Structured Prediction with Partial Labelling through the Infimum Loss

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

Abstract

Annotating datasets is one of the main costs in nowadays supervised learning. The goal of weak supervision is to enable models to learn using only forms of labelling which are cheaper to collect, as partial labelling. This is a type of incomplete annotation where, for each datapoint, supervision is cast as a set of labels containing the real one. The problem of supervised learning with partial labelling has been studied for specific instances such as classification, multi-label, ranking or segmentation, but a general framework is still missing. This paper provides a unified framework based on structured prediction and on the concept of *infimum loss* to deal with partial labelling over a wide family of learning problems and loss functions. The framework leads naturally to explicit algorithms that can be easily implemented and for which proved statistical consistency and learning rates. Experiments confirm the superiority of the proposed approach over commonly used baselines.

1. Introduction

Fully supervised learning demands tight supervision of large amounts of data, a supervision that can be quite costly to acquire and constrains the scope of applications. To overcome this bottleneck, the machine learning community is seeking to incorporate weaker sources of information in the learning framework. In this paper, we address those limitations through partial labelling: *e.g.*, giving only partial ordering when learning user preferences over items, or providing the label "flower" for a picture of Arum Lilies[1], instead of spending a consequent amount of time to find the exact taxonomy.

[1]INRIA - Département d'Informatique de l'École Normale Supérieure - PSL Research University, Paris, France. Correspondence to: Vivien Cabannes <vivien.cabannes@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]https://en.wikipedia.org/wiki/Arum

Partial labelling has been studied in the context of classification (Cour et al., 2011; Nguyen & Caruana, 2008), multi-labelling (Yu et al., 2014), ranking (Hüllermeier et al., 2008; Korba et al., 2018), as well as segmentation (Verbeek & Triggs, 2008; Papandreou et al., 2015), or natural language processing tasks (Fernandes & Brefeld, 2011; Mayhew et al., 2019), however a generic framework is still missing. Such a framework is a crucial step towards understanding how to learn from weaker sources of information, and widening the spectrum of machine learning beyond rigid applications of supervised learning. Some interesting directions are provided by Cid-Sueiro et al. (2014); van Rooyen & Williamson (2017), to recover the information lost in a corrupt acquisition of labels. Yet, they assume that the corruption process is known, which is a strong requirement that we want to relax.

In this paper, we make the following contributions:

- We provide a principled framework to solve the problem of learning with partial labelling, via *structured prediction*. This approach naturally leads to a variational framework built on the *infimum loss*.
- We prove that the proposed framework is able to recover the original solution of the supervised learning problem under identifiability assumptions on the labelling process.
- We derive an explicit algorithm which is easy to train and with strong theoretical guarantees. In particular, we prove that it is consistent and we provide generalization error rates.
- Finally, we test our method against some simple baselines, on synthetic and real examples. We show that for certain partial labelling scenarios with symmetries, our infimum loss performs similarly to a simple baseline. However in scenarios where the acquisition process of the labels is more adversarial in nature, the proposed algorithm performs consistently better.

2. Partial labelling with infimum loss

In this section, we introduce a statistical framework for partial labelling, and we show that it is characterized naturally in terms of risk minimization with the infimum loss. First, let's recall some elements of fully supervised and weakly supervised learning.

Fully supervised learning consists in learning a function $f \in \mathcal{Y}^{\mathcal{X}}$ between a input space $\mathcal{X}$ and a output space $\mathcal{Y}$, given a joint distribution $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$, and a loss function $\ell \in \mathbb{R}^{\mathcal{Y}\times\mathcal{Y}}$, that minimizes the risk

$$\mathcal{R}(f;\rho) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X),Y)\right], \tag{1}$$

given observations $(x_i, y_i)_{i\leq n} \sim \rho^{\otimes n}$. We will assume that the loss ℓ is proper, *i.e.* it is continuous non-negative and is zero on, and only on, the diagonal of $\mathcal{Y} \times \mathcal{Y}$, and strictly positive outside. We will also assume that $\mathcal{Y}$ is compact.

In *weakly supervised learning*, given $(x_i)_{i\leq n}$, one does not have direct observations of $(y_i)_{i\leq n}$ but weaker information. The goal is still to recover the solution $f \in \mathcal{Y}^{\mathcal{X}}$ of the fully supervised problem Eq. (1). In *partial labelling*, also known as *superset learning* or as *learning with ambiguous labels*, which is an instance of weak supervision, information is cast as closed sets $(S_i)_{i\leq n}$ in $\mathcal{S}$, where $\mathcal{S} \subset 2^{\mathcal{Y}}$ is the space of closed subsets of $\mathcal{Y}$, containing the true labels $(y_i \in S_i)$. In this paper, we model this scenario by considering a data distribution $\tau \in \Delta_{\mathcal{X}\times\mathcal{S}}$, that generates the samples (x_i, S_i). We will denote τ as *weak distribution* to distinguish it from ρ. Capturing the dependence on the original problem, τ must be compatible with ρ, a matching property that we formalize with the concept of eligibility.

Definition 1 (Eligibility)**.** *Given a probability measure τ on $\mathcal{X} \times \mathcal{S}$, a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$ is said to be eligible for τ (denoted by $\rho \vdash \tau$), if there exists a probability measure π over $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ such that ρ is the marginal of π over $\mathcal{X} \times \mathcal{Y}$, τ is the marginal of π over $\mathcal{X} \times \mathcal{S}$, and, for $y \in \mathcal{Y}$ and $S \in \mathcal{S}$*

$$y \notin S \qquad \Rightarrow \qquad \mathbb{P}_\pi\left(S \,|\, Y = y\right) = 0.$$

We will alternatively say that τ is a weakening *of ρ, or that ρ and τ are* compatible*.*

2.1. Disambiguation principle

According to the setting described above, the problem of partial labelling is completely defined by a loss and a weak distribution (ℓ, τ). The goal is to recover the solution of the original supervised learning problem in Eq. (1) assuming that the original distribution verifies $\rho \vdash \tau$. Since more than one ρ may be eligible for τ, we would like to introduce a guiding principle to identify a $\rho^\star$ among them. With this goal we define the concept of *non-ambiguity* for τ, a setting in which a natural choice for $\rho^\star$ appears.

