Approximately Bayes-Optimal Pseudo-Label Selection (Supplementary Material)

Julian Rodemann[1] **Jann Goschenhofer**[1,2,3] **Emilio Dorigatti**[1,2,4] **Thomas Nagler**[1,2] **Thomas Augustin**[1]

[1]Department of Statistics, Ludwig-Maximilians-Universität (LMU), Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
[3]Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
[4]Institute of Computational Biology, Helmholtz-Zentrum, Neuherberg, Germany

A PSEUDO-CODE FOR BPLS

We summarize the procedure of Bayesian Pseudo-Label Selection (BPLS) with approximate Pseudo Posterior Predictive (PPP) in Algorithm 1. Pseudo-code describing the proposed extensions can be found in section E of this supplementary material. Notation and mathematical symbols follow the main paper. Notably, the number of unobserved data $|\mathcal{U}|$ was denoted m in the main paper.

Algorithm 1: Bayesian Pseudo-Label Selection (BPLS) with approximate Pseudo Posterior Predictive (PPP)

Data: $\mathcal{D}$,$\mathcal{U}$
Result: $\mathcal{D}$, fitted model $\hat{y}^*(x)$
Fit model M on labeled data $\mathcal{D}$ to obtain prediction function $\hat{y}(x)$
while *stopping criterion not met* **do**
  for $i \in \{1, \dots, |\mathcal{U}|\}$ **do**
    predict $\mathcal{Y} \ni \hat{y}_i = \hat{y}(x_i)$
    approximate PPP $p(\mathcal{D} \cup (x_i, \hat{y}_i)\,|\mathcal{D})$
  end
  obtain $i^* = \arg\max_i \{p(\mathcal{D} \cup (x_i, \hat{y}_i)\,|\mathcal{D})\}$
  add $(x_i, \hat{y}_i)$ to labeled data: $\mathcal{D} \leftarrow \mathcal{D} \cup (x_i, \hat{y}_i)$
  update $\mathcal{U} \leftarrow \mathcal{U} \setminus (x_i, \mathcal{Y})_i$
end

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

B MISSING PROOFS

We present the proofs for Theorems 1-3 in section 2 of the main paper. For the sake of readability, we repeat the underlying theorems as well.

B.1 PROOF OF THEOREM 1

Theorem 1 *In the decision problem $(\mathbb{A}, \Theta, u(\cdot))$ with $\mathbb{A} = \mathcal{U}$ (definition 1), with the pseudo-label likelihood as utility function (definition 2), and a prior $\pi(\theta)$ on Θ, the standard Bayes criterion*

$$\begin{aligned}\Phi(\cdot, \pi) \colon \mathcal{U} &\to \mathbb{R} \\ a &\mapsto \Phi(a, \pi) = \mathbb{E}_\pi(u(a, \theta))\end{aligned}$$

corresponds to the pseudo marginal likelihood $p(\mathcal{D} \cup (x_i, \hat{y}_i))$.

Proof 1 *The definition of the expected value for measurable $u(\cdot, \cdot)$ directly delivers $\Phi(a, \pi) = \mathbb{E}_\pi(u(a, \theta)) = \int u(a, \theta) d\pi(\theta) = \int p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta) d\pi(\theta) = p(\mathcal{D} \cup (x_i, \hat{y}_i))$.*

B.2 PROOF OF THEOREM 2

Theorem 2 *In the decision problem $(\mathbb{A}, \Theta, u(\cdot))$, using the pseudo-label likelihood as utility function as in theorem 1 but with the prior updated by the posterior $\pi(\theta) = p(\theta \mid \mathcal{D})$ on Θ, the standard Bayes criterion $\Phi(\cdot, \pi) \colon \mathcal{U} \to \mathbb{R}$; $a \mapsto \Phi(a, \pi) = \mathbb{E}_\pi(u(a, \theta))$ corresponds to the pseudo posterior predictive $p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$.*

Proof 2 *Analogous to Proof 1, we have $\Phi(a, \pi) = \mathbb{E}_\pi(u(a, \theta)) = \int u(a, \theta) d\pi(\theta)$. Now with the updated prior $\pi(\theta) = p(\theta \mid \mathcal{D})$ it follows $\int u(a, \theta) d\pi(\theta) = \int p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta) dp(\theta \mid \mathcal{D}) = p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$.*

B.3 PROOF OF THEOREM 3

Theorem 3 *In the decision problem $(\mathbb{A}, \Theta, u(\cdot))$, using the pseudo-label likelihood as utility function as in theorem 1, the max-max criterion*

$$\begin{aligned}\Phi \colon \mathcal{U} &\to \mathbb{R} \\ a &\mapsto \Phi(a) = \max_\theta(u(a, \theta))\end{aligned}$$

corresponds to the (full) likelihood at $\hat{\theta}_{ML}$.

Proof 3 *Recall definition 2 of the pseudo-label likelihood as utility function: $u \colon \mathcal{U} \times \Theta \to \mathbb{R}$; $((x_i, \mathcal{Y}), \theta) \mapsto u((x_i, \mathcal{Y}), \theta) = p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta)$. Thus, it holds for the max-max criterion $\Phi(a) = \max_\theta(u(a, \theta)) = \max_\theta(p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta)) = p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \hat{\theta}_{ML})$, with $\hat{\theta}_{ML}$ the ML-estimator.*

The max-max criterion hence corresponds to direct optimization with regard to a of the likelihood, evaluated at $\hat{\theta}_{ML}$. The respective max-max-action is thus $a^*_{max-max} = \max_a \max_\theta p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \theta) = \max_a p(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \hat{\theta}_{ML})$.

C EXPERIMENTAL SETUP

We describe the setup for the experiments with both the simulated and the real-world data along with additional empirical results comparing our approximate PPP with predominant PLS methods in section D.

C.1 BENCHMARKS

Throughout our experiments, we compare our proposed approximate PPP with a set of baseline and competing approaches:

- *Likelihood (max-max)*: Self-training using the Likelihood max-max action as selection criterion
- *Predictive Variance*: Self-training using the predictive variance $\mathrm{Var}[\hat{y}] = \mathbb{E}[\hat{y} - \mathbb{E}[\hat{y}]]^2$ of the model predictions as a selection criterion
- *Probability Score*: Self-training using the predicted probabilities (scores) $\mathbb{P}(y = \hat{y})$ as a selection criterion
- *Supervised Learning*: regular supervised model fitting using the labeled training data only

All data sets reflect binary classification tasks with a fairly balanced class label distribution. Hence, we report and compare with model performance as measured in accuracy on the holdout test data sets.

C.1.1 Generalized Linear Models

We choose generalized linear models (GLMs) [Nelder and Wedderburn, 1972] as predictive models for BPLS with PPP as well as for all competing methods listed in section C.1. By considering the binomial distribution from the exponential family this yields logistic regression:

$$\mathrm{P}(Y = 1 \mid X = x_i) = \mathrm{P}(Y_i = 1) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})}, \tag{1}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ and $\mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ik}\beta_k$. Such a regression with additive linear predictor $\mathbf{x}_i^\top \boldsymbol{\beta}$ can be easily be extended to target variables that follow a multinomial distribution (i.e. multi-class problems). Our setup described in section C.1 can thus be extended in a straightforward manner to such learners for multi-class classification tasks.

C.1.2 Generalized Additive Models

We also use non-parametric generalized additive models (GAMs) [Fahrmeir et al., 2013, Hastie, 2017] as predictive models. Here, the response variable depends on unknown smooth functions of some feature variables:

$$g(\mathbb{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m). \tag{2}$$

As above, we assume Y to follow a binomial distribution in our experiments, since we only consider binary classification. Like GLMs, GAMs can be easily extended to multi-class problems.

C.1.3 Simulation Design

For the simulation study, we created a simulated dataset with n samples for a binary classification based on a varying amount of q features. This simulation follows the model equation

$$y_i \sim Bin(1, p_i), \text{ with } p_i = (1 + exp(-x_{i,0} + x_{i,1} + ...x_{i,p}))^{-1} \tag{3}$$

where $x_i \sim \mathcal{N}(\mu, \sigma^2)$ independently with varying μ and σ^2.

C.1.4 Pre-Processing and Gathering of Real-world Data

Detailed information on sources, features, and target variables of all data sets [Dua and Graff, 2017] that were used in the experiments can be found in section L. The data sets were selected randomly after filtering according to the following criteria:

- We only consider binary classification tasks, since we test the PLS methods based on semi-supervised logistic regression.
- We choose from datasets with a low number of missing values in order to minimize algorithm differences in missing value handling.
- We restrict ourselves to datasets with $q < 100$ to avoid massive overfitting and computational trouble.

