Disambiguation of Weak Supervision leading to Exponential Convergence rates

Vivien Cabannes [1] **Francis Bach** [1] **Alessandro Rudi** [1]

Abstract

Machine learning approached through supervised learning requires expensive annotation of data. This motivates weakly supervised learning, where data are annotated with incomplete yet discriminative information. In this paper, we focus on partial labelling, an instance of weak supervision where, from a given input, we are given a set of potential targets. We review a disambiguation principle to recover full supervision from weak supervision, and propose an empirical disambiguation algorithm. We prove exponential convergence rates of our algorithm under classical learnability assumptions, and we illustrate the usefulness of our method on practical examples.

1. Introduction

In many applications of machine learning, such as recommender systems, where an input x characterizing a user should be matched with a target y representing an ordering of a large number m of items, accessing fully supervised data (x, y) is not an option. Instead, one should expect weak information on the target y, which could be a list of previously taken (if items are online courses), watched (if items are plays), *etc.*, items by a user characterized by the feature vector x. This motivates *weakly supervised learning*, aiming at learning a mapping from inputs to targets in such a setting where tools from supervised learning can not be applied off-the-shelves.

Recent applications of weakly supervised learning showcase impressive results in solving complex tasks such as action retrieval on instructional videos (Miech et al., 2019), image semantic segmentation (Papandreou et al., 2015), salient object detection (Wang et al., 2017), 3D pose estimation (Dabral et al., 2018), text-to-speech synthesis (Jia et al., 2018), to name a few. However, those applications of weakly supervised learning are usually based on clever heuristics, and theoretical foundations of learning from weakly supervised data are scarce, especially when compared to statistical learning literature on supervised learning (Vapnik, 1995; Boucheron et al., 2005; Steinwart & Christmann, 2008). We aim to provide a step in this direction.

In this paper, we focus on partial labelling, a popular instance of weak supervision, approached with a structured prediction point of view (Ciliberto et al., 2020). We detail this setup in Section 2. Our contributions are organized as follows.

- In Section 3, we introduce a disambiguation algorithm to retrieve fully supervised samples from weakly supervised ones, before applying off-the-shelf supervised learning algorithms to the completed dataset.
- In Section 4, we prove exponential convergence rates of our algorithm, in term of the fully supervised excess of risk, given classical learnability assumptions.
- In Section 5, we explain why disambiguation algorithms are intrinsically non-convex, and provide guidelines based on well-grounded heuristics to implement our algorithm.

We end this paper with a review of literature in Section 6, before showcasing the usefulness of our method on practical examples in Section 7, and opening on perspectives in Section 8.

2. Disambiguation of Partial Labelling

In this section, we review the supervised learning setup, introduce the partial labelling problem along with a principle to tackle this instance of weak supervision.

Algorithms can be formalized as mapping an input x to a desired output y, respectively belonging to an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Machine learning consists in automating the design of the mapping $f : \mathcal{X} \to \mathcal{Y}$, based on a joint distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ over input/output pairings (x, y) and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, measuring the error cost of outputting $f(x)$ when one should have output y. The optimal mapping is defined as satisfying

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\mu}\left[\ell(f(X), Y)\right]. \qquad (1)$$

In *supervised learning*, it is assumed that one does not have

[1] Institut National de Recherche en Informatique et en Automatique – Département d'Informatique de l'École Normale Supérieure – PSL Research University. Correspondence to: Vivien Cabannes <vivien.cabannes@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

access to the full distribution μ, but only to independent samples $(X_i, Y_i)_{i\leq n} \sim \mu^{\otimes n}$. In practice, accessing such samples means building a dataset of examples. While input data (x_i) are usually easily accessible, getting output pairings (y_i) generally requires careful annotation, which is both time-consuming and expensive. For example, in image classification, (x_i) can be collected by scrapping images over the Internet. Subsequently a "data labeller" might be ask to recognize a rare feline y_i on an image x_i. While getting y_i will be hard in this setting, recognizing that it is a feline and describing elements of color and shape is easy, and already helps to determine what outputs $f(x_i)$ are acceptable. A second example is given when pooling a known population (x_i) to get estimation of their political orientation (y_i), one might get information from recent election of percentage of voters across the political landscape, leading to global constraints that (y_i) should verify. A supervision that gives information on $(y_i)_{i\leq n}$ without giving its precise value is called *weak supervision*.

Partial labelling, also known as "superset learning", is an instance of weak supervision, in which, for an input x, we do not access the precise label y but only a set s of potential labels, $y \in s \subset \mathcal{Y}$. For example, on a caracal image x, one might not get the label "caracal" y, but the set s "feline", containing all the labels y corresponding to felines. It is modelled through a distribution $\nu \in \Delta_{\mathcal{X}\times 2^{\mathcal{Y}}}$ over $\mathcal{X} \times 2^{\mathcal{Y}}$ generating samples (X, S), which should be compatible with the fully supervised distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ as formalized by the following definition.

Definition 1 (Compatibility, Cabannes et al. (2020)). *A fully supervised distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ is* compatible *with a weakly supervised distribution $\nu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, denoted by $\mu \vdash \nu$ if there exists an underlying distribution $\pi \in \Delta_{\mathcal{X}\times\mathcal{Y}\times 2^{\mathcal{Y}}}$, such that μ, and ν, are the respective marginal distributions of π over $\mathcal{X}\times\mathcal{Y}$ and $\mathcal{X}\times 2^{\mathcal{Y}}$, and such that $y \in s$ for any tuple (x, y, s) in the support of π (or equivalently $\pi|_s \in \Delta_s$, with $\pi|_s$ denoting the conditional distribution of π given s).*

This definition means that a weakly supervised sample $(X, S) \sim \nu$ can be thought as proceeding from a fully supervised sample $(X, Y) \sim \mu$ after loosing information on Y according to the sampling of $S \sim \pi|_{X,Y}$. The goal of partial labelling is still to learn f^* from Eq. (1), yet without accessing a fully supervised distribution $\mu \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ but only the weakly supervised distribution $\nu \in \Delta_{\mathcal{X}\times 2^{\mathcal{Y}}}$. As such, this is an ill-posed problem, since ν does not discriminate between all μ compatible with it. Following *lex parsimoniae*, Cabannes et al. (2020) have suggested to look for μ such that the labels are the most deterministic function of the inputs, which they measure with a loss-based "variance", leading to the disambiguation

$$\mu^* \in \operatorname*{arg\,min}_{\mu\vdash\nu} \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\mu}\left[\ell(f(X), Y)\right], \qquad (2)$$

and to the definition of the optimal mapping $f^* : \mathcal{X} \to \mathcal{Y}$

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\mu^*}\left[\ell(f(X), Y)\right]. \qquad (3)$$

This principle is motivated by Theorem 1 of Cabannes et al. (2020) showing that f^* in Eq. (3) is characterized by $f^* \in \operatorname{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,S)\sim\nu}\left[\inf_{y\in S} \ell(f(X), y)\right]$, matching a prior formulation based on infimum loss (Cour et al., 2011; Luo & Orabona, 2010; Hüllermeier, 2014). In practice, it means that if $(S|X = x)$ has probability 50% to be the set "feline" and 50% the set "orange with black stripes", $(Y|X = x)$ should be considered as 100% "tiger", rather than 20% "cat", 30% "lion" and 50% "orange car with black stripes", which could also explain $(S|X = x)$. In other terms, Eq. (2) creates consensus between the different information provided on a label. Similarly to supervised learning, partial labelling consists in retrieving f^* without accessing ν but only samples $(X_i, S_i)_{i\leq n} \sim \nu^{\otimes n}$.