Definition 2 (Non-ambiguity)**.** *For any $x \in \mathcal{X}$, denote by $\tau|_x$ the conditional probability of τ given x, and define the set S_x as*

$$S_x = \bigcap_{S\in\mathrm{supp}(\tau|_x)} S.$$

The weak distribution τ is said non-ambiguous *if, for every $x \in \mathcal{X}$, S_x is a singleton. Moreover, we say that τ is* strictly non-ambiguous *if it is non-ambiguous and there exists $\eta \in (0, 1)$ such that, for all $x \in \mathcal{X}$ and $z \notin S_x$*

$$\mathbb{P}_{S\sim\tau|_x}(z \in S) \leq 1 - \eta.$$

This concept is similar to the one by Cour et al. (2011), but more subtle because this quantity only depends on τ, and makes no assumption on the original distribution ρ describing the fully supervised process that we can not access. In this sense, it is also more general.

When τ is non-ambiguous, we can write $S_x = \{y_x\}$ for any x, where y_x is the only element of S_x. In this case it is natural to identify $\rho^\star$ as the one satisfying $\rho^\star|_x = \delta_{y_x}$. Actually, such a $\rho^\star$ is characterized without S_x as the only deterministic distribution that is eligible for τ. Because deterministic distributions are characterized as minimizing the minimum risk of Eq. (1), we introduce the following *minimum variability principle* to disambiguate between all eligible ρ's, and identify $\rho^\star$,

$$\rho^\star \in \operatorname*{arg\,min}_{\rho\vdash\tau} \mathcal{E}(\rho), \qquad \mathcal{E}(\rho) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f;\rho). \tag{2}$$

The quantity $\mathcal{E}$ can be identified as a variance, since if f_ρ is the minimizer of $\mathcal{R}(f;\rho)$, $f_\rho(x)$ can be seen as the mean of $\rho|_x$ and ℓ the natural distance in $\mathcal{Y}$. Indeed, when $\ell = \ell_2$ is the mean square loss, this is exactly the case. The principle above recovers exactly $\rho^\star|_x = \delta_{y_x}$, when τ is non-ambiguous, as stated by Prop. 1, proven in Appendix A.1.

Proposition 1 (Non-ambiguity determinism)**.** *When τ is non-ambiguous, the solution $\rho^\star$ of Eq. (2) exists and satisfies that, for any $x \in \mathcal{X}$, $\rho^\star|_x = \delta_{y_x}$, where y_x is the only element of S_x.*

Prop. 1 provides a justification for the usage of the minimum variability principle. Indeed, under non-ambiguity assumption, following this principle will allow us to build an algorithm that recover the original fully supervised distribution. Therefore, given samples (x_i, S_i), it is of interest to test if τ is non-ambiguous. Such tests should leverage other regularity hypothesis on τ, which we will not address in this work.

Now, we characterize the minimum variability principle in terms of a variational optimization problem that we can tackle in Sec. 3 via empirical risk minimization.

2.2. Variational formulation via the infimum loss

Given a partial labelling problem (ℓ, τ), define the solutions based on the minimum variablity principle as the functions minimizing the recovered risk

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f;\rho^\star). \tag{3}$$

for $\rho^\star$ a distribution solving Eq. (2). As shown in Thm. 1 below, proven in Appendix A.2, the proposed disambiguation paradigm naturally leads to a variational framework involving the *infimum loss*.

Theorem 1 (Infimum loss (*IL*))**.** *The functions f^* defined in Eq. (3) are characterized as*

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}_S(f),$$

where the risk $\mathcal{R}_S$ is defined as

$$\mathcal{R}_S(f) = \mathbb{E}_{(X,S)\sim\tau}\left[L(f(X),S)\right], \tag{4}$$

and L is the infimum loss

$$L(z,S) = \inf_{y\in S} \ell(z,y). \tag{5}$$

The infimum loss, also known as the ambiguous loss (Luo & Orabona, 2010; Cour et al., 2011), or as the optimistic superset loss (Hüllermeier, 2014), captures the idea that, when given a set S, this set contains the good label y but also a lot of bad ones, that should not be taken into account when retrieving f. In other terms, f should only match the best guess in S. Indeed, if ℓ is seen as a distance, L is its natural extension to sets.

2.3. Recovery of the fully supervised solutions

In this subsection, we investigate the setting where an original fully supervised learning problem ρ_0 has been weakened due to incomplete labelling, leading to a weak distribution τ. The goal here is to understand under which conditions on τ and ℓ it is possible to recover the original fully supervised solution based with the infimum loss framework. Denote f_0 the function minimizing $\mathcal{R}(f;\rho_0)$. The theorem below, proven in Appendix A.3, shows that under non-ambiguity and deterministic conditions, it is possible to fully recover the function f_0 also from τ.

Theorem 2 (Supervision recovery)**.** *For an instance (ℓ,ρ_0,τ) of the weakened supervised problem, if we denote by f_0 the minimizer of Eq. (1), we have the under the conditions that (1) τ is not ambiguous (2) for all $x\in\mathcal{X}$, $S_x = \{f_0(x)\}$; the infimum loss recovers the original fully supervised solution,* i.e. *the f^* defined in Eq. (3) verifies $f^* = f_0$.*

Futhermore, when ρ_0 is deterministic and τ not ambiguous, the $\rho^\star$ defined in Eq. (2) verifies $\rho^\star = \rho_0$.