In order to benchmark BPLS against classical PLS methods, we split the data sets into train and test data first, before removing labels from a pre-defined share of training data. Our detailed splitting procedure for the real-world datasets with a total size of n samples each is the following:

1. draw n_{test} samples to create the holdout test set D_{test} where the remainder constitutes the training set D_{train} of size n_{train} such that $n_{train} = n_{test}$ (share of test data thus 50%).
2. draw $n_{labeled}$ samples from D_{train} to create the labeled training data $D_{train}^{labeled}$
3. Remove labels from remaining samples in D_{train} and treat them as unlabeled data $D_{train}^{unlabeled}$

Throughout our experiments, we repeat self-training R times and use varying shares of labeled data $\frac{n_{unlabeled}}{n_{train}}$.

C.1.5 Hypotheses

For interpretation purposes, recall our hypotheses that we specified before running the experiments:

Hypothesis 1 ***(a)*** *PPP with uninformative prior outperforms traditional PLS on data prone to initial overfitting (i.e., with high ratio of features to data $\frac{p}{n}$ and poor initial generalization).* ***(b)*** *For low $\frac{p}{n}$ and high initial generalization, BPLS is outperformed by traditional PLS.*

Hypothesis 2 ***(a)*** *Among all PLS methods, the pseudo-label likelihood (max-max-action) reinforces the initial model fit the most and* ***(b)*** *hardly improves generalization.*

Hypothesis 3 *PPP with informative prior outperforms traditional PLS methods universally.*

D FURTHER RESULTS

In this section, we present additional results. Section D.1 has the complete results for simulated data with $q = 60$ features.[1] In section D.2, we show additional results for smaller $q \in \{10, 15, 20, 30\}$ with varying $n \in \{300, 400, 800, 1000\}$.

D.1 RESULTS ON SIMULATED DATA WITH $q = 60$

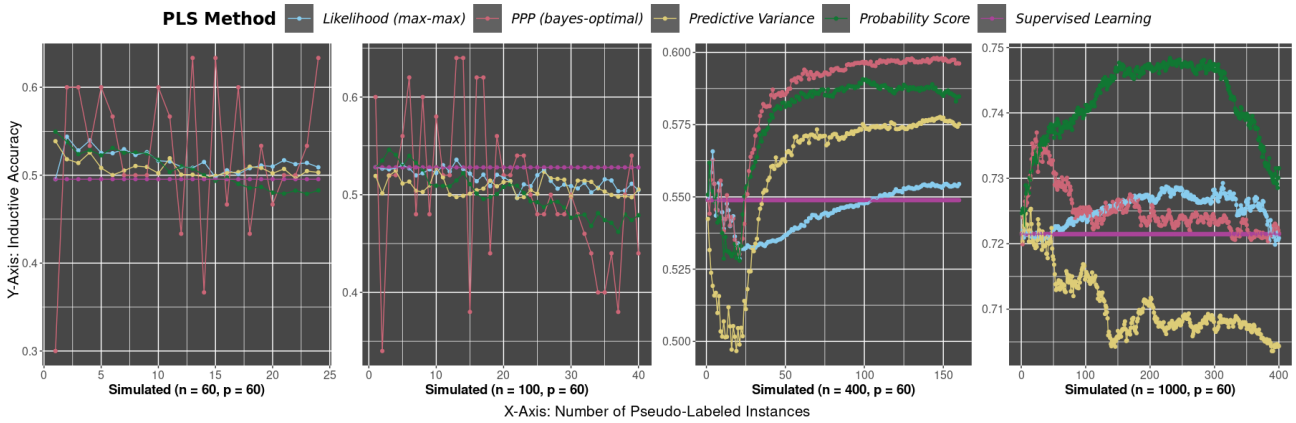

Figure 1: Complete Results on Simulated Data for $q = 60$. $R = 100$; $\frac{n_{unlabeled}}{n_{train}} = 0.8$.

D.2 FURTHER RESULTS ON SIMULATED DATA WITH $q \in \{10, 15, 20, 30\}$

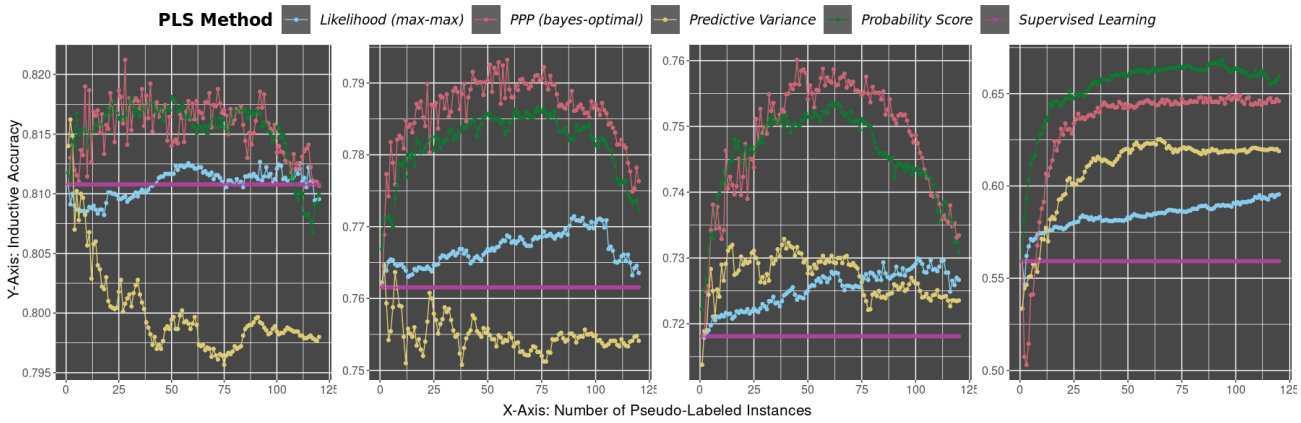

Figure 2: Results on Simulated Data, $n = 300$ and (from left to right) $q \in \{10, 15, 20, 30\}$. $R = 100$; $\frac{n_{unlabeled}}{n_{train}} = 0.8$.

[1] Results for $n = 100$ and $n = 400$ were already included in the paper, but are also shown here for the sake of completeness of the setup with $q = 60$. (Note that this is an exception; all other results presented herein have not been included in the paper.)

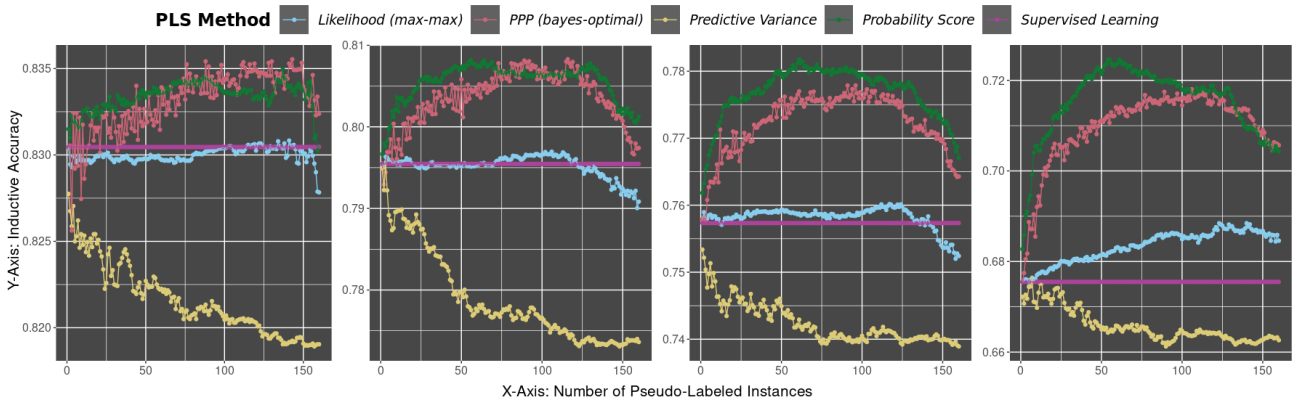

Figure 3: Results on Simulated Data, $n = 400$ and (from left to right) $q \in \{10, 15, 20, 30\}$. $R = 100$; $\frac{n_{unlabeled}}{n_{train}} = 0.8$.

