

# Semi-Supervised Learning guided by the Generalized Bayes Rule under Soft Revision

Stefan Dietrich, Julian Rodemann and Christoph Jansen

Department of Statistics, Ludwig-Maximilians-Universität (LMU), Munich, Germany

**Abstract.** We provide a theoretical and computational investigation of the Gamma-Maximin method with soft revision, which was recently proposed as a robust criterion for pseudo-label selection (PLS) in semi-supervised learning. Opposed to traditional methods for PLS we use credal sets of priors (“generalized Bayes”) to represent the epistemic modeling uncertainty. These latter are then updated by the Gamma-Maximin method with soft revision. We eventually select pseudo-labeled data that are most likely in light of the least favorable distribution from the so updated credal set. We formalize the task of finding optimal pseudo-labeled data w.r.t. the Gamma-Maximin method with soft revision as an optimization problem. A concrete implementation for the class of logistic models then allows us to compare the predictive power of the method with competing approaches. It is observed that the Gamma-Maximin method with soft revision can achieve very promising results, especially when the proportion of labeled data is low.<sup>1</sup>

**Keywords:** Semi-Supervised Learning, Self-Training, Pseudo-Labeling, Generalized Bayes Rule, Imprecise Probabilities, Credal Sets, Soft Revision, Classification, Machine Learning

## 1 Introduction

In this paper, we investigate Gamma-Maximin with soft revision for robust selection of pseudo-labeled data in semi-supervised learning (SSL) problems, as recently proposed in [32]. We use this introduction to collect those concepts from SSL that are most relevant to our article. For an in-depth motivation and extensive hints to relevant literature, we refer the reader to [32, Section 1].

### 1.1 Semi-Supervised Learning (SSL)

The fundamental idea of SSL is to use both labeled data  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$  and unlabeled data  $\mathcal{U} = \{(x_i, \mathcal{Y})\}_{i=n+1}^m$  for training a classification model.<sup>2</sup> In this approach, a limited set of labeled data points is combined with an often considerably larger set of unlabeled data points  $m \gg n$  to train a model  $g(x) = \hat{y} \in \mathcal{Y}$ ,

---

<sup>1</sup>This work is partly based on Stefan Dietrich’s Bachelor thesis “ $\alpha$ -cuts as robust method of pseudo label selection within the context of semi-supervised learning”.

<sup>2</sup>As such, SSL can be perceived as a special case of superset learning [11, 12, 33].

see [8]. The incorporation of unlabeled data when training a model can intuitively be useful only if the fundamental SSL assumption is met, i.e., if the unlabeled data contains information about the relationship between the input data and the label of the assigned class [39]. In this paper, we will address *classification problems* with self-training methods and incremental addition of pseudo-labeled data as categorized by [36]. A self-training algorithm will, in the first phase, estimate a model from the labeled training data using supervised learning. Based on this model, labels are then predicted for the unlabeled data. These labels are referred to as *pseudo-labels*. In the second phase, the training data is enlarged by adding a selected pseudo-labeled data point according to some criterion. The same point is then removed from the unlabeled dataset, resulting in a new unlabeled dataset. This process is repeated until a stopping criterion is met. This paper will focus on the selection of a potentially favorable pseudo-labeled data point. That is, we take the predicted pseudo-labels as given and only deal with selecting one pseudo-labeled data point from the set of pseudo-labeled data.