Remark 2 (Measure of determinism). *Eq.* (2) *is not the only variational way to push towards distribution where labels are deterministic function of the inputs. For example, one could minimize entropy (*e.g., Berthelot et al., 2019; Lienen & Hüllermeier, 2021*). However, a loss-based principle is appreciable since the loss usually encodes structures of the output space (Ciliberto et al., 2020), which will allow sample and computational complexity of consequent algorithms to scale with an intrinsic dimension of the space rather than the real one, e.g., m rather than $m!$ when $\mathcal{Y} = \mathfrak{S}_m$ and ℓ is a suitable ranking loss (see Section 5.4 or Nowak-Vila et al., 2019).*

3. Learning Algorithm

In this section, given weakly supervised samples, we present a disambiguation algorithm to retrieve fully supervised samples based on an empirical expression of Eq. (2), before learning a mapping from $\mathcal{X}$ to $\mathcal{Y}$ based on those fully supervised samples, according to Eq. (3).

Given a partially labelled dataset $\mathcal{D}_n = (x_i, s_i)_{i\leq n}$, sampled accordingly to $\nu^{\otimes n}$, we retrieve fully supervised samples, based on the following empirical version of Eq. (2), with $C_n = \prod_{i\leq n} s_i \subset \mathcal{Y}^n$

$$(\hat{y}_i)_{i\leq n} \in \operatorname*{arg\,min}_{(y_i)_{i\leq n}\in C_n} \inf_{(z_i)_{i\leq n}\in\mathcal{Y}^n} \sum_{i,j=1}^{n} \alpha_j(x_i)\ell(z_i, y_j), \qquad (4)$$

where $(\alpha_i(x))_{i\leq n}$ is a set of weights measuring how much one should base its prediction for x on the observations made at x_i. This formulation is motivated by the Bayes approximate rule proposed by Stone (1977), which can be seen as the approximation of μ by $n^{-1}\sum_{i,j=1}^{n} \alpha_j(x_i)\delta_{x_i}\otimes\delta_{y_j}$ in Eq. (2). In essence, z_i (which corresponds to $f(x_i)$) is likely to be y_i, although Eq. (4) allows for flexibility to avoid rigid interpolation. As such Eq. (4) should be understood as constraining $y_i \in s_i$ to be similar to $y_j \in s_j$ if x_i and x_j are close, with the measure of similarity defined by $\alpha_i(x_j)$.

Once fully supervised samples $(x_i, \hat{y}_i)$ have been recollected, one can learn $f_n : \mathcal{X} \to \mathcal{Y}$, approximating f^*, with classical supervised learning techniques. In this work, we will consider the structured prediction estimator introduced by Ciliberto et al. (2016), defined as

$$f_n(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x)\ell(z, \hat{y}_i). \tag{5}$$

Weighting scheme α. For the weighting scheme α, several choices are appealing. Laplacian diffusion is one of them as it incorporates a prior on low density separation to boost learning (Zhu et al., 2003; Zhou et al., 2003; Bengio et al., 2006; Hein et al., 2007). Kernel ridge regression is another due to its theoretical guarantees (Ciliberto et al., 2020). In the theoretical analysis, we will use nearest neighbors. Assuming $\mathcal{X}$ is endowed with a distance d, and assuming, for readability sake, that ties to define nearest neighbors do not happen, it is defined as

$$\alpha_i(x) = \begin{cases} k^{-1} & \text{if} \quad \sum_{j=1}^{n} \mathbf{1}_{d(x,x_j) \le d(x,x_i)} \le k \\ 0 & \qquad \text{otherwise,} \end{cases}$$

where k is a parameter fixing the number of neighbors. Our analysis, leading to Theorem 4, also holds for other local averaging methods such as partitioning or Nadaraya-Watson estimators.

4. Consistency Result

In this section, we assume $\mathcal{Y}$ finite, and prove the convergence of f_n towards f^* as n, the number of samples, grows to infinity. To derive such a consistency result, we introduce a surrogate problem that we relate to the risk through a calibration inequality. We later assume that weights are given by nearest neighbors and review classical assumptions, that we work to derive exponential convergence rates.

In the following, we are interested in bounding the expected generalization error, defined as

$$\mathcal{E}(f_n) = \mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*), \tag{6}$$

where $\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\mu^*}\left[\ell(f(X), Y)\right]$, by a quantity that goes to zero, when n goes to infinity. This implies convergence in probability (the randomness being inherited from $\mathcal{D}_n$) of $\mathcal{R}(f_n)$ towards $\inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f)$, which is referred as *consistency* of the learning algorithm. We first introduce a few objects.

Disambiguation ground truth (y_i^*). Introduce $\pi^* \in \Delta_{\mathcal{X}\times\mathcal{Y}\times 2^{\mathcal{Y}}}$ expressing the compatibility of μ^* and ν as in Definition 1. Given samples $(x_i, s_i)_{i \le n}$ forming a dataset $\mathcal{D}_n$, we enrich this dataset by sampling $y_i^* \sim \pi^*|_{x_i, s_i}$, which build an underlying dataset (x_i, y_i^*, s_i) sampled accordingly $(\pi^*)^{\otimes n}$. Given $\mathcal{D}_n$, while *a priori*, y_i^* are random variables, sampled accordingly to $\pi^*|_{x_i,s_i}$, because of the definition of μ^* (2), under basic definition assumptions, they are actually deterministic, defined as $y_i^* = \arg\min_{y \in s_i} \ell(f^*(x_i), y)$. As such, they should be seen as ground truth for $\hat{y}_i$.

Surrogate estimates. The approximate Bayes rule was successfully analyzed recently through the prism of plug-in estimators by Ciliberto et al. (2020). While we will not cast our algorithm as a plug-in estimator, we will leverage this surrogate approach, introducing two mappings φ and ψ from $\mathcal{Y}$ to an Hilbert space $\mathcal{H}$ such that

$$\forall\, z, y \in \mathcal{Y}, \qquad \ell(z, y) = \langle \psi(z), \varphi(y) \rangle, \tag{7}$$

Such mappings always exist when $\mathcal{Y}$ is finite, and have been used to encode "problem structure" defined by the loss ℓ (Nowak-Vila et al., 2019). Note that many losses (*e.g.* Hamming, Spearman, Kendall in ranking) can be written as correlation losses which corresponds to $\psi = -\varphi$, yet Eq. (7) allows to model much more losses, especially asymmetric losses (*e.g.* discounted cumulative gain). We introduce three surrogate quantities that will play a major role in the following analysis, they map $\mathcal{X}$ to $\mathcal{H}$ as

$$g^*(x) = \mathbb{E}_{\mu^*}\left[\varphi(Y) \,|\, X = x\right], \qquad g_n(x) = \sum_{i=1}^{n} \alpha_i(x)\varphi(\hat{y}_i),$$
$$g_n^*(x) = \sum_{i=1}^{n} \alpha_i(x)\varphi(y_i^*). \tag{8}$$

It is known that f^* and f_n are retrieved from g^* and g_n, through the decoding, retrieving $f : \mathcal{X} \to \mathcal{Y}$ from $g : \mathcal{X} \to \mathcal{H}$ as

$$f(x) = \operatorname*{arg\,min}_{z \in \mathcal{Y}} \langle \psi(z), g(x) \rangle, \tag{9}$$

which explains the wording of *plug-in* estimator (Ciliberto et al., 2020). We now introduce a *calibration inequality*, that relates the error between f_n and f^* with surrogate error quantities.