At a comprehensive levels, this theorem states that under non-ambiguity of the partial labelling process, if the labels are a deterministic function of the inputs, the infimum loss framework make it possible to recover the solution of the original fully supervised problem while only accessing weak labels. In the next subsection, we will investigate which is the relation between the two problems when dealing with an estimator f of f^*.

2.4. Comparison inequality

In the following, we want to characterize the error performed by $\mathcal{R}(f;\rho^\star)$ with respect to the error performed by $\mathcal{R}_S(f)$. This will be useful since, in the next section, we will provide an estimator for f^* based on structured prediction, that minimize the risk $\mathcal{R}_S$. First, we introduce a measure of discrepancy for the loss function.

Definition 3 (Discrepancy of the loss ℓ)**.** *Given a loss function ℓ, the* discrepancy degree *ν of ℓ is defined as*

$$\nu = \log \sup_{y,z'\neq z} \frac{\ell(z,y)}{\ell(z,z')}.$$

$\mathcal{Y}$ will be said discrete for ℓ when $\nu < +\infty$, which is always the case when $\mathcal{Y}$ is finite.

Now we are ready to state the comparison inequality that generalizes to arbitrary losses and output spaces a result on $0-1$ loss on classification from Cour et al. (2011).

Proposition 2 (Comparison inequality)**.** *When $\mathcal{Y}$ is discrete and τ is strictly non-ambiguous for a given $\eta\in(0,1)$, then the following holds*

$$\mathcal{R}(f;\rho^\star) - \mathcal{R}(f^*;\rho^\star) \leq C(\mathcal{R}_S(f) - \mathcal{R}_S(f^*)), \tag{6}$$

for any measurable function $f\in\mathcal{Y}^\mathcal{X}$, where C does not depend on τ, f, and is defined as follows and always finite

$$C = \eta^{-1}e^{\nu}.$$

When ρ_0 is deterministic, since we know from Thm. 2 that $\rho^\star = \rho_0$, this theorem allows to bound the error made on the original fully supervised problem with the error measured with the infimum loss on the weakly supervised one.

Note that the constant presented above is the product of two independent terms, the first measuring the ambiguity of the weak distribution τ, and the second measuring a form of discrepancy for the loss. In the appendix, we provide a more refined bound for C, that is $C = C(\ell,\tau)$, that shows a more elaborated interaction between ℓ and τ. This may be interesting in situations where it is possible to control the labelling process and may suggest strategies to active partial labelling, with the goal of minimizing the costs of labelling while preserving the properties presented in this section and reducing the impact of the constant C in the learning process. An example is provided in the Appendix A.5.

3. Consistent algorithm for partial labelling

In this section, we provide an algorithmic approach based on structured prediction to solve the weak supervised learning problem expressed in terms of infimum loss from Thm. 1. From this viewpoint, we could consider different structured prediction frameworks as structured SVM (Tsochantaridis

et al., 2005), conditional random fields (Lafferty et al., 2001) or surrogate mean estimation (Ciliberto et al., 2016). For example, Luo & Orabona (2010) used a margin maximization formulation in a structured SVM fashion, Hüllermeier & Cheng (2015) went for nearest neighbors, and Cour et al. (2011) design a surrogate method specific to the 0-1 loss, for which they show consistency based on Bartlett et al. (2006).

In the following, we will use the structured prediction method of Ciliberto et al. (2016); Nowak-Vila et al. (2019), which allows us to derive an explicit estimator, easy to train and with strong theoretical properties, in particular, consistency and finite sample bounds for the generalization error. The estimator is based on the pointwise characterization of f^* as

$$f^*(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Y}} \mathbb{E}_{S\sim\tau|_x} \left[\inf_{y\in S} \ell(z,y) \right],$$

and weights $\alpha_i(x)$ that are trained on the dataset such that $\hat{\tau}_{|x} = \sum_{i=1}^n \alpha_i(x)\delta_{S_i}$ is a good approximation of $\tau|_x$. Plugging this approximation in the precedent equation leads to our estimator, that is defined explicity as follows

$$f_n(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Y}} \inf_{y_i\in S_i} \sum_{i=1}^n \alpha_i(x)\ell(z,y_i). \tag{7}$$

Among possible choices for α, we will consider the following kernel ridge regression estimator to be learned at training time

$$\alpha(x) = (K + n\lambda)^{-1} v(x),$$

with $\lambda > 0$ a regularizer parameter and $K = (k(x_i,x_j))_{i,j} \in \mathbb{R}^{n\times n}, v(x) = (k(x,x_i))_i \in \mathbb{R}^n$ where $k \in \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a positive-definite kernel (Scholkopf & Smola, 2001) that defines a similarity function between input points (*e.g.*, if $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$ a commonly used kernel is the Gaussian kernel $k(x,x') = e^{-\|x-x'\|^2}$). Other choices can be done to learn α, beyond kernel methods, a particularly appealing one is harmonic functions, incorporating a prior on low density separation to boost learning (Zhu et al., 2003; Zhou et al., 2003; Bengio et al., 2006). Here we use the kernel estimator since it allows to derive strong theoretical results, based on kernel conditional mean estimation (Muandet et al., 2017).

3.1. Theoretical guarantees

In this following, we want to prove that f_n converges to f^* as n goes to infinity and we want to quantify it with finite sample bounds. The intuition behind this result is that as the number of data points tends toward infinity, $\hat{\tau}$ concentrates towards τ, making our algorithm in Eq. (7) converging to a minimizer of Eq. (4) as explained more in detail in Appendix A.6.

