb{R}$ to a generalized loss function $l_A^{IP} : X \times \mathcal{C}(Y) \times \mathcal{H} \to \mathbb{R}$, defined as

$$l_A^{IP}(x, C, h) := A(\{l(x, p, h)\}_{p\in C}).$$

The GRM algorithm, for hypothesis class $\mathcal{H}$ and a training set $S = \{(x_i, C_i)\}_{i=1}^m \sim \mathcal{D}^m$ is defined through the following minimization problem:

$$\begin{aligned} GRM_{l,A}(\mathcal{H}, S) &:= \arg\min_{h\in\mathcal{H}} L_S^A(h) \\ &= \arg\min_{h\in\mathcal{H}} \sum_{i=1}^n l_A^{IP}(x_i, C_i, h), \qquad (1) \end{aligned}$$

that is, we simply search the space of hypotheses in order to find one that minimizes the *imprecise empirical risk*, L_S^A.

In the following, we will require the technical condition that $\forall C \in \mathcal{C}(Y)$, $\forall A$ aggregation operator, it exists $p^* \in C$ s.t. $l_A^{IP}(x, C, h) = l(x, p^*, h)$, that is, the value of l_A^{IP} on a credal set is attained at (at least) one point of the credal set itself. Under the weak assumption that the credal sets C are compact and that A is an aggregation operators A s.t. generalized loss function l_A^{IP} is *convex* the above condition is always satisfied. Under the assumptions defined in the previous section, it is an easy consequence of basic facts from convex analysis [2] that this holds in particular when $A \in \{\min, \max\}$[4]. These two instances of the generalized risk minimization, in particular, have been widely studied in the literature under the names of, respectively, *optimistic* [15] and *pessimistic* [14] risk minimization (abbreviated, respectively, as ORM and PRM), and are the most commonly adopted instances of the framework [16].

Finally, we will also introduce the two following definitions, the *true empirical risk* w.r.t. to the unknown ground truth:

$$L_S(h) = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^n l(x_i, y_i, h),$$

and the *annotator-relative empirical risk*:

$$L_S^E(h) = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^n l(x_i, \mathcal{E}(x), h),$$

where, as before, $\mathcal{E}(x) = \arg\min_{p\in C_i} d(\mathcal{D}(x), p)$. Intuitively, the true empirical risk L_S represents the error that a given hypothesis h makes w.r.t. the unknown, true label; by contrast, the annotator-relative empirical risk L_S^E represents the error that h makes in comparison with the probability distribution, compatible with the credal set annotation C given by the annotator agent, that is closest to the unknown, true one.

2.2 Learning-Theoretic Properties

Having cleared the definitions of credal learning and of the GRM approach, in this section we study the learning-theoretic properties of GRM. In particular, we will be interested in quantifying the *generalization error* of GRM, that is, in bounding the quantity:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[|L_{\mathcal{D}}(GRM(\mathcal{H}, S)) - L_S^A(GRM(\mathcal{H}, S))|\right] \qquad (2)$$

in terms of the sample size m, of some property of $\mathcal{H}$, as well as of the parameters α, ϵ of the data generating distribution $\mathcal{D}$. Intuitively, such a bound would provide an estimate of the *excess risk* of any hypothesis h found by applying GRM and hence, having computed $L_S^A(h)$, allows to upper bound the true risk of h.

Through the following result we provide two such generalization bounds. We first present a generalization bound that is algorithm independent (hence, it applies to any learning algorithm, not only to GRM), and characterizes the generalization error of a class of models for credal learning as a function of the degree of ambiguity α and calibration $1 - \eta$ of the annotator. Then, we also provide a bound that is specific for GRM and only depends on the calibration $1 - \eta$ and the expected deviation between two different settings of the imprecise empirical risk. In particular, we first present a bound on the expected values, from which we derive a finite sample tail bound:

Theorem 2.1. *Let $\mathcal{H}$ be a convex hypothesis space, l a convex, L-Lipschitz loss function. Let $\mathcal{R}(\mathcal{H}, m, l)$ be the expected Rademacher complexity of $\mathcal{H}$ w.r.t. the true empirical risk L_S. Assume also that $\mathcal{D}$ is $(1-\eta)$- calibrated and has degree of ambiguity α. Then, uniformly over $h \in \mathcal{H}$, it jointly holds that $\mathbb{E}_{S\sim\mathcal{D}^m}\left[|L_{\mathcal{D}}(h) - L_S^A(h)|\right]$ can be upper bounded by:*

$$2\mathcal{R}(\mathcal{H}, m, l) + L(\eta + \alpha), \qquad (3)$$

$$2\mathcal{R}(\mathcal{H}, m, l) + L\eta + \mathbb{E}\left[|l_A^{IP}(x, C, h) - l_{A^*}^{IP}(x, C, h)|\right], \qquad (4)$$

where A^* is defined as

$$A^* = \arg \sup_{A' \text{ aggregation operator}} |l_A^{IP}(x, C, h) - l_{A'}^{IP}(x, C, h)|.$$

