In all Likelihoods: How to Reliably Select Pseudo-Labeled Data for Self-Training in Semi-Supervised Learning

Julian Rodemann JULIAN@STAT-UNI-MUENCHEN.DE
Christoph Jansen CHRISTOPH.JANSEN@STAT.UNI-MUENCHEN.DE
Georg Schollmeyer GEORG.SCHOLLMEYER@STAT.UNI-MUENCHEN.DE
Thomas Augustin THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, Ludwig-Maximilians-Universität (LMU) Munich

Abstract

Self-training is a simple yet effective method within semi-supervised learning. The idea is to iteratively enhance training data by adding pseudo-labeled data. Its generalization performance heavily depends on the selection of these pseudo-labeled data (PLS). In this paper, we aim at rendering PLS more robust towards the involved modeling assumptions. To this end, we propose to select pseudo-labeled data that maximize a multi-objective utility function. The latter is constructed to account for different sources of uncertainty, three of which we discuss in more detail: model selection, accumulation of errors and covariate shift. In the absence of second-order information on such uncertainties, we furthermore consider the generic approach of the generalized Bayesian α-cut updating rule for credal sets. As a practical proof of concept, we spotlight the application of three of our robust extensions on simulated and real-world data. Results suggest that in particular robustness w.r.t. model choice can lead to substantial accuracy gains.[1]

Keywords: Semi-Supervised Learning, Self-Training, Generalized Bayes, Model Selection, Covariate Shift, Generalized Updating Rules

1. Introduction

Labels for observations are burdensome to obtain in a myriad of applied learning tasks ranging from image classification [69] over financial econometrics [66] to genomics [25]. This scarcity of labeled data has given rise to the paradigm of *semi-supervised learning* (SSL). Within SSL, *self-training* (also called pseudo-labeling) is often considered the most straight-forward approach [65, 37, 48]. Self-training follows the general rationale of iteratively assigning pseudo-labels to unlabeled data according to the model's predictions. More precisely, the idea is to predict classes of unlabeled data by means of a model trained on labeled data and include some of the predictions as pseudo-labeled data in the training data, before predicting on the remaining unlabeled data again. This process requires a criterion (called confidence measure) for *pseudo-label selection* (PLS), that is, the selection of pseudo-labeled instances to be added to the training data. What most of these confidence measures have in common is the fact of stemming uniquely and exclusively from one sole model. The paper at hand aims at a selection of pseudo-labeled data with regard to a variety of (fitted) models, rendering PLS robust with regard to model imprecision. The latter can have multiple sources. Section 3 discusses how to deal with three of them in detail: model selection, accumulation of errors and covariate shift. In case such sources are not identifiable, we propose a generic robust approach to PLS in section 4, building on the rich literature on credal sets and generalized Bayesian inference. The remainder of this section discusses related work and introduces semi-supervised learning formally, leaning on [61]. The paper concludes with a brief real world application and some concluding remarks in chapters 5 and 6.

1.1. Semi-supervised Learning

The vast majority of SSL methods is concerned with classification tasks [73, 13]. Loosely leaning on [71], we formalize SSL as follows. Consider labeled data

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \in (\mathcal{X} \times \mathcal{Y})^n \tag{1}$$

and unlabeled data

$$\mathcal{U} = \{(x_i, \mathcal{Y})_i\}_{i=n+1}^{m} \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^{m-n} \tag{2}$$

from the same data generation process, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the categorical target space. The aim of SSL is to learn a predictive classification function $\hat{y}_\theta(x)$ parameterized by θ utilizing both $\mathcal{D}$ and $\mathcal{U}$. The objective can be twofold [71]. On the one hand, one simply aims at labeling $\mathcal{U}$ (transductive learning). On the other hand, and more commonly, both $\mathcal{D}$ and $\mathcal{U}$ can be used to learn a prediction function to predict any unseen test

[1] **Open Science:** Implementations of the proposed methods and reproducible scripts for the experimental analysis are available at: www.github.com/rodemann/reliable-pls.

data (inductive learning) in a more accurate way than only relying on $\mathcal{D}$ as in classical supervised learning.

1.2. Self-training

According to [53] and [73], SSL can be broadly categorized into self-training and co-training. We will focus on the former, whose general idea is commonly described as fitting a model on $\mathcal{D}$ by empirical risk minimization and then exploiting this model's predictions to label $\mathcal{U}$. Typically, those instances from $\mathcal{U}$ are added whose predictions are most confident according to some confidence measure. The predicted probability (probability score) is among the most popular ones [71]. Besides, the predictions' variance as well as a linear combination of variance and probability score are used [55]. Regarding the inclusion of pseudo-labeled data from $\mathcal{U}$ to $\mathcal{D}$, [71] and [35] distinguish between incremental, batch-wise, and amending mechanisms. The incremental approaches label instances one-by-one in a sequential fashion, whereas batch-wise and amending techniques allow for adding of multiple data points or removal of data, respectively. Moreover, [71] differentiate self-training methods into single- and multi-classifier ones, depending on the number of learned classifiers $\hat{y}(x)$ used while labeling. If multiple classifiers are used, they can either be based on the same model class or a variety of models. This is known as single- versus multi-learning, see [71] for instance. Combining and aggregating the predictions and confidence measures of multiple classifiers can be done in various ways. This is slightly related to our proposed model-robust PLS, see sections 3 and 4. The difference is of course that we *select* pseudo-labeled data in the light of multiple models, while multi-learning deploys multiple models for *predicting* pseudo-labels.

Additionally, self-training algorithms may have different stopping criteria [71]. A naive option is to label and add the entire set $\mathcal{U}$. Alternatively, one could stop when $\hat{y}(x)$, the predictive classifier, no longer changes due to $\mathcal{U}$, leaving the remaining data in $\mathcal{U}$ unlabeled. In this paper, we propose both incremental and batch-wise approaches that can be used with any stopping criterion for the purpose of inductive learning.

1.3. Superset Learning

The notion of superset learning is a generalization of semi-supervised learning. Instead of completely unlabeled (i.e. fully ambiguos) data $\mathcal{U} = \{(x_i, \mathcal{Y})_i\}_{i=1}^m \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^m$, superset learning considers $\{(x_i, Y_i)\}_{i=1}^m \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^m$, where $Y_i \subseteq \mathcal{Y}$. In this context, Y_i is regarded a superset of a "true" underlying singleton y_i, thus the name. There exist optimistic as well as pessimistic variants of superset learning and approaches to balance these extreme cases [27, 28, 29, 60]. The general idea is to find a singleton representation (often called instantiation) of the supersets that corresponds to the most predictive (optimistic) or least predictive (pessimistic) model when trained and evaluated on it. In the optimistic case, this can be achieved by minimizing an optimistic version of the empirical risk, the generalized empirical risk: $\frac{1}{n}\sum_{i=1}^n L^*(\hat{y}_i, Y_i) = \frac{1}{n}\sum_{i=1}^n \min_{y \in Y_i} L(\hat{y}_i, y)$ with L^* the *optimistic superset* or *infimum loss* [9].

1.4. Robust Semi-Supervised Learning

The robustness of SSL and in particular of self-training has been widely discussed. [3] propose an information-theoretic approach to pseudo-label prediction which is resistant to covariate shift. [74] worked to make self-training more robust to modeling assumptions by allowing model selection through the deviance information criterion. Coming close to our use of credal sets in section 4, [41, 42] suggest identifying pseudo-labels as sets of probability distributions ("credal self-supervised learning"). Inspired by consistency regularization [8, 67, 76], superset learning [27, 28, 29, 60] and distributional alignment [36], "credal self-supervised learning" aims at decreasing the reliance on a single distributional assumption. Our work follows the same rationale, while being conceptually different: [41, 42] start by imprecisiation of the training data by means of soft labels through data augmentation, thus obtaining set-valued predictions. In this paper, we exploit the expressiveness of credal sets only in the selection phase. Generally, there appears to be a large body of research on robustifying *predictions* in SSL by means of Bayesian techniques [22, 51, 1], weighted likelihood [68], conditional likelihood [23], and joint mixture likelihood [2]. On the other hand, there is only limited (Bayesian) or hardly any (likelihood-based) work regarding robust versions of Bayesian or likelihood-based *selection* of pseudo-labels, which is the very idea of the paper at hand. The authors of [40] quantify the uncertainties of pseudo-labels by mixtures of predictive distributions of a neural net, utilizing MC dropout, thus simulating a Bayesian setup without explicitly considering the posterior predictive. More recently, [52] proposed PLS with respect to the entropy of the pseudo-labels' posterior predictive distribution.