Theorem 3 (Consistency). *Let $\mathcal{Y}$ be finite and τ be a non-ambiguous probability. Let k be a bounded continuous universal kernel,* e.g. *the Gaussian kernel (see Micchelli et al., 2006, for details), and f_n the estimator in Eq. (7) trained on $n \in \mathbb{N}$ examples and with $\lambda = n^{-1/2}$. Then, holds with probability* 1

$$\lim_{n\to\infty} \mathcal{R}(f_n;\rho^\star) = \mathcal{R}(f^*;\rho^\star).$$

In the next theorem, instead we want to quantify how fast f_n converges to f^* depending on the number of examples. To obtain this result, we need a finer characterization of the infimum loss L as:

$$L(z,S) = \langle \psi(z), \varphi(S) \rangle,$$

where $\mathcal{H}$ is a Hilbert space and $\psi : \mathcal{Y} \to \mathcal{H}, \varphi : 2^{\mathcal{Y}} \to \mathcal{H}$ are suitable maps. Such a decomposition always exists in finite case (as for the infimum loss over $\mathcal{Y}$ finite) and many explicit examples for losses of interest are presented by Nowak-Vila et al. (2019). We now introduce the conditional expectation of $\varphi(S)$ given x, defined as

$$\begin{array}{rccc} g: & \mathcal{X} & \to & \mathcal{H} \\ & x & \to & \mathbb{E}_\tau\left[\varphi(S) \,|\, X = x\right]. \end{array}$$

The idea behind the proof is that the distance between f_n and f is bounded by the distance of g_n an estimator of g that is implicitly computed via α. If g has some form of regularity, *e.g.* $g \in \mathcal{G}$, with $\mathcal{G}$ the space of functions representable by the chosen kernel (see Scholkopf & Smola, 2001), then it is possible to derive explicit rates, as stated in the following theorem.

Theorem 4 (Convergence rates). *In the setting of Thm. 3, if τ is η-strictly non ambiguous for $\eta \in (0,1)$, and if $g \in \mathcal{G}$, then there exists a $\tilde{C}$, such that, for any $\delta \in (0,1)$ and $n \in \mathbb{N}$, holds with probability at least $1-\delta$,*

$$\mathcal{R}(f_n;\rho^\star) - \mathcal{R}(f^*;\rho^\star) \leq \tilde{C} \log\left(\frac{8}{\delta}\right)^2 n^{-1/4}. \tag{8}$$

Those last two theorem are proven in Appendix A.6 and combines the consistency and learning results for kernel ridge regression (Caponnetto & De Vito, 2006; Smale & Zhou, 2007), with a comparison inequality of Ciliberto et al. (2016) which relates the excess risk of the structured prediction problem with the one of the surrogate loss $\mathcal{R}_S$, together with our Prop. 2, which relates the error $\mathcal{R}$ to $\mathcal{R}_S$.

Thoses results make our algorithm the first algorithm for partial labelling, that to our knowledge is applicable to a generic loss ℓ and has strong theoretical guarantees as consistency and learning rates. In the next section we will compare with the state of the art and other variational principles.

4. Previous works and baselines

Partial labelling was first approached through discriminative models, proposing to learn $(Y \mid X)$ among a family of parameterized distributions by maximizing the log likelihood based on expectation-maximization scheme (Jin & Ghahramani, 2002), eventually integrating knowledge on the partial labelling process (Grandvalet, 2002; Papandreou et al., 2015). In the meanwhile, some applications of clustering methods have involved special instances of partial labelling, like segmentation approached with spectral method (Weiss, 1999), semi-supervision approached with max-margin (Xu et al., 2004). Also initially geared towards clustering, Bach & Harchaoui (2007) consider the infimum principle on the mean square loss, and this was generalized to weakly supervised problems (Joulin et al., 2010). The infimum loss as an objective to minimize when learning from partial labels was introduced by Cour et al. (2011) for the classification instance and used by Luo & Orabona (2010); Hüllermeier (2014) in generic cases. Comparing to those last two, we provide a framework that derives the use of infimum loss from first principles and from which we derive an explicit and easy to train algorithm with strong statistical guarantees, which were missing in previous work. In the rest of the section, we will compare the infimum loss with other variational principles that have been considered in the literature, in particular the supremum loss (Guillaume et al., 2017) and the average loss (Denoeux, 2013).

Average loss (*AC*). A simple loss to deal with uncertainty is to average over all potential candidates, assuming S discrete,

$$L_{ac}(z, S) = \frac{1}{|S|} \sum_{y \in S} \ell(z, y).$$

It is equivalent to a fully supervised distribution ρ_{ac} by sampling Y uniformly at random among S

$$\rho_{ac}(y) = \int_{\mathcal{S}} \frac{1}{|S|} \mathbf{1}_{y \in S} \, \mathrm{d}\tau(S).$$

This directly follows from the definition of L_{ac} and of the risk $\mathcal{R}(z; \rho_{ac})$. However, as soon as the loss ℓ has discrepancy, *i.e.* $\nu > 0$, the average loss will implicitly advantage some labels, which can lead to inconsistency, even in the deterministic not ambiguous setting of Prop. 2 (see Appendix A.7 for more details).

Supremum loss (*SP*). Another loss that have been considered is the supremum loss (Wald, 1945; Madry et al., 2018), bounding from above the fully supervised risk in Eq. (1). It is widely used in the context of robust risk minimization and reads

$$R_{sp}(f) = \sup_{\rho \vdash \tau} \mathbb{E}_{(X,Y) \sim \rho} \left[\ell(f(x), S) \right].$$

Similarly to the infimum loss in Thm. 1, this risk can be written from the loss function

$$L_{sp}(z, S) = \sup_{y \in S} \ell(z, y).$$

Yet, this adversarial approach is not consistent for partial labelling, even in the deterministic non ambiguous setting of Prop. 2, since it finds the solution that best agrees with *all* the elements in S and not only the true one (see Appendix A.7 for more details).