Lemma 3 (Calibration inequality). *When $\mathcal{Y}$ is finite, and the labels are a deterministic function of the input,* i.e.*, when $\mu^*|_x$ is a Dirac for all $x \in \operatorname{supp} \nu_{\mathcal{X}}$, for any weighting scheme that sums to one,* i.e.*, $\sum_{i=1}^{n} \alpha_i(x) = 1$ for all $x \in \operatorname{supp} \nu_{\mathcal{X}}$,*

$$\begin{aligned} \mathcal{R}(f_n) - \mathcal{R}(f^*) \le\ & 4c_\psi \left\| g_n^* - g_n \right\|_{L^1} \\ & + 8c_\psi c_\varphi \mathbb{P}_X\left(\left\| g_n^*(X) - g^*(X) \right\| > \delta\right), \end{aligned} \tag{10}$$

with $c_\psi = \sup_{z\in\mathcal{Y}} \|\psi(z)\|$, $c_\varphi = \sup_{z\in\mathcal{Y}} \|\varphi(y)\|$, and δ a parameter that depend on the geometry of ℓ and its decomposition through φ.

This lemma, proven in Appendix A.1, separates a part reading in $\left\|g_n - g_n^*\right\|$, due to the *disambiguation error* between $(\hat{y}_i)$ and (y_i^*) together with the *stability* of the learning algorithm when substituting $(\hat{y}_i)$ for (y_i^*), and a part in $\left\|g_n^* - g^*\right\|$ due to the *consistency* of the fully supervised learning algorithm. The expression of the first part relates to Theorem 7 in Ciliberto et al. (2020) while the second part relates to Theorem 6 in Cabannes et al. (2021).

4.1. Classical learnability assumptions

In the following, we suppose that the weights α are given by nearest neighbors, that $\mathcal{X}$ is a compact metric space endowed with a distance d, that $\mathcal{Y}$ is finite and that ℓ is proper in the sense that it strictly positive except on the diagonal of $\mathcal{Y}\times\mathcal{Y}$ diagonal where it is zero. We now review classical assumptions to prove consistency. First, assume that $\nu_{\mathcal{X}}$ is regular in the following sense.

Assumption 1 ($\nu_{\mathcal{X}}$ well-behaved)**.** *Assume that $\nu_{\mathcal{X}}$ is such that there exists $h_1, c_\mu, q > 0$ satisfying, with $\mathcal{B}$ designing balls in $\mathcal{X}$,*

$$\forall x \in \operatorname{supp}\nu_{\mathcal{X}},\ \forall r < h_1, \qquad \nu_{\mathcal{X}}(\mathcal{B}(x,r)) > c_\mu r^q.$$

Assumption 1 is useful to make sure that neighbors in $\mathcal{D}_n$ are closed with respect to the distance d, it is usually derived by assuming that $\mathcal{X}$ is a subset of $\mathbb{R}^q$; that $\nu_{\mathcal{X}}$ has a density p against the Lebesgue measure λ with *minimal mass* $p_{\min}$ in the sense that for any $x \in \operatorname{supp}\nu_{\mathcal{X}}$, $p(x) > p_{\min}$; and that $\operatorname{supp}\nu_{\mathcal{X}}$ has regular boundary in the sense that $\lambda(\mathcal{B}(x,r) \cap \operatorname{supp}\nu_{\mathcal{X}}) \geq c\lambda(\mathcal{B}(x,r)$ for any $x \in \operatorname{supp}\nu_{\mathcal{X}}$ and $r < h$ (*e.g.,* *Audibert & Tsybakov, 2007*).

We now switch to a classical assumption in partial labelling, allowing for population disambiguation.

Assumption 2 (Non ambiguity, Cour et al. (2011))**.** *Assume the existence of $\eta \in [0,1)$, such that for any $x \in \operatorname{supp}\nu_{\mathcal{X}}$, there exists $y_x \in \mathcal{Y}$, such that $\mathbb{P}_\nu\left(y_x \in S \,|\, X = x\right) = 1$, and*

$$\forall z \neq y_x, \quad \mathbb{P}_\nu\left(z \in S \,|\, X = x\right) \leq \eta.$$

Assumption 2 states that when given the full distribution ν, there is one, and only one, label that is coherent with every observable sets for a given input. It is a classical assumption in literature about the learnability of the partial labelling problem (*e.g.,* *Liu & Dietterich, 2014*). When ℓ is proper, this implies that $\mu^*|_x = \delta_{y_x}$, and $f^*(x) = y_x$.

Finally, we assume that g^* is regular. As we are considering local averaging method, we will use Lipschitz-continuity, which is classical in such a setting.[1]

Assumption 3 (Regularity of g^*)**.** *Assume that there exists $c_g > 0$, such that for any $x, x' \in \mathcal{X}$, we have*

$$\|g^*(x) - g^*(x')\|_{\mathcal{H}} \leq c_g d(x,x').$$

It should be noted that regularity of g^*, Assumption 3, together with determinism of $\mu^*|_x$ inherited from Assumption 2 implies that classes $\mathcal{X}_y = \{x \,|\, f^*(x) = y\}$ are separated in $\mathcal{X}$, in the sense that there exists $h_2 > 0$, such that for any $y, y' \in \mathcal{Y}$ and $(x,x') \in \mathcal{X}_y \times \mathcal{X}_{y'}$, $d(x,x') > h_2$, which is a classical assumption to derive consistency of semi-supervised learning algorithm (*e.g.,* *Rigollet, 2007*). Those implications results from the fact that separation in $\mathcal{Y}$ (hard Tsybakov condition) plus Lipschitzness of g^* implies separation of classes in $\mathcal{X}$, as we details in Appendix A.2.

4.2. Exponential convergence rates

We are now ready to state our convergence result. We introduce $h = \min(h_1, h_2)$ and $p = c_\mu h^q$, so that for any $x \in \operatorname{supp}\nu_{\mathcal{X}}$, $\nu_{\mathcal{X}}(\mathcal{B}(x,h)) > p$.

Theorem 4 (Exponential convergence rates)**.** *When the weights α are given by nearest neighbors, under Assumptions 1, 2 and 3, the excess of risk in Eq.* (6) *is bounded by*

$$\mathcal{E}(f_n) \leq 8c_\psi c_\varphi (n+1)\exp\left(-\frac{np}{16}\right) + 8c_\psi c_\varphi m \exp\left(-k\left|\log(\eta)\right|\right), \qquad (11)$$

as soon as $k < np/4$, with $m = \#\mathcal{Y}$. By taking $k_n = k_0 n$, for $k_0 < p/4$, this implies exponential convergence rates $\mathcal{E}(f_n) = O(n\exp(-n))$.