Proof Sketch. The results follows from the fact that the expected generalization error $\mathbb{E}_{S\sim\mathcal{D}^m}\left[|L_D(h) - L_S^A(h)|\right]$ of h can be upper bounded by the sum of three terms, namely

[4] Indeed, max is generally convex. By contrast, even though min is not generally convex in its domain, under the assumptions in Section 2 (i.e., l is convex in the pair (p, h)), minimization over the credal set C can be interpreted as a convex projection operation.

$\mathbb{E}_{S\sim\mathcal{D}^m}\left[|L_D(h) - L_S(h)|\right]$, $\mathbb{E}_{S\sim\mathcal{D}^m}\left[|L_S(h) - L_S^E(h)|\right]$, $\mathbb{E}_{S\sim\mathcal{D}^m}\left[|L_S^E(h) - L_S^A(h)|\right]$, and then bounding the three terms above. The first term can be bounded through expected Rademacher complexity of $\mathcal{H}$, while the second term can be bounded by noting that the data-generating $\mathcal{D}$ is $(1-\eta)$-calibrated. Then, for the bound in Eq. 3, we can upper bound the third term by noting that $\mathcal{D}$ has degree of ambiguity α. By contrast, for the bound in Eq. 4 we note that the third term can be bounded by noting that $|L_S^E(h) - L_S^A(h)| \leq |L_S^{A*}(h) - L_S^A(h)|$. □

Corollary 2.1. *Let $\rho_A(x,C,h) = |l_A^{IP}(x,C,h) - l_{A^*}^{IP}(x,C,h)|$. Assume that there exists η' such that, for all $\eta \leq \eta'$, the data-generating distribution $\mathcal{D}$ is $(1-\eta)$-calibrated. Then, under the same conditions of Theorem 2.1, and further assuming that l is B-bounded (see Section 2), with probability greater than $1-\delta$ over the sampling of $S \sim \mathcal{D}^m$ and uniformly over $h \in \mathcal{H}$, it holds that $L_D(h) - L_S^A(h)$ can be jointly upper bounded by:*

$$\Psi_{\delta,l}(\mathcal{H},S) + L\left(\eta' + \frac{1}{m}\sum_{(x,C)\in S} diam_d(C)\right) \tag{5}$$

$$\Psi_{\delta,l}(\mathcal{H},S) + L\eta' + \frac{1}{m}\sum_{(x,C)\in S}\rho_A(x,C,h), \tag{6}$$

where $\Psi_{\delta,l}(\mathcal{H},S) = 2R(\mathcal{H},S,l) + 6B\sqrt{\frac{2\log(\frac{2}{\delta})}{2m}}$. In particular, let $h = GRM_{l,A}(\mathcal{H},S)$, and let T be the aggregation operator s.t. $T = \frac{\min + \max}{2}$. Let $h^* = \arg\min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h')$. Then:

- *If $A \leq T$[5], then $\rho_A(x,C,h) = \max_{p\in C} l(x,p,h) - l_A^{IP}(x,p,h)$;*
- *If $A \geq T$ then $\rho_A(x,C,h) = l_A^{IP}(x,p,h) - \min_{p\in C} l(x,p,h)$;*

Furthermore, with probability $1-\delta$, $L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*)$ can be upper bounded by:

$$2R(\mathcal{H},S,l) + L\eta' + \frac{1}{m}\sum_{(x,C)\in S}\rho_A(x,C,h) + 7B\sqrt{\frac{2\log(\frac{2}{\delta})}{2m}} \tag{7}$$

Proof Sketch. The first three bounds follow from applying McDiarmid's inequality to Equations 3 and 4. The result on ρ_A directly stems from its definition and the monotonicity of A. □