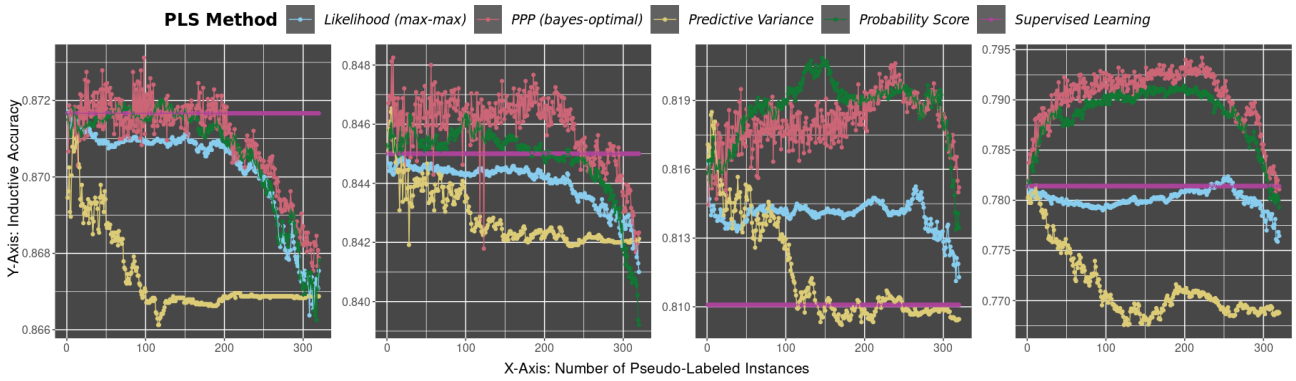

Figure 4: Results on Simulated Data, $n = 800$ and (from left to right) $q \in \{10, 15, 20, 30\}$. $R = 100$; $\frac{n_{unlabeled}}{n_{train}} = 0.8$.

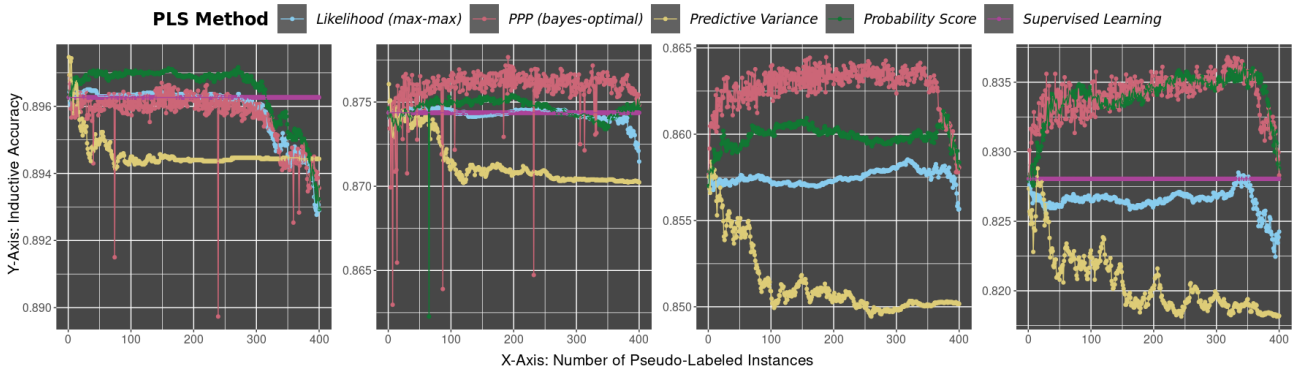

Figure 5: Results on Simulated Data, $n = 1000$ and (from left to right) $q \in \{10, 15, 20, 30\}$. $R = 100$; $\frac{n_{unlabeled}}{n_{train}} = 0.8$.

D.3 INFORMATIVE PRIOR: FURTHER RESULTS ON SIMULATED DATA

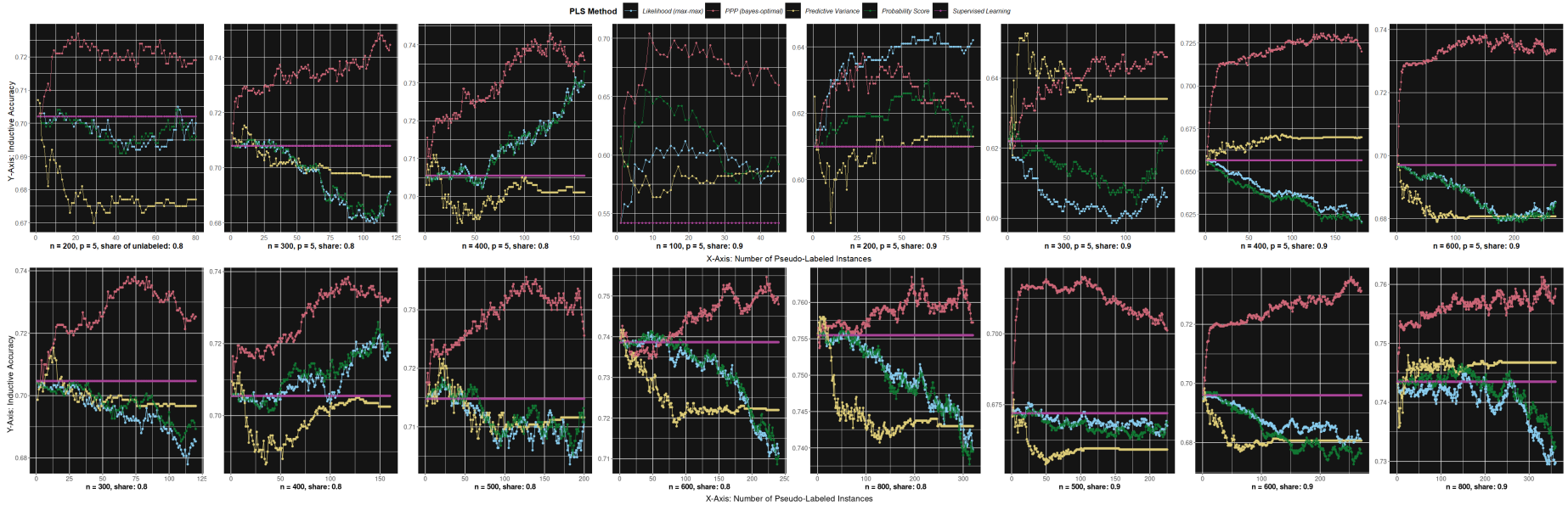

Figure 6: Results from simulated data in case of informative priors with simple GLMs (logistic regression, first row) and more complex non-parametric GAMs (second row). Note that resolution allows zooming in.

D.4 SUMMARY OF RESULTS ON SIMULATED DATA

Table 1 summarizes the results on simulated data in an ordinal manner. That is, it shows the best-performing method on the different setups. As in the main paper, "Oracle stopping" in table 1 refers to comparing PLS methods with regard to their overall best accuracy as opposed to "final" comparisons after the whole data set was labeled.

Table 1: Best performing PLS on Simulated Data

n	p	ORACLE STOPPING	FINAL
60	60	PPP	PPP
100	60	PPP	Supervised Learning
400	60	PPP	PPP
1000	60	Probability Score	Probability Score
300	30	Probability Score	Probability Score
300	20	PPP	PPP
300	15	PPP	PPP
300	10	PPP/Probability Score	PPP/Probability Score
400	30	Probability Score	PPP/Probability Score
400	20	Probability Score	Probability Score
400	15	PPP/Probability Score	Probability Score
400	10	PPP/Probability Score	PPP/Probability Score
800	30	PPP	PPP
800	20	PPP/Probability Score	PPP/Probability Score
800	15	PPP	PPP
800	10	PPP	PPP/Probability Score
1000	30	PPP	PPP/Probability Score
1000	20	PPP	PPP
1000	15	PPP	PPP
1000	10	Predictive Variance	Predictive Variance

E EXTENSIONS

We provide further details on the suggested extensions in section 6. Besides, we briefly discuss other potential extensions.

E.1 EXTENSIONS PROPOSED IN THE PAPER

We summarize the proposed extensions' procedure from section 6 in the paper by pseudo-code as follows.

E.1.1 Bivariate Pseudo-Label Selection

The idea of bivariate BPLS would be to touch the model class M. When comparing PPPs, one could then take into account the required model size q. The rough idea would be to prefer pseudo-labels that have high plausibility (high likelihood) even with simpler models (small q).