Sketch for Theorem 4. In essence, based on Lemma 3, Theorem 4 can be understood as two folds.

- A fully supervised error between g_n^* and g^*. This error can be controlled in $\exp(-np)$ as the non-ambiguity assumption implies a hard Tsybakov margin condition, a setting in which *the fully supervised estimate g_n^* is known to converge to the population solution g^* with such rates* (Cabannes et al., 2021).
- A weakly disambiguation error, that is exponential too, since, based on Assumption 2, disambiguating between $z \in \mathcal{Y}$ and y_x from k sets S sampled accordingly to $\nu|_x$ can be done in η^k, and disambiguating between all $z \neq y_x$ and y_x in $m\eta^k = m\exp(-k\left|\log(\eta)\right|)$.

Appendix A.3 provides details. □

Theorem 4 states that under a non-ambiguity assumption and a regularity assumption implying no-density separation, one can expect exponential convergence rates of f_n learned with weakly supervised data to f^* the solution of the fully supervised learning problem, measured with excess of fully supervised risk. Because of the exponential convergence, we could derive polynomial convergence rates for a broader class of problems that are approximated by problems satisfying assumptions of Theorem 4. *The derived rates in $n\exp(-n)$ should be compared with rates*

[1] Its generalization through Hölder-continuity would work too.

in $n^{-1/2}$ and $n^{-1/4}$, respectively derived, *under the same assumptions*, by Cour et al. (2011); Cabannes et al. (2020).

4.3. Discussion on Assumptions

While we have retaken classical assumptions from literature, those assumptions are quite strong, which allows us, by understanding their strength, to derive exponential convergence rates. Assumptions 1 and 3 are classical in the nearest neighbor literature with full supervision. If we were using (reproducing) kernel methods to define the weighting scheme α, those assumptions would be mainly replaced with "g^* belonging to the RKHS". Assumption 2 is the strongest assumption in our view, that we will now discuss.

How to check it in practice ? First, for Assumption 2 to hold, the labels have to be a deterministic function of the inputs. In other words, a zero error is achievable. Finally, Assumption 2 is related to dataset collection. If dealing with images, weak supervision could take the form of some information on shape, color, or texture, etc., Assumption 2 supposes that the weak information potentially given on a specific image x, allows to retrieve the unique label y of the image (*e.g.*, a "pig" could be recognized from its shape and its color). This is a reasonable assumption, if, for a given x, we ask at random a data labeller to provide us information on shape, color, or texture, etc. However, it will not be the case, if for some reasons (*e.g.* the dataset is built from several weakly annotated datasets), in some regions of the input space, we only get shape information, and in other regions, we only get color information. In particular, it is not verified for semi-supervised learning when the support of the unlabelled data distribution is not the same as the support of the labelled input data distribution.

How to relax it and what results to expect? Previous works used Assumption 2 to derive a calibration inequality between the infimum loss to the original loss (*e.g.*, see Proposition 2 by Cabannes et al., 2020). In contrast, we relate the surrogate and original problem through a refined calibration inequality (10). This technical progress allows us to derive exponential convergence rates similarly to the work of Cabannes et al. (2021). Importantly, in comparison with previous work, our calibration inequality Lemma 3 can easily be extended without the determinism assumption provided by Assumption 2. Essentially, in our work, Assumption 2 is used to simplify the study of $(\hat{y}_i)_{i\leq n}$ given by the disambiguation algorithm (4), and therefore the study of the disambiguation error in Eq. (10). The study of $(\hat{y}_i)_{i\leq n}$ without Assumption 2 would requires other tools than the one presented in this paper. It could be studied in the realm of graphical model and message passing algorithm, or with Wasserstein distance and topological considerations on measures. With much milder forms of Assumption 2, we expect the rates to degrade smoothly with respect to a parameter defining the hardness of the problem, similarly to the works of Audibert & Tsybakov (2007); Cabannes et al. (2021).

5. Optimizaton Considerations

In this section, we focus on implementations to solve Eq. (4). We explain why disambiguation objectives, such as Eq. (2) are intrinsically non-convex and express a heuristic strategy to solve Eq. (4) besides non-convexity in classical well-behaved instances of partial labelling. Note that we do not study implementations to solve Eq. (5) as this study has already been done by Nowak-Vila et al. (2019). We end this section by considering a practical example to make derivations more concrete.

5.1. Non-convexity of disambiguation objectives

For readability, suppose that $\mathcal{X}$ is a singleton, justifying to remove the dependency on the input in the following. Consider $\nu \in \Delta_{2^\mathcal{Y}}$ a distribution modelling weak supervision. While the domain $\{\mu \in \Delta_\mathcal{Y} \,|\, \mu \vdash \nu\}$ is convex, a disambiguation objective $\mathcal{E} : \Delta_\mathcal{Y} \to \mathbb{R}$ defining $\mu^* \in \arg\min_{\mu\vdash\nu} \mathcal{E}(\mu)$, similarly to Eq. (2), that is minimized for deterministic distributions, which correspond to μ a Dirac, *i.e.*, minimized on vertices of its definition domain $\Delta_\mathcal{Y}$, can not be convex. In other terms, any disambiguation objective that pushes toward distributions where targets are deterministic function of the input, as mentioned in Remark 2, can not be convex.

Indeed, smooth disambiguation objectives such as entropy and our piecewise linear loss-based principle (2), reading pointwise $\mathcal{E}(\mu) = \inf_{z\in\mathcal{Y}} \mathbb{E}_{Y\sim\mu}[\ell(z, Y)]$, are concave. Similarly, its quadratic variant $\mathcal{E}'(\mu) = \mathbb{E}_{Y,Y'\sim\mu}[\ell(Y, Y')]$, is concave as soon as $(\ell(y, y'))_{y,y'\in\mathcal{Y}}$ is semi-definite negative. We illustrate those considerations on a concrete example with graphical illustration in Appendix C. We should see how this translates on generic implementations to solve the empirical objective (4).

5.2. Generic implementation for Eq. (4)

Depending on ℓ and on the type of observed set (s_i), Eq. (4) might be easy to solve. In the following, however, we will introduce optimization considerations to solve it in a generic structured prediction fashion. To do so, we recall the decomposition of ℓ (7) and rewrite Eq. (4) as

$$(\hat{y}_i)_{i\leq n} \in \underset{(y_i)\in C_n}{\arg\min} \inf_{(z_i)\in\mathcal{Y}^n} \sum_{i,j=1}^{n} \alpha_j(x_i)\psi(z_i)^\top \varphi(y_j).$$

Since, given (y_j), the objective is linear in $\psi(z_j)$, the constraint $\psi(z_j) \in \psi(\mathcal{Y})$ can be relaxed with $\zeta_i \in \operatorname{Conv}\psi(\mathcal{Y})$.[2]

[2] The minimization pushes towards extreme points of the definition domain.

