Conformal Credal Self-Supervised Learning

Julian Lienen*   JULIAN.LIENEN@UPB.DE
Caglar Demir   CAGLAR.DEMIR@UPB.DE
Paderborn University, Germany

Eyke Hüllermeier   EYKE@LMU.DE
University of Munich (LMU), Germany

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

In semi-supervised learning, the paradigm of self-training refers to the idea of learning from pseudo-labels suggested by the learner itself. Recently, corresponding methods have proven effective and achieve state-of-the-art performance, e.g., when applied to image classification problems. However, pseudo-labels typically stem from ad-hoc heuristics, relying on the quality of the predictions though without guaranteeing their validity. One such method, so-called credal self-supervised learning, maintains pseudo-supervision in the form of sets of (instead of single) probability distributions over labels, thereby allowing for a flexible yet uncertainty-aware labeling. Again, however, there is no justification beyond empirical effectiveness. To address this deficiency, we make use of conformal prediction, an approach that comes with guarantees on the validity of set-valued predictions. As a result, the construction of credal sets of labels is supported by a rigorous theoretical foundation, leading to better calibrated and less error-prone supervision for unlabeled data. Along with this, we present effective algorithms for learning from credal self-supervision. An empirical study demonstrates excellent calibration properties of the pseudo-supervision, as well as the competitiveness of our method on several image classification benchmark datasets.

Keywords: Credal sets, semi-supervised learning, inductive conformal prediction, self-training, label relaxation

1. Introduction

In the recent years, machine learning applications, particularly deep learning models, have benefited greatly from the extensive amount of available data. While traditional supervised learning methods commonly rely on curated, precise target information, there is an essentially unlimited availability of unlabeled data, e.g., crowd-sourced images in computer vision applications. Semi-supervised learning (Chapelle et al., 2006) aims to leverage this source of information for further optimizing models, as successfully demonstrated by a wide range of algorithms (Berthelot et al., 2019, 2020; Sohn et al., 2020; Xie et al., 2020a). Among other approaches, so-called *self-training* (Lee, 2013) emerged as a simple yet effective paradigm to facilitate learning from additional unlabeled data. To this end, the model to be trained is used to predict "pseudo-labels" for the unlabeled data, which are then added to the labeled training data.

When learning probabilistic classifiers, e.g., to classify images, pseudo-labels can come in different forms. Traditionally, such supervision is typically in the form of a probability

---

* Corresponding author

distribution, either degenerate ("one-hot" encoded, hard) (Lee, 2013; Sohn et al., 2020) or non-degenerate (soft) (Berthelot et al., 2019, 2020). As pointed out by Lienen and Hüllermeier (2021), this form can be questioned from a data modeling perspective: Not only is a single distribution unlikely to match the true ground-truth, and may hence to bias the learner, but also incapable of expressing uncertainty about the ground-truth distribution. While uncertainty-aware approaches address these shortcomings by accompanying the targets with additional qualitative information (e.g., as in (Rizve et al., 2021)), so-called *credal self-supervised learning* (CSSL) (Lienen and Hüllermeier, 2021) replaces the targets by *credal sets*, i.e., sets of probability distributions assumed to guarantee (or at least likely) covering the true distribution. Such set-valued targets, for which CSSL employs a generalized risk minimization to enable learning, relieve the learner from committing to a single distribution, while facilitating uncertainty-awareness through increased expressivity.

Although all of the aforementioned approaches have demonstrated their effectiveness to the problem of semi-supervised learning, none of them entails meaningful guarantees about the validity of the pseudo-labels—their quality is solely subject to the model confidence itself without an objective error probability. Moreover, rather ad-hoc heuristics are considered to control the influence of vague beliefs on the learning process (e.g., by selecting pseudo-labels surpassing a confidence threshold). For instance, CSSL derives credal pseudo-supervision based on the predicted probability scores. Although it exhibits uncertainty-awareness by using entropy minimization (less entropy results in smaller sets), this may be misleading for overconfident yet poorly generalizing models. Again, no objective guarantee can be given for the validity of the credal set. This raises the quest for a pseudo-labeling procedure providing exactly that kind of guarantee, i.e., validating pseudo-supervision beyond mere empirical effectiveness.

Fortunately, the framework of conformal prediction (CP) (Vovk et al., 2005; Shafer and Vovk, 2008) provides a tool-set to satisfy such demands. As such, CP induces prediction regions quantifying the uncertainty for candidate values as outcome for a query instance by measuring their non-conformity ("strangeness") with previously observed data with known outcomes. Much like in classical hypothesis testing, it computes p-values for candidate outcomes (e.g., for each class y') as a criterion for hypothesis rejection, i.e., whether to include a value y' in the prediction set with a certain amount of confidence. Based on very mild technical assumptions, CP provides formal guarantees for the validity of prediction regions and is able to control the probability of an invalid prediction (not covering the true label).

In our work, we combine the technique of inductive conformal prediction, an efficient variant of CP, with the idea of credal self-supervised learning, leading to a method we dub *conformal credal self-supervised learning* (CCSSL). Instead of deriving credal sets based on an ad-hoc estimate of the model confidence, we proceed from a possibilistic interpretation of conformal predictions (Cella and Martin, 2022) to construct conformal credal labels. By this, pseudo-labels of that kind provide the validity guarantees implied by conformal predictors, laying a rigorous theoretical foundation for a set-valued pseudo-labeling strategy and thus paving the way for more thorough theoretical analyses of self-training in semi-supervised learning. To enable learning from conformal credal self-supervision, we further provide an effective algorithmic solution for generalized empirical risk minimization on set-valued targets. An exhaustive empirical study on several image classification

datasets demonstrates the usefulness of our method for effective and reliable semi-supervised learning — leveraging well calibrated pseudo-supervision, it is shown to improve the state-of-the-art in terms of generalization performance.

2. Related Work

Semi-supervised learning describes the learning paradigm where the aim is to leverage the potential of unlabeled in addition to labeled data to improve learning and generalization. As it is easy to lose track of related work due to the wide range of proposed approaches in the recent years, we focus here on classification methods applied to computer vision applications that are closely related to our work. For a more comprehensive overview, we refer to (Chapelle et al., 2006) and (van Engelen and Hoos, 2020).

Nowadays, so-called *self-training* (or self-supervised) methods protrude among these methods by providing a simple yet effective learning methodology to make use of unlabeled data. Such methods can be found in a wide range of domains, including computer vision (Rosenberg et al., 2005; Doersch et al., 2015; Godard et al., 2019; Xie et al., 2020b) and natural language processing (Du et al., 2021). As self-training can be considered as a general learning paradigm, where a model suggests itself labels to learn from, it has been wrapped around various model types, e.g., support vector machines (Lienen and Hüllermeier, 2021), decision trees (Tanha et al., 2017) and most prominently with neural networks (Oliver et al., 2018). Notably, it lays the foundation for so-called distillation models, e.g., in self-distillation (Kim et al., 2021) or student-teacher setups (Xie et al., 2020b; Pham et al., 2021). It is further popular for unsupervised pretraining, e.g., as described in (Grill et al., 2020; Caron et al., 2021).