[61] tackle the problem of pseudo-label selection (PLS) in semi-supervised learning from the viewpoint of decision theory, proposing Bayes optimal pseudo-label selection (BPLS). The idea is to make PLS more robust towards the initial fit by marginalizing over the parameters' posterior instead of considering the predictive distribution of a single best parameter vector. While this allows for selecting pseudo-labeled data in light of more than one fit of a given model, BPLS is still restricted to the assumed (type of) model and the distributional assumptions that come with it. This is the very starting point for several robust extensions of BPLS, that will be presented in the main part of this paper, namely

sections 3 and 4. To begin with, we introduce the conditional view on PLS in section 2.1. This allows our understanding of (B)PLS as decision problems, as explained in section 2.2.

2. Pseudo-Label Selection

2.1. Conditional Pseudo-Label Selection

As in standard self-training, we start by fitting a parametric model M with unknown parameter vector $\theta \in \Theta$ on labeled data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. In this work, we assume Θ to be compact and denote $\dim(\Theta) = q$. Note that Bayesian inference smoothly integrates in this setup, since we might state a prior function over Θ for a given parametric model M as $\pi(\theta \mid M)$. We aim at learning the conditional distribution of $p(\boldsymbol{y} \mid \boldsymbol{x})$ through θ from observing features $\boldsymbol{x} = (x_1, \ldots, x_n), x_i \in \mathcal{X}$, and classes $\boldsymbol{y} = (y_1, \ldots, y_n), y_i \in \mathcal{Y}$ in $\mathcal{D}$. As touched upon in sections 1.1 and 1.2, we start by estimating $\hat{\theta} \in \Theta$ from M through the labeled data $\mathcal{D}$, predict on unlabeled data $\mathcal{U}$ and select those predicted (pseudo-labeled) data points that we are most confident in according to some selection criterion and add them to $\mathcal{D}$.

Most importantly, throughout this paper, we do not deal with predicting unknown labels of $\mathcal{U} = \{(x_i, \mathcal{Y})_i\}_{i=n+1}^m$ by the fitted model on $\mathcal{D}$. Rather, we are primarily concerned with the problem of *selecting* from those already predicted. That is, we identify each element in $\mathcal{U} = \{(x_i, \mathcal{Y})_i\}_{i=n+1}^m \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^{m-n}$ with its corresponding prediction $\{(x_i, \hat{y})\}_i$, obtaining $\hat{\mathcal{U}} = \{(x_i, \hat{y})\}_{i=n+1}^m \in (\mathcal{X} \times \mathcal{Y})^{m-n}$. However, we will stick with $\mathcal{U}$ in the following to emphasize that our reasoning holds for any functional $\hat{y}(x)$. This is not to say that we completely abstain from any specifications of the prediction method, see remark 7, where we will rely on maximum-likelihood estimation.

2.2. PLS as Decision Problem

Following [61], we formalize pseudo-label selection as a canonical decision problem with likelihood utility and thus lay the groundwork for several robust extensions of classical decision criteria.

Definition 1 (Canonical Decision Problem) *Define $(\mathbb{A}, \Theta, u(\cdot))$ as decision-theoretic triple with an action space $\mathbb{A}$, an unknown set of states of nature Θ and a utility function $u : \mathbb{A} \times \Theta \to \mathbb{R}$.*

Throughout this section, we are concerned with the decision of selecting pseudo-labeled data, where an action corresponds to the selection of an instance from the unlabeled data $\mathbb{A}_{\mathcal{U}} = \{(z, \mathcal{Y}) \mid \exists i \in \{n+1, \ldots, m\} : (z, \mathcal{Y}) = (x_i, \mathcal{Y})_i \in \mathcal{U}\}$, i.e., instances as actions $\mathbb{A}_{\mathcal{U}} \ni a = (z, \mathcal{Y})$. This is in stark contrast to statistical decision theory, where estimators instead of data are to be selected. The decision for an action is guided by a utility function. Closely following [61] and loosely inspired by [10, 11], we proceed by defining the utility of a selected data point $(z, \mathcal{Y}) = (x_i, \mathcal{Y})_i$ as the plausibility of being generated jointly with $\mathcal{D}$ by a model M with states (parameters) $\theta \in \Theta$ if we include it with its predicted pseudo-label $\hat{y}(z) = \hat{y}(x_i) = \hat{y}_i \in \mathcal{Y}$ in $\mathcal{D} \cup (x_i, \hat{y}_i)$, see definition 2.

Definition 2 (Pseudo-Label Likelihood as Utility) *Given $\mathcal{D}$ and the prediction functional $\hat{y} : \mathcal{X} \to \mathcal{Y}$, we define the following utility function*

$$u \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}$$
$$((z, \mathcal{Y}), \theta) \mapsto u((z, \mathcal{Y}), \theta) = p(\mathcal{D} \cup (z, \hat{y}(z)) \mid \theta, M),$$

which is said to be the pseudo-label likelihood. In the following, for ease of exposition, we will write $\ell(i) := p(\mathcal{D} := p(i \mid \theta, M) \cup (x_i, \hat{y}(x_i)) \mid \theta, M)$ for the pseudo-label likelihood.

Based on this embedding of PLS in decision theory, classical decision criteria such as max-max or the Bayes criterion can be derived. [61, chapter 2.2] shows that the former corresponds to optimistic superset learning [28] and the latter to the posterior predictive of data to be pseudo-labeled $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D}, M)$, subsequently called pseudo posterior predictive (PPP). The *max-max-criterion* is defined by $\Phi_m \colon \mathbb{A}_{\mathcal{U}} \to \mathbb{R}; a \mapsto \max_\theta u(a, \theta)$. Each element of $\arg\max \Phi_m$ is then called a *max-max-action*. The *Bayes-criterion* given π is defined by $\Phi_\pi \colon \mathbb{A}_{\mathcal{U}} \to \mathbb{R}; a \mapsto \mathbb{E}_\pi(u(a, \theta))$. Each element of $\arg\max \Phi_\pi$ is then called *Bayes-action*.

3. Robust PLS: In All Likelihood*s*

Within common approaches to self-training in SSL, it might well be possible to generalize and robustify models used for *predicting* pseudo-labels. In the following, however, we aim at robust *selection* of pseudo-labeled data, see section 2.1. To this end, we will modify the generic utility function (definition 2) and the respective Bayes criterion [61, chapter 2.2] to account for three frequent sources of uncertainty and imprecision: model selection, accumulation of errors and covariate shift. Instead of relying on likelihood utilities from models that are assumed to be correct "in all likelihood", we suggest relying on all likelihood**s** from multiple models.

3.1. Model Selection: Reversing Occam's razor

An obvious and ubiquitous source of imprecision is the model choice: The likelihood under which distributional assumption (and corresponding model) should be taken into account? So far, we have defined the pseudo-label likelihood as the one under the model M that we have

used for predicting pseudo-labels. Albeit, this is far from necessary. As discussed above, our conditional approach to choosing pseudo-labeled data renders this selection completely orthogonal to predicting pseudo-labels. Instead of defining the utility function (see 2) as the likelihood of observing the pseudo-labeled data under the assumptions of model M, we might as well consider $\tilde{M}$ or a weighted sum of likelihoods under several models. In what follows, we start with the generic case of any finite number of different models that can be parameterized in a meaningful way and work our way through nested models, ending with nested generalized linear models, and discuss how to account for their specifications in PLS.