Similarly, with respect to $\varphi(y_j)$, this objective is the infimum of linear functions, therefore is concave, and the constraint $\varphi(y_j) \in \varphi(s_j)$, could be relaxed with $\xi_i \in \operatorname{Conv}\varphi(s_j)$. Hence, with $\mathcal{H}_0 = \operatorname{Conv}\psi(\mathcal{Y})$ and $\Gamma_n = \prod_{j\le n}\operatorname{Conv}\varphi(s_j)$, the optimization is cast as

$$(\hat{\xi}_i)_{i\le n} \in \operatorname*{arg\,min}_{(\xi_i)\in\Gamma_n} \inf_{(\zeta_i)\in\mathcal{H}_0^n} \sum_{i,j=1}^{n} \alpha_j(x_i)\zeta_i^\top \xi_j. \qquad (12)$$

Because of concavity, $(\hat{\xi}_i)$ will be an extreme point of Γ_n, that could be decoded into $\hat{y}_i = \varphi^{-1}(\hat{\xi}_i)$. However, it should be noted that if only interested in f_n and not in the disambiguation $(\hat{y}_i)$, this decoding can be avoided, since Eq. (5) can be rewritten as $f_n(x) \in \operatorname{arg\,min}_{z\in\mathcal{Y}} \psi(z)^\top \sum_{i=1}^n \alpha_i(x)\hat{\xi}_i$.

5.3. Alternative minimization with good initialization

To solve Eq. (12), we suggest to use an alternative minimization scheme. The output of such an scheme is highly dependent to the variable initialization. In the following, we introduce well-behaved problem, where $(\xi_i)_{i\le n}$ can be initialized smartly, leading to an efficient implementation to solve Eq. (12).

Definition 5 (Well-behaved partial labelling problem)**.** *A partial labelling problem (ℓ, ν) is said to be well-behaved if for any $s \in \operatorname{supp}\nu_{2^\mathcal{Y}}$, there exists a signed measure μ_s on $\mathcal{Y}$ such that the function from $\mathcal{Y}$ to $\mathbb{R}$ defined as $z \to \int_\mathcal{Y} \ell(z, y)\,\mathrm{d}\mu_s(y)$ is minimized for, and only for, $z \in s$.*

We provide a real-world example of a well-behaved problem in Section 5.4 as well as a synthetic example with graphical illustration in Appendix C. On those problems, we suggest to solve Eq. (12) by considering the initialization $\xi_i^{(0)} = \mathbb{E}_{Y\sim\mu_{s_i}}[\varphi(Y)]$, and performing alternative minimization of Eq. (12), until attaining $\xi^{(\infty)}$ as the limit of the alternative minimization scheme (which exists since each step decreases the value of the objective in Eq. (12) and there is a finite number of candidates for (ξ_i)). It corresponds to a disambiguation guess $\tilde{y}_i = \varphi^{-1}(\xi_i^{(\infty)})$. Then we suggest to learn $\hat{f}_n$ from $(x_i, \tilde{y}_i)$ based on Eq. (5), and existing algorithmic tools for this problem (Nowak-Vila et al., 2019). To assert the well-groundedness of this heuristic, we refer to the following proposition, proven in Appendix A.4.

Proposition 6. *Under the non-ambiguity hypothesis, Assumption 2, the solution of Eq. (3) is characterized by $f^* \in \operatorname{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,S)\sim\nu}\left[\mathbb{E}_{Y\sim\mu_S}[\ell(f(X), Y)]\right]$. Moreover, if the surrogate function $g_n^\circ : \mathcal{X} \to \mathcal{H}$ defined as $g_n^\circ(x) = \sum_{i=1}^n \alpha_i(x)\xi_{s_i}$, with $\xi_s = \mathbb{E}_{Y\sim\mu_s}[\varphi(Y)]$, converges towards $g^\circ(x) = \mathbb{E}_{S\sim\nu|_x}[\xi_S]$ in L^1, f_n° defined through the decoding Eq. (9) converges in risk towards f^*.*

Given that our algorithm scheme is initialized for $\xi_i^{(0)} = \xi_{s_i}$ and $\zeta_i^{(0)} = f_n^\circ(x_i)$ and stopped once having attained $\xi_i^{(\infty)}$ and $\zeta_i^{(\infty)} = \hat{f}_n(x_i)$, $\hat{f}_n$ is arguably better than f_n°, which given consistency result exposed in Proposition 6, is already good enough.

Remark 7 (IQP implementation for Eq. (4))**.** *Other heuristics to solve Eq. (4) are conceivable. For example, considering $z_i = y_i$ in this equation, we remark that the resulting problem is isomorphic to an integer quadratic program (IQP). Similarly to integer linear programming, this problem can be approached with relaxation of the "integer constraint" to get a real-valued solution, before "thresholding" it to recover an integer solution. This heuristic can be seen as a generalization of the Diffrac algorithm (Bach & Harchaoui, 2007; Joulin et al., 2010). we present it in details in Appendix B.*

Remark 8 (Link with EM, (Dempster et al., 1977))**.** *Arguably, our alternative minimization scheme, optimizing respectively the targets $\xi_i = \varphi(y_i)$ and the function estimates $\zeta_i = \psi(f_n(x_i))$ can be seen as the non-parametric version of the Expectation-Maximization algorithm, popular for parametric model (Dempster et al., 1977).*

5.4. Application: Ranking with partial ordering

Ranking is a problem consisting, for an input x in an input space $\mathcal{X}$, to learn a total ordering y, belonging to $\mathcal{Y} = \mathfrak{S}_m$, modelling preference over m items. It is usually approach with the Kendall loss $\ell(y, z) = -\varphi(y)^\top\varphi(z)$, with $\varphi(y) = (\operatorname{sign}(y(i) - y(j)))_{i,j\le m} \in \{-1, 1\}^{m^2}$ (Kendall, 1938). Full supervision corresponds, for a given x, to be given a total ordering of the m items. This is usually not an option, but one could expect to be given partial ordering that y should follow (Cao et al., 2007; Hüllermeier et al., 2008; Korba et al., 2018). Formally, this equates to the observation of some, but not all, coordinates $\varphi(y)_i$ of the vector $\varphi(y)$ for some $i \in I \subset [\![1, m]\!]^2$.

In this setting, $s \subset \mathcal{Y}$ is a set of total orderings that match the given partial ordering. It can be represented by a vector $\xi_s \in \mathcal{H}$, that satisfies the partial ordering observation, $(\xi_s)_I = \varphi(y)_I$, and that is agnostic on unobserved coordinates, $(\xi_s)_{^cI} = 0$. This vector satisfies that $z \to \psi(z)^\top\xi_s$ is minimized for, and only for, $z \in s$. Hence, it constitutes a good initialization for the alternative minimization scheme detailed above. We provide details in Appendix A.5, where we also show that ξ_s can be formally translated in a μ_s to match the Definition 5, proving that ranking with partial labelling is a well-behaved problem.

