Learning from Imprecise Data: Adjustments of Optimistic and Pessimistic Variants

Eyke Hüllermeier, Sébastien Destercke, Ines Couso

Learning from Imprecise Data: Adjustments of Optimistic and Pessimistic Variants

Eyke Hüllermeier[1], Sébastien Destercke[2], and Ines Couso[3]

[1] Paderborn University
Heinz Nixdorf Institute and Department of Computer Science
Intelligent Systems and Machine Learning Group
eyke@upb.de
[2] UMR CNRS 7253 Heudiasyc, Sorbonne Universités,
Université de Technologie de Compiègne, France
sebastien.destercke@hds.utc.fr
[3] Department of Statistics and Operations Research
University of Oviedo
couso@uniovi.es

Abstract. The problem of learning from imprecise data has recently attracted increasing attention, and various methods to tackle this problem have been proposed. In this paper, we discuss and compare two quite opposite approaches, an "optimistic" one that interprets imprecise data in a way that is most favourable for a candidate model, and a "pessimistic" one in which model choice is guided by the most unfavourable interpretation. To avoid an overly extreme behaviour, a modified version of the latter has recently been proposed, which we complement by an adjusted version of the optimistic approach. By presenting the various methods within a common (loss minimization) framework and discussing illustrative examples, we hope to provide some insight into important properties and differences, thereby paving the way for a more formal analysis.

1 Introduction

Superset learning is a specific type of learning from weak supervision, in which the outcome (response) associated with a training instance is only characterized in terms of a set of possible candidates. There are numerous applications in which supervision is partial in that sense [9]. Correspondingly, the superset learning problem has received increasing attention in recent years, and has been studied under various names, such as *learning from ambiguously labelled examples* or *learning from partial labels* [10, 2]. The contributions so far also differ with regard to their assumptions on the incomplete information being provided, and how it has been produced. In this paper, we only assume the actual outcome to be covered by the subset—hence the name *superset* learning.

In spite of the ambiguous, set-valued training data, the goal that is commonly considered in superset learning is to induce a *unique* model, or a set of models that are all deemed optimal (in the sense of fitting the observed data equally

well) and not differentiated any further. This differs from approaches that allow for a set of incomparable, undominated models, resulting for instance from the interval order induced by set-valued loss functions [3], or by the application of conservative, imprecise Bayesian updating rules [11].

In this paper, we reconsider the principle of generalized loss minimization based on the so-called *optimistic superset loss* (OSL) as introduced in [7]. To better understand its nature and possible deficiencies, we contrast the latter with another, in a sense diametral approach based on a "pessimistic" inference principle. Moreover, to compensate for a bias that might be caused by an overly optimistic attitude, we propose an adjustment of the OSL, which can be seen as a counterpart of a corresponding modification of the pessimistic approach [6]. Presenting the various methods within a common framework of loss minimization in supervised learning allows us to highlight some important properties and differences through illustrative examples.

2 Preliminaries

2.1 Setting and Notation

The OSL was introduced in a standard setting of supervised learning with an input (instance) space $\mathcal{X}$ and an output space $\mathcal{Y}$. The goal is to learn a mapping from $\mathcal{X}$ to $\mathcal{Y}$ that captures, in one way or the other, the dependence of outputs (responses) on inputs (predictors). The learning problem essentially consists of choosing an optimal model (hypothesis) h^* from a given model space (hypothesis space) $\mathcal{H}$, based on a set of training data

$$\mathcal{D} = \big\{ (\boldsymbol{x}_n, y_n) \big\}_{n=1}^{N} \in (\mathcal{X} \times \mathcal{Y})^N \ . \tag{1}$$

More specifically, optimality typically refers to optimal prediction accuracy, i.e., a model is sought whose expected prediction loss or *risk*

$$\mathcal{R}(h) = \int L\big(y, h(\boldsymbol{x})\big)\, d\,\mathbf{P}(\boldsymbol{x}, y) \tag{2}$$

is minimal; here, $L : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ is a loss function, and $\mathbf{P}$ is an (unknown) probability measure on $\mathcal{X} \times \mathcal{Y}$ modeling the underlying data generating process.

In the following, we assume hypotheses to be uniquely defined in terms of a parameter θ from an underlying parameter space Θ: $\mathcal{H} = \{h_\theta \,|\, \theta \in \Theta\}$, where h_θ is the hypothesis associated with θ. Selecting an optimal hypothesis $h^* \in \mathcal{H}$ thus reduces to estimating an optimal parameter $\theta^* \in \Theta$.