Algorithm 2: Bivariate Bayesian Pseudo-Label Selection (BPLS)

Data: $\mathcal{D}$,$\mathcal{U}$
Result: $\mathcal{D}$, fitted model $\hat{y}^*(x)$
Fit model M on labeled data $\mathcal{D}$ to obtain prediction function $\hat{y}(x)$
while *stopping criterion not met* **do**
  for $i \in \{1, \dots, |\mathcal{U}|\}$ **do**
    predict $\mathcal{Y} \ni \hat{y}_i = \hat{y}(x_i)$ with models of varying $dim(\Theta)$
    evaluate PPP $p(\mathcal{D} \cup (x_i, \hat{y}_i)\,|\mathcal{D})$ with predictions from the best performing (on training data) model and save respective $dim(\Theta)$
  end
  obtain $i^* = \arg\max_i \{f(p(\mathcal{D} \cup (x_i, \hat{y}_i)\,|\mathcal{D}), dim(\Theta))\}$, with $f(\cdot,\cdot)$ some linear combination of the PPP and the model size $dim(\Theta)$
  retrain M on $\mathcal{D} \cup (x_i, \hat{y}_{i^*})$
  predict $\mathcal{Y} \ni \hat{y}_i^*(\mathbf{x} \cup x_i), \mathbf{x} \in \mathcal{D}$
  add $(x_i, \hat{y}_i)$ to labeled data: $\mathcal{D} \leftarrow \mathcal{D} \cup (x_i, \hat{y}_i)$
  update $\mathcal{U} \leftarrow \mathcal{U} \setminus (x_i, \mathcal{Y})_i$
end

E.2 ADDITIONAL EXTENSIONS

E.2.1 Robust PPP

We further propose a robust extension of PPP based on generalized Bayesian inference [Dempster, 1968, Walley, 1991, Ruggeri et al., 2005, Augustin et al., 2014]. Recall that for the *robust* PPP, now denoted as $p^*(\hat{y} \mid x, \boldsymbol{y}, \boldsymbol{x})$, we consider the prior $\pi^*(\theta)$ among all priors from a convex set of priors Π that has the smallest value $\pi^*(\hat{\theta})$ at the ML-estimator $\hat{\theta}$. Recall that $\Pi \subseteq \{\pi(\theta) \mid \pi(\cdot) \text{ a probabilty measure on } (\Theta, \sigma(\Theta))\}$ with Θ compact as throughout the paper and $\sigma(\cdot)$ an appropriate σ-algebra.

More formally and encapsulating the notion of Γ-Maximin as in [Guo and Tanaka, 2010], for instance, we have the decision problem $(\mathbb{A}, \Theta, u(\cdot))$ with $\mathbb{A} = \mathcal{U}$ (definition 1 in paper) with the pseudo-label likelihood as utility function (definition 2) and a set of priors Π as above. Then the Γ-maximin criterion

$$\Phi(\cdot, \Pi) \colon \mathcal{U} \to \mathbb{R};\ a \mapsto \Phi(a, \pi) = \underline{\mathbb{E}}_{\Pi}(u(a, \theta)) \tag{4}$$

with $\underline{\mathbb{E}}_{\Pi}(u(a,\theta)) = \inf_{\pi \in \Pi} \mathbb{E}(u(a,\theta))$ corresponds to the *robust pseudo posterior predictive* $p^*(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$ that results from updating the prior $\pi^*(\cdot) \in \Pi$ that has the lowest value in $\hat{\theta}$. Action $a^*_{\Gamma} = \arg\max_i p^*(\mathcal{D} \cup (x_i, \hat{y}_i) \mid \mathcal{D})$ is Γ-maximin-optimal for prior $\pi^*(\cdot)$.

In practice, the proposed extension heavily depends on the exact nature of Π. For illustrative purposes, suppose that we can specify Π such that the most contradicting prior is such that the resulting posterior is uniform. Effectively, we then end up

with the same situation as with the marginal likelihood when the prior is uniform in case of independent observations, see the end of section 3.1 in the main paper: We randomly select pseudo-labeled instances. Quite intuitively, the selection that is most robust toward the initial fit given no other information is just such a random selection.

E.2.2 Bayesian Pseudo-Label Selection without predictions

The idea here would be to directly assign all possible q classes in $\mathcal{Y}$ to the unlabeled data points with $q = |\mathcal{Y}|$. The following pseudo-code lines out the procedure. Note that the inner loop thus requires $|\mathcal{U}| \cdot |\mathcal{Y}|$ assignments and respective PPP evaluations.

Algorithm 3: Bayesian Pseudo-Label Selection (BPLS) without predictions

Data: $\mathcal{D}$,$\mathcal{U}$
Result: $\mathcal{D}$, fitted model $\hat{y}^*(x)$
Fit model M on labeled data $\mathcal{D}$ to obtain prediction function $\hat{y}(x)$
while *stopping criterion not met* **do**
  for $i \in \{1, \ldots, |\mathcal{U}|\}$ **do**
    assign all possible $\hat{y}_i \in \mathcal{Y}$ to $(x_i, \hat{y}_i)$
    evaluate all possible PPP $p(\mathcal{D} \cup (x_i, \hat{y}_i)\,|\mathcal{D})$
  end
  obtain $i^* = \arg\max_i \{p(\mathcal{D} \cup (x_i, \hat{y}_i)\,|\mathcal{D})\}$
  retrain M on $\mathcal{D} \cup (x_i, \hat{y}_{i^*})$
  predict $\mathcal{Y} \ni \hat{y}_i^*(\mathbf{x} \cup x_i), \mathbf{x} \in \mathcal{D}$
  add $(x_i, \hat{y}_i)$ to labeled data: $\mathcal{D} \leftarrow \mathcal{D} \cup (x_i, \hat{y}_i)$
  update $\mathcal{U} \leftarrow \mathcal{U} \setminus (x_i, \mathcal{Y})_i$
end

E.2.3 Fantasy PPP

In complete analogy to the proposed extension in section E.2.2, we consider assignment of all possible classes instead of predictions of single classes. As opposed to selecting from all possible pseudo-labels directly, we could also combine the PPPs from pseudo-labels for each instance to a fantasy PPP by a weighted sum. See the following pseudo-code for details. The formulation allows for different ways of how to define the weighted sum Σ. Regarding one instance, we would have a PPP for each class $y \in \mathcal{Y}$. One way to define Σ would be to consider the maximal and minimal PPP only and compute a weighted sum thereof, leaning on the Hurwicz-criterion in decision theory [Hurwicz, 1951]. The weight assigned to the maximal PPP is then regarded the decision-maker's degree of optimism.

Algorithm 4: Bayesian Pseudo-Label Selection (BPLS) with fantasy PPPs

Data: $\mathcal{D}$,$\mathcal{U}$
Result: $\mathcal{D}$, fitted model $\hat{y}^*(x)$
Fit model M on labeled data $\mathcal{D}$ to obtain prediction function $\hat{y}(x)$
while *stopping criterion not met* **do**
  for $i \in \{1, \ldots, |\mathcal{U}|\}$ **do**
    assign all possible $y_i \in \mathcal{Y}$ to $(x_i, y_i)_i$
    evaluate weighted sum Σ of respective PPPs $p(\mathcal{D} \cup (x_i, y_i)\,|\mathcal{D})$
  end
  obtain $i^* = \arg\max_i \Sigma$
  retrain M on $\mathcal{D} \cup (x_i, \hat{y}_{i^*})$
  predict $\mathcal{Y} \ni \hat{y}_i^*(\mathbf{x} \cup x_i), \mathbf{x} \in \mathcal{D}$
  add $(x_i, \hat{y}_i)$ to labeled data: $\mathcal{D} \leftarrow \mathcal{D} \cup (x_i, \hat{y}_i)$
  update $\mathcal{U} \leftarrow \mathcal{U} \setminus (x_i, \mathcal{Y})_i$
end

F NUMERICAL EXPERIMENTS VERIFYING THE SIMPLIFIED APPROXIMATION

F.1 SIMPLIFIED APPROXIMATION

We test the equivalence of PLS with regard to the approximate PPP criterion (Equation (6) in main paper)

$$\tilde{\ell}(\tilde{\theta}) - \frac{1}{2}\log|\mathcal{I}(\tilde{\theta})| + \log\pi(\tilde{\theta})$$

with $\tilde{\ell}(\tilde{\theta}) = \ell_{\mathcal{D}\cup(x_i,\hat{y}_i)}(\tilde{\theta}) + \ell_{\mathcal{D}}(\tilde{\theta})$, and our simplified version thereof (Equation (7) in main paper):

$$\ell_{\mathcal{D}\cup(x_i,\hat{y}_i)}(\tilde{\theta}) - \frac{1}{2}\log|\mathcal{I}(\tilde{\theta})| + \log\pi(\tilde{\theta}).$$

Recall that these terms are approximately equivalent when comparing pseudo-samples $(x_i, \hat{y}_i)$ and $(x_j, \hat{y}_j)$. We expanded $\ell_{\mathcal{D}}$ around its maximizer $\hat{\theta}$, so that $\ell_{\mathcal{D}}(\tilde{\theta}) = \ell_{\mathcal{D}}(\hat{\theta}) + O(\|\hat{\theta} - \tilde{\theta}\|^2)$. Since $\mathcal{D} \cup (x_i, \hat{y}_i)$ and $\mathcal{D}$ differ in only one sample, the difference $\hat{\theta} - \tilde{\theta}$ is of order $O(n^{-1})$. Thus,

$$\tilde{\ell}(\theta) = \ell_{\mathcal{D}\cup(x_i,\hat{y}_i)}(\theta) + \ell_{\mathcal{D}}(\hat{\theta}) + O(n^{-2}).$$

The remainder is negligible compared to the other terms in Equation (6) and $\ell_{\mathcal{D}}(\hat{\theta})$ does not depend on the pseudo-sample $(x_i, \hat{y}_i)$. This suggests the simplified *informative BPLS criterion* : $\ell_{\mathcal{D}\cup(x_i,\hat{y}_i)}(\tilde{\theta}) - \frac{1}{2}\log|\mathcal{I}(\tilde{\theta})| + \log\pi(\tilde{\theta})$.