Uncertainty-awareness (Hüllermeier and Waegeman, 2021) is a critical aspect for a cautious pseudo label selection to learn from, without exposing the model to the risk of confirmation biases (Arazo et al., 2020). This aspect has been considered in previous works by different means. (Rizve et al., 2021) employs Bayesian sampling techniques, such as MC-Dropout (Gal and Ghahramani, 2016) or DropBlock (Ghiasi et al., 2018), to estimate uncertainty of a prediction, which is then used as an additional filter criterion. (Ren et al., 2020) uses an adaptive instance weighting approach to control the influence of individual pseudo labels. Moreover, credal self-supervised learning (Lienen and Hüllermeier, 2021) expresses certainty by the size of maintained credal sets as pseudo-supervision. Also, learning from softened probability distribution can also be considered as a way to suppress overconfidence tendencies to learn in a more cautious way (Berthelot et al., 2019; Arazo et al., 2020). Lastly, several domain-specific adaptions have been proposed, e.g., for text classification (Mukherjee and Awadallah, 2020) or semantic segmentation (Zheng and Yang, 2021).

As prominently used throughout this work, conformal prediction (Vovk et al., 2005; Shafer and Vovk, 2008; Balasubramanian et al., 2014) provides an elegant framework to naturally express model prediction uncertainty in a set-valued form. We refer the interested reader to (Zeni et al., 2020) for a more comprehensive overview beyond the covered literature here. Whereas conformal prediction initially proceeded from a transductive form in an online setting (Vovk et al., 2005), its high complexity called for more efficient variants. To this end, several variations have been proposed, where inductive conformal prediction

([Papadopoulos et al., 2007](#)) most prominently draw attention by assuming a separated calibration split available used for non-conformity measurement. Beyond this, several alternative approaches varying data assumption or algorithmic improvements have been proposed (e.g., as described in ([Lei et al., 2018](#); [Kim et al., 2020](#))). Recently, [Cella and Martin](#) ([2021](#), [2022](#)) suggest a reinterpretation of conformal transducer as plausibility contours, allowing to derive validity guarantees on predictive distributions rather than prediction regions, which is being used throughout our work.

3. Conformal Credal Pseudo-Labeling

In this section, we revisit credal pseudo-labeling and introduce conformal prediction as a framework for reliable prediction and an integral part of our conformal credal pseudo-labeling approach.

3.1. Credal Pseudo-Labeling

In supervised classification, one typically assumes instances $\boldsymbol{x} \in \mathcal{X}$ of an instance space $\mathcal{X}$ to be associated with a ground-truth in the form of a conditional probability distribution $p^*(\cdot \,|\, \boldsymbol{x}) \in \mathbb{P}(\mathcal{Y})$ over the class space $\mathcal{Y} = \{y_1, ..., y_K\}$. Training a probabilistic classifier $\hat{p} : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$ commonly involves the optimization of a probabilistic surrogate loss function $\mathcal{L} : \mathbb{P}(\mathcal{Y}) \times \mathbb{P}(\mathcal{Y}) \longrightarrow \mathbb{R}_+$ (e.g., cross-entropy) comparing the model prediction $\hat{p}$ to a proxy distribution p reflecting the true target $p^*(\cdot \,|\, \boldsymbol{x})$. Although it would be desirable, most classification datasets do not give direct access to p^*, but only to a realization $y \in \mathcal{Y}$ of the random variable $Y \sim p^*(\cdot \,|\, \boldsymbol{x})$, whence a degenerate distribution p_y with $p_y(y) = 1$ and $p_y(y') = 0$ for $y' \neq y$ is often considered as a proxy. Other data modeling techniques such as label smoothing ([Szegedy et al., 2016](#)) also consider "softened" versions of approaching p_y.

What methods of this kind share is their reliance on a single probability distribution as target information. As already said, this information does not allow for representing any uncertainty about the ground-truth distribution p^*: Predicting a precise distribution $\hat{p}$, the learner pretends a level of certainty that is not warranted. This is a sharp conflict with the observation that, in practice, such predictions tend to be poorly calibrated and overconfident ([Müller et al., 2019](#)), and are hence likely to bias the learner.

To overcome these shortcomings, [Lienen and Hüllermeier](#) ([2021](#)) suggest to the replace a single probabilistic prediction by a *credal set* $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$, i.e., as set of (candidate) probability distributions. This set $\mathcal{Q}$ is supposed to cover the ground-truth p^*, very much like a confidence interval is supposed to cover a ground-truth parameter in classical statistics. Thus, the learner is able to represent its uncertainty about the ground-truth p^* in a more faithful way. More specifically, the learner's (epistemic) uncertainty is in direct correspondence with the size of the credal set. It can represent complete ignorance about p^* (by setting $\mathcal{Q} = \mathbb{P}(\mathcal{Y})$), complete certainty ($\mathcal{Q} = \{p\}$), but also epistemic states in-between these extremes.

One way to specify credal sets is by means of so-called *possibility distributions* ([Dubois and Prade, 2004](#)) $\pi : \mathcal{Y} \longrightarrow [0, 1]$, which induce a possibility measure $\Pi : 2^{\mathcal{Y}} \longrightarrow [0, 1]$ by virtue of $\Pi(Y) = \max_{y \in Y} \pi(y)$ for all $Y \subseteq \mathcal{Y}$. Possibility degrees can be interpreted as upper probabilities, so that a possibility distribution π specifies the following set of probability

distributions:

$$\mathcal{Q}_\pi := \Big\{ p \in \mathbb{P}(\mathcal{Y}) \mid \forall Y \subseteq \mathcal{Y} :$$
$$P(Y) = \sum_{y \in Y} p(y) \leq \max_{y \in Y} \pi(y) = \Pi(Y) \Big\} \ . \tag{1}$$

To guarantee $\mathcal{Q}_\pi \neq \emptyset$, the distribution π is normalized such that $\max_{y \in \mathcal{Y}} \pi(y) = 1$, i.e., there is at least one label y considered fully plausible.

When considering the setting of semi-supervised learning, credal labeling is employed as a pseudo labeling technique within the framework of credal self-supervised learning (CSSL) (Lienen and Hüllermeier, 2021): For each unlabeled instance, a credal set is maintained to express the current belief about p^* in terms of a confidence region $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$. To this end, CSSL derives a possibility distribution π by an ad-hoc heuristic that assigns full plausibility $\pi(\hat{y}) = 1$ to the class $\hat{y}$ for which the predicted probability is highest, and determines a constant plausibility degree α for all other classes $y' \neq \hat{y}$; the latter depends on the learner's confidence as well as the class prior and the prediction history.