We are interested in the case where parts of the data are not observed precisely. More specifically, focusing on the output values[4] $y_n \in \mathcal{Y}$, we assume that only supersets $Y_n \subseteq \mathcal{Y}$ are observed. Thus, the learning algorithm does not

[4] The principles of optimistic (and likewise pessimistic) loss minimization also extend to the case of imprecision in the instance features.

have direct access to the (precise) data (1), but only to the (imprecise, coarse, ambiguous) observations

$$\mathcal{O} = \big\{(\boldsymbol{x}_n, Y_n)\big\}_{n=1}^{N} \in (\mathcal{X} \times 2^{\mathcal{Y}})^N \ . \tag{3}$$

In the following, we denote by $\mathbf{Y} = Y_1 \times Y_2 \times \cdots \times Y_N$ the (Cartesian) product of the supersets observed for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. Moreover, each $\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathbf{Y}$ is called an *instantiation* of the imprecisely observed data. More generally, we call a sample $\mathcal{D}$ in (1) an instantiation of $\mathcal{O}$ if the instances $\boldsymbol{x}_n$ coincide and $y_n \in Y_n$ for all $n \in [N] := \{1, \ldots, N\}$.

2.2 Optimistic and Pessimistic learning

According to [7], a candidate $\theta \in \Theta$ is evaluated optimistically in terms of

$$\mathcal{R}^{OPT}_{emp}(\theta) := \min_{\boldsymbol{y} \in \mathbf{Y}} \frac{1}{N} \sum_{n=1}^{N} L\big(y_n, h_\theta(\boldsymbol{x}_n)\big) \ , \tag{4}$$

i.e., in terms of the empirical risk of h_θ in the case of a most favourable selection of the outcomes y_n. Moreover, given a loss L that is decomposable (over examples), the "optimism" can be moved into the loss:

$$\theta^* := \operatorname*{argmin}_{\theta \in \Theta} \mathcal{R}^{OPT}_{emp}(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} L_O\big(Y_n, h_\theta(\boldsymbol{x}_n)\big) \ , \tag{5}$$

with the *optimistic superset loss* (OSL)

$$L_O(Y, \hat{y}) = \min\big\{L(y, \hat{y}) \,|\, y \in Y\big\} \ , \tag{6}$$

which compares (precise) predictions with set-valued observations. A key motivation of the OSL is the idea of *data disambiguation*, i.e., the idea of simultaneously inducing the true model (parameter θ) and reconstructing the values of the underlying precise data.

A completely opposite principle is to replace the optimistic minimum in (4) by a pessimistic maximum [5]. More specifically, this principle was introduced in the realm of statistical inference (instead of supervised learning) with L the logistic loss, i.e., in the setting of maximum likelihood inference. The idea is to evaluate each candidate θ in terms of the worst likelihood it can achieve over all instantiations $\boldsymbol{y} \in \mathbf{Y}$, and to pick the best among these pessimistic evaluations. Expressed in terms of generic loss functions (possibly but not necessarily the logistic loss), this principle would amount to considering

$$\mathcal{R}^{PESS}_{emp}(\theta) := \max_{\boldsymbol{y} \in \mathbf{Y}} \frac{1}{N} \sum_{n=1}^{N} L\big(y_n, h_\theta(\boldsymbol{x}_n)\big) \ , \tag{7}$$

and (again assuming the loss to be decomposable) choosing

$$\theta_* := \operatorname*{argmin}_{\theta \in \Theta} \mathcal{R}_{emp}^{PESS}(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} L_P\big(Y_n, h_\theta(\boldsymbol{x}_n)\big) \tag{8}$$

as a presumably best model, with the *pessimistic superset loss* (PSL)

$$L_P(Y, \hat{y}) = \max\big\{L(y, \hat{y}) \,|\, y \in Y\big\} \ . \tag{9}$$

3 Illustrative Examples

Which of the two approaches to superset learning is more reasonable, the optimistic or the pessimistic one? This question is difficult (or actually impossible) to answer without further assumptions on the coarsening process, i.e., the process that turns precise data into imprecise observations. In the following, to get a better idea of the nature of the two approaches, we illustrate them by some simple examples. We shall refer to the optimistic approach (based on the OSL) as OPT and to the pessimistic one (based on the PSL) and PESS.

3.1 Linear Regression

In linear regression, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and the goal is to learn a linear predictor $h(\boldsymbol{x}) = \boldsymbol{x}^\top \theta = \langle \boldsymbol{x}, \theta \rangle$. Training data is typically assumed to be noisy observations $y_n = \boldsymbol{x}^\top \theta_0 + \epsilon$, where θ_0 is the *ground-truth* parameter and ϵ a noise term (with zero expectation). Correspondingly, in the setting of superset learning, we assume observations $Y_n \ni y_n$. Note, therefore, that Y_n does not necessarily cover the ideal outcome (e.g., the expected value $\mathbb{E}(y \,|\, \boldsymbol{x}_n) = \boldsymbol{x}^\top \theta_0$); instead, just like the precise observation y_n itself, it might be shifted by the noise.