4.1. Instance showcasing superiority of our method

In the rest of this section, we consider a pointwise example to showcase the underlying dynamics of the different methods. It is illustrated in Fig. 1. Consider $\mathcal{Y} = \{a, b, c\}$ and a proper symmetric loss function such that $\ell(a, b) = \ell(a, c) = 1$, $\ell(b, c) = 2$. The simplex $\Delta_{\mathcal{Y}}$ is naturally split into decision regions, for $e \in \mathcal{Y}$,

$$R_e = \left\{ \rho \in \Delta_{\mathcal{Y}} \,\middle|\, e \in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \mathbb{E}_{\rho}[\ell(z, Y)] \right\}.$$

Both *IL* and *AC* solutions can be understood geometrically by looking at where $\rho^\star$ and ρ_{ac} fall in the partition of the simplex $(R_e)_{e \in \mathcal{Y}}$. Consider a fully supervised problem with distribution δ_c, and a weakening τ of ρ defined by $\tau(\{a, b, c\}) = \frac{5}{8}$ and $\tau(\{c\}) = \tau(\{a, c\}) = \tau(\{b, c\}) = \frac{1}{8}$. This distribution can be represented on the simplex in terms of the region $R_\tau = \{\rho \in \Delta_{\mathcal{Y}} \mid \rho \vdash \tau\}$. Finding $\rho^\star$ correspond to minimizing the piecewise linear function $\mathcal{E}(\rho)$ (Eq. (2)) inside R_τ. On this example, it is minimized for $\rho^\star = \delta_c$, which we know from Prop. 2. Now note that if we use the average loss, it disambiguates ρ as

$$\rho_{ac}(c) = \frac{11}{24} = \frac{1}{3}\frac{5}{8} + \frac{1}{8} + 2 \cdot \frac{1}{2}\frac{1}{8}, \quad \rho_{ac}(b) = \rho_{ac}(a) = \frac{13}{48}.$$

This distribution falls in the decision region of a, which is inconsistent with the real label $y = c$. For the supremum loss, one can show, based on $\mathcal{R}_{sp}(a) = \ell(a, c) = 1$, $\mathcal{R}_{sp}(b) = \ell(b, c) = 2$ and $\mathcal{R}_{sp}(c) = 3/2$, that the supremum loss is minimized for $z = a$, which is also inconsistent. Instead, by using the infimum loss, we have $f^* = f_0 = c$, and moreover that $\rho^\star = \rho_0$ that is the optimal one.

4.2. Algorithmic considerations for AC, SP

The averaging candidates principle, approached with the framework of quadratic surrogates (Ciliberto et al., 2016), leads to the following algorithm

$$\begin{aligned} f_{ac}(x) &\in \operatorname*{arg\,min}_{z \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \frac{1}{|S_i|} \sum_{y \in S_i} \ell(z, y) \\ &= \operatorname*{arg\,min}_{z \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \left(\sum_{i=1}^{n} \mathbf{1}_{y \in S_i} \frac{\alpha_i(x)}{|S_i|} \right) \ell(z, y). \end{aligned}$$

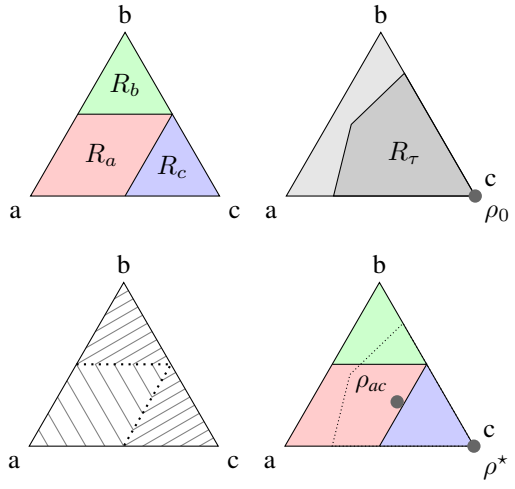

Figure 1. Simplex $\Delta_\mathcal{Y}$. (Left) Decision frontiers. (Middle left) Full and weak distributions. (Middle right) Level curves of the piece-wise linear objective $\mathcal{E}$ (Eq. (2)), to optimize when disambiguating τ into $\rho^\star$. (Right) Disambiguation of *AC* and *IL*.

This estimator is computationally attractive because the inference complexity is the same as the inference complexity of the original problem when approached with the same structured prediction estimator. Therefore, one can directly reuse algorithms developed to solve the original inference problem (Nowak-Vila et al., 2019). Finally, with a similar approach to the one in Sec. 3, we can derive the following algorithm for the supremum loss

$$f_{sp}(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Y}} \sup_{y_i\in S_i} \sum_{i=1}^{n} \alpha_i(x)\ell(z, y_i).$$

In the next section, we will use the average candidates as baseline to compare with the algorithm proposed in this paper, as the supremum loss consistently performs worth, as it is not fitted for partial labelling.

5. Applications and experiments

In this section, we will apply Eq. (7) to some synthetic and real datasets from different prediction problems and compared with the average estimator presented in the section above, used as a baseline. Code is available online.[2]

5.1. Classification

Classification consists in recognizing the most relevant item among m items. The output space is isomorphic to the set of indices $\mathcal{Y} = [\![1, m]\!]$, and the usual loss function is the 0-1 loss

$$\ell(z, y) = \mathbf{1}_{y\neq z}.$$

[2] https://github.com/VivienCabannes/partial_labelling

It has already been widely studied with several approaches that are calibrated in non ambiguous deterministic setting, notably by Cour et al. (2011). The infimum loss reads $L(z, S) = \mathbf{1}_{z\notin S}$, and its risk in Eq. (4) is minimized for

$$f(x) \in \operatorname*{arg\,max}_{z\in\mathcal{Y}} \mathbb{P}\left(z \in S \,|\, X = x\right).$$

Based on data $(x_i, S_i)_{i\leq n}$, our estimator Eq. (7) reads

$$f_n(x) = \operatorname*{arg\,max}_{z\in\mathcal{Y}} \sum_{i;z\in S_i} \alpha_i(x).$$

For this instance, the supremum loss is really conservative, only learning from set that are singletons $L_{sp}(z, S) = \mathbf{1}_{S\neq\{z\}}$, while the average loss is similar to the infimum one, adding an evidence weight depending on the size of S, $L_{ac}(z, S) \simeq \mathbf{1}_{z\notin S}/\,|S|$.