Although CSSL has proven to provide competitive generalization performance, the credal set construction heuristic lacks a solid theoretical foundation and does not provide any quality guarantees. Especially in the case of overconfident models, the credal sets may not reflect uncertainty properly and misguide the learner. Moreover, the class prior and the prediction history employed in the specification of possibility distribution constitute yet another potential source of bias in the learning process. This raises the question whether credal pseudo-labeling can be based on a more solid theoretical foundation and equipped with validity guarantees. An affirmative answer to this question is offered by the framework of conformal prediction.

3.2. Inductive Conformal Prediction

Conformal prediction (CP) (Vovk et al., 2005) is a distribution-free uncertainty quantification framework that judges a prediction for a query by its "non-conformity" with data observed before. While it has been originally introduced and extensively analyzed in an online setting, we refer to an offline ("batch-wise") variant as typically considered in supervised learning settings. Several alternative variants of conformal prediction have been proposed in the past, most notably transductive and inductive methods. In its original formulation, the former requires a substantial amount of training and is clearly unsuitable in large-scale learning scenarios such as typical deep learning applications. This issue has been addressed in so-called *inductive conformal prediction* (ICP) (Papadopoulos, 2008) by alleviating the computational demands through additional calibration data.

Assume we are given training data $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} \subset (\mathcal{X} \times \mathcal{Y})^N$ and calibration data $\mathcal{D}_{\text{calib}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{L} \subset (\mathcal{X} \times \mathcal{Y})^L$ comprising N resp. L i.i.d. observations. For a given query (resp. test) instance $\boldsymbol{x}_{N+1}$, the goal of conformal prediction is to provide an uncertainty quantification with provable guarantees about the likeliness of each possible candidate $\hat{y} \in \mathcal{Y}$ being the true outcome associated with $\boldsymbol{x}_{N+1}$. This is done by measuring the conformity of $(\boldsymbol{x}_{N+1}, \hat{y})$ with $\mathcal{D}_{\text{calib}}$ for each candidate label $\hat{y}$.

To this end, conformal prediction calculates a *non-conformity measure* $\alpha : (\mathcal{X} \times \mathcal{Y})^m \times (\mathcal{X} \times \mathcal{Y}) \longrightarrow \mathbb{R}$, which indicates how well a pair of a query instance and a candidate label

conforms to an observed sequence of m pairs. Typically, the non-conformity $\alpha(\mathcal{D}, (\boldsymbol{x}, y))$ involves the training of an underlying predictor (a model $\hat{p}$ in our case) on $\mathcal{D}$, whose prediction $\hat{p}(\boldsymbol{x})$ is then compared to y. In a probabilistic learning scenario, common choices are

$$\alpha(\mathcal{D}, (\boldsymbol{x}, y)) = \max_{y_j \neq y} \hat{p}_{\mathcal{D}}(\boldsymbol{x})(y_j) - \hat{p}_{\mathcal{D}}(\boldsymbol{x})(y) \tag{2}$$

and

$$\alpha(\mathcal{D}, (\boldsymbol{x}, y)) = \frac{\max_{y_j \neq y} \hat{p}_{\mathcal{D}}(\boldsymbol{x})(y_j)}{\hat{p}_{\mathcal{D}}(\boldsymbol{x})(y) + \gamma} \ , \tag{3}$$

where $\hat{p}_{\mathcal{D}}$ is a probabilistic classifier trained on $\mathcal{D}$, $\hat{p}(\cdot)(y')$ denotes the output probability for class y' and $\gamma \geq 0$ is a sensitivity parameter (Papadopoulos et al., 2007). In our case, we set $\mathcal{D}$ to the training data $\mathcal{D}_{\text{train}}$.

In ICP, non-conformity scores α_i are calculated for all $(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{calib}}$, i.e., it is determined for each calibration instance how well it conforms with the underlying training data $\mathcal{D}_{\text{train}}$. For the uncertainty quantification of the prediction for a query instance $\boldsymbol{x}_{N+1}$, the non-conformity $\alpha_{N+1}^{\hat{y}} := \alpha(\mathcal{D}_{\text{train}}, (\boldsymbol{x}_{N+1}, \hat{y}))$ is calculated for each candidate label $\hat{y} \in \mathcal{Y}$. Given the non-conformity values $\alpha_1, \ldots, \alpha_L$ associated with the calibration data, these non-conformity values can be used to construct *p-values* for each candidate label $\hat{y}$, comparable to the notion of p-values in traditional statistics:

$$\pi_{N+1}(\hat{y}) = \frac{|\{\alpha_i \geq \alpha_{N+1}^{\hat{y}} \,|\, i \in \{1, ..., L\}\}| + 1}{L + 1} \ . \tag{4}$$

Consequently, one can define a conformal predictor Γ_δ with confidence $1 - \delta$ by

$$\Gamma_\delta(\mathcal{D}_{\text{calib}}, \boldsymbol{x}_{N+1}) = \Big\{ \hat{y} \in \mathcal{Y} : \pi_{N+1}(\hat{y}) \geq \delta \Big\}. \tag{5}$$

Such a Γ_δ can be shown to cover the true label y_{N+1} associated with $\boldsymbol{x}_{N+1}$ with high probability:

$$\Pr\Big(y_{N+1} \in \Gamma_\delta(\mathcal{D}_{\text{calib}}, \boldsymbol{x}_{N+1}) \Big) \geq 1 - \delta \ , \tag{6}$$

where the probability is taken over $\boldsymbol{x}_{N+1}$ and $\mathcal{D}_{\text{calib}}$. This property is typically referred to as (marginal) *validity*, however, following (Cella and Martin, 2022), we refer to (6) as *weak validity*. Note that this holds for any underlying probability distribution, any $\delta \in (0, 1)$, and $N \in \mathbb{N}_+$. Nevertheless, practically speaking, small calibration datasets $\mathcal{D}_{\text{calib}}$ may be problematic if a certain granularity in the confidence degree is desired (Johansson et al., 2015).

3.2.1. CONFORMALIZED PREDICTIVE DISTRIBUTIONS

While the notion of weak validity applies to set-valued prediction regions, its application as quantification of *predictive distributions* is not obvious. In the realm of probabilistic classification (as considered here), one may wonder how CP can be applied to provide a meaningful uncertainty quantification. Recently, Cella and Martin (2021) proposed an interpretation of p-values π (cf. Eq. (4)) in terms of possibility degrees, i.e., upper probabilities of the event

that the respective candidate label is the true outcome. Consequently, these possibilities π induce possibility measures Π, such that

$$\Pi(A) := \sup_{\hat{y} \in A} \pi(\hat{y}), \qquad A \subseteq \mathcal{Y} . \tag{7}$$

With $\sup_{\hat{y} \in \mathcal{Y}} \pi(\hat{y}) = 1$ for all $\mathcal{D}_{\mathrm{train}}$ and assuming that π is stochastically no smaller than the uniform distribution $\mathcal{U}(0, 1)$ under any underlying (ground-truth) probability distribution, probabilistic predictors satisfy the *strong validity* property defined as follows:

$$\mathrm{Pr}\left(\Pi(A) \le \delta, y_{N+1} \in A\right) \le \delta \tag{8}$$

holds for any $\delta \in (0, 1)$, $A \subseteq \mathcal{Y}$, training data $\mathcal{D}_{\mathrm{train}}$ and any underlying true probability distribution.