3.1.1. Generic Case

Start by considering any $M_1, \ldots, M_K$, $K < \infty$, different parametric models specified on respective parameter spaces $\Theta_1, \ldots, \Theta_K$. Denote by $\tilde{\Theta} = \times_{k=1}^{K} \Theta_k$ their Cartesian product and by $f_k : \tilde{\Theta} \to \Theta_k$, $k \in \{1, \ldots, K\}$ the projections from the Cartesian product to each Θ_k. We can easily extend the pseudo-label likelihood utility (definition 2) to account for several models, inducing a multiobjective decision problem.

Definition 3 (Multi-Model Likelihood Utility) *As in definition 2 consider $\mathcal{D}$ and pseudo-labels $\hat{y} \in \mathcal{Y}$ from $\hat{y} : \mathcal{X} \to \mathcal{Y}$ as given. The K-dimensional utility function*

$$\begin{aligned} u &: \mathbb{A}_{\mathcal{U}} \times \tilde{\Theta} \to \mathbb{R}^K \\ &((x_i, \mathcal{Y})_i, \theta) \mapsto (\ell(i,1), \ldots, \ell(i,K))' \end{aligned}$$

shall be called multi-model likelihood. We write $\ell(i,k) = p(i \mid f_k(\theta), M_k) = p(\mathcal{D} \cup (z, \hat{y}(z)) \mid f_k(\theta), M_k)$ with $\theta_k \in \Theta_k$ for brevity. Let K again denote the number of models under consideration.

For the optimization of such a multiobjective utility considered in definition 3 one is faced with a multicriteria decision problem. For such decision problems there are lots of solution strategies. One modern way to deal with a multidimensional utility function was recently proposed in [32]. The idea is – utilizing that each *single* dimension considered is perfectly cardinal – to embed the image of the utility function into a *preference system* $\mathcal{A}$, i.e. into a specific order-theoretic structure allowing for modeling spaces with locally cardinal scale of measurement.[2] Each such preference system is then describable by a set of functions $\mathcal{N}_{\mathcal{A}}$, where each element of this set is of the form $\phi : [0,1]^K \to [0,1]$.

[2] A preference system is a triplet $\mathcal{A} = [A, R_1, R_2]$ consisting of a non-empty set $A \neq \emptyset$, a pre-order $R_1 \subseteq A \times A$ on A, and a pre-order $R_2 \subseteq R_1 \times R_1$ a on R_1. Intuitively, the relation R_1 captures the available ordinal information, whereas R_2 encodes the information's cardinal part.

The selection of the optimal unlabeled data would then consequently be based on this same set $\mathcal{N}_{\mathcal{A}}$. To generalize the already mentioned Bayes criterion to this set of utility functions, there are a lot of possibilities (for a compilation of these see in particular [31]). We will only briefly discuss here the one among them that does not need to make any additional assumptions and is a consequential generalization of first-order stochastic dominance to our partial cardinal setting. The idea of this generalization is straightforward: If still π denotes the prior distribution on the set Θ of states of nature (= parameters), then now – instead of choosing unlabeled data that maximize expected utility w.r.t. some fixed utility function – we exclude all unlabeled data which is expectation-dominated by some other data for all compatible functions $\phi : [0,1]^K \to [0,1]$. More formally, the solution to the decision problem from definition 3 with respect to this generalized stochastic dominance criterion is then given by the set $\mathbb{A}_{\mathcal{U}}^{\pi}$ defined by $\{a \mid \nexists a' : d_\pi(a', a) \geq 0 \wedge d_\pi(a, a') < 0\}$, where, for $a_1, a_2 \in \mathcal{U}_{\mathbb{A}}$, we set $d_\pi(a_1, a_2) = \inf_{\phi \in \mathcal{N}_{\mathcal{A}}} \left[\mathbb{E}_\pi(\phi \circ u(a_1, \cdot)) - \mathbb{E}_\pi(\phi \circ u(a_2, \cdot))\right]$.

Importantly, note that all elements remaining in the above set are *incomparable* with respect to the considered criterion of optimality, that is, each of them is an equally plausible candidate for the best next unlabeled data point. In case domain-specific knowledge induces a preference for some of the models under consideration that can be expressed by weights, one might as well simply scalarize the single likelihoods as follows.

Definition 4 (Weighted Sum of Likelihoods) *The utility function $u : \mathbb{A}_{\mathcal{U}} \times \tilde{\Theta} \to \mathbb{R}$;*

$$((x_i, \mathcal{Y})_i, \theta) \mapsto \sum_k w_k \cdot \ell(i,k),$$

with weights $w_k \in (0,1)$, $k \in \{1, \ldots, K\}$ summing up to 1, shall be called weighted sum of likelihoods.

The respective Bayes criterion (cf. section 2.2) with multi-model likelihood utility is a weighted sum of posterior predictives of pseudo-labeled data (cf. ibid.). This fact follows directly from theorem 2 in [61] as well as from the additivity and homogeneity of the expected value.

Remarkably, the following should be noted: The Bayes-optimal pseudo-labeled data, i.e. the optimal solutions of the decision problem for selecting pseudo-labeled data according to the Bayes criterion, are always elements of the set $\mathbb{A}_{\mathcal{U}}^{\pi}$ considered before. This means in particular that the aforementioned generalized stochastic dominance and the Bayes criterion based on multi-model likelihood utility are compatible in the sense that the latter – independent of the concrete weights – ensures that no labels excluded by the former are chosen. This suggests the following recommendation for criterion selection in concrete application

situations: If no content-motivated way of choosing the weights of the multi-model likelihood utility is available, further analysis should rely on the set $\mathbb{A}^{\pi}_{\mathcal{U}}$ alone. If, on the contrary, there is the possibility to determine the weights informed by the content, the Bayes criterion based on the multi-model likelihood utility provides more precise and – then also non-arbitrary – results. We now consider a case where a natural choice of weights via penalization of model complexity is appropriate, namely nested models.

3.1.2. NESTED MODELS

Now let the models under consideration $M_1, \dots, M_K$, $K < \infty$, be nested in the sense of $\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K$. We can interpret the so-induced hierarchy on the parameter space such that the lower $k \in \{1, \dots, K\}$, the simpler the hypothesis space. Aiming at regularization of PLS, we could penalize the respective likelihood utilities of more complex models. In definition 4, this could imply e.g. setting $w_k = \frac{\dim(\Theta_k)}{\dim(\Theta_K)}$ for all $k \in \{1, \dots, K\}$.[3]

However, we will opt for a safer approach that guarantees plausibility of at least some pre-specified level τ under all models $M_1, \dots, M_K$. We therefore draw on the common practice of thresholding selection criteria when selecting pseudo-labeled data in self-training. That is, not only one data point with highest selection function value but all above a threshold are to be selected. We propose to extent this to an intersection of thresholds resulting from likelihood utilities from different models.

Definition 5 (Bayesian Multi-Model Threshold Criterion) *As in definition 2, let $(x_i, \mathcal{Y})_i$ be any decision (selection) from $\mathbb{A}_{\mathcal{U}}$. We assign utility to each $(x_i, \mathcal{Y})_i$ given $\mathcal{D}$ and pseudo-labels $\hat{y} \in \mathcal{Y}$ by the multi-model likelihood utility function from definition 3. Now consider the following thresholding Bayes criterion $\Phi_{\tau,\xi,\pi} \colon \mathbb{A}_{\mathcal{U}} \to \mathbb{R}$*

$$a \mapsto \Phi_{\tau,\xi,\pi}(a) = \begin{cases} 0, & \exists k : \mathbb{E}_{\pi}(\ell(i,k)) < \tau \\ 0.5, & \forall k : \tau < \mathbb{E}_{\pi}(\ell(i,k)) < \xi, \\ 1, & \textit{else.} \end{cases}$$

again with $\ell(i,k) = p(i \mid f_k(\theta), M_k)$, $k \in \{1, \dots, K\}$, and $\xi > \tau$ some pre-specified thresholds.