## 1.2 Bayesian Analysis

In the entire process of PLS we encounter several sources of uncertainty. Traditionally, only the model used for predicting pseudo labels is utilized. Thus, the selection is performed based on the estimated model parameters, without considering the uncertainty associated with those. In the endeavor to establish a selection criterion that is robust against the initial model, i.e., more robust against the choice of model parameters, but at the same time efficiently utilizes the information from labeled data, the choice of a Bayesian approach appears fruitful [30, 24]. Here, it is possible to decompose epistemic uncertainty into statistical uncertainty due to a lack of data and modeling uncertainty due to parametric assumptions. The fundamental idea of Bayesian pseudo-label selection (BPLS) is to choose an instance with a pseudo-label that is probable given the training data but not necessarily probable concerning the estimated model parameters. In the spirit of [29, 25, 34], we formalize PLS as a decision problem. The utility of a selected data point  $(x_i, \mathcal{Y})$  shall now be the likelihood of being generated jointly with  $\mathcal{D}$  by a model with parameters  $\theta \in \Theta$  if we include it with the pseudo-label  $\hat{y}_i \in \mathcal{Y}$  (obtained through any predictive model) in  $\mathcal{D} \cup (x_i, \hat{y}_i)$ .

**Definition 1.** *Let  $(x_i, \mathcal{Y})$  be any decision (selection) from  $\mathcal{U}$ . We assign utility to each  $(x_i, \mathcal{Y})$  given  $\mathcal{D}$  and the pseudo-label  $\hat{y}_i \in \mathcal{Y}$  by the following measurable utility function which is said to be the pseudo-label likelihood.*

$$u : \mathcal{U} \times \Theta \rightarrow \mathbb{R}, ((x_i, \mathcal{Y}), \theta) \mapsto p(\mathcal{D} \cup (x_i, \hat{y}_i) | \theta) \quad (1)$$

The function  $u$  is a natural choice to assign utilities to pseudo-labels given the predicted pseudo-labels. Moreover, under a prior  $\pi$ , it can be shown that maximizing expected utility, i.e., applying the standard Bayes-criterion  $\Phi(\cdot, \pi) : \mathcal{U} \rightarrow \mathbb{R}; a \mapsto \Phi(a, \pi) = \mathbb{E}_\pi(u(a, \cdot))$ , with respect to  $u$  and  $\pi$ , corresponds to maximizing the pseudo marginal likelihood  $p(\mathcal{D} \cup (x_i, \hat{y}_i))$  [30]. For any prior, the action

which results in the maximal marginal likelihood is Bayes optimal. If we now update the prior with data to a posterior, the standard Bayes criterion corresponds to the pseudo posterior predictive (PPP), i.e., the posterior predictive distribution of the predicted pseudo-labels.

### 1.3 Uncertainty in BPLS

An important source of uncertainty is the choice of the model, i.e., the distributional assumption under which the likelihood is calculated. So far, this distributional assumption is the same as the one used for pseudo-label prediction. For addressing uncertainty regarding model choice, a multi-model approach can be applied, which computes a weighted sum of likelihoods of multiple models, see [32]. Additionally, in the same paper, uncertainty regarding the prior selection was briefly discussed. The uncertainty associated with the prior will be thoroughly examined in the remainder of the paper through an implementation via an approximation and experiments.

## 2 Related work

The robustness of pseudo-label selection has been extensively debated in the semi-supervised learning literature. Aminian et al. [1] propose an information theory based PLS approach adaptable to distributional shifts. Vandewalle et al. [38] aim to enhance model robustness against presumptive biases by facilitating model choice via the deviance information criterion within a semi-supervised learning framework. Rizve et al. [23] introduce an uncertainty-aware pseudo-label selection technique that competes with popular semi-supervised learning (SSL) methods anchored in consistency regularization. This technique involves selecting pseudo-labels by assessing both their probability scores and predictive uncertainty, the latter gauged by the prediction variance. Principled approaches, however, that hedge against misspecification of prior assumptions on which the PLS is based upon, are lacking so far. In the context of self-supervised learning, there is the notable exception of Lienen et al. [19]. Contrary to our approach, they use credal sets for labeling directly rather than for selecting pseudo-labels. Similar to us, they minimize the reliance on distributional assumptions. Finally, algorithms for computing optimal  $\Gamma$ -maximin decisions (or optimal decisions with respect to other imprecise decision criteria) have also been discussed, e.g., in [37, 13, 14, 15] in the context of decision making under ambiguity or in [21, 17, 16] in the context of robust (Bayesian) statistics. However, these approaches are not (directly) applicable here, as they assume finite state spaces or strictly restricted action spaces of their underlying decision problems, respectively.