3.3. Conformal Credal Pseudo-Labeling

Strongly valid possibility distributions π of this kind can be directly employed in the credal set formulation in (1) to bound the space of probability distributions. Relating to the self-training paradigm we consider in a semi-supervised setting, where a model suggests itself credal sets as pseudo-supervision to learn from, we refer to such credal sets as *conformal credal pseudo-labels*. Note that such pseudo-labels do not necessarily need to be constructed by an inductive conformal prediction procedure, but can used any CP methodology with the same guarantees.

To guarantee property (8) for the possibility distribution π defined in (4), this distribution needs to be normalized such that $\max_{\hat{y} \in \mathcal{Y}} \pi(\hat{y}) = 1$ to ensure strong validity and that $\mathcal{Q}_\pi \ne \emptyset$. A commonly applied normalization is suggested in (Cella and Martin, 2021):

$$\pi(\hat{y}) = \frac{\pi(\hat{y})}{\max_{y' \in \mathcal{Y}} \pi(y')} . \tag{9}$$

By the described conformal credal labeling method, we replace the ad-hoc construction of sets in CSSL by a more profound technique with formal guarantees. More precisely, we build the learner's uncertainty-awareness on top of an *objective* quality criterion deduced from its accuracy on the calibration data. This not only provides strong validity guarantees, but also paves the path to a more rigorous theoretical analysis of self-supervised credal learning. Nevertheless, these guarantees come at a cost: By applying a conformal prediction procedure, one either provokes an increased computational overhead (for instance in a transductive CP setting) or a decrease in data efficiency by requiring additional calibration data. However, as will be seen in the empirical evaluation, the improved pseudo-label quality makes up for this in terms of generalization performance.

4. Conformal Credal Self-Supervised Learning

In the following, we provide an overview of our methodology for effective semi-supervised learning from conformal credal pseudo-labels.
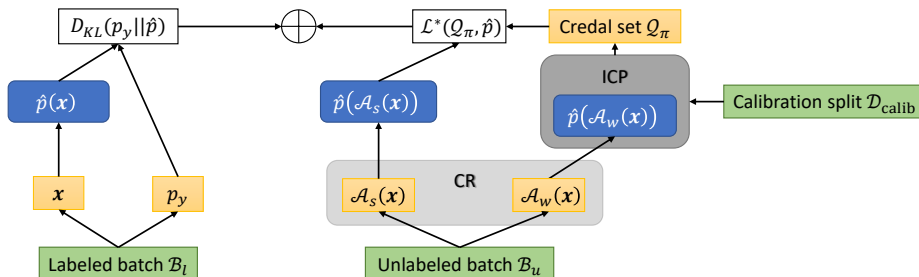
Figure 1: Overview over CCSSL: While learning on the labeled batch $\mathcal{B}_l$ is performed conventionally, the labels for the unlabeled batch $\mathcal{B}_u$ are constructed by an inductive conformal prediction (ICP) procedure provided calibration data $\mathcal{D}_{\text{calib}}$, employing consistency regularization (CR) for confirmation bias mitigation.

4.1. Overview

In our approach, which we call *conformal credal self-supervised learning* (CCSSL), we combine credal self-supervised learning as proposed in (Lienen and Hüllermeier, 2021) with conformal credal labels. We thus replace the credal set construction in CSSL to ensure guarantees of the pseudo-label's validity. Although both CSSL and CCSSL are not specifically tailored to a particular domain, we adopt the framework of the former with a focus on image classification here. Fig. 1 gives an overview of the algorithm, its pseudo-code can be found in the appendix.[1]

In each iteration, we observe batches of labeled $\mathcal{B}_l = \{(\boldsymbol{x}_i, p_{y_i})\}_{i=1}^{B}$ and unlabeled instances $\mathcal{B}_u = \{\boldsymbol{x}_i\}_{i=1}^{\mu B}$, where $\mu \geq 1$ is the multiplicity of unlabeled over labeled instances in each batch. Here, we consider degenerate distributions p_{y_i} as target information in $\mathcal{B}_l$. For the instances $(\boldsymbol{x}, p_y) \in \mathcal{B}_l$, we compute the Kullback-Leibler divergence $D_{KL}(p_y \,||\, \hat{p}(\boldsymbol{x}))$ as loss $\mathcal{L}_l$.

To calculate the loss on unlabeled instances $\boldsymbol{x} \in \mathcal{B}_u$ leveraging our conformal credal labeling approach, we adopt the *consistency regularization* (CR) framework (Bachman et al., 2014; Sajjadi et al., 2016), which aims to mitigate the so-called *confirmation bias* (Arazo et al., 2020) that describes the manifestation of misbeliefs of an accurate learner. CR enforces consistent predictions for different perturbed appearances of an instance, and has been successfully combined with pseudo-labeling, whereby Sohn et al. (2020) propose a particularly simple scheme: An unlabeled instance $\boldsymbol{x} \in \mathcal{X}$ is augmented by a weak (e.g., horizontal flipping) and a strong (e.g., RandAugment (Cubuk et al., 2020) or CTAugment (Berthelot et al., 2020)) augmentation policy $\mathcal{A}_w : \mathcal{X} \longrightarrow \mathcal{X}$ and $\mathcal{A}_s : \mathcal{X} \longrightarrow \mathcal{X}$, respectively. The weakly-augmented representation $\mathcal{A}_w(\boldsymbol{x})$ is then used to construct a pseudo-label based on $\hat{p}(\mathcal{A}_w(\boldsymbol{x}))$, which is finally compared to $\hat{p}(\mathcal{A}_s(\boldsymbol{x}))$ to compute the loss.

Transferred to our methodology, we derive conformal predictions for the weakly-augmented instance $\mathcal{A}_w(\boldsymbol{x})$ in an inductive manner based on a previously separated calibration dataset $\mathcal{D}_{\text{calib}}$. To this end, a (strongly valid) possibility distribution π over all possible

---

1. The appendix, along with the official implementation, is externally available at https://github.com/julilien/C2S2L.

labels $y \in \mathcal{Y}$ is determined for the prediction $\hat{p}(\mathcal{A}_w(\boldsymbol{x}))$, which is used to construct a credal target set $\mathcal{Q}_\pi$ according to (1). Finally, $\mathcal{Q}_\pi$ is compared to the prediction $\hat{p}(\mathcal{A}_s(\boldsymbol{x}))$ in terms of a generalized probabilistic loss $\mathcal{L}^* : 2^{\mathbb{P}(\mathcal{Y})} \times \mathbb{P}(\mathcal{Y}) \longrightarrow \mathbb{R}_+$, which we shall detail in Section 4.3.