To evaluate predictions $\hat{y} = h(\boldsymbol{x})$, the loss function most commonly used in linear regression is the squared error loss. For the case of interval-valued data $Y = [y_{min}, y_{max}]$, the OSL (6) is then given as follows (cf. Fig. 1):

$$L_O\big([y_{min}, y_{max}], \hat{y}\big) = \begin{cases} (y_{min} - \hat{y})^2 & \text{if } \hat{y} < y_{min} \\ 0 & \text{if } y_{min} \leq \hat{y} \leq y_{max} \\ (\hat{y} - y_{max})^2 & \text{if } y_{max} < \hat{y} \end{cases} \tag{10}$$

Thus, the loss is 0 if the prediction is inside the interval, i.e., if the regression function intersects with the interval, and grows quadratically with the distance from the interval outside. A small one-dimensional example of a set of interval-valued data together with a regression line minimizing (5) is shown in Fig. 2 (left).

The PSL version (9) of the squared error loss is given as follows (cf. Fig. 1):

$$L_P\big([y_{min}, y_{max}], \hat{y}\big) = \begin{cases} (y_{max} - \hat{y})^2 & \text{if } \hat{y} < \frac{1}{2}(y_{min} + y_{max}) \\ (\hat{y} - y_{min})^2 & \text{if } \hat{y} \geq \frac{1}{2}(y_{min} + y_{max}) \end{cases} \tag{11}$$

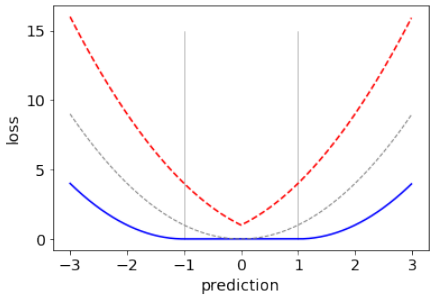

Fig. 1. The OSL (solid line in blue) and PSL (dashed line in red) as extensions of the squared error loss (gray line) in the case of an interval-valued observation (here the interval $[-1, 1]$, indicated by the vertical lines).

As can be seen in Fig. 1, the PSL targets the midpoint of the interval as an optimal "compromise value"; this point minimizes the maximal prediction error possible, and hence the loss function. Moreover, the larger the interval, the stronger the loss function increases. Therefore, PESS is very similar to *weighted linear regression*, where the weight of an example increases with the width of the corresponding interval. The OSL behaves in a quite different way: the larger the interval, the smaller the loss function. Moreover, OSL does not prefer any values inside the interval (e.g., the midpoint) to any other values. Note that, if the data is completely coherent with a (noise-free) linear model, i.e., if there is a regression function intersecting all intervals, then any such function will be optimal for OPT, while this is not necessarily the case for PESS, as PESS may prefer a function not intersecting all intervals (see Fig. 2 (right) for an illustration). Obviously, since the OSL is no longer strictly convex (in contrast with PSL), the optimisation problem solved by OPT may no longer have a unique solution.

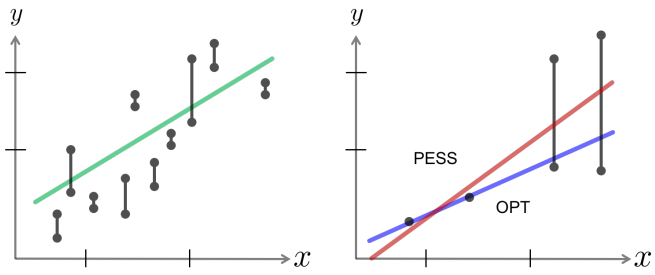

Fig. 2. Left: Linear regression with interval-valued data. Right: Comparison between PESS and OPT for linear regression.

We can also compare OPT and PESS from the point of view of *model updating* or *revision* in the case where new data is observed. Imagine, for example, that a new data point $(\boldsymbol{x}_{N+1}, Y_{N+1})$ is added to the data seen so far. OPT will check for how compatible its current model is with the interval Y_{N+1} and make adjustments only if necessary. In particular, if $\hat{y}_{N+1} = h_\theta(\boldsymbol{x}_{N+1}) \in Y_{N+1}$, i.e., the interval includes the current prediction, the model will not be changed at all, as it is considered fully coherent with the new observation. This also implies that an extremely wide interval will be ignored as being completely uninformative. PESS, on the other side, will always change its current estimate θ, unless $\hat{y}_{N+1} = h_\theta(\boldsymbol{x}_{N+1})$ corresponds exactly to the midpoint of Y_{N+1}; this is because any deviation from this "perfect" prediction is considered as a mistake (or at least a suboptimal choice) that ought to be mitigated.