Many real world problems can be formalized as a ranking problem with partial ordering observations. For example, x could be a social network user, and the m items could be posts of her connection that the network would like to order on her feed accordingly to her preferences. One might be told that the user x prefer posts from her close rather than from her distant connections, which translates formally as

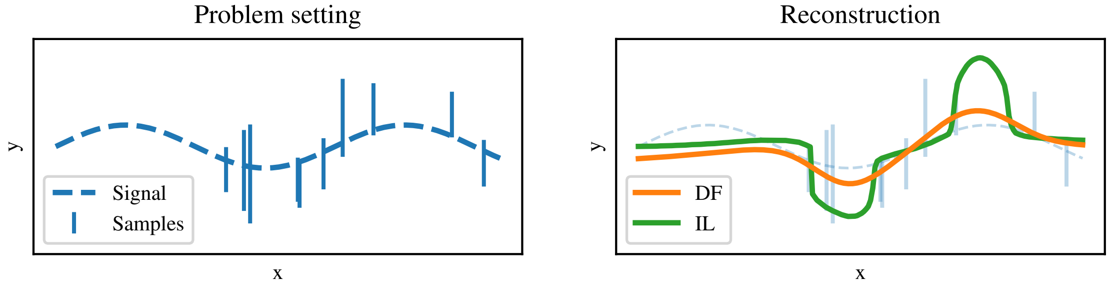

Figure 1. Interval regression. See Appendix D for the exact reproducible experimental setup (Left) Setup. The goal is to learn $f^* : \mathcal{X} \to \mathbb{R}$ represented by the dashed line, given samples (x_i, s_i), where (s_i) are intervals represented by the blue segments. (Right) We compare the Infimum Loss (IL) baseline (13) shown in green, with our Disambiguation Framework (DF), Eqs. (4) and (5), shown in orange; with weights α given by kernel ridge regression. (DF) retrieves $\hat{y}_i$ before learning a smooth f_n based on $(x_i, \hat{y}_i)$, while (IL) implicitly retrieves $\hat{y}_i(x)$ differently for each input, leading to irregularity of the consequent estimator of f^*.

the constraint that for any i corresponding to a post of a close connection and j corresponding to a post of a distant connection, we have $\varphi(y)_{ij} = 1$. Nonetheless, designing non-parametric structured prediction models that scale well when the intrinsic dimension m of the space $\mathcal{Y}$ is very large (such as the number of post on a social network) remains an open problem, that this paper does not tackle.

6. Related work

Weakly supervised learning has been approached through parametric and non-parametric methods. Parametric models are usually optimized through maximum likelihood (Heitjan & Rubin, 1991; Jin & Ghahramani, 2002). Hüllermeier (2014) show that this approach, as formalized by Denoeux (2013), equates to disambiguating sets by averaging candidates, which was shown inconsistent by Cabannes et al. (2020) when data are *not missing at random*. Among non-parametric models, Xu et al. (2004); Bach & Harchaoui (2007) developed an algorithm for clustering, that has been cast for weakly supervised learning problem (Joulin et al., 2010; Alayrac et al., 2016), leading to a disambiguation algorithm similar than ours, yet without consistency results. More recently, half-way between theory and practice, Gong et al. (2018) derived an algorithm geared towards classification, based on a disambiguation objective, incorporating several heuristics, such as class separation, and Laplacian diffusion. Those heuristics could be incorporated formally in our model.

The infimum loss principle has been considered by several authors, among them Cour et al. (2011); Luo & Orabona (2010); Hüllermeier (2014). It was recently analyzed through the prism of structured prediction by Cabannes et al. (2020), leading to a consistent non-parametric algorithm that will constitute the baseline of our experimental comparison. This principle is interesting as it does not assume knowledge on the corruption process $(S|Y)$ contrarily to the work of Cid-Sueiro et al. (2014) or van Rooyen & Williamson (2017).

The non-ambiguity assumption has been introduced by Cour et al. (2011) and is a classical assumption of learning with partial labelling (Liu & Dietterich, 2014). Assumptions of Lipschitzness and minimal mass are classical assumptions to prove convergence of local averaging method (Audibert & Tsybakov, 2007; Biau & Devroye, 2015). Those assumptions imply class separation in $\mathcal{X}$, which has been leverage in semi-supervised learning, justifying Laplacian regularization (Rigollet, 2007; Zhu et al., 2003).

Note that those assumptions might not hold on raw representation of the data, but with appropriate metrics, which could be learned through unsupervised (Duda et al., 2000) or self-supervised learning (Doersch & Zisserman, 2017). Indeed, Wei et al. (2021) provide an analysis akin ours based on such an assumption. As such, the practitioner might consider weights α given by similarity metrics derived through such techniques, before computing the disambiguation (4) and learning f_n from the recollected fully supervised dataset with deep learning.

7. Experiments

In this section, we review a baseline, and experiments that showcase the usefulness of our algorithm Eqs. (4) and (5).

Baseline. We consider as a baseline the work of Cabannes et al. (2020), which is a consistent structured prediction approach to partial labelling through the infimum loss. It is arguably the state-of-the-art of partial labelling approached through structured prediction. It follow the same loss-based variance disambiguation principle, yet in an implicit fashion, leading to the inference algorithm, $f_n : \mathcal{X} \to \mathcal{Y}$,

$$f_n(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \inf_{(y_i) \in C_n} \sum_{i=1}^{n} \alpha_i(x) \ell(z, y_i). \qquad (13)$$

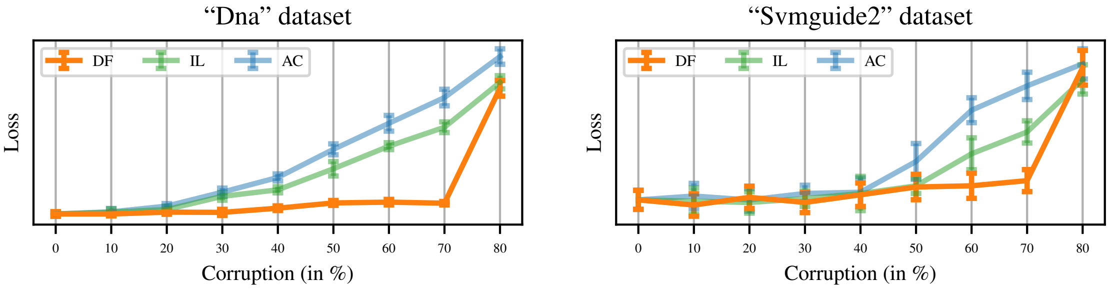

Figure 2. Testing errors as function of the supervision corruption on real dataset corresponding to classification with partial labels. We split fully supervised LIBSVM datasets into training and testing dataset. We corrupt training data in order to get partial labels. Corruption is managed through a parameter, represented by the x-axis, that relates to the ambiguity degree η of Assumption 2. For each methods (our algorithm (DF), the baseline (IL), and the baseline of the baseline (AC, consisting of averaging candidates y_i in sets S_i)), we consider weights α given by kernel ridge regression with Gaussian kernel, for which we optimized hyperparameters with cross-validation on the training set. We then learn an estimate f_n that we evaluate on the testing set, represented by the y-axis, on which we have full supervision. The figure show the superiority of our method, that achieves error similar to baseline when full supervision ($x = 0$) or no supervision ($x = 100\%$) is given, but performs better when only in presence of partial supervision. See Appendix D for reproducibility specifications, where we also provide Figure 6 showcasing similar empirical results in the case of ranking with partial ordering.