F.2 EXPERIMENTAL SETUP

In addition to this theoretical argument, we provide empirical evidence for this equivalence. It is verified numerically for small n by experiments on the ionosphere data [Sigillito et al., 1989a], EEG data [Zhang et al., 1995], banknote data [Dua and Graff, 2017], abalone data [Waugh, 1995] as well as on simulated binomially distributed data, see section C.1. For all data sets, we compare semi-supervised GLM performance of BPLS with simplified criterion ("rough PPP", eq. 7) and unsimplified criterion ("fine PPP", eq. 6) with regard to test accuracy averaged over 40 repetitions.

F.3 RESULTS

Figure 11 shows the results for EEG data, Figure 13 for abalone data, while Figure 14 displayes results for the simulated binomially distributed data. In order to assess the *ceteris paribus* effect of growing n, we take random subsamples of the ionosphere data with varying size $n \in \{220, 260, 300\}$ and the full data set with $n = 350$. Figures 7 through 10 show the respective results.

It becomes apparent that with growing n the differences between the performances of the two approximations diminishes. Already for small n

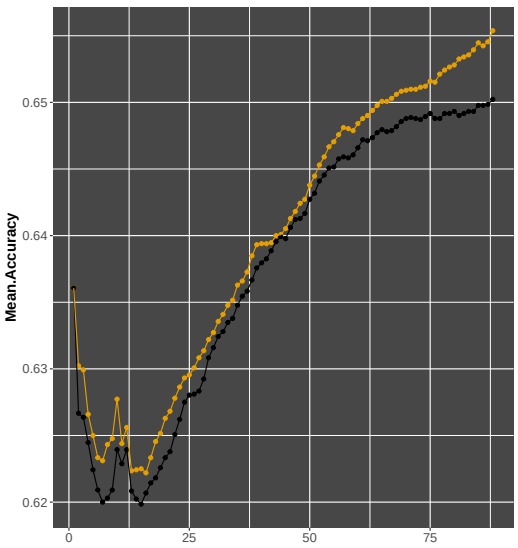

Figure 7: Approximations' performances on ionosphere subsample of size $n = 220$.

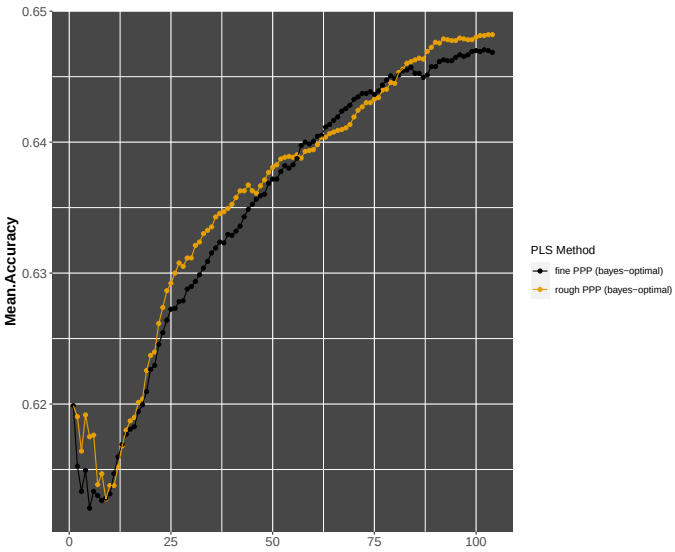

Figure 8: Approximations' performances on ionosphere subsample of size $n = 260$.

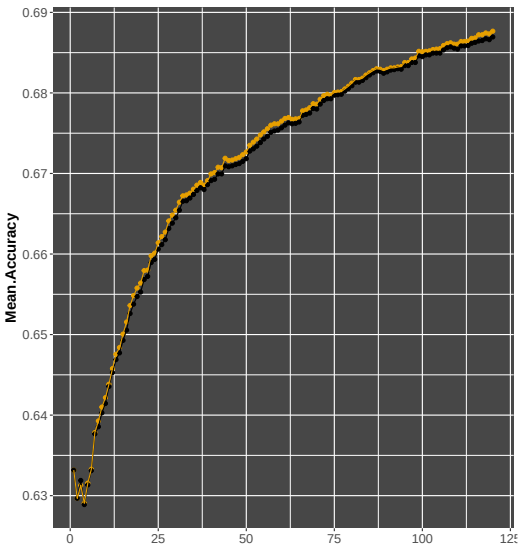

Figure 9: Approximations' performances on ionosphere subsample of size $n = 300$.

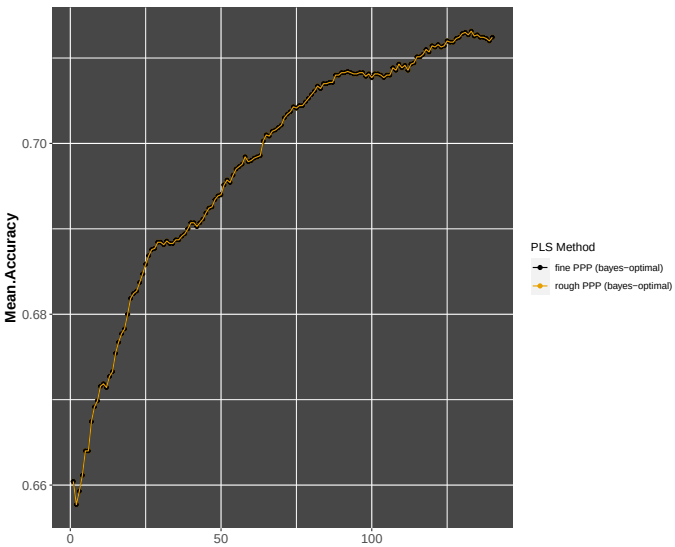

Figure 10: Approximations' performances on ionosphere data set of size $n = 350$.