We can make some observations about the previous results. First, we note that, conditional on the calibration error η and ambiguity degree α not being too large, Theorem 2.1 shows that the credal learning task is learnable: this implies that, if the information available is sufficiently accurate (i.e., the data-generating distribution is close to being calibrated) and sufficiently unambiguous (i.e., the ambiguity degree is small), the hypothesis given as output by a learning algorithm for solving the credal learning task would have a generalization gap close to the one that could be obtained from completely supervised data. Obviously, in general, the generalization gap for a credal learning task upper bounds the gap for the corresponding supervised learning task: this is to be expected, as the lack of information that is implicit in credal learning will in general make the learning problem harder. Also, we can observe that the upper bounds shown in Theorem 2.1 are sharp, in the sense that there exist learning problems for which the generalization gap almost matches the mentioned bounds: as an example, take the case where, given $(x, \mathcal{D}(x))$ and an aggregation operator A, an adversary explicitly constructs a credal set C_x s.t. $d(\mathcal{E}(x), \mathcal{D}(x)) = \eta$ and such that $\inf_{h\in\mathcal{H}} |l(x,\mathcal{E}(x),h) - l_A^{IP}(x,C_x,h)| = \inf_{h\in\mathcal{H}} |l_{A^*}^{IP}(x,C_x,h) - l_A^{IP}(x,C_x,h)| = \alpha$. As a second point, we note that, even though Theorem 2.1 refers to some parameters of the data-generating distribution $\mathcal{D}$ (namely, the calibration error η and the ambiguity α), the bounds in Corollary 2.1 only refer to a fixed constant η' and on quantities that can (in principle) be computed based only on the given finite sample of data S: this implies that, if one knows an upper bound η' on the calibration error η, the bounds given in Theorem 2.1 provide a way to upper bound the true error $L_D(GRM_{l,A}(\mathcal{H},S))$ of $h_S^A = GRM_{l,A}(\mathcal{H},S)$ as:

$$L_S^A(h_S^A) + \frac{1}{|S|}\sum_{(x,C)\in S}\rho_A(x,C,h_S^A) + L\eta' + \Psi_{\delta,l}(\mathcal{H},S) \tag{8}$$

From a practical point of view, we note, however, that the above bound requires that the Rademacher complexity $R(\mathcal{H},S,L)$, as well as the quantity ρ_A, can be computed efficiently. Unfortunately, it is believed that, for general hypotheses classes, computing the Rademacher complexity (or even approximating it) is an NP-HARD problem [19]. By contrast, in the next section we will study the complexity of computing ρ_A and show conditions under which this quantity can be computed efficiently. Before getting to the complexity-theoretic analysis of GRM, however, we conclude with a final observation about the previous results. One key issue in the implementation of GRM is the specification of the aggregation operator A to be used for evaluating hypotheses in $\mathcal{H}$: indeed, multiple such aggregation rules have been proposed in the literature and, as a simple consequence of Corollary 2.1, the selection of A may strongly influence the generalization error of GRM. Since Eq. (8) provides an upper bound for the true error of GRM, a sensible choice is to select A so as to minimize this bound. The following result identifies such an instantation of GRM and, surprisingly, shows that this optimal instantiation of GRM differs from those commonly adopted in the weakly supervised learning literature (namely, ORM and PRM).

Theorem 2.2. *Let $\rho_A(x,C,h)$ and T be defined as in Corollary 2.1. Let $U(h,A) = L_S^A(h) + \frac{1}{|S|}\sum_{(x,C)\in S}\rho_A(x,C,h)$. Let $\mathcal{A}_T = \{A \text{ aggregation operator} \,|\, A = T \vee A \geq T \vee A \leq T\}$. Then $T \in \arg\min_{A\in\mathcal{A}}\rho_A(x,C,h)$ and $U(h,T) = \frac{L_S^{min}(h) + L_S^{max}(h)}{2}$. Furthermore:*

- *If $A \leq T$, then $U(h,A) = L_S^{max}(h) \geq U(h,T)$;*
- *If $A > T$, then $U(h,A) = 2L_S^A(h) - L_S^{min}(h) \geq L_S^{max}(h) \geq U(h,T)$.*

Thus, $T \in \arg\min_{A\in\mathcal{A}} U(h,A)$.

Proof Sketch. The results is a consequence of Corollary 2.1. □

If we denote the instantiation of GRM based on the aggregation operator T as *trade-off risk minimization* (TRM), then Theorem 2.2 states that for all fixed hypotheses in $\mathcal{H}$, TRM minimizes the upper bound on the true risk given by $U(h,A)$, across a large class of aggregation operators and *uniformly* over $\mathcal{H}$. We remark, however, that the Theorem does not imply that $U(GRM_{l,T}(\mathcal{H},S)) \in \arg\min_{A\in\mathcal{A}} U(GRM_{l,A}(\mathcal{H},S))$: indeed, in general there may be a complex interplay among the optimization process (implicit in the definition of GRM) and the value of the upper bound U. Notably, however, TRM still offers two remarkable advantages. First, as mentioned above, TRM is guaranteed to minimize the upper bound U uniformly across all hypotheses: as the aggregation operator for GRM has to be selected a-priori (to enable optimization

[5] Let $A_1, A_2 \in \mathcal{A}$. $A_1 \leq A_2$ if $A_1(I) \leq A_2(I)\ \forall I \in \mathbb{I}_B$.

of the empirical risk), uniform minimization provides a useful criterion for this selection. Second, and most relevant, we note that $L_S^T(h) = U(h, T)$, and T is the only aggregation operator in $\mathcal{A}$ for which this property holds: this means that TRM is the only instantiation of GRM for which minimizing the empirical risk L_S^T jointly minimizes also the upper bound on the true risk given by $U(h, T)$.