From the above comments, it is clear that the two strategies may behave quite differently on the same data. OPT assumes that Y_n is a set of candidate values, one of which corresponds to the true measurement. Therefore, fitting one of these candidates, namely the one that is maximally coherent with the model assumption and the rest of the data, is enough. As opposed to this, PESS seeks to fit all values $y_n \in Y_n$ simultaneously, i.e., to find a good compromise prediction $\hat{y}_n$ that is not in conflict with any of the candidates.

It appears that OPT proceeds from a *disjunctive* interpretation of the set Y_n, and considers that the true data will not be chosen so as to systematically put the assumed model in default. In contrast, PESS is more in line with a *conjunctive* interpretation, which makes sense if all the candidates are indeed guaranteed to be possible measurements. One could imagine, for example, that $\boldsymbol{x}_n$ actually characterizes a whole set of entities, and that Y_n is the collection of outputs associated with these entities. As an illustration, suppose we would like to learn a control rule that prescribes an autonomous car the strength of braking depending on its current speed x. Since the optimal strength will also depend on other factors (such as weather conditions), which are ignored (or "integrated out") here, training examples might be interval-valued. For example, depending on further unknown conditions, the optimal strength could be in-between y_{min} and y_{max} for a speed of x Km/h. Adopting a "cautious" model, which minimizes the worst mistake it can make, may look like a reasonable strategy then.

3.2 Logistic Regression

In logistic regression, the goal is to learn a probabilistic classifier

$$h_\theta(\boldsymbol{x}) = \frac{1}{1 + \exp(-\langle \theta, \boldsymbol{x} \rangle)}\,, \tag{12}$$

where $h_\theta(\boldsymbol{x})$ is an estimate of the (conditional) probability $\mathbf{p}(y = 1 \,|\, \boldsymbol{x})$ of the positive class. Inference is done on the basis of the maximum likelihood principle, wich is equivalent to minimizing the log-loss on the training data:

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \sum_{n=1}^{N} L(y_n, h_\theta(\boldsymbol{x}_n))$$

with

$$L(y,p) = -\log\big(py + (1-p)(1-y)\big) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{if } y = 0 \end{cases}$$

Using the representation (12) for the probability p, and the class encoding $\mathcal{Y} = \{-1, +1\}$ instead of $\mathcal{Y} = \{0, 1\}$, the loss can also be written as follows:

$$L(y,s) = \log\big(1 + \exp(-ys)\big) \ ,$$

where $s = \langle \theta, \boldsymbol{x} \rangle$ is the predicted score and ys is the *margin*, i.e., the distance from the decision boundary (to the right side).

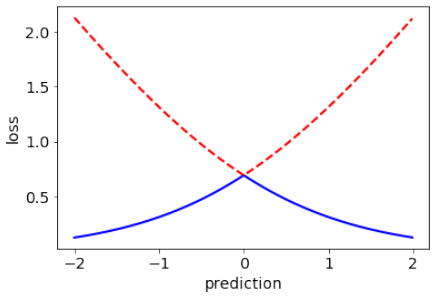

Fig. 3. OSL (blue, solid line) and PSL (red, dashed line) for the logistic loss function.

Since $\mathcal{Y} = \{-1, +1\}$ contains only two elements, there is only one imprecise observation that can be made, namely $Y = \{-1, +1\} = \mathcal{Y}$, and the setting reduces to so-called semi-supervised learning (with a part of the data being precisely labeled, and another part without any supervision). Thus, the OSL is given by

$$L_O(Y,s) = \begin{cases} L(-1,s) & \text{if } Y = \{-1\} \\ L(+1,s) & \text{if } Y = \{+1\} \\ \min\{L(-1,s), L(+1,s)\} & \text{if } Y = \{-1,+1\} \end{cases} \ ,$$

and the pessimistic version L_P by the same expression with min in the third case replaced by max. As a consequence, if an imprecise observation is made, OPT will try to *disambiguate*, i.e., to choose θ such that $ys = y\langle \theta, \boldsymbol{x} \rangle$ is large (and hence p is close to 0 or close to 1); this is in line with a large margin approach, i.e., the learner tries to move the decision boundary away from the data points. Indeed, the generalized loss L_O can be seen as the logistic version of the "hat loss" that is used in semi-supervised learning of support vector machines [1].