Altogether, the training objective is given by

$$
\mathcal{L} = \overbrace{\frac{1}{|\mathcal{B}_l|} \sum_{(\boldsymbol{x}_i, p_{y_i}) \in \mathcal{B}_l} D_{KL}(p_{y_i} \,||\, \hat{p}(\boldsymbol{x}_i))}^{\mathcal{L}_l}
$$
$$
+ \lambda_u \underbrace{\frac{1}{|\mathcal{B}_u|} \sum_{\boldsymbol{x}_i \in \mathcal{B}_u} \mathcal{L}^*(Q_{\pi_i}, \hat{p}((\mathcal{A}_s(\boldsymbol{x})_i))}_{\mathcal{L}_u} \,, \tag{10}
$$

where $\lambda_u \geq 0$ weights the importance of the unlabeled loss part $\mathcal{L}_u$.

4.2. Validity Mitigates Confirmation Biases

Many of the recent SSL approaches, including CSSL, leverage techniques such as consistency regularization to mitigate confirmation biases. Intuitively speaking, it enforces a smooth decision boundary in the neighborhood of (augmented) unlabeled instances, also referred to as *local consistency* (Wei et al., 2021). For instance, classifiers trained on images showing cars provided in the training dataset predict consistent labels for similar cars that differ in the color, perspective changes or in an alternative scenery. If a certain proximity of the pseudo-labeled instances to the known labeled instances in the latent feature space is preserved, known as the *expansion assumption*, CR can lead to a global class-wise consistency, which can "de-noise" wrong pseudo-labels. Hence, it serves as a means to alleviate the problem of misguidance through mislabels.

However, the expansion assumption often appears too unrealistic in practice. For instance, violations happen in cases where the neighborhood of an (unlabeled) instance is not homogeneously populated by instances of the same true class. Typical examples for such situations are highly uncertain classification problems in which individual classes are hard to separate, e.g., distinguishing model variants of a car based on subtle visual differences such as badges, or data with imbalanced class frequencies. In the latter case, some classes dominate underrepresented ones, such that the neighborhood of unlabeled instances may be mostly populated by instances from different classes, thereby violating the expansion assumption. Consequently, from an empirical risk minimization point of view, it might be more reasonable to attribute larger regions populated by unlabeled instances from the minority class to a majority classes. This leaves the former overlooked, so that the correction of wrong pseudo-labels is not possible anymore.

Credal labels as constructed in CSSL are incapable of solving the issue of the expansion violation assumption and hence ineffective CR: Similar to single-point probabilistic pseudo-labeling methods, the pseudo-labels have no error guarantees that the actual true class is adequately incorporated in the target. As opposed to this, conformal credal pseudo-labels can provide such, which is why their validity guarantee can be regarded as a second fallback when combined with CR. When the expansion assumption is violated, a higher validity of

9

credal sets oppose committing to a misbelief in a too premature manner by a more cautious learning behavior. As will be seen in the experimental evaluation on commonly used semi-supervised image classification benchmarks, conformal (and hence valid) credal sets indeed show a more robust behavior towards confirmation biases when facing highly uncertain neighborhoods (cf. Sec. 5.2). In addition, aiming to isolate the contributions of CR and validity in a different setting, we provide further experiments on imbalanced data in the appendix.

4.3. Generalized Credal Learning

As set-valued targets are provided as (pseudo-)supervision for the unlabeled instances, a generalization of a single-point probability loss $\mathcal{L}$ is required. Here, we follow (Lienen and Hüllermeier, 2021) and leverage a generalized empirical risk minimization approach by minimizing the *optimistic superset loss* (Hüllermeier and Cheng, 2015; Cabannes et al., 2020)

$$\mathcal{L}^*(\mathcal{Q}_\pi, \hat{p}) := \min_{p \in \mathcal{Q}_\pi} \mathcal{L}(p, \hat{p}) \ , \tag{11}$$

with $\mathcal{L} = D_{KL}$ in our case. This generalization is motivated by the idea of *data disambiguation* (Hüllermeier and Cheng, 2015), in which targets are disambiguated within imprecise sets according to their plausibility in the overall loss minimization context over all data points, leading to a minimal empirical risk with respect to the original loss $\mathcal{L}$.

While credal sets as considered in CSSL are of rather simplistic nature, credal sets with arbitrary (but normalized) possibility distributions as introduced in Section 3.3 impose a more complex optimization problem. More precisely, consider a possibility distribution π and denote $\pi_i := \pi(y_i)$. Without loss of generality, let the possibilities be ordered, i.e., $0 \leq \pi_1 \leq \ldots, \leq \pi_K = 1$. Then, it has been shown that the following holds for any distribution $p \in \mathcal{Q}_\pi$ with $p_i := p(y_i)$ (Delgado and Moral, 1987):

$$p \in \mathcal{Q}_\pi \Leftrightarrow \sum_{k=1}^{i} p_k \leq \pi_i \ , \qquad i \in \{1, ..., K\} \tag{12}$$

The resulting set of inequality constraints induces a convex polytope (Kroupa, 2006), such that the optimization problem in (11) becomes the problem of finding the closest point in a convex polytope (here $\mathcal{Q}_\pi$) with minimal distance to a given query point ($\hat{p}$). We provide examples illustrating such convex credal polytopes in the appendix.

Optimization problems of this kind are not new, and many approximate solutions have been proposed in the past, including projected gradient algorithms (Bahmani and Raj, 2011), conditional gradient descent (alias Frank-Wolfe algorithms) (Frank and Wolfe, 1956; Jaggi, 2013), and other projection-free stochastic methods (Li et al., 2020). However, we found that one can take advantage of the structure of the problem (12) to induce a precise and average-case efficient algorithmic solution to (11).

Algorithm 1 lists the procedure to determine the loss $\mathcal{L}^*(\mathcal{Q}_\pi, \hat{p})$. The key idea is to consider each face of the convex polytope being associated with a particular possibility constraint π_i. The algorithm then determines the optimal face on which the prediction $\hat{p}$ can be projected without violating the constraint for any $\pi_j \leq \pi_i$, which is guaranteed to provide the optimal solution for the classes y_j. In an iterative scheme, this procedure is

---

Algorithm 1 Generalized Credal Learning Loss

Require: Predicted distribution $\hat{p} \in \mathbb{P}(\mathcal{Y})$, (normalized) possibility distribution $\pi : \mathcal{Y} \longrightarrow [0,1]$

  if $\hat{p} \in \mathcal{Q}_\pi$ **then return** $D_{KL}(\hat{p} \,||\, \hat{p}) = 0$

  Initialize set of unassigned classes $Y = \mathcal{Y}$

  while Y is not empty **do**

    Determine $y^* \in Y$ with highest $\pi(y^*)$, such that the probabilities

$$\bar{p}(y) = \left(\pi(y^*) - \sum_{y' \notin Y} p^r(y') \right) \cdot \frac{\hat{p}(y)}{\sum_{y' \in Y'} \hat{p}(y')}$$

    for all $y \in Y' := \{y \in Y \mid \pi(y) \leq \pi(y^*)\}$ do not violate the constraints in Eq. (12)

    Assign $p^r(y) = \bar{p}(y)$ for all $y \in Y'$

    $Y = Y \setminus Y'$

  end while

  return $D_{KL}(p^r \,||\, \hat{p})$

---

applied to all classes, leading to the distribution $p \in \mathcal{Q}_\pi$ with minimal D_{KL} to $\hat{p}$. Hence, we can state the following theorem.