2.3 Complexity-Theoretic Properties

Based on the results obtained for GRM in the previous section, in this section we study the computational complexity of GRM. Our results, in particular, show that GRM is not only of theoretical, but also practical interest, as it provides computationally efficient algorithms to solve large classes of credal learning problems. In the rest of this section we will require an additional assumption on the hypothesis space $\mathcal{H}$, namely, that it is a convex set in a Reproducing Kernel Hilbert Space (RKHS) of functions. This assumption is not too strong: indeed, many standard ML approaches (e.g., kernel methods) satisfy the above mentioned assumption. We denote with $K_{\mathcal{H}} : X \times X \rightarrow \mathbb{R}^Y$ the kernel function associated with $\mathcal{H}$.

First of all, since the TRM algorithm can be expressed as $\arg\min_{h \in \mathcal{H}} \frac{L_S^{\max}(h) + L_S^{\min}(h)}{2}$, we note that the TRM problem can be reduced to ORM and PRM problems[6]. Thus, we study the complexity of these two latter approaches. In particular, we will focus on learning algorithms that are based on the *stochastic (sub)gradient descent* approach, due to their popularity in modern machine learning.We note that, since the loss function l is jointly convex in its second and third arguments, then its subgradient set is non-empty: we denote the subgradient set of l as ∂l, and with $\partial_p l, \partial_h l$ the subgradient sets w.r.t. to its second (resp., third) argument.

For the case of ORM, the following result shows that a simple[7] stochastic coordinate descent algorithm (see Algorithm 1) is able to approximately compute the GRM hypothesis:

Theorem 2.3. *Let $\mathcal{H}$ be a convex subset of a vector-valued Reproducing Hilbert Kernel Space (RKHS). Let l be a loss function satisfying the assumptions stated in Section 2. Assume that* $\max_{h \in \mathcal{H}} ||h|| \leq B$. *Then, for any $S \sim \mathcal{D}^m$, if Algorithm 1 is executed for T steps, returning hypothesis $h = \sum_{i=1}^m \alpha_i K_{\mathcal{H}}(\cdot, x_i)$, it holds that, with probability greater than $1 - \delta$ over the randomization in the Algorithm:*

$$L_S^{\min}(h) - L_S^A(GRM_{l,\min}(\mathcal{H}, S)) \leq \frac{4B^2 L}{\delta \sqrt{T}}. \tag{9}$$

Thus, if T_G denotes the time required to evaluate a subgradient of l, and $T_P(|Y|)$ denotes the time required to compute the d-projection at line 8 of Algorithm 1, then the ORM problem can be ϵ-approximated (with probability of error smaller than δ) within time complexity $O\left(\frac{B^4 L^2}{\delta^2 \epsilon^2}\left[T_G + T_P(|Y|)\right]\right)$. In particular, if, with probability 1 over $(x, y, C) \sim \mathcal{D}$ it holds that C is a probability interval (resp., a possibility distribution, a comparative probability, a linear credal set), then the ORM problem can be ϵ-approximated (with probability of error smaller than δ) within time complexity $O\left(poly(\frac{1}{\epsilon}, \frac{1}{\delta}, |Y|)\right)$.

Proof Sketch. The main result follows from an analysis of the block-coordinate descent strategy adopted in Algorithm 1. Indeed, since l is convex in the pair $(p, h) \in \Delta(Y) \times \mathcal{H}$, then the functions f, g defined pointwise by $f_C(h) = \min_{p \in C} l(x, p, h), g_h(p) = \min_{h \in \mathcal{H}} l(x, p, h)$ are both convex: hence, the alternating minimization of g_h and f_C will converge to the optimal solution for $l_{\min}^{IP}$ [39]. The other results follow by noting that the running time of Algorithm 1 scales linearly in the number of iterations, and each iteration takes time $T_G + T_P(|Y|)$, whereas $T_P(|Y|)$ is polynomial in Y [3]. □