As opposed to this, PESS will try to choose θ such that $s \approx 0$ and hence $p \approx \frac{1}{2}$. Obviously, this may lead to drastically different solutions. An example

is shown in Fig. 4, where a few labeled training examples are given (positive in blue and negative in red) and many unlabeled. OPT seeks to maximize the margin of the decision boundary, and hence puts it in-between the two clusters. This is in line with the goal of disambiguation: ideally, the unlabeled examples are far from the decision boundary, which means they are clearly identified as positive or negative. PESS is doing exactly the opposite and tries to have the unlabeled examples close to the decision boundary.

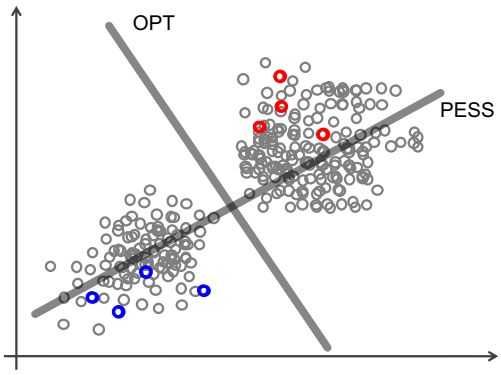

Fig. 4. Logistic regression in a semi-supervised setting: Solutions for OPT and PESS.

This example suggests that PESS is not really appropriate for tackling discriminative learning tasks. To be fair, however, one has to acknowledge that PESS may produce more reasonable results in other scenarios. For example, if the unlabeled examples are not chosen arbitrarily but indeed correspond to those cases that are very close to the true decision boundary, i.e., for which the posterior probability is indeed close to $\frac{1}{2}$, and which could hence be hard to label, then PESS is just doing the right thing.

As another rather extreme example, suppose that the precise observations in Fig. 4 are just the "noisy" cases, whereas all "normal" cases are hidden (the blue class is actually in the upper right and the red class in the lower left). One can imagine, for example, an "adversarial" coarsening process that coarsens all normal cases and only reveals the noise in the data. In this scenario, it is clear that OPT will be completely misled and produce exactly the opposite of the right model. In such adversarial settings [8], PESS (and more generally minimax approaches) may indeed be considered a more reasonable strategy, as it may provide some guarantees in terms of protection with regard to the coarsening process. Anyway, what all these examples are showing is that the reasonableness of an approach strongly depends on which assumptions about the coarsening process can be considered as plausible.

3.3 Statistical Parameter Estimation

As already said, OPT and PESS have been introduced in different contexts. While generalized loss minimization with the OSL was mainly motivated by problems of supervised machine learning, PESS has mostly been considered in a setting of statistical parameter estimation, such as the estimation of the parameter θ of a Bernoulli distribution in coin tossing. In these cases, OPT may tend to produce rather extreme estimates. For example, consider a sample such as

$$1, 0, ?, 0, ?, 1, 1, 1, ?, ? \ ,$$

with p positive outcomes indicated by a 1 (e.g., a coin toss landing heads up), n negative outcomes indicated by a 0, and u unknowns indicated by a ?. One can check that, in the case where $p > n$, OPT will produce the estimate $\theta^* = {}^{p+u}/_{p+u+n}$, based on a corresponding disambiguation in which each unknown is replaced by a positive outcome. More generally, in a multinomial case, all unknowns are supposed to belong to the majority of the precise part of the data. This estimate maximizes the likelihood or, equivalently, minimizes the log-loss

$$L(\theta) = -\sum_{n=1}^{N} X_i \log(\theta) + (1 - X_i) \log(1 - \theta) \ .$$

Such an estimate may appear somewhat implausible. Why should all the unknowns be positive? Of course, one may not exclude that the coarsening process is such that only positives are hidden. In that case, OPT will exactly do the right thing. Still, the estimate remains rather extreme and hence arguable.

In contrast, PESS would try to maximize the entropy of the estimated distribution [4, Corollary 1], which is equivalent to having $\theta^* = 1/2$ in the example given above. While such an estimate may seem less extreme and more reasonable, there is again no compelling reason to consider it more (or less) legitimate than the one obtained by POSS, unless further assumptions are made about the coarsening process. Finally, note that neither POSS nor PESS can produce the estimate obtained by the classical coarsening-at-random (CAR) assumption, which would give $\theta^* = 2/3$.