Note that this corresponds to thresholding all pseudo posterior predictive, respectively, see section 2.2. For parametric models like additive regressions with $K = dim(\Theta)$ we can exploit the hierarchy among models induced by the number of parameters K. Before running the procedure (see algorithm 1), we start by thresholding pseudo-labeled data based on the full model (K covariates). We refit with lower threshold if no data is selected. If a positive number of data is selected, we kick off the algorithm. We begin decreasing k in a step-wise manner and terminate the process if none of the pseudo-labeled data that were selected in all previous rounds makes it past the threshold. The pseudo-code in algorithm 1 describes the procedure.

Algorithm 1 Reversed Occam's Razor

```
Data: 𝒟, 𝒰, set S_{K+1} = ∅
Result: 𝒟
for k ∈ {K, ..., 1} do
    for i ∈ {1, ..., |𝒰|} do
        predict 𝒴 ∋ ŷ_i = ŷ(x_i)
        evaluate 𝔼_π(ℓ(i, k))
    end
    select 𝒰 ⊇ S_k = {(x_i, ŷ_i)_i | Φ_{τ,ξ,π}(a) = 1, a ~ i}
    if S_k ∩ S_{k+1} ≠ ∅ : update 𝒟 = 𝒟 ∪ S_l
    else stop
end
```

We thus ensure not only $\forall k \in \{1, \dots, K\} : \mathcal{S}_k \neq \emptyset$, but also $\bigcap_{k=1,\dots,K} \mathcal{S}_k \neq \emptyset$. That is, among those elements that can be explained equally well by a fixed model, we opt for those that are explained similarly well by simpler models. This can be viewed as reversing Occam's razor, since we are concerned with selecting data instead of hypotheses. Occam's time-honored razor advocates selecting the hypothesis with least assumptions among competing hypotheses that have the same explanatory power regarding a single phenomenon. Conversely, we consider multiple phenomena and choose those ones which can still be explained by the simplest hypothesis from a set of competing ones. Occam's razor can be operationalized by Bayesian statistics through the marginal likelihood (or Bayesian evidence), see [33, 54, 44, 43] for instance. Recall that the Bayesian selection of pseudo-labeled data corresponds to selection with regard to the posterior predictive which is nothing but a marginalized version of the pseudo-labeled data's likelihood.[4] Just like in model selection by Bayesian evidence or Bayes factors (i.e., ratios of marginal likelihoods), we are concerned with how well data can be explained by a model. The only difference is that we are interested in comparing (pseudo-)data by how likely it is given a model and not vice versa.

3.2. Accumulation of Errors: In All Posteriors?

The most inherent uncertainty in PLS is caused by the fact that pseudo-labeled data are treated as ground truths in subsequent iterations.

[3] One could also weight the likelihoods $\ell(i,1), \dots, \ell(i,K)$ directly in the general case of the multiobjective utility from definition 3.

[4] Marginalized with regard to the posterior.

Definition 6 (Multi-Label Likelihood as Utility) *As in definition 2, let $(z, \mathcal{Y})$ be any decision (selection) from $\mathbb{A}_{\mathcal{U}}$. Conversely to definition 2, we now consider not only the predicted pseudo-labels $\hat{y}_i \in \mathcal{Y}$, but also all other hypothetical labels $\tilde{y}_i \in \mathcal{Y} \setminus \{\hat{y}_i\}$. Denote by $\tilde{y}_{i,j} \in \mathcal{Y}$ all possible labels for $(x_i, \mathcal{Y})_i$ with $j \in \{1, \ldots, J\}$ and $J = |\mathcal{Y}|$. We assign utility to each $(x_i, \mathcal{Y})_i$ by the following utility function $u \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}$;*

$$((z, \mathcal{Y}), \theta) \mapsto \sum_{j=1}^{J} w_j \cdot p(\mathcal{D} \cup (z, \tilde{y}_{i,j}) \mid \theta, M)$$

with weights $w_j \in (0, 1)$ summing up to 1. This utility function shall be called multi-label likelihood.

Again, the respective Bayes criterion is a weighted sum of posterior predictives of pseudo-labeled data (cf. section 2.2), because of theorem 2 in [61] and the additivity and homogeneity of the expected value. A logical choice for the weights $w_j \in (0, 1)$ would be the predicted probability of the respective j-th label, i.e. $p((z, \mathcal{Y}) = (z, \tilde{y}_{j,i}))$. This appears quite intuitive. However, while allowing to characterize the unlabeled data points by their plausibilty with hypothetically assigned labels one is still forced to add them with their actually predicted label.

As of now, we loosen this restriction. Notably, definition 2 and thus all subsequent deliberations depended on a model M as well as on already predicted labels. We have relaxed the former dependency, while having left the latter untouched. The following remark calls this into question.

Remark 7 (Sub-Optimal Labels Are Not Redundant) *Consider $u \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}$ from definition 2 with $\hat{y}_i = \hat{y}_{\hat{\theta}_{ML}}(x_i)$ and $\hat{\theta}_{ML} = \arg\max_\theta p(\mathcal{D} \mid \theta, M)$ the maximum-likelihood estimator. Furthermore, consider $u \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}$;*

$$((z, \mathcal{Y}), \theta) \mapsto u((z, \mathcal{Y}), \theta) = p(\mathcal{D} \cup (z, \tilde{y}(z)) \mid \theta, M),$$

where $z = x_i$ and $\tilde{y}(z) = \tilde{y}_i = \hat{y}_{\tilde{\theta}}(x_i)$ with any sub-optimal $\tilde{\theta} \in \Theta$ such that $p(\mathcal{D} \mid \tilde{\theta}, M) \leq p(\mathcal{D} \mid \hat{\theta}_{ML}, M)$. It holds that the max-max-action $a^*_m = \max_a \max_\theta u(a, \theta)$ w.r.t. u does generally not have lower utility than the max-max-action $\tilde{a}^*_m = \max_a \max_\theta u(a, \theta)$ w.r.t. $\tilde{u}$. To see this, let a^*_m be the max-max action under u as above. It holds $a^*_m = \max_a \max_\theta (p(\mathcal{D} \cup (z, \hat{y}_i) \mid \theta, M)) = \max_a p(\mathcal{D} \cup (z, \hat{y}_i) \mid \hat{\theta}_{ML}, M)$. Analogously, $\tilde{a}^*_m$ maximizes $p(\mathcal{D} \cup (z, \tilde{y}_i) \mid \hat{\theta}_{ML}, M)$. As both $(z, \tilde{y}_i)$ and $(z, \hat{y}_i)$ were not considered in ML estimation, we cannot make any statement about the relation of $u(a^*_m)$ to $\tilde{u}(\tilde{a}^*_m)$. The same holds for the Bayes criterion, as also the posterior of θ does not include $(z, \tilde{y}_i)$ and $(z, \hat{y}_i)$ either.

Motivated by this remark, let us now consider the standard utility (definition 2) on a different action space $\tilde{\mathbb{A}}_{\mathcal{U}} = \{(z, y_j) \mid y_j \in \mathcal{Y} \text{ and } \exists i \in \{n+1, \ldots, m\} : (z, \mathcal{Y}) = (x_i, \mathcal{Y})_i \in \mathcal{U}\}$ and a modified (full) Bayes criterion that accounts for a prior ρ on $\mathcal{Y}$ that weights labels proportional to the predictive distribution from the prediction step before, i.e., $\mathbb{E}_\rho(\Phi_\pi(a)) = \mathbb{E}_\rho \mathbb{E}_\pi(u(a, \theta))$.