Algorithm 1 Stochastic Gradient Descent procedure for Optimistic Risk Minimization

```
 1: procedure SGD-OPTIMISTIC-GRM($\mu$: learning rate, $T$: number of iterations,
    $S$: training set)
 2:     $\alpha^{(1)} \leftarrow \mathbf{0} \in \mathbb{R}^{|S|}$
 3:     $P^{(1)} \leftarrow \mathbf{0} \in \mathbb{R}^{|S| \times |Y|}$
 4:     for $t = 1, \ldots, T$ do
 5:         Select $(x_i, C_i)$ uniformly from $S$
 6:         $h^{(t)} \leftarrow \sum_{j=1}^{|S|} \alpha_j^{(t)} K_{\mathcal{H}}(\cdot, x_j)$
 7:         Let $v_p^{(t)} \in \partial_p g(P^{(t)}[i, :], h^{(t)}(x_i))$
 8:         $P^{(t+\frac{1}{2})}[i, :] \leftarrow P^{(t)}[i, :] - \mu v_p^{(t)}$
 9:         $P^{(t+1)}[i, :] \leftarrow \arg\min_{p \in C_i} d(P^{(t+\frac{1}{2})}[i, :], p)$
10:         Let $v_h^{(t)} \in \partial_h l(x_i, P^{(t+1)}[i, :], h^{(t)})$
11:         $\alpha^{(t+1)} \leftarrow \alpha^{(t)} - \mu v_h^{(t)}$
12:     end for
13:     return $\frac{1}{T} \sum_{t=1}^{T} \alpha^{(t)}$
14: end procedure
```

Thus, the previous theorem shows that, if one adopts the ORM algorithm, the credal learning problem is not only (conditionally) learnable, but also efficiently so: this holds, in particular, when the class of possible credal sets is restricted to one of the classes defined in Section 2. Interestingly, most practical applications of the ORM algorithm [6, 15, 22] have indeed considered such restricted classes of credal sets (in particular, we refer to the problem of learning from fuzzy labels and the superset learning problem, see Section 2): thus, the above result applies, as a special case, to the above mentioned applications, showing that the problems therein studied, when reformulated as credal learning problems, can be solved efficiently. This results is also of practical interest. Indeed, while it is known that ORM for the former learning problems is NP-HARD [5], Theorem 2.3, by contrast, shows if we relax the above mentioned problems as credal learning problems, then these latter can be solved efficiently (see also Corollary A.1 in the Appendix).

Next, we study the complexity of PRM (i.e., GRM with $A = \max$). Through the following result, we show that, in general, the PRM problem cannot be solved efficiently: nonetheless, we propose a SGD-style algorithm, and provide conditions for its efficient convergence.

Theorem 2.4. *Let $\mathcal{H}, l, K_{\mathcal{H}}$ be as in Theorem 2.3. Then, for any $S \sim \mathcal{D}^m$, if Algorithm 1 is executed for T steps, returning hypothesis $h = \sum_{i=1}^m \alpha_i K_{\mathcal{H}}(\cdot, x_i)$, it holds that, with probability greater than $1 - \delta$ over the randomization in the Algorithm:*

$$L_S^{\max}(h) - L_S^{\max}(GRM_{l,\max}(\mathcal{H}, S)) \leq \frac{BL}{\delta \sqrt{T}}. \tag{10}$$

In particular, the problem of solving PRM cannot be solved efficiently for arbitrary credal sets.

Proof Sketch. The proof for the first statement can be derived similarly to the proof of Theorem 2.3, as any convex maximization problem can be reduced to extrema enumeration on the supporting convex set [31]. For the second statement, it suffices to note that the problem of maximizing a convex function (i.e., the problem of computing

[6] Indeed, since the loss function for TRM is defined as the average of those for ORM and PRM, it follows that a subgradient for the former can directly be obtained from subgradients for the latter.

[7] Algorithm 1 can be easily implemented using any numerical computing library (e.g., TensorFlow), as it only involves standard convex optimization routines such as the computation of sub-gradients or convex projections.