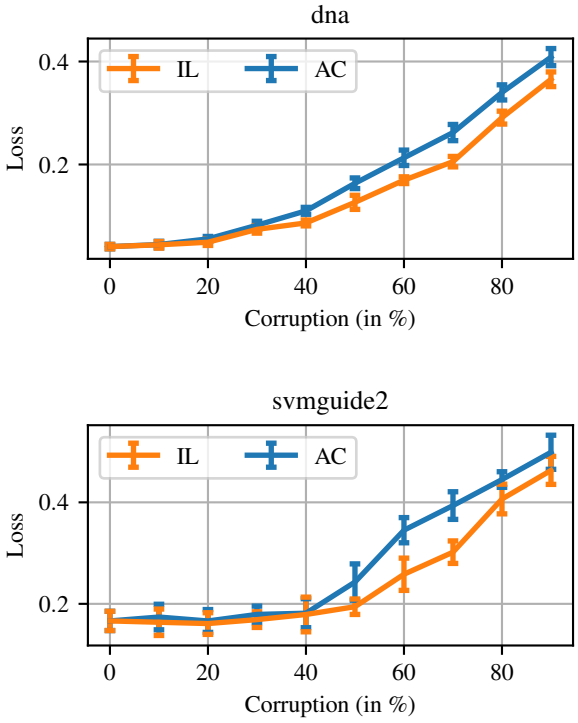

Figure 2. Classification. Testing risks (from Eq. (1)) achieved by *AC* and *IL* on the "dna" and "svmguide2" datasets from *LIBSVM* as a function of corruption parameter c, when the corruption is as follows: for y being the most present labels of the dataset, and $z' \neq z$, $\mathbb{P}\left(z' \in S \,|\, Y = z\right) = c \cdot \mathbf{1}_{z=y}$. Plotted intervals show the standard deviation on eight-fold cross-validation. Experiments were done with the Gaussian kernel. See all experimental details in Appendix B.

Real data experiment. To compare *IL* and *AC*, we used *LIBSVM* datasets (Chang & Lin, 2011) on which we corrupted labels to simulate partial labelling. When the corruption is uniform, the two methods perform the same.

Yet, when labels are unbalanced, such as in the "dna" and "svmguide2" datasets, and we only corrupt the most frequent label $y \in \mathcal{Y}$, the infimum loss performs better as shown in Fig. 2.

5.2. Ranking

Ranking consists in ordering m items based on an input x that is often the conjunction of a user u and a query q, ($x = (u, q)$). An ordering can be thought as a permutation, that is, $\mathcal{Y} = \mathfrak{S}_m$. While designing a loss for ranking is intrinsincally linked to a voting system (Arrow, 1950), making it a fundamentally hard problem; Kemeny (1959) suggested to approach it through pairwise disagreement, which is current machine learning standard (Duchi et al., 2010), leading to the Kendall embedding

$$\varphi(y) = (\operatorname{sign}(y_i - y_j))_{i<j\leq m},$$

and the Kendall loss (Kendall, 1938), with $C = m(m-1)/2$

$$\ell(y, z) = C - \varphi(y)^T \varphi(z).$$

Supervision often comes as partial order on items, *e.g.*,

$$S = \{y \in \mathfrak{S}_m \,|\, y_i > y_j > y_k, y_l > y_m\}.$$

It corresponds to fixing some coordinates in the Kendall embedding. In this setting, *AC* and *SP* are not consistent, as one can recreate a similar situation to the one in Sec. 4, considering $m = 3$, $a = (1, 2, 3)$, $b = (2, 1, 3)$ and $c = (1, 3, 2)$ (permutations being represented with $(\sigma^{-1}(i))_{i\leq m}$), and supervision being most often $S = (1 > 3) = \{a, b, c\}$ and sometimes $S = (1 > 3 > 2) = \{c\}$.

Minimum feedback arc set. Dealing with Kendall's loss requires to solve problem of the form,

$$\arg\min_{y\in S} \langle c, \varphi(y)\rangle,$$

for $c \in \mathbb{R}^{m^2}$, and constraints due to partial ordering encoded in $S \subset \mathcal{Y}$. This problem is an instance of the constrained minimum feedback arc set problem.We provide a simple heuristic to solve it in Appendix B.5, which consists of approaching it as an integer linear program. Such heuristics are analyzed and refined for analysis purposes by Ailon et al. (2005); van Zuylen et al. (2007).

Algorithm specification. At inference, the infimum loss requires to solve:

$$f_n(x) = \arg\max_{z\in\mathcal{Y}} \sup_{(y_i)\in S_i} \sum_{i=1}^{n} \alpha_i(x) \langle \varphi(z), \varphi(y_i)\rangle. \quad (7)$$

It can be approached with alternate minimization, initializing $\varphi(y_i) \in \operatorname{Conv}(\varphi(S_i))$, by putting 0 on unseen observed pairwise comparisons, then, iteratively, solving a minimum feedback arc set problem in z, then solving several minimum feedback arc set problems with the same objective, but different constraints in (y_i). This is done efficiently using warmstart on the dual simplex algorithm.