As a first remark, let us repeat that generalized loss minimization based on OSL was actually not intended, or at least not motivated, by this sort of problem. To explain this point, let us compare the above (statistical estimation) example of coin tossing with the previous (machine learning) example of logistic regression. In fact, the former can be seen as a special case of the latter, with an instance space $\mathcal{X} = \{\boldsymbol{x}_0\}$ consisting of a single instance, such that $\theta = \mathbf{p}(y = 1 \,|\, \boldsymbol{x}_0)$. Correspondingly, since $\mathcal{X}$ has no structure, it is impossible to leverage any *structural assumptions* about the sought model $h : \mathcal{X} \longrightarrow \mathcal{Y}$, which is the basis of the idea of data disambiguation as performed by OPT.

In particular, in the case of coin flipping, each ? can be replaced by any (hypothetical) outcome, independently of all others and without violating any model assumptions. In other words, every instantiation of the coarse data is as plausible as any other. This is in sharp contrast with the case of logistic

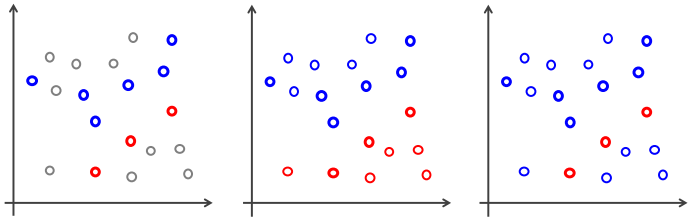

Fig. 5. Coarse data (left) together with two instantiations (middle and right).

regression, where the assumption of a linear model, i.e., the assumption that the probability of success for an input $\boldsymbol{x}$ depends on the spatial position of that point, lets many disambiguations appear implausible. For example, in Fig. 5, the instantiation in the middle, where half of the unlabeled examples are disambiguated as positive and the other half as negative, is clearly more coherent with the assumption of (almost) linearly separable classes than the instantiation on the right, where all unknowns are assigned to the positive class.

In spite of this, examples like the one of coin tossing are indeed suggesting that OSL might be overly optimistic in certain cases. Even in discriminative learning, OSL makes the assumption that the chosen model class is the right one, which may lead to overly confident results should the model choice be wrong. This motivates a reconsideration of the optimistic inference principle and perhaps a suitable adjustment.

4 Adjustments of OSL and PSL

A noticeable property of the previous coin tossing example is a bias of the estimation (or learning) process, which is caused by the fact that a higher likelihood can principally be achieved with a more extreme θ. For example, with $\theta \in \{0, 1\}$, the probability of an "ideal" sample is 1, whereas for $\theta = 1/2$, the highest probability achievable on a sample of size N is $(1/2)^N$. Thus, it seems that, from the very beginning, the candidate estimate $\theta = 1/2$ is put at a systematic disadvantage.

This can also be seen as follows: Consider any sample produced by $\theta = 1$, i.e., a sequence of tosses with heads up. When coarsening the data by covering a subset of the sample, OPT will still produce $\theta = 1$ as an estimate. Roughly speaking, $\theta = 1$ is "robust" toward coarsening. As opposed to this, when coarsening a sample produced with $\theta = 1/2$, OPT will diverge and either produce a smaller or a larger estimate.

4.1 Regularized OSL

One way to counter a systematic bias in disfavour of certain parameters or hypotheses is to adopt a Bayesian approach. Instead of looking at the highest likelihood value $\max_{\boldsymbol{y} \in \mathbf{Y}} \mathbf{p}(\boldsymbol{y} \,|\, \theta)$ of θ across different instantiations of the imprecise

data[5], one may start with a prior π on θ and look at the highest posterior[6]

$$\max_{\boldsymbol{y}\in\mathbf{Y}} \frac{\mathbf{p}\big(\boldsymbol{y}\,|\,\theta\big)\,\pi(\theta)}{\mathbf{p}\big(\boldsymbol{y}\big)}\,,$$

or, equivalently,

$$\max_{\boldsymbol{y}\in\mathbf{Y}} \Big\{ \log\,\mathbf{p}\big(\boldsymbol{y}\,|\,\theta\big) - H(\theta,\boldsymbol{y})\Big\} = \max_{\boldsymbol{y}\in\mathbf{Y}} \left\{ \sum_{i=1}^{N} \log\,\mathbf{p}(y_n\,|\,\theta) - H(\theta,\boldsymbol{y})\right\} \tag{13}$$

with

$$H(\theta,\boldsymbol{y}) \coloneqq \log\,\mathbf{p}(\boldsymbol{y}) - \log\,\pi(\theta) \tag{14}$$

At the level of loss minimization, when ignoring the role of $\boldsymbol{y}$ in (14), this approach essentially comes down to adding a regularization term to the empirical risk, and hence to minimizing the *regularized* OSL

$$\mathcal{R}^{OPT}_{reg}(\theta) \coloneqq \frac{1}{N}\sum_{n=1}^{N} L_O\big(Y_n, h_\theta(\boldsymbol{x}_n)\big) + F(h_\theta)\ \ , \tag{15}$$

where $F(h_\theta)$ is a suitable penalty term.