## 3 Generalized Bayesian PLS under Soft Revision

In real-world applications, the unique choice of an adequate prior as required for Bayesian PLS [30] (and any other Bayesian method in machine learning

and statistics [10, 26, 28, 31, 20, 6, 27]) is often not straightforward. Instead, several different candidate priors occur equally plausible. In order to make the selection of the pseudo-labeled data more robust against a potentially bad choice of prior in such situations, *credal sets* can be used. A credal set  $\Pi$ , as first introduced in [18] and heavily been applied in the imprecise probability (IP) community since then, is simply a set of probability measures. Each element  $\pi \in \Pi$  is interpreted as an equally plausible candidate prior for the application at hand. Consequently, in the case of modeling with credal sets, any method for pseudo label selection should also include the entire credal set instead of artificially reducing uncertainty to classical probability.

The transition from a precise prior to a credal set (also referred to as “generalized Bayes” [3]) naturally also requires a decision criterion that can process this information. We choose the Gamma-Maximin criterion, whose fundamental idea is to compute the expected utility under a prior distribution which contrasts the likelihood of the estimated model the most. Then, the instance corresponding to the highest expected utility with respect to its *least favorable prior* should be chosen. Even though Gamma-Maximin can be criticized for being rather conservative, one can also argue that this very conservatism is often preferable as it safeguards against taking risky actions on the basis of unjustified assumptions [3]. In our case, it is better to choose a prior that may not align well with the likelihood rather than selecting one that aligns too closely with the likelihood.

The Gamma-maximin method in turn heavily depends on the choice of the credal set. The larger it is, the higher is the tendency to make a decision with complete disregard of the likelihood. On the other hand, a small credal set might not be expressive enough to sufficiently attenuate the likelihood. Therefore, in the following, we aim to restrict the set of all priors over which we calculate an infimum dynamically, i.e., for each iteration separately [32]. We adopt a credal set updating rule based on Cattaneo’s  $\alpha$ -cut rule with  $\alpha \in (0, 1)$ , also known as “soft revision” [7, 2]. This updating strategy aims to update only those priors whose marginal likelihood is greater than or equal to  $\alpha$  times the corresponding maximum marginal likelihood. Let, therefore, for the likelihood  $\mathcal{L}$  and all  $\pi \in \Pi$  be  $m(\pi) = \int \mathcal{L}(\theta) f_\pi(\theta) d\theta$  the marginal likelihood, where  $f_\pi(\theta)$  is the density of  $\pi$  with respect to the Lebesgue measure. Then  $\Pi_\alpha = \{\pi \in \Pi \mid m(\pi) \geq \alpha \cdot \sup_\pi m(\pi)\}$  is the credal set of priors that we update to posteriors. The Gamma-Maximin decision criterion with soft revision is now defined as follows:

$$\Phi_\alpha : \mathcal{U} \rightarrow \mathbb{R} \quad , \quad ((x_i, \mathcal{Y})_i) \mapsto \inf_{\pi \in \Pi_\alpha} \mathbb{E}_\pi(u(a, \cdot)). \quad (2)$$

$\Phi_\alpha$  addresses uncertainty in a general way without completely ignoring the likelihood. The steeper the likelihood, the smaller the credal set at a constant  $\alpha$ .