Proposition 8 (Full Bayes Equates Weighted Utility) *In case of $w_j = \rho(y_j)$ the Bayes criterion under multi-label utility (definition 6) defined on $\tilde{\mathbb{A}}_{\mathcal{U}}$ instead of $\mathbb{A}_{\mathcal{U}}$ equals the (full) Bayes criterion $\mathbb{E}_\rho(\Phi_\pi(a))$ on $\mathbb{A}_{\mathcal{U}}$.*

Proof $\mathbb{E}_\rho \mathbb{E}_\pi(u(a, \theta)) = \int_{\mathcal{Y}} \int_\Theta u(a, \theta)\, d\pi(\theta)\, d\rho(y_j) = \int_{\mathcal{Y}} p(\mathcal{D} \cup (z, y_j) \mid \theta, M)\, d\rho(y_j) = \sum_{\mathcal{Y}} p(\mathcal{D} \cup (z, y_j) \mid M)\, \rho(y_j) = \sum_j p(\mathcal{D} \cup (z, y_j) \mid M)\, \rho(y_j)$ ■

3.3. Covariate Shift

Selection criteria typically render some unlabeled data more likely to be added than others [59]. In the course of self-training, this can lead to a distributional shift of X, often referred to as covariate shift. Depending on the stopping criterion, this covariate shift can be propagated to the final model, potentially harming the model's interpretability by techniques from the realm of interpretable machine learning (IML). For instance, regions in the covariate space where data is scarce are detrimental to reliable estimates of partial dependencies [18]. Notably, this distributional shift affects all previously discussed selection criteria for PLS. In this subsection, we discuss possible extensions that aim at selecting pseudo-labeled data that are optimal with regard to both the *de facto* selected data $\mathcal{D}$ and a hypothetical *i.i.d.* sample $\mathcal{D}'$ that we generate by drawing pseudo-labeled data randomly. In the spirit of the multi-model likelihood utility (definition 3) and in complete analogy to the previously discussed generalizations, we can define a multi-data likelihood utility, rendering PLS robust with regard to covariate shift. The above discussed decision criteria apply as well. Further note that in this special case of a bi-objective, one might also proceed with an interval-valued utility (loss) function as e.g. in [63, section 3.2].

Definition 9 (Multi-Data Likelihood Utility) *We assign utility to each $(x_i, \mathcal{Y})_i$ given $\mathcal{D}$, $\mathcal{D}'$ and the prediction functional $\hat{y} : \mathcal{X} \to \mathcal{Y}$ by the following bi-objective utility function $u \colon \mathbb{A}_{\mathcal{U}} \times \Theta \to \mathbb{R}^2$;*

$$((z, \mathcal{Y}), \theta) \mapsto (\ell_{\mathcal{D}}(i), \ell_{\mathcal{D}'}(i))',$$

with $\ell_{\mathcal{D}}(i) = p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta, M)$ and $\ell_{\mathcal{D}'}(i) = p(\mathcal{D}' \cup (x_i, \hat{y}_i) \mid \theta, M)$.

4. Updating by α-cuts

All robust extensions of PLS discussed in section 3 require some second-level information about the involved uncertainties (e.g., model choice, previous confidence, covariate shift). Aiming at an agnostic and universally robust approach to PLS, we turn to imprecise probabilities [75, 5] and credal sets [38, 39], more specifically to the fruitful frameworks of convex sets of priors [62], Γ-maximin [64] and α-cut updating [11, 12].

4.1. Updating Credal Sets

Due to our aforementioned general skepticism regarding the initial model fit $\hat{\theta}$, we would like to weaken the influence of the likelihood on the posterior in a general way. This can be achieved by means of generalizing Bayesian analysis [75, 62, 5]. Again, we can avail ourselves of rich decision theoretical literature dating back to [16, 34, 7]. We will borrow from the theory on Max-E-Min [34] or equivalently Γ-maximin, see for instance [64, 7, 19, 72, 26]. To this end, we introduce a convex set of priors $\Pi \subseteq \{\pi(\theta) \mid \pi(\cdot)$ a probabilty measure on $(\Theta, \sigma(\Theta))\}$ with Θ compact as above and $\sigma(\cdot)$ an appropriate σ-algebra.[5]. The rough idea now is this: After observing data, we base our selection (action) on the prior from Π that corresponds to the lowest posterior from the set of resulting posteriors. In other words, we hedge against the worst-case prior. In a nutshell, we select the pseudo-labeled instance that would have had the highest expected utility (likelihood) if we had specified the prior in such a way that it contradicted the (potentially overfitted) model's likelihood the most. The respective decision criterion would be the Γ-maximin criterion $\Phi_\Pi : \mathbb{A}_\mathcal{U} \to \mathbb{R}; a \mapsto \Phi(a) = \underline{\mathbb{E}}_\Pi(u(a,\theta))$ with $\underline{\mathbb{E}}_\Pi(u(a,\theta)) = \inf_{\pi\in\Pi} \mathbb{E}(u(a,\theta))$ the lower expectation, which we assume to be affinely superadditive (thus equating coherent lower previsions) in the following. This will allow us to exploit the α-cut updating rule introduced by [12] for lower previsions. The lower expectation corresponds to the posterior predictive with regard to the posterior that results from updating the prior $\pi^*(\cdot) \in \Pi$ that has the lowest value in the maximum-likelihood estimator $\hat{\theta}_{ML}$.

Such an approach, however, might be too much of a good thing, since its respective decisions can completely disregard the likelihood, not to mention its high sensitivity towards Π. Instead, we opt for an updating rule of credal sets leaning on [11, 12][6]: Cattaneo's α-cut updating rule with $\alpha \in (0,1)$, also referred to as "soft revision" [4]. Its rough idea is to only update those priors whose respective marginal likelihood (evidence) is larger or equal than α times the corresponding maximum marginal likelihood. In other words, the priors whose (relative) likelihood is below α are discarded from the set of lower expectations, before updating all prior lower expectations to posterior lower expectations in this set. This implies restricting the set of alle posteriors to

$$\{\pi \in \Pi \mid m(\pi) \geq \alpha \cdot \max_\pi m(\pi)\}, \tag{3}$$

with $m(\ell, \pi) = \int_\Theta \ell(\theta)\pi(\theta)d\theta$ the marginal likelihood. This way, we can make sure no decision is made in complete disregard of the likelihood, i.e., based on a θ with tiny likelihood.

What is more, the α-cut updating rule allows for a dynamically adaptive selection of pseudo-labelled data. Note that each predicted pseudo-label $\hat{y}$ comes with a predicted probability $\hat{p}_{\hat{y}} \in [0,1]$ for $\hat{y}$ to be the true label. After selecting $(x_i, \mathcal{Y})_i$ with respective $(x_i, \hat{y}_i)$, the probability $\hat{p}_{\hat{y}}$ represents our belief in the data $\mathcal{D} \cup (x_i, \hat{y}_i)$ under which the subsequent model's likelihood is specified.[7] More generally, in iteration t of SSL, our belief in the pseudo-labeled data is $\prod_{t=1}^T \hat{p}_{\hat{y},t}$. We thus could update Π in iteration t by α-cuts such that $\alpha_t = \prod_{t=1}^T \hat{p}_{\hat{y},t}$. The interpretation of such an adaptive α-cut rule is this: The less we trust the pseudo-labeled data, the wider the cuts should be, since we want to make sure not to down-weight a θ only because our possibly flawed data says so. Vice versa, if we trust the pseudo-labeled data, we can be more restrictive with regard to the cuts. While providing this strong intuition, we could not find any guarantees for an updating rule of this kind so far. Hence, in what follows, we will motivate an updating rule for SSL based on the expected regret of having considered specific predictions from one specific model in PLS.

4.2. A Regret-Based Updating Rule

The previous deliberations on model selection (section 3.1) and non-redundancy of sub-optimal labels (remark 7) motivate our modification of the α-cut updating rule for PLS: We update Π such that our Bayes action has some quantifiable guarantee with regard to a regret (as ratio, see definitions 10 through 12) that stems from both the possibly wrong labels and the possibly wrong models.[8] Thus, we start by quantifying these two regrets as random variables on Θ, before defining the total regret as a (posterior) expectation of a function of the two regrets.

[5] The priors in Π can reflect uncertainty regarding prior information, but might as well represent priors near ignorance, see e.g. [6, 46, 45, 56, 57, 58]

[6] Updating rules of similar nature have already been introduced by [50, 49, 20]. Notably, [21, p. 46f] introduced the special case of $\alpha = 1$ as "type 2 maximum likelihood", see also [7, section 3.5.4].