Coming back to our original motivation, namely that some parameters can principally achieve a higher likelihood than others, one instantiation of F one may think of is the maximal (log-)likelihood conceivable for θ (where the sample can be chosen freely and does not depend on the actual imprecise observations):

$$F(\theta) = -\max_{\boldsymbol{y}\in\mathcal{Y}^N} \log \mathbf{p}\big(\boldsymbol{y}\,|\,\theta\big) \tag{16}$$

In this case, $F(\theta)$ can again be moved inside the loss function L_O in (15):

$$\mathcal{R}^{OPT}_{reg}(\theta) \coloneqq \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_O\big(Y_n, h_\theta(\boldsymbol{x}_n)\big) \tag{17}$$

with

$$\mathcal{L}_O(Y,\hat{y}) \coloneqq \min_{y\in Y} L(y,\hat{y}) - \min_{y\in\mathcal{Y}} L(y,\hat{y})\,. \tag{18}$$

For some losses, such as squared error loss in regression, the adjustment (18) has no effect, because $L(y,\hat{y}) = 0$ can always be achieved for at least one $y \in \mathcal{Y}$. For others, however, $\mathcal{L}_O$ may indeed differ from L_O. For the log-loss in binary

[5] We assume the $\boldsymbol{x}_n$ in the data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ to be fixed.

[6] The obtained bound are similar to the upper expectation bound obtained by the updating rule discussed by Zaffalon and Miranda [11] in the case of a completely unknown coarsening process and precise prior information. However, Zaffalon and Miranda discussed generic robust updating schemes leading to sets of probabilities or sets of models, which is not the intent of the methods discussed in this paper.

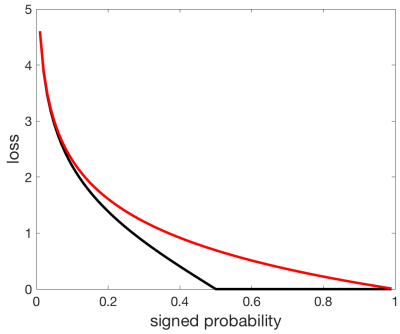

Fig. 6. The adjusted OSL version (19) of the logistic loss (black line) compared to the original version (red line).

classification, for example, the normalizing term in (18) is $\min\{L(0,p), L(1,p)\}$, which means that

$$\mathcal{L}_O(Y,p) = \begin{cases} \log(1-p) - \log(p) & \text{if } Y = \{1\},\ p < 1/2 \\ \log(p) - \log(1-p) & \text{if } Y = \{0\},\ p > 1/2 \\ 0 & \text{otherwise} \end{cases} \quad . \tag{19}$$

A graphical representation of this loss function, which can be seen as a combination of the 0/1 loss (it is 0 for signed probabilities $\geq 1/2$) and the log-loss, is shown in Fig. 6.

4.2 Adjustment of PSL: Min-Max Regret

Interestingly, a similar adjustment, called *min-max regret* criterion, has recently been proposed for PESS [6]. The motivation of the latter, namely to assess a parameter θ in a *relative* rather than *absolute* way, is quite similar to ours. Adopting our notation, a candidate θ is evaluated in terms of

$$\max_{\boldsymbol{y} \in \mathbf{Y}} \left\{ \log \mathbf{p}\big(\boldsymbol{y} \,|\, \theta\big) - \max_{\hat{\theta}} \log \mathbf{p}\big(\boldsymbol{y} \,|\, \hat{\theta}\big) \right\} \quad . \tag{20}$$

That is, θ is assessed on a concrete instantiation $\boldsymbol{y} \in \mathbf{Y}$ by comparing it to the best estimation $\hat{\theta}_{\boldsymbol{y}}$ on that data, which defines the regret, and then the worst comparison over all possible instantiations (the maximum regret) is considered. Like in the case of OSL, this can again be seen as an approximation of (14) with

$$F(\boldsymbol{y}) = \max_{\hat{\theta}} \log \mathbf{p}\big(\boldsymbol{y} \,|\, \hat{\theta}\big) \,,$$

which now depends on $\boldsymbol{y}$ but not on θ (whereas the F in (15) depends on θ but not on $\boldsymbol{y}$). Obviously, the min-max regret principle is less pessimistic than

the original PSL, and leads to an adjustment of PESS that is even somewhat comparable to OPT: The loss of a candidate θ on an instantiation $\boldsymbol{y}$ is corrected by the minimal loss $F(\boldsymbol{y})$ that can be achieved on this instantiation. Obviously, by doing so, the influence of instantiations that necessarily cause a high loss is reduced. But these instantiations are exactly those that are considered as "implausible" and down-weighted by OPT (cf. Section 3.3). See Fig. 7 for an illustrative comparison in the case of coin tossing as discussed in Section 3.3. Note