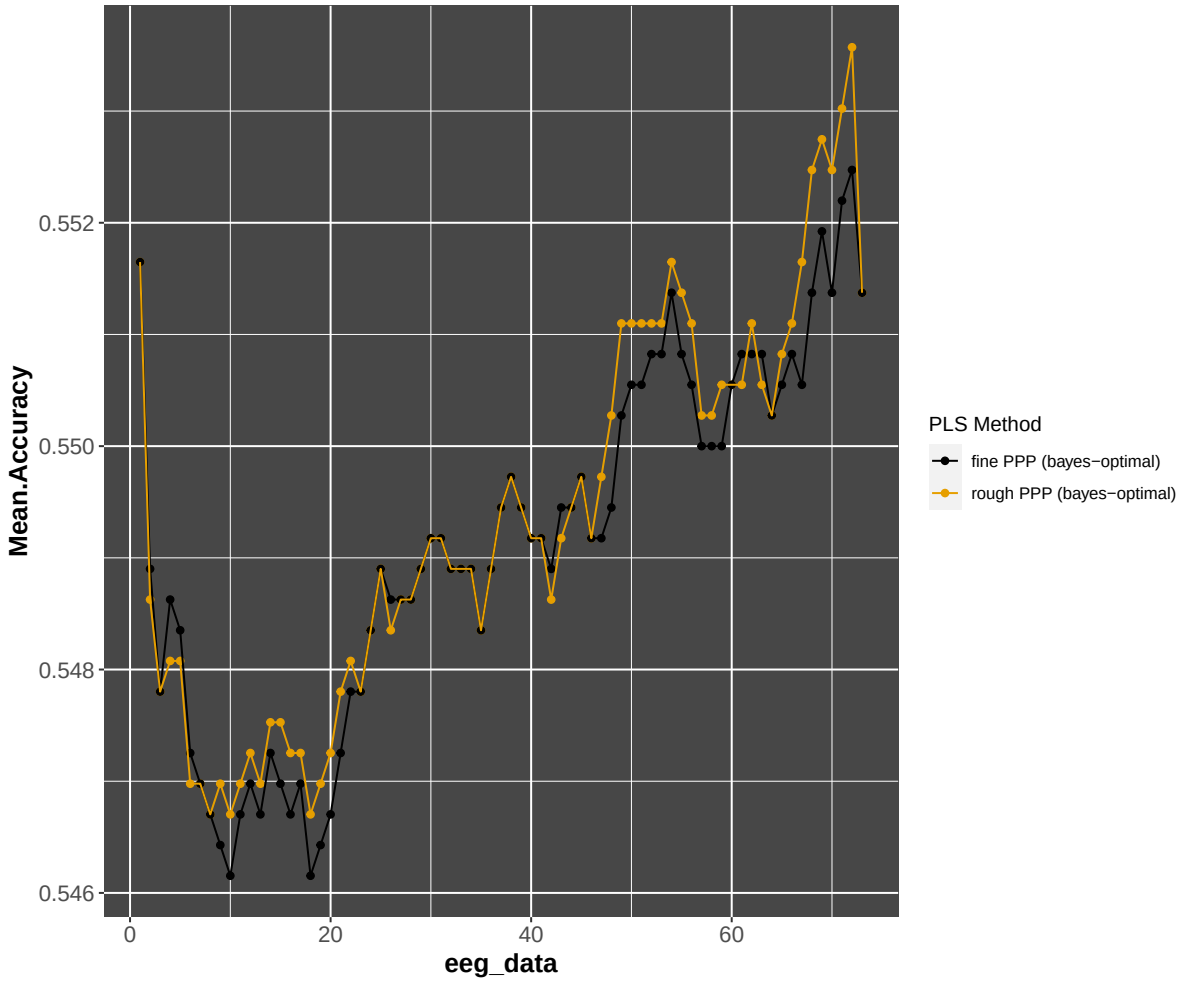

Figure 11: Approximations' performances on EEG data set ($n = 185$, $q = 13$).

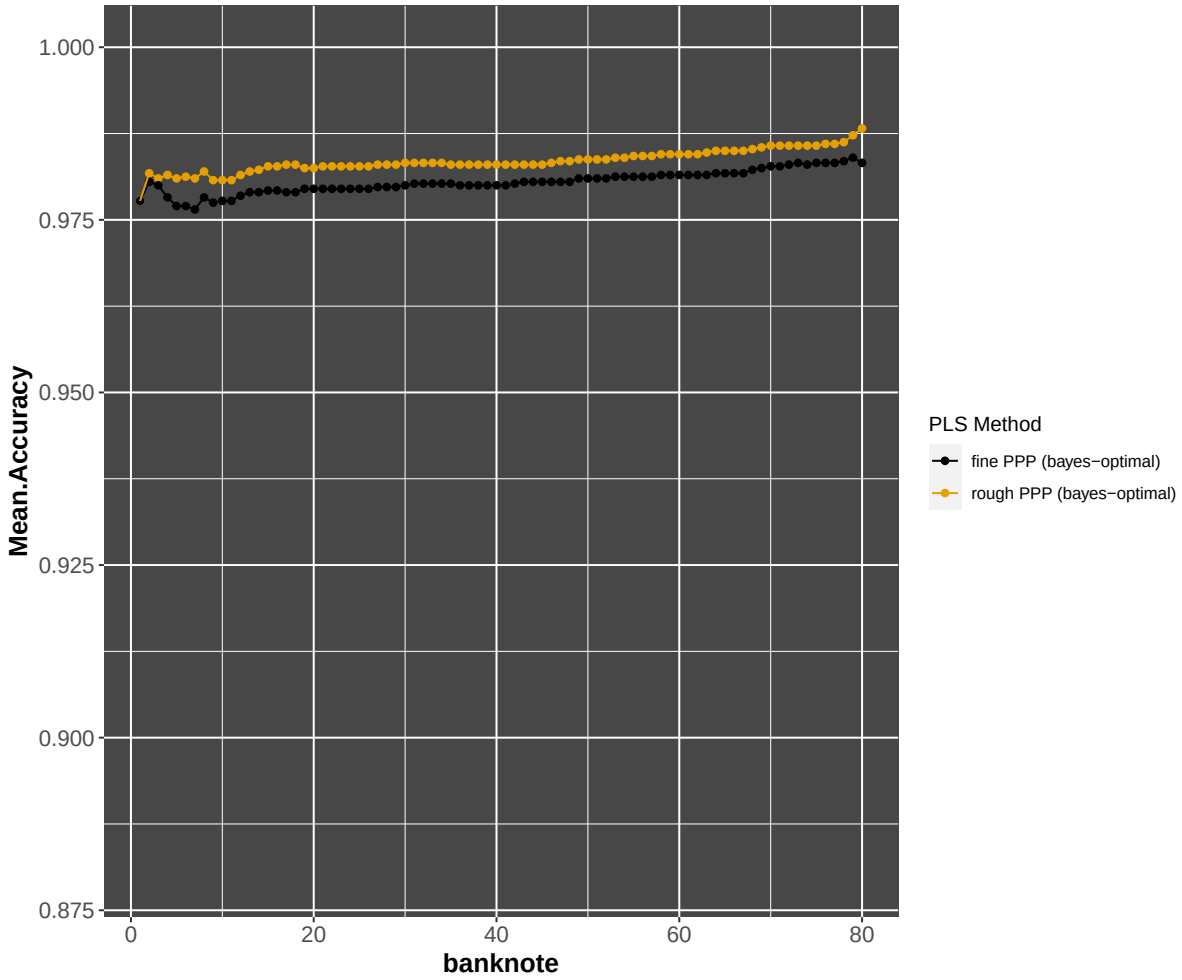

Figure 12: Approximations' performances on banknote data set ($n = 200$, $q = 3$).

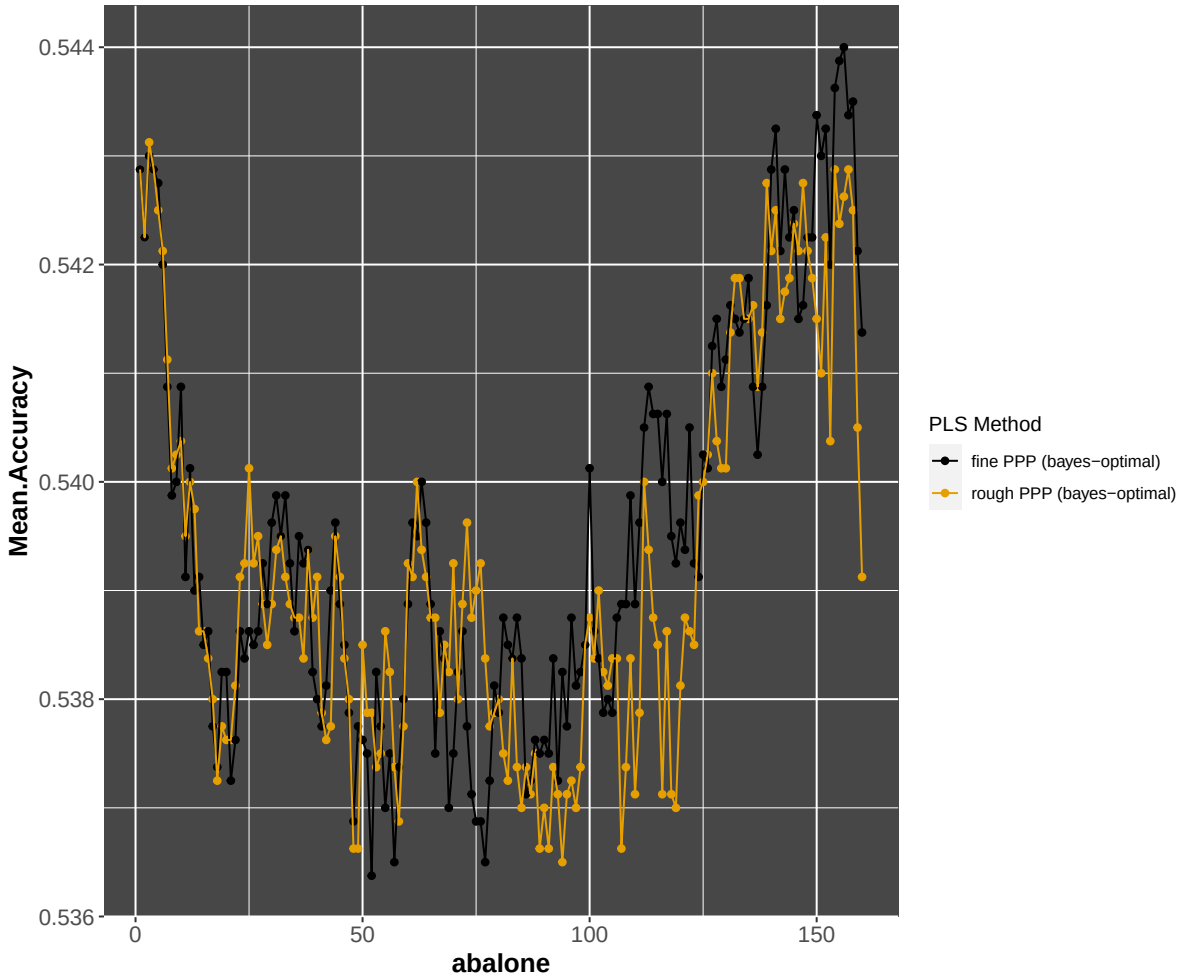

Figure 13: Approximations' performances on abalone data set ($n = 400$, $q = 4$).