Theorem 1 (Optimality) *Given a credal set $\mathcal{Q}_\pi$ induced by a normalized possibility distribution $\pi : \mathcal{Y} \longrightarrow [0,1]$ with $\max_{y \in \mathcal{Y}} \pi(y) = 1$ according to (1), Algorithm 1 returns the solution of $\mathcal{L}^*(\mathcal{Q}_\pi, \hat{p})$ as defined in (11) for an arbitrary distribution $\hat{p} \in \mathbb{P}(\mathcal{Y})$.*

Due to space limitations, we provide the proof of Theorem 1 in the appendix. We further provide a discussion of the computational complexity of Algorithm 1 therein, showing that the worst case complexity is cubic in the number of labels — the average case complexity is much lower, however, making the method amenable to large data sets as demonstrated in our experimental evaluation.

5. Experiments

To demonstrate the effectiveness of our method, we present empirical results for the domain of image classification as an important and practically relevant application. We refer to the appendix for additional results, including a study on knowledge graphs to demonstrate generalizability of our approach, as well as a more comprehensive overview over experimental settings for reproducibility.

5.1. Experimental Setting

Following previous semi-supervised learning evaluation protocols (Sohn et al., 2020; Lienen and Hüllermeier, 2021), we performed experiments on CIFAR-10/-100 (Krizhevsky and Hinton, 2009), SVHN (Netzer et al., 2011) (without the extra data split) and STL-10 (Coates et al., 2011) with various numbers of sub-selected labels. To this end, we trained Wide ResNet-28-2 (Zagoruyko and Komodakis, 2016) models for CIFAR-10, SVHN and

STL-10, whereas we considered Wide ResNet-28-8 as architecture for CIFAR-100. For each combination, we conducted a Bayesian optimization to tune the hyperparameters with 20 runs each on a separate validation split, while we report the final test performances per model trained with the tuned parameters. Each model was trained for 2^{18} iterations. We repeated each experiment for 3 different seeds, whereby we re-used the best hyperparameters for all seeds due to the high computational complexity.

For CCSSL, we distinguish two variants employing either the non-conformity measure (2) or (3), which we refer to as *CCSSL-diff* and *CCSSL-prop* respectively. As baselines, we consider FixMatch (Sohn et al., 2020) and its distribution alignment version as hard and soft probabilistic pseudo-labeling technique, respectively. Moreover, we compare our method to UDA (Xie et al., 2020a) as another soft variant. Recently, FlexMatch (Zhang et al., 2021) has been proposed as an advancement of FixMatch by adding curriculum learning encompassing uncertainty-awareness. Finally, we report results of CSSL (Lienen and Hüllermeier, 2021) as methodically closest related work to our approach. In all cases, we employ RandAugment as strong augmentation policy to realize consistency regularization. As all of the mentioned approaches were embedded in the basic FixMatch framework, we achieve a fair comparison alleviating side-effects.

5.2. Generalization Performance

Table 1: Averaged accuracies over 3 seeds for different numbers of labels. **Bold** entries indicate the best performing method per column.

	CIFAR-10			CIFAR-100			SVHN			STL-10
	40 lab.	250 lab.	4000 lab.	400 lab.	2500 lab.	10000 lab.	40 lab.	250 lab.	1000 lab.	1000 lab.
UDA	86.44 ±2.70	**94.81** ±0.22	95.31 ±0.10	49.41 ±2.96	68.33 ±0.31	75.98 ±0.45	84.82 ±10.6	96.41 ±1.04	97.14 ±0.24	83.94 ±1.49
FixMatch	87.14 ±2.61	93.81 ±1.02	95.08 ±0.18	47.73 ±1.88	66.82 ±0.26	76.66 ±0.18	86.26 ±14.2	96.23 ±1.50	**97.32** ±0.09	85.34 ±0.92
FixMatch DA	89.54 ±5.90	94.00 ±0.56	95.13 ±0.21	51.31 ±2.67	69.98 ±0.30	76.67 ±0.08	86.00 ±16.3	95.85 ±1.62	97.02 ±0.16	85.59 ±1.21
FlexMatch	91.21 ±3.46	94.08 ±0.64	94.62 ±0.27	49.99 ±0.39	**71.47** ±1.06	77.01 ±0.17	85.78 ±1.37	96.52 ±0.29	96.54 ±0.30	85.24 ±1.49
CSSL	**91.70** ±4.77	94.59 ±0.15	95.41 ±0.04	52.54 ±1.60	67.81 ±0.64	77.56 ±0.22	**87.27** ±5.69	95.54 ±1.63	96.69 ±0.75	85.07 ±1.11
CCSSL-diff	90.13 ±3.33	93.56 ±0.21	95.43 ±0.06	**54.13** ±1.97	69.33 ±0.78	77.40 ±0.18	86.52 ±6.84	95.67 ±1.51	96.97 ±0.23	85.26 ±0.79
CCSSL-prop	89.24 ±2.56	94.34 ±0.27	**95.48** ±0.06	53.48 ±2.75	67.90 ±0.37	**77.72** ±0.08	85.38 ±7.67	**96.87** ±0.20	97.04 ±0.38	**85.67** ±1.14

Table 1 shows the generalization performance of all methods with respect to the accuracy for various amounts of labeled data. Note that the calibration set for the conformal credal variant is taken from the labeled instances, i.e., the effective number of instances to learn from is further decreased. In the appendix, we provide a quantification of the network calibration, i.e., the quality of the predicted probability distributions.

As can be seen, CCSSL leads to competitive generalization when a sufficient amount of labeled instances is provided, often even performing best among the compared methods. In the label-scarce settings, the separation of a calibration set (which is taken from the labeled data part) has a stronger effect, and the performance is slightly inferior to CSSL. However, the performance still does not drop, which confirms the effectiveness of the improved pseudo-label quality over CSSL. In most other cases, CCSSL appears to be superior compared to CSSL.