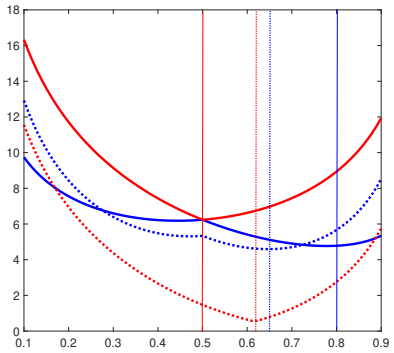

Fig. 7. Loss functions and optimal predictions of θ (minima of the losses indicated by vertical lines) in the case of coin tossing with observations 0, 0, 1, 1, 1, 1, ?, ?, ?: solid blue line for OSL, dashed blue for the regularized OSL version (14) with π the beta(5,5) distribution, solid red for PSL, and dashed red for the adjusted PSL (20).

that (20) does not permit an additive decomposition into losses on individual training examples, because the regret is defined on the entire set of data. Instead, a generalization of (20) to loss functions other than log-loss suggests evaluating each θ in terms of the maximal regret

$$\mathrm{MReg}(\theta) := \max_{\boldsymbol{y} \in \mathbf{Y}} \left(\mathcal{R}_{emp}(\theta, \boldsymbol{y}) - \min_{\hat{\theta}} \mathcal{R}_{emp}(\hat{\theta}, \boldsymbol{y}) \right) , \tag{21}$$

where $\mathcal{R}_{emp}(\theta, \boldsymbol{y})$ denotes the empirical risk of θ on the data obtained for the instantiation $\boldsymbol{y}$. Computing the maximal regret (21), let alone finding the minimizer $\theta^* = \mathrm{argmin}_\theta \, \mathrm{MReg}(\theta)$, appears to be intractable except for trivial cases. In particular, the problem will be hard in cases like logistic regression, where the empirical risk minimizer $\min_{\hat{\theta}} \mathcal{R}_{emp}(\hat{\theta}, \boldsymbol{y})$ cannot be obtained analytically, because then even the evaluation of a single candidate θ on a single instantiation $\boldsymbol{y}$ requires the solution of a complete learning task — not to mention that the minimization over all instantiations $\boldsymbol{y}$ comes on top of this.

5 Concluding Remarks

The goal of our discussion was to provide some insight into the basic nature of the "optimistic" and the "pessimistic" approach to learning from imprecise data. To this end, we presented both of them in a unified framework and highlighted important properties and differences through illustrative examples.

As future work, we plan a more thorough comparison going beyond anecdotal evidence. Even if both approaches deliberately refrain from specific assumptions about the coarsening process, it would be interesting to characterize situations in which they are likely to produce accurate results, perhaps even with formal guarantees, and situations in which they may fail. In addition to a formal analysis of that kind, it would also be interesting to compare the approaches empirically. This is not an easy task, however, especially due to a lack of suitable (real) benchmark data. Synthetic data can of course be used as well, but as our examples have shown, it is always possible to create the data in favour of the one and in disfavour of the other approach.

References

1. O. Chapelle, V. Sindhwani, and S. S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9(Feb):203–233, 2008.
2. T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.
3. I. Couso and L. Sánchez. Machine learning models, epistemic set-valued data and generalized loss functions: an encompassing approach. *Information Sciences*, 358:129–150, 2016.
4. R. Guillaume, I. Couso, and D. Dubois. Maximum likelihood with coarse data based on robust optimisation. In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pages 169–180, 2017.
5. R. Guillaume and D. Dubois. Robust parameter estimation of density functions under fuzzy interval observations. In *9th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA'15)*, pages 147–156, 2015.
6. R. Guillaume and D. Dubois. A maximum likelihood approach to inference under coarse data based on minimax regret. In *Proc. SMPS 2018, Soft Methods in Probability and Statistics*, pages 99–106. Springer, 2019.
7. E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.
8. P. Laskov and R. Lippmann. Machine learning in adversarial environments. *Machine Learning*, 81(2):115–119, 2010.
9. L.P. Liu and T.G. Dietterich. A conditional multinomial mixture model for superset label learning. In *Proc. NIPS*, 2012.
10. N. Nguyen and R. Caruana. Classification with partial labels. In *Proc. KDD 2008, 14th Int. Conf. on Knowledge Discovery and Data Mining*, Las Vegas, USA, 2008.
11. M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821, 2009.