$l^{IP}_{\max}(x, C, h)$, for fixed h) cannot be solved efficiently for arbitrary convex sets C, in particular it is NP-HARD [40]. □

Corollary 2.2. *Let $m \in \mathbb{N}$ be a training set size and $S \sim \mathcal{D}^m$. Define $Y_S = \{y \in Y : (x, y, C) \in S\}$. Assume that there exists $c \in \mathbb{N}$ s.t., with probability 1 over the sampling of a training set S, $|Y_S|! \leq O\left(|S|^c\right)$. Assume that there exists a* polynomial delay *[18] algorithm for enumerating the extreme points of C, where $(x, y, C) \sim \mathcal{D}$. Assume, further, that, with probability 1 over the sampling of $(x, y, C) \sim \mathcal{D}$, C is a lower coherent prevision. Then Algorithm 2 ϵ-approximates PRM (with probability of error smaller than δ) within time complexity $O\left(poly(\frac{1}{\epsilon}, \frac{1}{\delta}, |S||Y|\right)$.*

Proof. We first note that an arbitrary lower coherent prevision[8] C over Y has at most $|Y|!$ extreme points [38]. Therefore, if it exists a polynomial delay algorithm for enumerating the extreme points of C, the maximum time complexity of the main for loop in Algorithm 2 is $O\left(poly(|S||Y|)\right)$. The result follows, since at most $poly(\frac{1}{\epsilon}, \frac{1}{\delta})$ iterations are needed to ϵ-approximate PRM. □

Algorithm 2 Stochastic Gradient Descent procedure for Pessimistic Risk Minimization

```
procedure SGD-PESSIMISTIC-GRM($\mu$: learning rate, $T$: number of iterations, $S$: training set)
    $\alpha^{(1)} \leftarrow \mathbf{0} \in \mathbb{R}^{|S|}$
    $P^{(1)} \leftarrow \mathbf{0} \in \mathbb{R}^{|S| \times |Y|}$
    for $t = 1, \dots, T$ do
        Select $(x_i, C_i)$ uniformly from $S$
        $h^{(t)} \leftarrow \sum_{j=1}^{|S|} \alpha_j^{(t)} K_{\mathcal{H}}(\cdot, x_j)$
        Let $p_1, \dots, p_r$ be the extremes of $C_i$
        $k \leftarrow \arg\max_{j \in \{1, \dots, r\}} g(p_j, h^{(t)}(x_i))$
        Let $v_h^{(t)} \in \partial_h l(x_i, p_k, h^{(t)})$
        $\alpha^{(t+1)} \leftarrow \alpha^{(t)} - \mu v_h^{(t)}$
    end for
    return $\frac{1}{T} \sum_{t=1}^{T} \alpha^{(t)}$
end procedure
```

Theorem 2.4 and Corollary 2.2 (see also the Appendix for additional results that apply to specific classes of credal sets) show that, even if in general the problem of approximating PRM is NP-HARD, in several relevant cases it can be solved efficiently through Algorithm 2, which is a simple variant[9] of stochastic (sub)gradient descent combined with an explicit extrema enumeration algorithm. By combining Algorithms 1 and 2 it is easy to illustrate a general algorithm for TRM, as shown in Algorithm 3: it is similarly easy to see that such an algorithm has the same computational complexity (asymptotically) as Algorithm 2 and that Corollary 2.2 applies equivalently also for TRM. Thus, even though in Section 2.2 we showed that TRM enjoys favorable generalization bounds as compared with other approaches based on GRM (such as ORM and PRM), by contrast, it may be computationally harder to train a ML model through TRM, in general cases. Indeed, while Algorithm 3 is guaranteed to efficiently ϵ-approximate the theoretical optimal hypothesis found by TRM for several relevant classes of credal sets, in general its running time may be exponential. By contrast, ORM can always be approximated in polynomial time, an insight that may explain the popularity of ORM in practical applications [22, 24]. In light of this computational advantage of ORM in comparison with TRM, it is of theoretical and practical interest to understand better the behaviour of the former approach, especially as it regards its possible use as an approximation to TRM (as, intuitively, TRM can be expressed as the combination of ORM and PRM): we leave this question as future work. Nonetheless, as a consequence of Corollary 2.2, it is easy to note that TRM can be applied to efficiently solve the credal learning problem whenever the size of the target space Y is not too large and the credal sets are constrained to lower coherent previsions, while enjoying the theoretical guarantees on the generalization error studied in Section 2.2.

Algorithm 3 Stochastic Gradient Descent procedure for Trade-Off Risk Minimization

```
procedure SGD-TRADEOFF-GRM($\mu$: learning rate, $T$: number of iterations, $S$: training set, $\epsilon$: tolerance)
    $\alpha^{(1)} \leftarrow \mathbf{0} \in \mathbb{R}^{|S|}$
    $P^{(1)} \leftarrow \mathbf{0} \in \mathbb{R}^{|S| \times |Y|}$
    for $t = 1, \dots, T$ do
        Select $(x_i, C_i)$ uniformly from $S$
        $h^{(t)} \leftarrow \sum_{j=1}^{|S|} \alpha_j^{(t)} K_{\mathcal{H}}(\cdot, x_j)$
        Let $p_*$ s.t. $l(x_i, p_*, h^{(t)}) - \min_{p \in C_i} l(x_i, p, h^{(t)}) \leq \epsilon$
        Let $v_{\min}^{(t)} \in \partial_h l(x_i, p_*, h^{(t)})$
        Let $p_1, \dots, p_r$ be the extremes of $C_i$
        $k \leftarrow \arg\max_{j \in \{1, \dots, r\}} g(p_j, h^{(t)}(x_i))$
        Let $v_{\max}^{(t)} \in \partial_h l(x_i, p_k, h^{(t)})$
        $\alpha^{(t+1)} \leftarrow \alpha^{(t)} - \mu \frac{v_{\min}^{(t)} + v_{\max}^{(t)}}{2}$
    end for
    return $\frac{1}{T} \sum_{t=1}^{T} \alpha^{(t)}$
end procedure
```