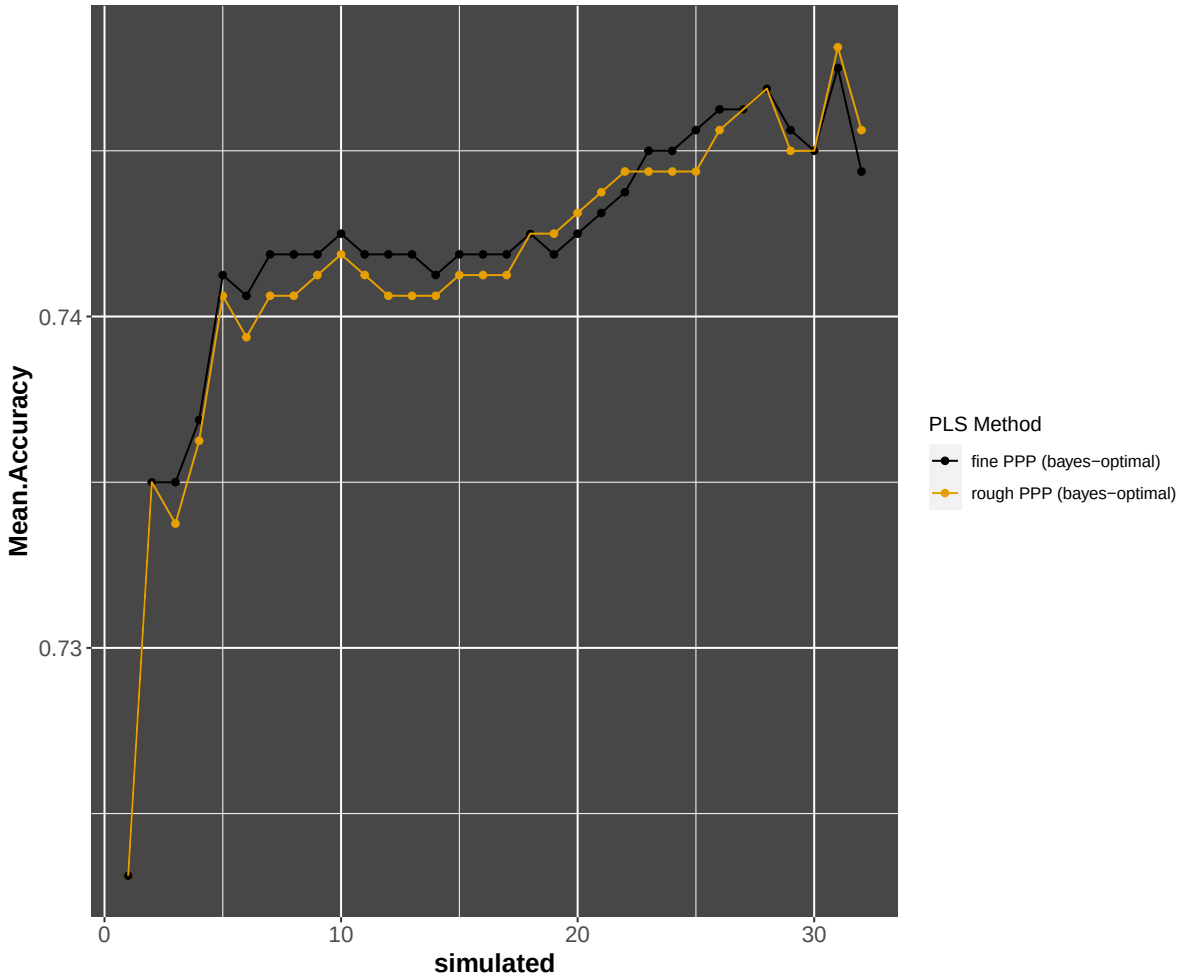

Figure 14: Approximations' performances on simultated data set ($n = 120$, $q = 4$).

G MCMC BASELINE

In order to compare our method against MCMC approximations of the pseudo posterior predictive (PPP), we compare BPLS with our approximation against selecting pseudo-samples according to an MCMC-approximation of the PPP.

The latter is reminiscent of [Li et al., 2020]: They propose to use mixtures of predictive distributions of a neural net (applying MC dropout) as a selection criterion. Essentially, this approach thus considers (MC)MC-based approximations of the posterior predictive of single pseudo-labeled data. Our approach differs by considering the joint posterior predictive but is similar with regard to the (Bayesian) concept. The main difference, however, is our analytical approximation of the posterior predictive. Li et al., 2020, thus propose a valid alternative to our approximate Bayes optimal criteria iBPLS (eq. 7 in main paper) and uBPLS (eq. 8 in main paper).

We benchmark this PLS method against our uBPLS on the smallest data set (cars) from the 8 UCI data sets in the paper as well as on a balanced subsample of another small data set (Pima diabetes data, see [Chang et al., 2022]), since the MC-based approximation is computationally very costly for high n. The following tables show the final (after all data were pseudo-labeled) mean accuracy on unseen test data, averaged over 100 replications.

The experiments demonstrate that our analytical approximations can compete with MC-sampling based approximations of the pseudo posterior predictive. The results can be smoothly added to the existing results, as they simply entail an additional baseline. All in all, we emphasize these additional results do not change the takeaway from the empirical evaluation of our method in the main paper.

The fact that our Laplace-based approximation outperforms MC-bases ones is slightly reminiscent of recent interesting trends in Bayesian uncertainty quantification in deep learning, where Laplace-based analytical approximations of the Hessian matrix were shown to outperform sampling-based (MC), see [Benzing, 2022, Daxberger et al., 2021, Izmailov et al., 2021, Wenzel et al., 2020].

Cars Data	
PLS Method	Mean Accuracy (Final)
MC-based approximation of PPP (Li et al.)	0.658
uBPLS approximation of PPP (our paper)	0.760
Likelihood (max-max)	0.719
Predictive Variance	0.691
Probability Score	0.733
Supervised Learning	0.727

Table 2: Comparison of BPLS with uBPLS approximation of PPP against MC-based approximation and other baselines on cars data.

Pima Data	
PLS Method	Mean Accuracy (Final)
MC-based approximation of PPP (Li et al.)	0.603
uBPLS approximation of PPP (our paper)	0.670
Likelihood (max-max)	0.667
Predictive Variance	0.663
Probability Score	0.677
Supervised Learning	0.675

Table 3: Comparison of BPLS with uBPLS approximation of PPP against MC-based approximation and other baselines on Pima data.

Cervical Cancer Data	
PLS Method	Mean Accuracy (Final)
MC-based approximation of PPP (Li et al.)	0.556
uBPLS approximation of PPP (our paper)	0.701
Likelihood (max-max)	0.611
Predictive Variance	0.644
Probability Score	0.688
Supervised Learning	0.611

Table 4: Comparison of BPLS with uBPLS approximation of PPP against MC-based approximation and other baselines on cervical cancer data.

EEG Data	
PLS Method	Mean Accuracy (Final)
MC-based approximation of PPP (Li et al.)	0.549
uBPLS approximation of PPP (our paper)	0.551
Likelihood (max-max)	0.544
Predictive Variance	0.541
Probability Score	0.547
Supervised Learning	0.537

Table 5: Comparison of BPLS with uBPLS approximation of PPP against MC-based approximation and other baselines on cervical cancer data.

Sonar Data	
PLS Method	Mean Accuracy (Final)
MC-based approximation of PPP (Li et al.)	0.521
uBPLS approximation of PPP (our paper)	0.550
Likelihood (max-max)	0.534
Predictive Variance	0.535
Probability Score	0.521
Supervised Learning	0.52

Table 6: Comparison of BPLS with uBPLS approximation of PPP against MC-based approximation and other baselines on EEG data.