[7] Not without a dash of impudence, we might as well borrow from frequentist reasoning and interpret $1 - \hat{p}_{\hat{y}}$ as frequency of error.

[8] Note that reasoning with both sets of priors and model imprecision is reminiscent of [75, chapter 8]

Definition 10 (Label-Induced Regret) *Consider $\tilde{y}_{i,j} \in \mathcal{Y}$ all possible labels for $(x_i, \mathcal{Y})_i$ with $j \in \{1, \dots, J\}$ and $J = |\mathcal{Y}|$. As in remark 7, let $\tilde{u}_j$ be the pseudo-label likelihood $\tilde{u}_j(\cdot, \cdot)$ (definition 2) with $\tilde{y}_i = \tilde{y}_{i,j}$. Furthermore, set $\hat{y}_{i,h} = \hat{y}_{\hat{\theta}}(x_i)$ as actually predicted label, see remark 7. For any given decision a^* and any θ, the function $r_l(\cdot, a^*) \colon \Theta \to \mathbb{R}$;*

$$\theta \mapsto r_l(\theta, a^*) = \sup_j \frac{u_j(\theta, a^*)}{\tilde{u}_h(\theta, a^*)}$$

is said to be the label-induced regret.

Definition 11 (Model-Induced Regret) *Let $M_1, \dots, M_K$ and $\Theta_1, \dots, \Theta_K$ denote all models under consideration and their parameter spaces, respectively, as well as $\tilde{\Theta} = \times_{k=1}^K \Theta_k$ their Cartesian product. As in definition 3, consider as $\ell(i, k) = p(i \mid f_k(\theta), M_K)$ the likelihood utility of selecting $(x_i, \mathcal{Y})_i$ given model M_k with the projection on Θ_k. In analogy to definition 10, denote by M_h the actually used model. For any decision $a^* \hat{=} i^*$ and any $\theta \in \tilde{\Theta}$, the function $r_m(\cdot, a^*) \colon \tilde{\Theta} \to \mathbb{R}$;*

$$\theta \mapsto r_m(\theta, a^*) = \sup_k \frac{\ell(i^*, k)}{\ell(i^*, h)}$$

is said to be the model-induced regret.

Definition 12 (Total Prediction Regret in SSL) *Denote by $\tilde{u}_{j,k}(\theta, a^*)$ the utility of $a^* \hat{=} i^*$ with prediction $\tilde{y}_{i^*,j}$ under model M_k. The function $r(\cdot, a^*) \colon \tilde{\Theta} \to \mathbb{R}$*

$$\theta \mapsto r(\theta, a^*) = \frac{\sup_{j,k} \tilde{u}_{j,k}(\theta, a^*)}{\tilde{u}_{h,h}(\theta, a^*)}$$

shall be be called total (prediction) regret.

Definition 13 (Expected Total Regret Functional) *Based on definition 12, the expectation functional $\Theta \times \Pi \to \mathbb{R}; (\theta, \pi) \mapsto \mathbb{E}_\pi(r(\theta, a^*))$ for given $a^* \in \mathbb{A}_\mathcal{U}$ with posterior $\pi \in \Pi$ is said to be the expected total regret functional.*

We can now define an α-cut updating rule such that the posterior credal set is

$$\Pi_\alpha = \{\pi \in \Pi \mid m(\ell_{h,h}, \pi) \geq \alpha \cdot \sup_{j,k} m(\ell_{j,k}, \pi)\}. \quad (4)$$

Note that this is just a robustified version of the generic α-cut updating according to equation 3, such that it gives us the following guarantee with regard to the expected regret.

Proposition 14 (Myopic Regret-Guarantee of α-Cuts) *Bayes optimal selections a^* of pseudo-labeled data under the above α-cut updating rule have expected total regret $\mathbb{E}_\pi(r(\theta, a^*)) \leq \frac{1}{\alpha}$ for any posterior $\pi \in \Pi$.*

Proof Consider any $\pi \in \Pi_\alpha$. It holds $\forall a \in \mathbb{A}_\mathcal{U} : m(\ell_{h,h}, \pi) \geq \alpha \cdot \sup_{j,k} m(\ell_{j,k}, \pi)$. With $m(\ell, \pi)$ the marginal likelihood w.r.t. to π we get: $\forall a \in \mathbb{A}_\mathcal{U} : \int_\Theta \ell_{h,h}(\theta)\pi(\theta)d\theta \geq \alpha \cdot \sup_{j,k} \int_\Theta \ell_{j,k}(\theta)\pi(\theta)d\theta \implies \forall a \in \mathbb{A}_\mathcal{U} : \mathbb{E}_\pi(\ell_{h,h}(\theta)) \geq \alpha \cdot \mathbb{E}_\pi(\sup_{j,k} \ell_{j,k}(\theta)) \geq \alpha \cdot \sup_{j,k} \mathbb{E}_\pi(\ell_{j,k}(\theta))$. In particular for $a^* \in \mathbb{A}_\mathcal{U}$ we have $\frac{1}{\alpha} \geq \frac{\sup_{j,k} \mathbb{E}_\pi(\tilde{u}_{j,k}(a^*, \theta))}{\mathbb{E}_\pi(\tilde{u}_{h,h}(a^*, \theta))} \geq \frac{\mathbb{E}_\pi(\sup_{j,k} \tilde{u}_{j,k}(a^*, \theta))}{\mathbb{E}_\pi(\tilde{u}_{h,h}(a^*, \theta))} = \mathbb{E}_\pi(r(\theta, a^*))$ with $\ell_{j,k}(\theta) = \tilde{u}_{j,k}(a, \theta)$. ■

The α-cut updating rule was motivated as continuous updating rule by [12]. This continuity still holds tor the regret-based α-cut updating, as follows directly from [12, theorem 3].

4.3. Generalized Stochastic Dominance under IP

In the case of using the multi-model likelihood utility from definition 3 (rather than a weighted-sum of its components) together with a credal-prior Π, the criterion of generalized stochastic dominance addressed in section 3.1.1 can also be easily adapted. Instead of the solution set $\mathbb{A}_\mathcal{U}^\pi$ used under precise π, here we would move to the solution set $\mathbb{A}_\mathcal{U}^\Pi$ robustified under the IP model and defined by

$$\{a \mid \nexists a' : D(a', a) \geq 0 \wedge D(a, a') < 0\}, \quad (5)$$

where, for $a_1, a_2 \in \mathcal{U}_\mathbb{A}$, we set $D(a_1, a_2) = \inf_{\pi \in \Pi} d_\pi(a_1, a_2)$. The interpretation of the set $\mathbb{A}_\mathcal{U}^\Pi$ robustified by Π is similar to the interpretation of the set $\mathbb{A}_\mathcal{U}^\pi$ under precise π: It contains all pseudo-labeled data a which are not strictly dominated with respect to generalized stochastic dominance by another pseudo-labeled data a' for no matter which prior $\pi \in \Pi$. Put formally, we thus have that $\mathbb{A}_\mathcal{U}^\Pi = \bigcap_{\pi \in \Pi} \mathbb{A}_\mathcal{U}^\pi$. Again, similar to the α-cuts method, there are ways to reduce the set of non-excludable pseudo-labels by transitioning from sets $\mathcal{N}_\mathcal{A}$ and Π to (reasonably chosen) subsets $\mathcal{N} \subset \mathcal{N}_\mathcal{A}$ and $\tilde{\Pi} \subset \Pi$ in the definition of the set $\mathbb{A}_\mathcal{U}^\Pi$.

Such a reduction of the set might be desirable, since – depending on the richness of sets $\mathcal{N}_\mathcal{A}$ and Π – set $\mathbb{A}_\mathcal{U}^\Pi$ might contain too many (possibly even all) available options. In the case of the set $\mathcal{N}_\mathcal{A}$, a natural way of reduction is discussed in [31] and further deepened in [30]: Instead of considering all possible representatives ϕ of the underlying preference system, here it is proposed to consider only those that evaluate strict comparability in the underlying partial order above some pre-specified threshold $\xi \in [0, 1]$. Also for the reduction of the set Π a completely natural possibility offers itself: One can simply shrink the set Π by transitioning to the set $\tilde{\Pi} = \Pi_\alpha$ from equation (4) for some reasonable value of α. Of course, also combinations of both reduction methods can be used.