On CIFAR-100 with 400 labels, the learner often faces neighborhoods populated by instances with heterogeneous class distributions. As a result, consistency regularization be-

comes less effective as the rich class space harms the continuity of regions covering instances of a particular class, leading to the manifestation of misbeliefs in these regions. As consistently observed in the learning curves for CIFAR-100 shown in Fig. 2, CSSL with CR suffers from this issue with increased confidence. In the terminal phase of the training, credal sets are kept rather vague, which is why misbeliefs do not affect the overall performance too much. However, later phases show a performance degradation, which can be attributed to the fact that (potentially mislabeled) credal pseudo-labels become smaller and have thus a higher weight in the overall loss minimization. As opposed to that, CCSSL learns more cautiously thanks to the conformal construction of credal sets and can continuously improve the generalization performance.
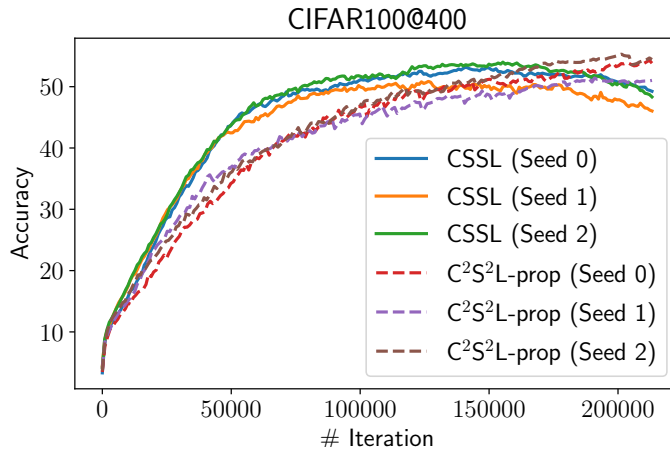

Figure 2: Test accuracies over the course of the training on CIFAR-100 with 400 labels.

5.3. Pseudo-Label Quality

To assess the quality of the (credal) pseudo-labels, we report the validity according to (8) by measuring the error rate in terms of $\mathbb{1}(\pi(y) \leq \delta)$ given the pseudo-supervisions π for the unlabeled training instances with true labels y (for significance levels $\delta \in \{0.05, 0.1, 0.25\}$). To ensure fairness, we compute these scores for trained models so as not to give an advantage to CCSSL, which can in contrast to CSSL rely on the strong validity guarantee throughout the training.

Table 2: The final validity as specified in Eq. (8) of all credal pseudo-labels for different significance levels δ averaged over 3 random seeds. **Bold** entries indicate the best method per column, the standard deviation is a factor of $1e^{-3}$.

| | CIFAR-10 | | | | | | SVHN | | | | | |
| | 250 lab. | | | 1000 lab. | | | 250 lab. | | | 1000 lab. | | |
δ	0.05	0.1	0.25	0.05	0.1	0.25	0.05	0.1	0.25	0.05	0.1	0.25
CSSL	0.027 ±0.6	0.033 ±1.1	**0.042** ±0.7	0.033 ±1.1	0.037 ±4.8	0.042 ±1.1	0.038 ±0.6	0.044 ±1.4	0.053 ±0.8	0.030 ±1.6	0.034 ±6.1	0.040 ±2.1
CCSSL-diff	0.026 ±2.9	0.032 ±1.9	0.047 ±4.7	0.029 ±1.5	0.034 ±1.5	0.042 ±0.6	0.028 ±1.0	0.035 ±1.1	0.042 ±1.3	0.024 ±1.2	0.028 ±1.6	0.032 ±1.2
CCSSL-prop	**0.022** ±2.1	**0.031** ±1.5	0.043 ±1.7	**0.024** ±2.7	**0.029** ±3.4	**0.039** ±1.9	**0.014** ±1.1	**0.020** ±1.1	**0.024** ±0.3	**0.021** ±1.1	**0.023** ±2.7	**0.029** ±1.5

In the context of self-training, a lower error rate is desirable to obtain less noisy and more informative self-supervision. As shown in Table 2, the two variants of CCSSL indeed achieve a consistent improvement in the validity of the pseudo-labels over CSSL in terms of the error rate, although the supervision provided by the latter is already quite accurate and does not violate the strong validity property for these instances either. Moreover, CCSSL-prop leads to more accurate sets compared to CCSSL-diff. Note that in standard conformal prediction one is typically interested in error rates that match the respective significance levels. In the setting considered here, this is undermined by relatively small calibration sets $\mathcal{D}_{\text{calib}}$, as well as the fact that the unlabeled instances, whose validity is presented here, have already been observed in previous training iterations, thus leading to optimistic error rates. Nevertheless, this optimism turns out to be beneficial for effective self-training, as our empirical results confirm.
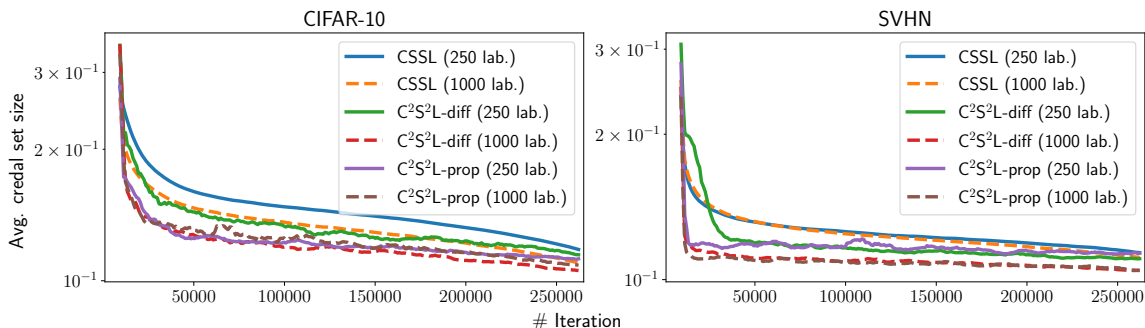

Figure 3: The credal set efficiency in terms of the mean possibilities π for each class, averaged over all pseudo-labels in each iteration for all 3 seeds.

Furthermore, the efficiency of the credal sets, i.e., their sizes, has an effect on the learning behavior. Fig. 3 compares the credal set sizes of CSSL and CCSSL over the course of the training. As can be seen, CCSSL constructs smaller credal sets, which in combination with the improved validity leads to a more effective supervision. Moreover, the credal sets sizes decrease with higher numbers of labels, which is due to a higher prediction quality involved in the credal set construction. Apart from the results on CIFAR-10 with 250 labels, one can further see that the credal set size deviates only slightly between the two non-conformity measures employed in CCSSL.

6. Conclusion

In the context of semi-supervised learning, previous pseudo-labeling approaches lack validity guarantees for the quality of the pseudo-supervision. Such methods often suffer from misleading supervision, leaving much potential unused. In our work, we address these shortcomings by a conformal credal labeling approach leveraging the framework of conformal prediction, which entails validity guarantees for the constructed credal pseudo-labels. Our empirical study confirms the adequacy of this approach when combined with consistency regularization in terms of generalization performance, as well as the calibration and effi-

ciency of pseudo-labels compared to previous methods. At the same time, the combination of a rigorously studied uncertainty quantification framework with pseudo-labeling paves the way for more thorough theoretical analyses in the field of self-training in semi-supervised learning.

In future work, we plan to investigate approaches to achieve *conditional* (per-instance) validity, i.e., conformal credal labels specifically tailored to (conditioned on) the query instance x_{N+1}, which would lead to even stronger guarantees for the quality of pseudo-labels. Approximations of this type of validity for conformal prediction have already been provided (Vovk, 2013; Bellotti, 2021). Moreover, a more rigorous analysis of the efficiency of alternative non-conformity scores needs to be performed, including their robustness to label noise as typically present in real-world data.