Note that with our proof technique, which overcome the suboptimality of calibration inequality (Audibert & Tsybakov, 2007; Cabannes et al., 2021), exponential convergence rates similar to Theorem 4 could be derived for the baseline. Yet, as we will see, our algorithm outperforms this state-of-the-art baseline. This could be explained by the fact that our algorithm introduce an intrinsically smaller surrogate space (in essence, Cabannes et al. (2020) introduced surrogate functions from inputs in $\mathcal{X}$ to powersets represented in $\mathbb{R}^{2^{\mathcal{Y}}}$, while we look at functions from input in $\mathcal{X}$ to output represented in $\mathbb{R}^{\mathcal{Y}}$).

Disambiguation coherence - Interval regression. The baseline Eq. (13) implicitly requires to disambiguate $(\hat{y}_i(x))$ differently for every $x \in \mathcal{X}$. This is counter intuitive since (y_i^*) does not depend on x. It means that $(\hat{y}_i)$ could be equal to some $(\hat{y}_i^{(0)})$ on a subset $\mathcal{X}_0$ of $\mathcal{X}$, and to another $(\hat{y}_i^{(1)})$ on a disjoint subset $\mathcal{X}_1 \subset \mathcal{X}$, leading to irregularity of f_n between $\mathcal{X}_0$ and $\mathcal{X}_1$. We illustrate this graphically on Figure 1. This figure showcases an interval regression problem, which corresponds to the regression setup ($\mathcal{Y} = \mathbb{R}$, $\ell(y, z) = |y - z|^2$) of partial labelling, where one does not observed $y \in \mathbb{R}$ but an interval $s \subset \mathbb{R}$ containing y. Among others, this problem appears in physics (Sheppard, 1897) and economy (Tobin, 1958).

Computation attractiveness - Ranking. Computationally, the baseline requires to solve a disambiguation problem, recovering $(\hat{y}_i(x)) \in C_n$ for every $x \in \mathcal{X}$ for which we want to infer $f_n(x)$. This is much more costly, than doing the disambiguation of $(\hat{y}_i) \in C_n$ once, and solving the supervised learning inference problem Eq. (5), for every $x \in \mathcal{X}$ for which we want to infer $f_n(x)$. To illustrate the computation attractiveness of our algorithm, consider the case of ranking, defined in Section 5.4. Fully supervised inference scheme (5) corresponds to solving a NP-hard problem, equivalent to the minimum feedback arcset problem (Duchi et al., 2010). While disambiguation approaches with alternative minimization implied by Eq. (4) and Eq. (13) require to solve this NP-hard problem for each minimization step. In other terms, the baseline ask to solve multiple NP-hard problem every time one wants to infer f_n given by Eq. (13) on an input $x \in \mathcal{X}$. Meanwhile, our disambiguation approach asks to solve multiple NP-hard problem upfront to solve Eq. (4), yet only require to solve one NP-hard problem to infer f_n given by Eq. (5) on an input $x \in \mathcal{X}$.

Better empirical results - Classification. Finally, we compare our algorithm, our baseline (13) and the baseline considered by Cabannes et al. (2020) on real datasets from the LIBSVM dataset (Chang & Lin, 2011). Those datasets (x_i, y_i) correspond to fully supervised classification problem. In this setup, $\mathcal{Y} = [\![1, m]\!]$ for m a number of classes, and $\ell(y, z) = \mathbf{1}_{y \neq z}$. We "corrupt" labels in order to create a synthetic weak supervision datasets (x_i, s_i). We consider skewed corruption, in the sense that (s_i) is generated by a probability such that $\sum_{z \in \mathcal{Y}} \mathbb{P}_{S_i}(z \in S_i | y_i)$ depends on the value of y_i. This corruption is parametrized by a parameter that related with the ambiguity parameter η of Assumption 2. Results on Figure 2 show that, in addition to having a lower computation cost, our algorithm performs better in practice than the state-of-the-art baseline.[3]

Beyond Eq. (2) - Semi-supervised learning. The main limitation of Eq. (2) is that it is a pointwise principle that decorrelates inputs, in the sense that the optimization of

[3] All the code is available online - https://github.com/VivienCabannes/partial_labelling.

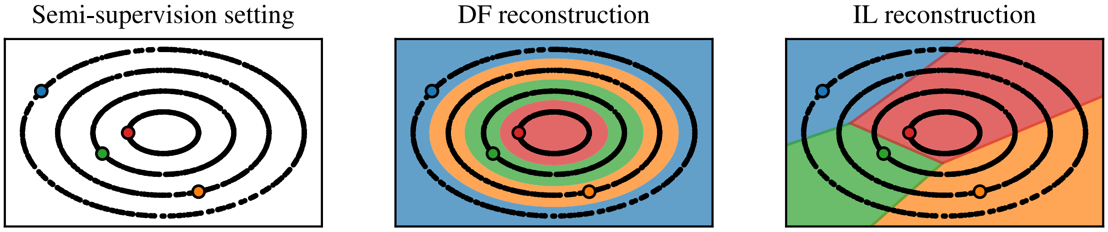

Figure 3. Semi-supervised learning, "concentric circle" instance with four classes (red, green, blue, yellow). Reproducibility details provided in Appendix D. (Left) We represent points $x_i \in \mathcal{X} \subset \mathbb{R}^2$, there is many unlabelled points (represented by black dots and corresponding to $S_i = \mathcal{Y}$), and one labelled point for each class (represented in color, corresponding to $S_i = \{y_i\}$). (Middle) Reconstruction $f_n : \mathcal{X} \to \mathcal{Y}$ given by our algorithm Eqs. (4) and (5). Our algorithm succeeds to comprehend the concentric circle structure of the input distribution and clusters classes accordingly. (Right) Reconstruction $f_n : \mathcal{X} \to \mathcal{Y}$ given by the baseline Eq. (13). The baseline performs as if only the four supervised data points where given.

$\mu^*|_x$, for $x \in \mathcal{X}$, only depends on $\nu|_x$ and not on what is happening on $\mathcal{X} \setminus \{x\}$. As such, this principle failed to tackle semi-supervised learning, where $\nu|_x$ is equal to $\mu|_x$ (in the sense that $\pi|_{x,y} = \delta_{\{y\}}$) for $x \in \mathcal{X}_l$ and is equal to $\delta_\mathcal{Y}$ for $x \in \mathcal{X}_u := \mathcal{X} \setminus \mathcal{X}_l$. In such a setting, for $x \in \mathcal{X}_u$, $\mu^*|_x$ can be set to any δ_y for $y \in \mathcal{Y}$. Interestingly, in practice, while the baseline suffer the same limitation, for our algorithm, *weighting schemes have a regularization effect*, that contrasts with those considerations. We illustrate it on Figure 3.

Real real-world applications. There is a real lack of clean datasets to experiment with partial labelling. Most theoretical papers consist in synthetic corruption of fully supervised dataset (*e.g.,* Korba et al., 2018) as we did. Empirical papers are built on highly complex datasets that require skilled pre-processing and tricks beside theoretically-grounded principle (*e.g.*, action recognition on Youtube videos). However, note that the state-of-the-art work of Miech et al. (2019) is built on heuristics from the Diffrac algorithm, which we generalized (see Alayrac et al., 2016, for details). We hope that, by providing theoretical understanding of the problem, our paper could help to design powerful heuristics in practice, even though this is out of scope of the present paper.