5. Application

Most of the above decision criteria require the computation of the pseudo posterior predictive (PPP) that involves a possibly intractable integral. MCMC sampling is the usual Bayesian way to circumvent such issues. This in turn usually comes at the cost of some computational hurdles. In order to avoid them, we lean on the analytical approximation of the PPP proposed [61, chapter 3]. For the sake of computational feasibility, we further approximate the log-likelihood given $\mathcal{D} \cup (x_i, \hat{y}_i)$ by the log-likelihood given $\mathcal{D}$, obtaining: $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D}, M) \approx 2\ell(\hat{\theta}_{ML}) - \frac{1}{2}\log|I(\hat{\theta}_{ML})|$ with $I(\hat{\theta}_{ML})$ the Fisher information-matrix. We use this approximation to implement three of the above proposed extensions of PLS: multi-label utility (def. 6) as both unweighted and weighted (see proposition 8) sum as well as multi-model utility (def. 3). We benchmark semi-supervised logistic regression with these robust PLS criteria against four common PLS criteria (probability score, posterior predictive (Bayes action), likelihood (max-max action) and predictive variance) as well as a supervised baseline. For the latter, we abstain from self-training and only use the labeled data for training. Experiments are run on simulated binomially distributed data and real world data sets from the UCI machine learning repository [15]. Since target classes are fairly balanced in all data sets, we compare the methods w.r.t. to (test) accuracy. We average the test accuracy for all data sets over a number of repeated self-training rounds each with a new random train-test split. The results are promising: For simulated data, PLS w.r.t multi-model PPP achieves accuracy gains of up to 15 percentage points.[9]

Here, we spotlight the application of our methods on the banknote data [17, 70] that contains measures (diagonal length, bottom margin, length of bill) of 100 genuine and 100 counterfeit Swiss franc banknotes. The learning task at hand is to classify banknotes based on these covariates. Figure 1 shows the average accuracy (evaluated on unseen test data, averaged over 40 repetitions) of different PLS methods for 80% unlabeled data. For the multi-model approach, all possible covariate combinations were considered. While multi-model PPP outperforms competing PLS methods, the multi-label extension fails to even beat the supervised baseline. Apparently, it is not worth considering alternative classifications given the initial supervised accuracy is that high (~ 0.966).

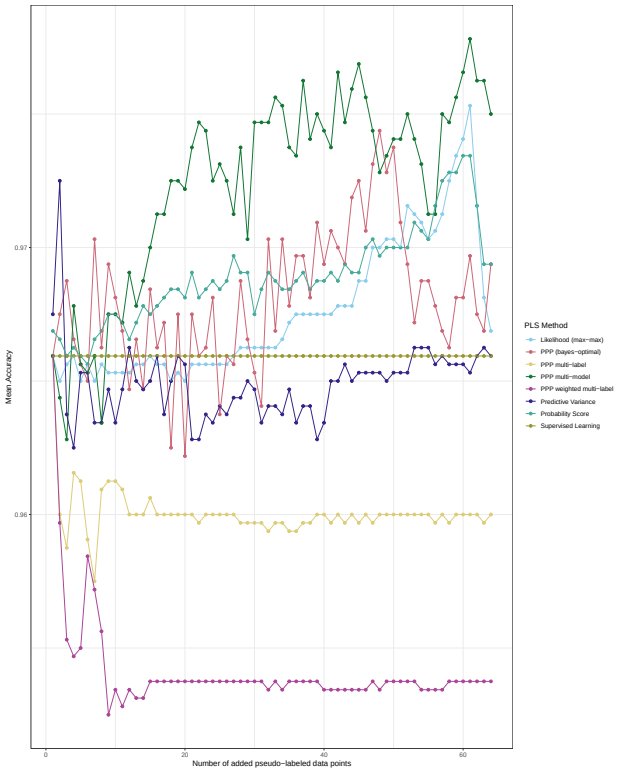

Figure 1: Results from Banknote Data.

6. Discussion

We have introduced a number of robust extensions of PLS, some of which in turn surfaced avenues for future work. For instance, the accumulated expected errors (section 3.2) could be used as adaptive learning rate in fractional Bayesian updating [77, 24, 14]. Future work might also focus on implementing and testing the generic generalizations based on α-cuts, as introduced in section 4. Conclusively, PLS appears to be a promising field for applying existing fruitful frameworks for robust statistical learning such as generalized Bayesian updating using credal sets or more specific multi-model and multi-label robustification. Most of them can potentially be easily transferred to PLS when taking the view on PLS as decision problem. This might not only increase the credibility of the inference by weakening the assumptions, leaning on Manski's "law of decreasing credibility" [47]. It can also, as preliminary evidence suggests, increase predictive performance substantially. In particular, our experiments indicate that considering alternative model specifications as well as non-predicted labels in PLS appears to be a promising and fruitful. Further research is also needed on clarifying interactions among different kind of robustifications (between multi-label and multi-model PLS, for instance).

[9] For **all results**, more details on the experiments and reproducible code, please refer to www.github.com/rodemann/reliable-pls.

Acknowledgments

Georg Schollmeyer would like to thank the LMU mentoring program for support.

Author Contributions

Julian Rodemann developed the main idea of robust PLS extensions that account for model selection, accumulation of error and covariate shift. He drafted and wrote the majority of the paper. Julian Rodemann further implemented robust PLS. He also conceived and conducted the experimental analyses. Christoph Jansen contributed the idea of deploying α-cut updating rules for robust PLS. Its regret-based adaption was developed by Julian Rodemann. Christoph Jansen contributed several passages on solving robust PLS problems w.r.t. generalized stochastic dominance. Georg Schollmeyer and Christoph Jansen also aided with making technical notations more concise. Thomas Augustin, Georg Schollmeyer, Christoph Jansen and Julian Rodemann further contributed by stimulating discussions and detailed proof-reading.

References

[1] Ryan Prescott Adams and Zoubin Ghahramani. Archipelago: nonparametric Bayesian semi-supervised learning. In *International Conference on Machine Learning*, pages 1–8, 2009.

[2] Massih-Reza Amini and Patrick Gallinari. Semi-supervised logistic regression. In *15th European Conference on Artificial Intelligence*, volume 2, page 11, 2002.

[3] Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. An information-theoretical approach to semi-supervised learning under covariate-shift. In *International Conference on Artificial Intelligence and Statistics*, pages 7433–7449. PMLR, 2022.

[4] Thomas Augustin and Georg Schollmeyer. Comment: on focusing, soft and strong revision of choquet capacities and their role in statistics. 2021.

[5] Thomas Augustin, Frank P. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. John Wiley, Chichester, 2014.

[6] Alessio Benavoli and Marco Zaffalon. Prior near ignorance for inferences in the k-parameter exponential family. *Statistics*, 49(5):1104–1140, 2015.

[7] James O. Berger. *Statistical decision theory and Bayesian analysis.* Springer, Berlin., 2nd edition, 1985.

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[9] Vivien Cabannnes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labelling through the infimum loss. In *International Conference on Machine Learning*, pages 1230–1239. PMLR, 2020.

[10] Marco EGV Cattaneo. *Statistical decisions based directly on the likelihood function*. PhD thesis, ETH Zurich, 2007.

[11] Marco EGV Cattaneo. Likelihood decision functions. *Electronic Journal of Statistics*, 7:2924–2946, 2013.

[12] Marco EGV Cattaneo. A continuous updating rule for imprecise probabilities. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 426–435. Springer, 2014.

[13] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning. Adaptive computation and machine learning series*. MIT Press, 2006.

[14] R. de Heide, A. Kirichenko, P. Grünwald, and N. Mehta. Safe-Bayesian generalized linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2623–2633. PMLR, 2020.