3 Conclusion

In this article we studied the problem of credal learning, a flexible and increasingly popular weakly supervised paradigm.We focused, in particular, on analyzing the theoretical properties of one of the most commonly adopted algorithmic methods in this setting, namely GRM.After providing generalization bounds for credal learning (and GRM in particular), we proposed a novel approach based on the GRM paradigm called trade-off risk minimization (TRM), and showed its desirable properties from the learning-theoretic point of view. We then proposed stochastic gradient descent algorithms for TRM and the two most common varieties of GRM proposed in the literature (namely, ORM and PRM) and showed that, despite the above mentioned positive results, solving the TRM problem is in general computationally hard, while we showed that, by contrast, ORM (the most popular algorithm based on GRM) can be solved efficiently. We also highlighted the advantage of credal learning in comparison with alternative weakly supervised learning paradigms [5, 15]. In light of the flexibility and rising popularity of credal learning, we believe that our work could pave the way for further exploration of this setting, from both the theoretical and practical point of views. In particular, we believe the following open problems to be of particular interest: 1) In Section 2.2 we provided upper bounds on the generalization error for credal learning: finding matching lower bounds could be useful to characterize the intrinstic complexity and resource-bounds for this problem; 2) In Section 2.3 we showed that, despite its intuitively appealing learning-theoretic properties, learning through TRM is in general NP-HARD, while, by contrast, ORM is computationally easy: it would be interesting to better characterize the properties of this latter approach, in particular in regards to its ability to approximate TRM; 3) In this article we focused on the theoretical side of credal learning, future work should analyze the empirical effectiveness of algorithms for this setting, focusing in particular on comparing TRM with other variants of GRM.

[8] We note that the mentioned property also holds for other specific classes of credal sets not considered in this paper, e.g. for belief functions [30].

[9] Indeed, Algorithm 2 combines a vertex enumeration step (for which efficient implementations exists, see e.g. https://pypi.org/project/pypoman/) with a standard application of stochastic gradient descent.

References

[1] Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes, *Introduction to imprecise probabilities*, John Wiley & Sons, 2014.

[2] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski, *Robust optimization*, volume 28, Princeton university press, 2009.

[3] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[4] Andrea Campagner, 'Learnability in "learning from fuzzy labels"', in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6. IEEE, (2021).

[5] Andrea Campagner, 'Learning from fuzzy labels: Theoretical issues and algorithmic solutions', *International Journal of Approximate Reasoning*, 108969, (2023). In Press.

[6] Andrea Campagner, Davide Ciucci, Carl-Magnus Svensson, Marc Thilo Figge, and Federico Cabitza, 'Ground truthing from multi-rater labeling with three-way decision and possibility theory', *Information Sciences*, **545**, 771–790, (2021).

[7] Timothee Cour, Ben Sapp, and Ben Taskar, 'Learning from partial labels', *The Journal of Machine Learning Research*, **12**, 1501–1536, (2011).

[8] Luis M De Campos, Juan F Huete, and Serafin Moral, 'Probability intervals: a tool for uncertain reasoning', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **2**(02), 167–196, (1994).

[9] Thierry Denoeux, 'Maximum likelihood estimation from uncertain data in the belief function framework', *IEEE Transactions on knowledge and data engineering*, **25**(1), 119–130, (2011).

[10] Thierry Denœux and Lalla Meriem Zouhal, 'Handling possibilistic labels in pattern classification using evidential reasoning', *Fuzzy sets and systems*, **122**(3), 409–424, (2001).

[11] Sébastien Destercke, 'Uncertain data in learning: challenges and opportunities', *Conformal and Probabilistic Prediction with Applications*, 322–332, (2022).

[12] Michel Grabisch et al., *Set functions, games and capacities in decision making*, volume 46, Springer, 2016.