Acknowledgments

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *Proc. of the International Joint Conference on Neural Networks, IJCNN, Glasgow, United Kingdom, July 19-24*, pages 1–8. IEEE, 2020.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, NIPS, Montreal, Quebec, Canada, December 8-13*, pages 3365–3373, 2014.

Sohail Bahmani and Bhiksha Raj. A unifying analysis of projected gradient descent for ℓ_p-constrained least squares. *CoRR*, abs/1107.4623, 2011.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2014. ISBN 0123985374.

Anthony Bellotti. Approximation to object conditional validity with inductive conformal predictors. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Conformal and Probabilistic Prediction and Applications, virtual, September 8-10*, volume 152 of *Proceedings of Machine Learning Research*, pages 4–23. PMLR, 2021.

David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS, Vancouver, BC, Canada, December 8-14*, pages 5050–5060, 2019.

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Proc. of the 8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30*. OpenReview.net, 2020.

Vivien Cabannes, Alessandro Rudi, and Francis R. Bach. Structured prediction with partial labelling through the infimum loss. In *Proceedings of the 37th International Conference on Machine Learning, ICML, virtual, July 13-18*, volume 119, pages 1230–1239. PMLR, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.

Leonardo Cella and Ryan Martin. Valid inferential models for prediction in supervised learning problems. In *International Symposium on Imprecise Probability: Theories and Applications, ISIPTA, Granada, Spain, July 6-9*, volume 147 of *Proceedings of Machine Learning Research*, pages 72–82. PMLR, 2021.

Leonardo Cella and Ryan Martin. Validity, consonant plausibility measures, and conformal prediction. *Int. J. Approx. Reason.*, 141:110–130, 2022.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 2006.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proc. of the 14th International Conference on Artificial Intelligence and Statistics, AISTATS, Fort Lauderdale, FL, USA, April 11-13*, volume 15 of *JMLR Proceedings*, pages 215–223. JMLR.org, 2011.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12*, 2020.

M. Delgado and S. Moral. On the concept of possibility-probability consistency. *Fuzzy Sets and Systems*, 21(3):311–318, 1987. ISSN 0165-0114.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proc. of the IEEE International Conference on Computer Vision, ICCV, Santiago, Chile, December 7-13*, pages 1422–1430. IEEE Computer Society, 2015.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5408–5418. Association for Computational Linguistics, 2021.

Didier Dubois and Henri Prade. Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of Mathematics and Artificial Intelligence*, 32:35–66, 2004.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proc. of the 33nd International Conference on Machine Learning, ICML, New York City, NY, USA, June 19-24*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. DropBlock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS, Montreal, Quebec, Canada, December 3-8*, pages 10750–10760, 2018.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proc. of the IEEE/CVF International Conference on Computer Vision, ICCV, Seoul, Korea (South), October 27 - November 2*, pages 3827–3837. IEEE, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12*, 2020.

Eyke Hüllermeier and Weiwei Cheng. Superset learning based on generalized loss minimization. In *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD, Porto, Portugal, September 7-11, Proc. Part II*, volume 9285 of *LNCS*, pages 260–275. Springer, 2015.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506, 2021.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proc. of the 30th International Conference on Machine Learning, ICML, Atlanta, GA, USA, June 16-21*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 427–435. JMLR.org, 2013.

Ulf Johansson, Ernst Ahlberg, Henrik Boström, Lars Carlsson, Henrik Linusson, and Cecilia Sönströd. Handling small calibration sets in mondrian inductive conformal regressors. In *Proc. of the Statistical Learning and Data Sciences - Third International Symposium, SLDS, Egham, UK, April 20-23*, volume 9047 of *Lecture Notes in Computer Science*, pages 271–280. Springer, 2015.

Byol Kim, Chen Xu, and Rina Foygel Barber. Predictive inference is free with the jackknife+-after-bootstrap. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12, 2020.

Kyungyul Kim, Byeongmoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6547–6556. IEEE, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Canada, 2009.

Tomáš Kroupa. How many extreme points does the set of probabilities dominated by a possibility measure have. In *Proc. of 7th Workshop on Uncertainty Processing WUPES*, volume 6, pages 89–95, 2006.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, International Conference on Machine Learning, ICML, Atlanta, GA, USA, June 16-21*, volume 3, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Yan Li, Xiaofeng Cao, and Honghui Chen. Fully projection-free proximal stochastic gradient method with optimal convergence rates. *IEEE Access*, 8:165904–165912, 2020.

Julian Lienen and Eyke Hüllermeier. Credal self-supervised learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-14*, pages 14370–14382, 2021.

Julian Lienen and Eyke Hüllermeier. From label smoothing to label relaxation. In *Proc. of the 35th AAAI Conference on Artificial Intelligence, virtual, February 2-9*, 2021.

Julian Lienen and Eyke Hüllermeier. Instance weighting through data imprecisiation. *Int. J. Approx. Reason.*, 134:1–14, 2021. ISSN 0888-613X.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. Uncertainty-aware self-training for few-shot text classification. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12*, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS, Vancouver, BC, Canada, December 8-14*, pages 4696–4705, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *Advances*

in *Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NIPS, Granada, Spain, November 12-17, Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS, Montreal, Quebec, Canada, December 3-8*, pages 3239–3250, 2018.

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Harris Papadopoulos, Volodya Vovk, and Alexander Gammerman. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence, ICTAI, Patras, Greece, October 29-31, Volume 2*, pages 388–395. IEEE Computer Society, 2007.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11557–11568. Computer Vision Foundation / IEEE, 2021.

Zhongzheng Ren, Raymond A. Yeh, and Alexander G. Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12*, 2020.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *Proc. of the 9th International Conference on Learning Representations, ICLR, virtual, May 3-7*. OpenReview.net, 2021.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *Proc. of the 7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing, WACV/MOTION, Breckenridge, CO, USA, January 5-7*, pages 29–36. IEEE Computer Society, 2005.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, NIPS, Barcelona, Spain, December 5-10*, pages 1163–1171, 2016.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, 2008.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12*, 2020.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA, June 27-30*, pages 2818–2826. IEEE Computer Society, 2016.

Jafar Tanha, Maarten van Someren, and Hamideh Afsarmanesh. Semi-supervised self-training for decision tree classifiers. *Int. J. Mach. Learn. Cybern.*, 8(1):355–370, 2017. ISSN 1868-808X.

Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Mach. Learn.*, 109(2):373–440, 2020.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92 (2-3):349–376, 2013.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *9th International Conference on Learning Representations, ICLR, Virtual Event, Austria, May 3-7*. OpenReview.net, 2021.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12*, 2020a.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. Computer Vision Foundation / IEEE, 2020b.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proc. of the British Machine Vision Conference, BMVC, York, UK, September 19-22*, 2016.

Gianluca Zeni, Matteo Fontana, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *CoRR*, abs/2005.07972, 2020.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-14*, pages 18408–18419, 2021.

Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *Int. J. Comput. Vis.*, 129(4):1106–1120, 2021.