[15] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.

[16] Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4): 643–669, 1961.

[17] Bernhard Flury. *Multivariate statistics: a practical approach*. Chapman & Hall, Ltd., 1988.

[18] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29: 1189–1232, 2001.

[19] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

[20] Itzhak Gilboa and David Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory*, 59(1): 33–49, 1993.

[21] Irving John Good. *Good Thinking: The Foundations of Probability and its Applications*. U of Minnesota Press, 1983.

[22] Jonathan Gordon and José Miguel Hernández-Lobato. Combining deep generative and discriminative models for Bayesian semi-supervised learning. *Pattern Recognition*, 100:107156, 2020.

[23] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

[24] P. Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

[25] Hüseyin Anil Gündüz, Martin Binder, Xiao-Yin To, René Mreches, Philipp C Münch, Alice C McHardy, Bernd Bischl, and Mina Rezaei. Self-genomenet: Self-supervised learning with reverse-complement context prediction for nucleotide-level genomics data. 2021.

[26] Peijun Guo and Hideo Tanaka. Decision making with interval probabilities. *European Journal of Operational Research*, 203(2):444–454, 2010.

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

[28] Eyke Hüllermeier and Weiwei Cheng. Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275. Springer, 2015.

[29] Eyke Hüllermeier, Sébastien Destercke, and Ines Couso. Learning from imprecise data: adjustments of optimistic and pessimistic variants. In *International Conference on Scalable Uncertainty Management (SUM)*, pages 266–279. Springer, 2019.

[30] C. Jansen, H. Blocher, T. Augustin, and G. Schollmeyer. Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty. *International Journal of Approximate Reasoning*, 144:69–91, 2022.

[31] Christoph Jansen, Georg Schollmeyer, and Thomas Augustin. Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *International Journal of Approximate Reasoning*, 98:112–131, 2018.

[32] Christoph Jansen, Georg Schollmeyer, and Thomas Augustin. Multi-target decision making under conditions of severe uncertainty. *arXiv preprint*, 2022.

[33] Harold Jeffreys. *The Theory of Probability*. Clarendon Press, Oxford, 1939.

[34] Eduard Kofler and Günter Menges. *Entscheidungen bei unvollständiger Information*. Springer, 1976.

[35] Georgios Kostopoulos, Stamatis Karlos, Sotiris Kotsiantis, and Omiros Ragos. Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems*, 35(2):1483–1500, 2018.

[36] Alex Kurakin, Colin Raffel, David Berthelot, Ekin Dogus Cubuk, Han Zhang, Kihyuk Sohn, and Nicholas Carlini. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations (ICLR)*, 2020.

[37] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, International Conference on Machine Learning*, number 2, page 896, 2013.

[38] Isaac Levi. On indeterminate probabilities. *J Philos*, 71:391–418, 1974.

[39] Isaac Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Rrobability, and Chance*. MIT press, 1980.

[40] Shuangshuang Li, Zhihui Wei, Jun Zhang, and Liang Xiao. Pseudo-label selection for deep semi-supervised learning. In *IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 1–5. IEEE, 2020.

[41] Julian Lienen and Eyke Hüllermeier. Credal self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 14370–14382, 2021.

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

[43] Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, and Andrew Gordon Wilson. Bayesian model selection, the marginal likelihood, and generalization. In *International Conference on Machine Learning*, pages 14223–14247, 2022.

[44] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[45] Francesca Mangili. A prior near-ignorance Gaussian process model for nonparametric regression. In *International Symposium on Imprecise Probabilities (ISIPTA) 2015*.

[46] Francesca Mangili and Alessio Benavoli. New prior near-ignorance models on the simplex. *International Journal of Approximate Reasoning*, 56:278–306, 2015.

[47] Charles Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

[48] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, 2006.

[49] Serafin Moral. Calculating uncertainty intervals from conditional convex sets of probabilities. In *Uncertainty in Artificial Intelligence*, pages 199–206. Elsevier, 1992.

[50] Serafín Moral and Luis M De Campos. Updating uncertain information. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 58–67. Springer, 1990.

[51] Yin Cheng Ng, Nicolò Colombo, and Ricardo Silva. Bayesian semi-supervised learning with graph gaussian processes. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[52] Gaurav Patel, Jan P Allebach, and Qiang Qiu. Seq-ups: Sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6180–6190, 2023.

[53] Nitin Namdeo Pise and Parag Kulkarni. A survey of semi-supervised learning methods. In *International conference on computational intelligence and security*, volume 2, pages 30–34. IEEE, 2008.

[54] Carl Rasmussen and Zoubin Ghahramani. Occam's razor. *Advances in neural information processing systems*, 13, 2000.

[55] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[56] Julian Rodemann. Robust generalizations of stochastic derivative-free optimization. Master's thesis, LMU Munich, 2021.

[57] Julian Rodemann and Thomas Augustin. Accounting for imprecision of model specification in Bayesian optimization. *Poster presented at International Symposium on Imprecise Probabilities (ISIPTA)*, 2021.

[58] Julian Rodemann and Thomas Augustin. Accounting for Gaussian process imprecision in Bayesian optimization. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*, pages 92–104. Springer, 2022.

[59] Julian Rodemann, Sebastian Fischer, Lennart Schneider, Malte Nalenz, and Thomas Augustin. Not all data are created equal: Lessons from sampling theory for adaptive machine learning. *Poster presented at International Conference on Statistics and Data Science (ICSDS), Institute of Mathematical Statistics (IMS)*, 2022.

[60] Julian Rodemann, Dominik Kreiss, Eyke Hüllermeier, and Thomas Augustin. Levelwise data disambiguation by cautious superset learning. In *International Conference on Scalable Uncertainty Management (SUM)*, page 263–276. Springer, 2022.

[61] Julian Rodemann, Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, and Thomas Augustin. Bayesian PLS! Approximate Bayes optimal pseudo-label selection (PLS). *arXiv preprint*, 2023.

[62] Fabrizio Ruggeri, David Ríos Insua, and Jacinto Martín. Robust bayesian analysis. *Handbook of statistics*, 25:623–667, 2005.

[63] Patrick Michael Schwaferts and Thomas Augustin. Imprecise hypothesis-based Bayesian decision making with simple hypotheses. In *International Symposium on Imprecise Probabilities: Theories and Applications*, pages 338–345. PMLR, 2019.

[64] Teddy Seidenfeld. A contrast between two decision rules for use with (convex) sets of probabilities: γ-maximin versus e-admissibility. *Synthese*, 140(1/2): 69–88, 2004.

[65] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *European Conference on Computer Vision*, pages 299–315, 2018.

[66] Hyunjung Shin, Tianya Hou, Kanghee Park, Chan-Kyoo Park, and Sunghee Choi. Prediction of movement

direction in crude oil prices based on semi-supervised learning. *Decision Support Systems*, 55(1):348–358, 2013.

[67] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608, 2020.

[68] Nataliya Sokolovska, Olivier Cappé, and François Yvon. The asymptotics of semi-supervised learning in discriminative probabilistic models. In *International Conference on Machine learning*, pages 984–991, 2008.

[69] Farzin Soleymani, Mohammad Eslami, Tobias Elze, Bernd Bischl, and Mina Rezaei. Deep variational clustering framework for self-labeling large-scale medical images. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 68–76. SPIE, 2022.

[70] Cristina Tortora, Ryan P Browne, Brian C Franczak, Paul D McNicholas, Maintainer Cristina Tortora, and Depends Bessel. Package 'mixghd'. 2014.

[71] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284, 2015.

[72] Matthias C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

[73] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109 (2):373–440, 2020.

[74] Vincent Vandewalle, Christophe Biernacki, Gilles Celeux, and Gérard Govaert. A predictive deviance criterion for selecting a generative model in semi-supervised classification. *Computational Statistics & Data Analysis*, 64:220–236, 2013.

[75] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman Hall, 1991.

[76] Bowen Zhang, Yidong Wang, Wenxin Hou, HAO WU, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, pages 18408–18419, 2021.

[77] T. Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.