To technically implement<sup>1</sup> the Gamma-Maximin criterion under soft revision, we first establish a model and a distribution assumption.<sup>3</sup> Here, we will consider datasets with bivariate labels and perform logistic regression on them. The labels are assumed to follow a Bernoulli distribution  $Y_i \sim \mathcal{B}(p_i)$ , where  $p_i$  represents the response function that connects the linear predictor to the expected value or the parameters of the distributional assumption:  $\mathbb{E}[Y_i] = p_i = \exp(\eta_i)/(1 + \exp(\eta_i))$ . The credal prior, which the Gamma-Maximin criterion is based on in our case, is a set of normal distributions that vary only in their location parameter, precisely

$$\Pi = \left\{ \pi_\mu : \pi_\mu = \mathcal{N}(\mu, \Sigma) \wedge \mu \in [\underline{\mu}_0, \overline{\mu}_0] \times \dots \times [\underline{\mu}_p, \overline{\mu}_p] \right\}. \quad (3)$$

To calculate the marginal likelihood, it is not possible to resort to an analytical method. In this paper, we aim to adapt a numerical method that allows us to approximate the marginal likelihood with the help of the Laplace approximation. This is possible because the function for which the integral is to be computed is twice continuously differentiable and most of its mass is around its maximum [35, 4]. We define  $-h(\theta) := \log(\mathcal{L}(\theta)f_{\pi_\mu}(\theta))$  and  $V(\theta)$  as the Hessian matrix of  $h(\theta)$ . Let  $\tilde{\theta}$  be the argument that minimizes  $h(\theta)$ . This minimum is unique and global because  $h(\theta)$  is convex. Now the Laplace approximation of the marginal likelihood<sup>4</sup> is:

$$m(\pi_\mu) \approx (2\pi)^{\frac{d}{2}} |V(\tilde{\theta})|^{-\frac{1}{2}} \mathcal{L}(\tilde{\theta})f_{\pi_\mu}(\tilde{\theta}), \quad (4)$$

see [35, 4]. To calculate  $\tilde{\theta}$  the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is used [5]. The BFGS method belongs to the quasi-Newton methods and is a numerical technique used to solve nonlinear optimization problems. This method appears particularly suitable for finding optima in our case, as our functions are twice differentiable and it saves computing resources by not computing the Hessian matrix in each step. To execute the soft revision we now want only the prior  $\pi_\mu$  in our credal set that fulfills

$$m(\pi_\mu) - \alpha \max_{\pi_\mu \in \Pi} m(\pi_\mu) \geq 0 \quad (5)$$

We now want to construct an approximation of the pseudo posterior predictive (PPP), which depends only on the given data and on the prior  $\pi_\mu$ . As described in [30], the PPP can be written as:

$$\frac{1}{p_{\pi_\mu}(\mathcal{D})} \int_{\Theta} p(\mathcal{D} \cup (x_i, \tilde{y}_i) | \theta) p(\mathcal{D} | \theta) f_{\pi_\mu}(\theta) d\theta \quad (6)$$

---

<sup>1</sup>Implementation of the proposed method and reproducible scripts for the experiments are available at: <https://github.com/Stefan-Maximilian-Dietrich/reliable-pls>.

<sup>3</sup>The specific choice of model and distribution assumption here is more for the purpose of demonstration. The following can also be applied to other situations. In case of non-normal priors, alternative approximation strategies need to be considered.

<sup>4</sup> $d$  is the number of dimensions. The  $\pi$  without a subscript is the number  $\pi$ .

Here, we can approximate the marginal likelihood  $p_{\pi_\mu}(\mathcal{D})$  again using the method described above because the difference is in  $\mathcal{O}(n^{-1})$  [30]. The remaining integral is now similar to the marginal likelihood and can be approximated by Laplace because it is also twice continuously differentiable and most of its mass is around its mode. The approach here is analogous to the one used in approximating the marginal likelihood. Let now  $q(\pi_\mu) \approx \int_{\Theta} p(\mathcal{D} \cup (x_i, \tilde{y}_i)|\theta)p(\mathcal{D}|\theta)f_{\pi_\mu}(\theta)d\theta$ . The PPP dependent on the prior  $\pi_\mu$  can now be approximated by the approximations of the marginal likelihood and the remaining integral:

$$p_{\pi_\mu}(\mathcal{D} \cup (x_i, \tilde{y}_i)|\mathcal{D}) \approx \frac{q(\pi_\mu)}{m(\pi_\mu)} \quad (7)$$