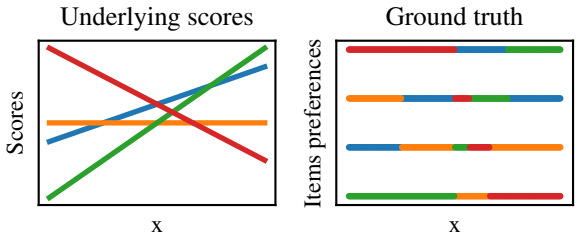

Figure 3. Ranking, experimental setting. Colors represent four different items to rank. Each item is associate to a utility function of x shown on the left figure. From those scores, is retrieved an ordering y of the items as represented on the right.

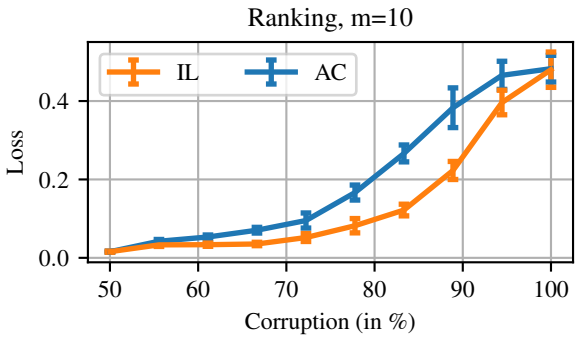

Figure 4. Ranking, results. Testing risks (from Eq. (1)) achieved by *AC* and *IL* as a function of corruption parameter c. When $c = 1$, both risks are similar at 0.5. The simulation setting is the same as in Fig. 2. The error bars are defined as for Fig. 2, after cross-validation over eight folds. *IL* clearly outperforms *AC*.

Synthetic experiments. Let us consider $\mathcal{X} = [0, 1]$ embodying some input features. Let $\{1, \dots, m\}$, $m \in \mathbb{N}$ be abstract items to order, each item being linked to a utility function $v_i \in \mathbb{R}^{\mathcal{X}}$, that characterizes the value of i for x as $v_i(x)$. Labels $y(x) \in \mathcal{Y}$ are retrieved by sorting $(v_i(x))_{i\leq m}$. To simulate a problem instance, we set v_i as $v_i(x) = a_i \cdot x + b_i$, where a_i and b_i follow a standard normal distribution. Such a setting is illustrated in Fig. 3.

After sampling x uniformly on $[0, 1]$ and retrieving the ordering y based on scores, we simulate partial labelling by randomly loosing pairwise comparisons. The comparisons are formally defined as coordinates of the Kendall's embed-

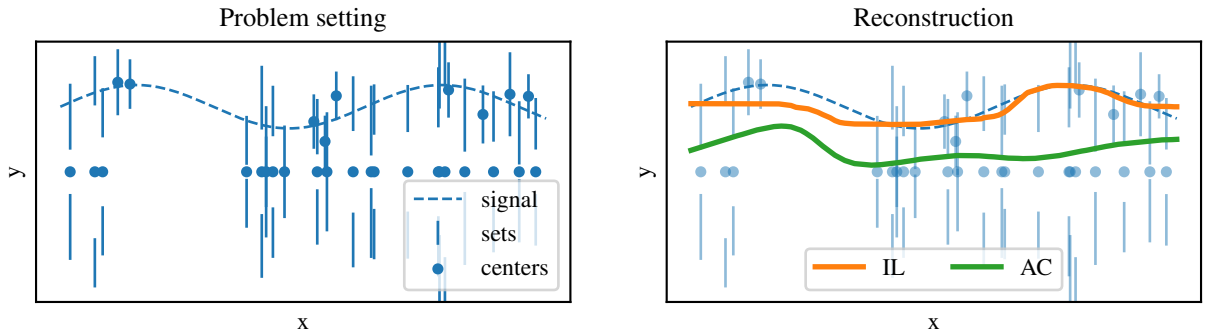

Figure 5. Partial regression on $\mathbb{R}$. In this setting we aim at recovering a signal $y(x)$ given upper and lower bounds on it amplitude, and in thirty percent of case, information on its phase, or equivalently in $\mathbb{R}$, its sign. *IL* clearly outperforms the baseline. Indeed *AC* is a particular ill-fitted method on such a problem, since it regresses on the barycenters of the resulting sets.

ding $(\varphi(y)_{jk})_{jk\leq m}$. To create non symmetric perturbations we corrupt more often items whose scores differ a lot. In other words, we suppose that the partial labelling focuses on pairs that are hard to discriminate. The corruption is set upon a parameter $c \in [0, 1]$. In fact, for $m = 10$, until $c = 0.5$, our corruption is fruitless since it can most often be inverted based on transitivity constraint in ordering, while the problem becomes non-trivial with $c \geq 0.5$. In the latter setting, *IL* clearly outperforms *AC* on Fig. 4.

5.3. Partial regression

Partial regression is an example of non discrete partial labelling problem, where $\mathcal{Y} = \mathbb{R}^m$ and the usual loss is the Euclidean distance

$$\ell(y, z) = \|y - z\|^2 .$$

This partial labelling problem consists of regression where observation are sets $S \subset \mathbb{R}^m$ that contains the true output y instead that y. Among others, it arises for example in economical models, where bounds are preferred over approximation when acquiring training labels (Tobin, 1958). As an example, we will illustrate how partial regression could appear for some phase problems arising with physical measurements. Suppose a physicist want to measure the law between a vectorial quantity Y and some input parameters X. Suppose that, while she can record the input parameters x, her sensors do not exactly measure y but render an interval in which the amplitude $\|y\|$ lays and only occasionally render its phase $y/\|y\|$, in a fashion that leads to a set of candidates S for y. The geometry over ℓ^2 makes it a perfect example to showcase superiority of the infimum loss as illustrated in Fig. 5.