8. Conclusion

In this work, we have introduced a structured prediction algorithm Eqs. (4) and (5), to tackle partial labelling. We have derived exponential convergence rates for the nearest neighbors instance of this algorithm under classical learnability assumptions. We provided optimization considerations to implement this algorithm in practice, and have successfully compared it with the state-of-the-art. Several open problems offer prospective follow-up of this works.

- *Semi-supervised learning and beyond.* While we only proved convergence in situation where μ^* of Eq. (2) is uniquely defined, therefore excluding semi-supervised learning, Figure 3 suggests that our algorithm (4) could be analyzed in a broader setting than the one considered in this paper. Among others, the non-ambiguity assumption could be replaced by a cluster assumption (Rigollet, 2007) together with a non-ambiguity assumption cluster-wise in Theorem 4.
- *Hard-coded weak supervision.* Variational principles Eqs. (2) and (3) could be extended beyond partial labelling to any type of hard-coded weak supervision, which is when weak supervision can be cast as a set of hard constraint that μ should satisfy, formally written as a set of fully supervised distributions compatible with weak information. Hard-coded weak supervision includes label proportion (Quadrianto et al., 2009; Dulac-Arnold et al., 2019), but excludes supervision of the type "80% of the experts say this nose is broken, and 20% say it is not". Providing a unifying framework for those problems would make an important step in the theoretical foundation of weakly supervised learning.
- *Missing input data.* While weak supervision assumes that only y is partially known, in many applications of machine learning, x is also only partially known, especially when the feature vector x is built from various source of information, leading to missing data. While we only considered a principle to fill missing output information, similar principles could be formalized to fill missing input information. This would be particularly valuable when data are not missing at random (Rubin, 1976; Muzellec et al., 2020).

Acknowledgements

We would like to thanks anonymous referees for helpful comments. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support of the European Research Council (grants SEQUOIA 724063, REAL 94790).

References

Alayrac, J., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., and Lacoste-Julien, S. Unsupervised learning from narrated instruction videos. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Audibert, J.-Y. and Tsybakov, A. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 2007.

Bach, F. and Harchaoui, Z. DIFFRAC: a discriminative and flexible framework for clustering. In *Neural Information Processing Systems 20*, 2007.

Bengio, Y., Delalleau, O., and Roux, N. L. Label propagation and quadratic criterion. In *Semi-Supervised Learning*. MIT Press, 2006.

Berthelot, D., Carlini, N., Goodfellow, I. J., Papernot, N., Oliver, A., and Raffel, C. Mixmatch: A holistic approach to semi-supervised learning. In *Neural Information Processing Systems*, 2019.

Biau, G. and Devroye, L. *Lectures on the Nearest Neighbor Method*. Springer International Publishing, 2015.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 2005.

Cabannes, V., Rudi, A., and Bach, F. Structured prediction with partial labelling through the infimum loss. In *International Conference on Machine Learning*, 2020.

Cabannes, V., Rudi, A., and Bach, F. Fast rates in structured prediction. In *Conference on Learning Theory*, 2021.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: from pairwise approach to listwise approach. In *International Conference of Machine Learning*, 2007.

Chang, C. and Lin, C. LIBSVM: A library for support vector machines. *ACM TIST*, 2011.

Cid-Sueiro, J., García-García, D., and Santos-Rodríguez, R. Consistency of losses for learning from weak labels. *Lecture Notes in Computer Science*, 2014.

Ciliberto, C., Rosasco, L., and Rudi, A. A consistent regularization approach for structured prediction. In *Neural Information Processing Systems 29*, 2016.

Ciliberto, C., Rosasco, L., and Rudi, A. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 2020.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *Journal of Machine Learning Research*, 2011.

Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., and Jain, A. Learning 3D human pose from structure and motion. In *European Conference on Computer Vision*, 2018.

Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977.

Denoeux, T. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineerin*, 2013.

Doersch, C. and Zisserman, A. Multi-task self-supervised visual learning. In *International Conference on Computer Vision*, 2017.

Duchi, J. C., Mackey, L. W., and Jordan, M. I. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, 2010.

Duda, R., Hart, P., and Stork, D. *Pattern Classification, 2nd Edition*. Wiley, 2000.

Dulac-Arnold, G., Zeghidour, N., Cuturi, M., Beyer, L., and Vert, J.-P. Deep multiclass learning from label proportions. In *ArXiv*, 2019.

Gong, C., Liu, T., Tang, Y., Yang, J., Yang, J., and Tao, D. A regularization approach for instance-based superset label learning. *IEEE Transactions on Cybernetics*, 2018.

Harris, C., Millman, J., van der Walt, S., et al. Array programming with NumPy. *Nature*, 2020.

Hein, M., Audibert, J.-Y., and von Luxburg, U. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 2007.

Heitjan, D. and Rubin, D. Ignorability and coarse data. *The Annals of Statistics*, 1991.

Hüllermeier, E. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 2014.

Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 2008.

IBM. *IBM ILOG CPLEX 12.7 User's Manual*. IBM ILOG CPLEX Division, 2017.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez-Moreno, I., and Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Neural Information Processing Systems*, 2018.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Neural Information Processing Systems*, 2002.

Joulin, A., Bach, F., and Ponce, J. Discriminative clustering for image co-segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2010.

Kendall, M. A new measure of rank correlation. *Biometrika*, 1938.

Korba, A., Garcia, A., and d'Alché-Buc, F. A structured prediction approach for label ranking. In *Neural Information Processing Systems*, 2018.

Lienen, J. and Hüllermeier, E. From label smoothing to label relaxation. In *AAAI Conference on Artificial Intelligence*, 2021.

Liu, L.-P. and Dietterich, T. Learnability of the superset label learning problem. In *International Conference on Machine Learning*, 2014.

Luo, J. and Orabona, F. Learning from candidate labeling sets. In *Neural Information Processing Systems*, 2010.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *International Conference on Computer Vision*, 2019.

Muzellec, B., Josse, J., Boyer, C., and Cuturi, M. Missing data imputation using optimal transport. In *International Conference of Machine Learning*, 2020.

Nowak-Vila, A., Bach, F., and Rudi, A. Sharp analysis of learning with discrete losses. In *Artificial Intelligence and Statistics*, 2019.

Papandreou, G., Chen, L.-C., Murphy, K., and Yuille, A. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *International Conference on Computer Vision*, 2015.

Perchet, V. and Quincampoix, M. On a unified framework for approachability with full or partial monitoring. *Mathematics of Operations Research*, 2015.

Quadrianto, N., Smola, A., Caetano, T., and Le, Q. V. Estimating labels from label proportions. *Journal of Machine Learning Research*, 2009.

Rigollet, P. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 2007.

Rubin, D. Inference and missing data. *Biometrika*, 1976.

Sheppard, W. On the calculation of the most probable values of frequency constants, for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society*, 1897.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Stone, C. Consistent nonparametric regression. *The Annals of Statistics*, 1977.

Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica*, 1958.

van Rooyen, B. and Williamson, R. C. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 2017.

Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., and Ruan, X. Learning to detect salient objects with image-level supervision. In *Conference on Computer Vision and Pattern Recognition*, 2017.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. Maximum margin clustering. *Neural Information Processing Systems*, 2004.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. Learning with local and global consistency. In *Neural Information Processing Systems*, 2003.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference of Machine Learning*, 2003.