Now we have both a PPP for which it holds  $\mathbb{E}_{\pi_\mu}(u(a, \theta)) = p_{\pi_\mu}(\mathcal{D} \cup (x_i, \tilde{y}_i)|\mathcal{D})$  dependent on a prior or its parameters and a soft revision ( $\alpha$ -cut) criteria that sorts out all priors that would overly restrict the influence of the likelihood. We now want to treat this as an optimization problem:

$$\begin{aligned} & \inf_{\pi_\mu \in \Pi} \mathbb{E}_{\pi_\mu}(u(a, \theta)) \\ & \text{s.t. } m(\pi_\mu) - \alpha \max_{\pi_\mu \in \Pi} m(\pi_\mu) \geq 0, \end{aligned} \quad (8)$$

which yields the Gamma-Maximin criterion with soft revision, see Equation 2. The optimization is implemented using the COBYLA (Constrained Optimization by Linear Approximations) algorithm [22]. The basic idea is forming linear approximations to the objective function via interpolation at the vertices of a simplex. This algorithm iteratively proposes new vertices for this simplex, similar to the Nelder-Mead algorithm. The COBYLA algorithm does not rely on gradients and handles inequality constraints well, which is useful for optimization of Equation (8). As termination criterion, we use the fractional tolerance.

## 4 Experiments

We will now conduct various experiments. For this purpose, we estimate a model in each iteration, which we then evaluate using a test dataset. Since in each round the respective criteria need to be estimated for all pseudo-labeled data, the complexity lies in  $\mathcal{O}((m-n)^2)$  with  $m-n$  the number of unlabeled data. We benchmark semi-supervised logistic regression with the Gamma-Maximin method under soft revision ( $\alpha$ -cut) (●) against the following methods: 1.) A standard supervised (●) baseline where only the labeled data is considered and no self-training takes place. 2.) three standard pseudo-label selection criteria: Probability Score (●), Predictive Variance (●) and Likelihood (max-max) (●). 3.) Four (robustified) Bayesian methods that depend on the pseudo posterior predictive: PPP (Bayes-optimal) (●), PPP multi-label (●), PPP multi-model (●) and PPP weighted multi-label (●).<sup>5</sup> Experiments are run on simulated binomially

<sup>5</sup>They incorporate uncertainty regarding model choice (multi-model and weighted multi-model (●)) and accumulation of errors (multi-label (●)). For a thorough introduction, please refer to [32].

distributed data as well as on three data sets (Banknote, MTcars, Mushrooms) for binary classification from the UCI repository [9]. The binomially distributed data was simulated through a linear predictor consisting of normally distributed features. Furthermore, the effect of varying  $\alpha$  is tested. Further results can be viewed in the repository.<sup>6</sup>

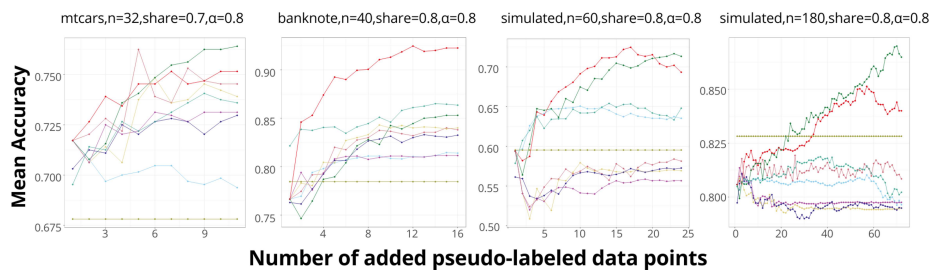


Fig. 1. Experiments on real-world data (left) and simulated data (right).