In this figure, we consider $\mathcal{Y} = \mathbb{R}$ and suppose that Y is a deterministic function of X as shown by the dotted blue line signal. If, for a given x_i, measurements only provides that $|y_i| \in [1, 2]$ without the sign of y_i, a situation where the phase is lost, this correspond to the set $S_i = [-2, -1] \cup [1, 2]$, explaining the shape of observed sets that are symmetric around the origin. Whenever the acquired data has no phase, which happen seventy percent of the time in our simulation, AC will target the set centers, explaining the green curve. On the other hand, IL is aiming at passing by each set, which explains the orange curve, crossing all blue bars.

6. Conclusions

In this paper, we deal with the problem of weakly supervised learning, beyond standard regression and classification, focusing on the more general case of arbitrary loss functions and structured prediction. We provide a principled framework to solve the problem of learning with partial labelling, from which a natural variational approach based on the infimum loss is derived. We prove that under some identifiability assumptions on the labelling process the framework is able to recover the solution of the original supervised learning problem. The resulting algorithm is easy to train and with strong theoretical guarantees. In particular we prove that it is consistent and we provide generalization error rates. Finally the algorithm is tested on simulated and real datasets, showing that when the acquisition process of the labels is more adversarial in nature, the proposed algorithm performs consistently better than baselines. This paper focuses on the problem of partial labelling, however the resulting mathematical framework is quite flexible in nature and it is interesting to explore the possibility to extend it to tackle also other weakly supervised problems, as imprecise labels from non-experts (Dawid & Skene, 1979), more general constraints over the set $(y_i)_{i\leq n}$ (Quadrianto et al., 2009) or semi-supervision (Chapelle et al., 2006).

Acknowledgements

The authors would like to thanks Alex Nowak-Vila for precious discussions, Yann Labbé for coding insights, as well

as the reviewers and Eyke Hüllermeier for their precious time and remarks. 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 (grant SEQUOIA 724063).

References

Ailon, N., Charikar, M., and Newman, A. Aggregating inconsistent information: ranking and clustering. In *37th Symposium on Theory of Computing*, 2005.

Arrow, K. J. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58, 1950.

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

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101, 2006.

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

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7, 2006.

Chang, C. and Lin, C. LIBSVM: A library for support vector machines. *ACM TIST*, 2, 2011.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. The MIT Press, 2006.

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.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *Journal of Machine Learning Research*, 12, 2011.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28, 1979.

Denoeux, T. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering*, 25, 2013.

Duchi, J. C., Mackey, L. W., and Jordan, M. I. On the consistency of ranking algorithms. In *27th International Conference on Machine Learning*, 2010.

Fernandes, E. and Brefeld, U. Learning from partially annotated sequences. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2011.

Grandvalet, Y. Logistic regression for partial labels. In *9th Information Processing and Management of Uncertainty*, 2002.

Guillaume, R., Couso, I., and Dubois, D. Maximum likelihood with coarse data based on robust optimisation. In *10th International Symposium on Imprecise Probability*, 2017.

Hüllermeier, E. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55, 2014.

Hüllermeier, E. and Cheng, W. Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2015.

Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172, 2008.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Neural Information Processing Systems 15*, 2002.

Joulin, A., Bach, F. R., and Ponce, J. Discriminative clustering for image co-segmentation. In *23th Conference on Computer Vision and Pattern Recognition*, 2010.

Kemeny, J. G. Mathematics without numbers. *Daedalus*, 88, 1959.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30, 1938.

Korba, A., Garcia, A., and d'Alché-Buc, F. A structured prediction approach for label ranking. In *Neural Information Processing Systems 31*, 2018.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th International Conference on Machine Learning*, 2001.

Luo, J. and Orabona, F. Learning from candidate labeling sets. In *Neural Information Processing Systems 23*, 2010.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*, 2018.

Mayhew, S., Chaturvedi, S., Tsai, C., and Roth, D. Named entity recognition with partially annotated training data. In *23rd Conference on Computational Natural Language Learning*, 2019.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7, 2006.

Muandet, K., Fukumizu, K., Sriperumbudur, B. K., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learnig*, 10, 2017.

Nguyen, N. and Caruana, R. Classification with partial labels. In *14th International Conference on Knowledge Discovery and Data Mining*, 2008.

Nowak-Vila, A., Bach, F., and Rudi, A. Sharp analysis of learning with discrete losses. In *22nd International Conference on Artificial Intelligence and Statistics*, 2019.

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

Quadrianto, N., Smola, A. J., Caetano, T. S., and Le, Q. V. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10, 2009.

Scholkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26, 2007.

Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica*, 26, 1958.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6, 2005.

van Rooyen, B. and Williamson, R. C. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18, 2017.

van Zuylen, A., Hegde, R., Jain, K., and Williamson, D. P. Deterministic pivoting algorithms for constrained ranking and clustering problems. In *18th Symposium on Discrete Algorithms*, 2007.

Verbeek, J. and Triggs, W. Scene Segmentation with CRFs Learned from Partially Labeled Images. In *Neural Information Processing Systems 20*, 2008.

Wald, A. Statistical decision functions which minimize the maximum risk. *The Annals of Mathematics*, 46, 1945.

Weiss, Y. Segmentation using eigenvectors: a unifying view. *7th International Conference on Computer Vision*, 1999.

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

Yu, H., Jain, P., Kar, P., and Dhillon, I. S. Large-scale multi-label learning with missing labels. In *31th International Conference on Machine Learning*, 2014.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. Learning with local and global consistency. In *Neural Information Processing Systems 16*, 2003.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In *20th International Conference of Machine Learning*, 2003.