[13] Michel Grabisch, Jean-Luc Marichal, Radko Mesiar, and Endre Pap, *Aggregation functions*, volume 127, Cambridge University Press, 2009.

[14] Romain Guillaume, Inés Couso, and Didier Dubois, 'Maximum likelihood with coarse data based on robust optimisation', in *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pp. 169–180. PMLR, (2017).

[15] Eyke Hüllermeier, 'Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization', *International Journal of Approximate Reasoning*, **55**(7), 1519–1534, (2014).

[16] Eyke Hüllermeier, Sébastien Destercke, and Ines Couso, 'Learning from imprecise data: adjustments of optimistic and pessimistic variants', in *Scalable Uncertainty Management: 13th International Conference, SUM 2019, Compiègne, France, December 16–18, 2019, Proceedings 13*, pp. 266–279. Springer, (2019).

[17] Rong Jin and Zoubin Ghahramani, 'Learning with multiple labels', in *Advances in neural information processing systems*, pp. 921–928, (2003).

[18] David S Johnson, Mihalis Yannakakis, and Christos H Papadimitriou, 'On generating all maximal independent sets', *Information Processing Letters*, **27**(3), 119–123, (1988).

[19] Matti Kääriäinen, 'Relating the rademacher and vc bounds', Technical report, Citeseer, (2004).

[20] Isaac Levi, *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*, MIT press, 1980.

[21] Julian Lienen, Caglar Demir, and Eyke Hüllermeier, 'Conformal credal self-supervised learning', *arXiv preprint arXiv:2205.15239*, (2022).

[22] Julian Lienen and Eyke Hüllermeier, 'Credal self-supervised learning', *Advances in Neural Information Processing Systems*, **34**, 14370–14382, (2021).

[23] Julian Lienen and Eyke Hüllermeier, 'From label smoothing to label relaxation', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8583–8591, (2021).

[24] Julian Lienen and Eyke Hüllermeier, 'Instance weighting through data imprecisiation', *International Journal of Approximate Reasoning*, **134**, 1–14, (2021).

[25] Liping Liu and Thomas Dietterich, 'Learnability of the superset label learning problem', in *International Conference on Machine Learning*, pp. 1629–1637. PMLR, (2014).

[26] Liping Liu and Thomas G Dietterich, 'A conditional multinomial mixture model for superset label learning', in *Advances in neural information processing systems*, pp. 548–556, (2012).

[27] Enrique Miranda, Inés Couso, and Pedro Gil, 'Extreme points of credal sets generated by 2-alternating capacities', *International Journal of Approximate Reasoning*, **33**(1), 95–115, (2003).

[28] Enrique Miranda and Sébastien Destercke, 'Extreme points of the credal sets generated by comparative probabilities', *Journal of Mathematical Psychology*, **64**, 44–57, (2015).

[29] Andrea Montanari and Basil N Saeed, 'Universality of empirical risk minimization', in *Conference on Learning Theory*, pp. 4310–4312. PMLR, (2022).

[30] Ignacio Montes and Sebastien Destercke, 'On extreme points of p-boxes and belief functions', *Annals of Mathematics and Artificial Intelligence*, **81**, 405–428, (2017).

[31] Lawrence Narici and Edward Beckenstein, *Topological vector spaces*, CRC Press, 2010.

[32] Lars Schmarje, Johannes Brünger, Monty Santarossa, Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch, 'Beyond cats and dogs: Semi-supervised classification of fuzzy labels with overclustering', *arXiv preprint arXiv:2012.01768*, (2020).

[33] Shai Shalev-Shwartz and Shai Ben-David, *Understanding machine learning: From theory to algorithms*, Cambridge university press, 2014.

[34] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan, 'Stochastic convex optimization.', in *COLT*, volume 2, p. 5, (2009).

[35] Nicholas Syring and Ryan Martin, 'Stochastic optimization for numerical evaluation of imprecise probabilities', in *International Symposium on Imprecise Probability: Theories and Applications*, pp. 289–298. PMLR, (2021).

[36] Matthias CM Troffaes and Gert De Cooman, *Lower previsions*, John Wiley & Sons, 2014.

[37] Peter Walley, *Statistical reasoning with imprecise probabilities*, volume 42, Springer, 1991.

[38] Anton Wallner, 'Extreme points of coherent probabilities in finite spaces', *International Journal of Approximate Reasoning*, **44**(3), 339–357, (2007).

[39] Stephen J Wright and Benjamin Recht, *Optimization for data analysis*, Cambridge University Press, 2022.

[40] Philip B Zwart, 'Global maximization of a convex function with linear inequality constraints', *Operations Research*, **22**(3), 602–609, (1974).