## 5 Discussion

At first glance, it is apparent that if any of the nine tested methods achieves an improvement over supervised learning, then Gamma-Maximin method with soft revision also achieves an improvement. In the conducted tests, it is evident that the Gamma-Maximin method with soft revision consistently ranks among the best out of all nine tested methods. This is particularly true for real-world data. In the case of simulated data, we see that besides the Gamma-Maximin method with soft revision, only the Multi-Model method consistently performs better than the Supervised Learning baseline. Upon closer comparison, two effects emerge: 1) The lower the total data volume, the better the performance of the Gamma-Maximin method with soft revision. 2) The further away the alternative models are to reality, the better the Gamma-Maximin method with soft revision performs relative to the Multi-Model method (in the example shown here, the alternative models are intentionally chosen to be very favorable to Multi-Model). Additionally, it was observed in the experiments that the Gamma-Maximin method with soft revision copes particularly well with a low proportion of labeled data compared to other methods. In the conducted experiments, no clear trend can be observed regarding which  $\alpha$  leads to particularly good performance.

<sup>6</sup><https://github.com/Stefan-Maximilian-Dietrich/reliable-pls>.

## References

- [1] G. Aminian et al. “An Information-theoretical Approach to Semi-supervised Learning under Covariate-shift”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 7433–7449.
- [2] T. Augustin and G. Schollmeyer. “Comment: on focusing, soft and strong revision of choquet capacities and their role in statistics”. In: *Statistical Science* (2021).
- [3] T. Augustin et al. *Introduction to imprecise probabilities*. Wiley, 2014.
- [4] N. Bleistein and R. A Handelsman. *Asymptotic expansions of integrals*. Ardent Media, 1975.
- [5] C. Broyden. “The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations”. In: *IMA Journal of Applied Mathematics* 6 (1970), pp. 76–90.
- [6] M. Caprio et al. “Imprecise Bayesian neural networks”. In: *arXiv preprint arXiv:2302.09656* (2023).
- [7] M. Cattaneo. “A Continuous Updating Rule for Imprecise Probabilities”. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Part III*. Springer, 2014, pp. 426–435.
- [8] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.
- [9] D. Dua and C. Graff. *UCI machine learning repository*. 2017.
- [10] V. Fortuin. “Priors in bayesian deep learning: A review”. In: *International Statistical Review* 90.3 (2022), pp. 563–591.
- [11] E. Hüllermeier. “Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization”. In: *International Journal of Approximate Reasoning* 55 (2014), pp. 1519–1534.
- [12] E. Hüllermeier, S. Destercke, and I. Couso. “Learning from imprecise data: adjustments of optimistic and pessimistic variants”. In: *International Conference on Scalable Uncertainty Management*. Springer, 2019, pp. 266–279.
- [13] C. Jansen, T. Augustin, and G. Schollmeyer. “Decision Theory Meets Linear Optimization Beyond Computation”. In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2017). Lecture Notes in Computer Science, vol 10369*. Springer, 2017, pp. 329–339.
- [14] C. Jansen, G. Schollmeyer, and T. Augustin. “Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences”. In: *International Journal of Approximate Reasoning* 98 (2018), pp. 112–131.
- [15] C. Jansen, G. Schollmeyer, and T. Augustin. “Quantifying degrees of E-admissibility in decision making with imprecise probabilities”. In: *Reflections on the Foundations of Statistics: Essays in Honor of Teddy Seidenfeld*. Ed. by Thomas Augustin, F. Cozman, and G. Wheeler. Theory and Decision Library A, Springer, 2022.
- [16] C. Jansen et al. “Robust statistical comparison of random variables with locally varying scale of measurement”. In: *Proceedings of the Thirty-Ninth*



- Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin J. Evans and Ilya Shpitser. PMLR, 2023, pp. 941–952.
- [17] C. Jansen et al. “Statistical Comparisons of Classifiers by Generalized Stochastic Dominance”. In: *Journal of Machine Learning Research* 24.231 (2023), pp. 1–37.
- [18] I. Levi. “On Indeterminate Probabilities”. In: *The Journal of Philosophy* 71 (1974), pp. 391–418.
- [19] J. Liene and E. Hüllermeier. “Credal Self-Supervised Learning”. In: *Neural Information Processing Systems*. Vol. 34. 2021, pp. 14370–14382.
- [20] A. Marquardt, J. Rodemann, and T. Augustin. “An empirical study of prior-data conflicts in Bayesian neural networks”. In: (2023).
- [21] R. Noubiap and W. Seidel. “An algorithm for calculating  $\Gamma$ -minimax decision rules under generalized moment conditions”. In: *Annals of Statistics* 29 (2001), pp. 1094–1116.
- [22] M. Powell. “A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation”. In: *Advances in Optimization and Numerical Analysis*. Springer, 1994, pp. 51–67.
- [23] M. Rizve et al. “In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning”. In: *International Conference on Learning Representations*. 2020.
- [24] J. Rodemann. “Bayesian Data Selection”. In: *arxiv* (2024). Preprint, under review.
- [25] J. Rodemann. “Pseudo Label Selection is a Decision Problem”. In: *Proc. of the 46th German Conference on Artificial Intelligence*. Springer, 2023.
- [26] J. Rodemann and T. Augustin. “Accounting for Gaussian process imprecision in Bayesian optimization”. In: *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*. Springer. 2022, pp. 92–104.
- [27] J. Rodemann and T. Augustin. “Imprecise Bayesian Optimization”. In: *Knowledge-based Systems* (2024). Acc. with min. rev.
- [28] J. Rodemann, T. Augustin, and R. de Heide. “Interpreting Generalized Bayesian Inference by Generalized Bayesian Inference”. In: *Poster presented at International Symposium on Imprecise Probabilities (ISIPTA)*. 2023.
- [29] J. Rodemann, C. Jansen, and G. Schollmeyer. “Reciprocal Learning”. In: *arxiv* (2024). Preprint, under review.
- [30] J. Rodemann et al. “Approximately Bayes-optimal pseudo-label selection”. In: *Proc. of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Vol. 216. PMLR, 2023, pp. 1762–1773.
- [31] J. Rodemann et al. “Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration”. In: *arXiv:2403.04629* (2024).
- [32] J. Rodemann et al. “In all likelihoods: robust selection of pseudo-labeled data”. In: *Proc. of the 13th International Symposium on Imprecise Probability: Theories and Applications*. Vol. 215. PMLR, 2023, pp. 412–425.

- [33] J. Rodemann et al. “Levelwise Data Disambiguation by Cautious Superset Learning”. In: *International Conference on Scalable Uncertainty Management*. Springer, 2022.
- [34] J. Rodemann et al. “Not All Data Are Created Equal: Lessons From Sampling Theory For Adaptive Machine Learning”. *Poster presented at the International Conference on Statistics and Data Science (ICSDS), Institute of Mathematical Statistics (IMS)*. 2022.
- [35] E. Ruli, N. Sartori, and L. Ventura. “Improved Laplace approximation for marginal likelihoods”. In: *Electronic Journal of Statistics* 10.2 (2016).
- [36] I. Triguero, S. García, and F. Herrera. “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study”. In: *Knowledge and Information Systems* 42 (2013), pp. 245–284.
- [37] L. Utkin and T. Augustin. “Powerful algorithms for decision making under partial prior information and general ambiguity attitudes”. In: *Proc. of the Forth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA '05)*. SIPTA, 2005, pp. 349–358.
- [38] V. Vandewalle et al. “A predictive deviance criterion for selecting a generative model in semi-supervised classification”. In: *Computational Statistics & Data Analysis* 64 (2013), pp. 220–236.
- [39] X. Zhu. “Semi-Supervised Learning Literature Survey”. In: *Comput Sci, University of Wisconsin-Madison* 2 (July 2008).