H APPLICATION ON BAYESIAN NEURAL NETWORKS

We have implemented BPLS on Bayesian neural nets (BNNs) and run experiments on simulated data to benchmark BPLS against other PLS methods (exactly the same setup as in the paper, just with BNNs used to predict pseudo-labels instead of GLMs and GAMs).

We opt for BNNs, because they come with out-of-the-box uncertainty quantification. As model architecture, we use a simple feed-forward neural network with one layer consisting of 128 hidden neurons with a tanh activation function and one output neuron with a sigmoid activation for the binary classification case. For computing iBPLS (eq. 7 in main paper) and uBPLS (eq. 8 in main paper), we simply access the log-likelihood of the trained network. As we use variational inference by posterior mean-field approximation in the BNN with a multivariate normal prior with covariance of 0 for all weights, the evaluation of the log-determinant of the Fisher-info matrix (being a diagonal matrix here) simplifies to summing up the weights' variances.

We present preliminary results in the following tables. They show the mean accuracies (on test data) for different PLS methods on simulated data with uninformative and informative prior and BNNs trained with 50 and 150 epochs each. For the informative setup, we simulate data from a BNN, while for the uninformative setup, we simulate from a simple binomial distribution, which makes the classification task easier (see the generally higher accuracies in the uninformative setup). The general simulation setup follows the one described in supplement C.

The results confirm those for GLMs and GAMs reported in the main paper: In scenarios of low initial generalization (tables

2-4) inducing a high risk of overfitting, our method clearly outperforms other PLS methods, see experiments on hypothesis 1 a) in the paper. This is particularly pronounced in settings with informative priors (tables 3 and 4), see hypothesis 3 in the paper. In scenarios of high initial generalization (table 1) with low risk of overfitting, other methods have higher mean accuracies than our method, in line with hypothesis 1 b) in the paper.

Uninformative Setup, 150 epochs	
PLS Method	Mean Accuracy
Likelihood (max-max)	0.889
PPP (bayes-optimal)	0.884
Predictive Variance	0.884
Probability Score	0.880
Supervised Learning	0.890

Table 7: Comparison of mean accuracies of different PLS methods on simulated data with uninformative priors.

Uninformative Setup, 50 epochs	
PLS Method	Mean Accuracy
Likelihood (max-max)	0.562
PPP (bayes-optimal)	0.677
Predictive Variance	0.657
Probability Score	0.557
Supervised Learning	0.662

Table 8: Comparison of mean accuracies of different PLS methods on simulated data with uninformative priors.

Informative Setup, 150 epochs	
PLS Method	Mean Accuracy
Likelihood (max-max)	0.671
PPP (bayes-optimal)	0.702
Predictive Variance	0.695
Probability Score	0.637
Supervised Learning	0.583

Table 9: Comparison of mean accuracies of different PLS methods on simulated data with uninformative priors.

Informative Setup, 50 epochs	
PLS Method	Mean Accuracy
Likelihood (max-max)	0.513
PPP (bayes-optimal)	0.587
Predictive Variance	0.520
Probability Score	0.564
Supervised Learning	0.578

Table 10: Comparison of mean accuracies of different PLS methods on simulated data with uninformative priors.

I EXPERIMENTS: STATISTICAL HYPOTHESIS TESTING

As mentioned in section 4 of the main paper, we perform several non-parametric hypothesis tests tailored to comparing classification accuracies of different ML methods across multiple data sets, see especially [Demšar, 2006]. All hypotheses formulated in the paper (1a), 1b), 2a), 2b), and 3)) were tested. For conducting the tests, we compare both the final and the oracle-stopping (best among all iterations) accuracies of all PLS methods across different classification tasks on both simulated and real-world data. We present the results in what follows.

Hypothesis 1 (a) *PPP with uninformative prior outperforms traditional PLS on data prone to initial overfitting (i.e., with high ratio of features to data $\frac{q}{n}$ and poor initial generalization).* **(b)** *For low $\frac{q}{n}$ and high initial generalization, BPLS is outperformed by traditional PLS.*

1 a) *Final:* Using the multiple comparison approaches from [Demšar, 2006], the Friedman-test [Friedman, 1937, Friedman, 1940] for overall differences in final accuracies indicates a significant ($\alpha = 0.05$) difference between performances of all PLS methods (likelihood, PPP with uninformative prior, predictive variance, probability score and supervised baseline) on tasks prone to initial overfitting (with high ratio of features to data). A post-hoc Nemenyi-test [Nemenyi, 1963] for pairwise comparisons indicates a statistically significant ($\alpha = 0.05$) difference between PPP and the supervised baseline and no statistically significant ($\alpha = 0.05$) difference between all other PLS methods.

1 b) *Final:* On tasks with low $\frac{q}{n}$ and high initial generalization, the Friedman-test suggests no significant difference among all the PLS methods. A post-hoc test for pairwise comparisons can thus not be conducted.

1 a) *Oracle-stopping:* Again, the Friedman-test [Friedman, 1937, Friedman, 1940] for overall differences in oracle-stopping accuracies indicates a significant ($\alpha = 0.05$) difference between performances of the PLS methods on tasks prone to initial overfitting (with high ratio of features to data). This time, however, the post-hoc Nemenyi-test [Nemenyi, 1963] for pairwise comparisons indicates a statistically significant ($\alpha = 0.05$) pairwise difference between PPP with uninformative prior and all other methods. This confirms our heuristic reasoning in the interpretation section in section 4 in the main paper.

1 b) *Oracle-stopping:* On tasks with a low ratio of features to data, the Friedman-test indicated a significant difference ($\alpha = 0.05$) between all PLS methods. The post-hoc Nemenyi-test for pairwise comparisons, however, does not indicate significant differences ($\alpha = 0.05$) between any of the PLS methods.

Hypothesis 2 (a) Among all PLS methods, the pseudo-label likelihood (max-max-action) reinforces the initial model fit the most and **(b)** hardly improves generalization.

For this hypothesis, we do not compare (final and oracle-stopping) accuracies but the differences of them to the initial model's test accuracy. Further note that for **2 b)** we do not need the multiple comparison approaches from [Demšar, 2006], because we only compare the likelihood (max-max) PLS method to the supervised baseline. A standard Wilcoxon rank sum test will do.

2 a) *Final:* The Friedman-test [Friedman, 1937, Friedman, 1940] for overall differences indicates a significant ($\alpha = 0.05$) difference between all PLS methods' improvements compared to the supervised baseline. The post-hoc Nemenyi-test [Nemenyi, 1963] indicates a statistically significant ($\alpha = 0.05$) difference between the pseudo-label likelihood's (max-max-action's) improvements and the improvements of our PPP method.

2 b) *Final:* The Wilcoxon rank sum test [Wilcoxon, 1992] does not reject the (one-sided) null hypothesis that likelihood performs better than the initial model. (Note that, in order to be able to control the error probability when searching for evidence for our hypothesis **2 b)**, we test the complementary hypothesis as null.) As mentioned in section 4 (paragraph on interpretation) of the paper, there seems to be not enough evidence that the likelihood method cannot improve generalization. **Oracle-Stopping:** The test results for **Oracle-Stopping** accuracies exactly match those for **final** accuracies regarding Hypotheses **2 a)** and **2 b)**.

Hypothesis 3 PPP with informative prior outperforms traditional PLS methods universally.

3) *Final:* The Friedman-test shows a significant ($\alpha = 0.05$) difference between performances of PLS methods, and, indeed, the post-hoc Nemenyi-test [Nemenyi, 1963] for pairwise comparisons indicates a statistically significant ($\alpha = 0.05$) pairwise difference between our PPP with informative prior and all other methods. **Oracle-Stopping:** The test decisions for final and oracle-stopping accuracy metrics do not differ.

J EXPERIMENTS ON PLS UNDER DISTRIBUTIONAL SHIFT

In order to check whether the robustness towards confirmation is also helpful in the presence of distributional shifts, see section 6 of the main paper, we have conducted some preliminary experiments: We simulated labeled and unlabeled data from two different binomial distributions with the test data from the same distribution as the unlabeled data (share of unlabeled: 0.8 and 0.9, train/test-ratio: $\frac{1}{9}$, q = 7, GAMs with non-parametric splines). We closely followed the experimental setup described in supplement C.1.3.

The preliminary results below support the intuition that our Bayesian approach robustifies PLS also towards distributional shift. The tables below depict mean accuracy after varying number of self-training iterations (colums of tables) of PPP with informative priors and concurring PLS methods with non-parametric GAMs on simulated binomially distributed (with mean shift from labeled to unlabeled) data of varying sizes, just like in the experiments presented in figure 3 of the main paper.

At first sight, the results closely resemble the results from the experimental setup without distributional shift (see figure 3 in the paper and figure 6 in supplement D.3 as well as supplement C, in particular C.1.3.). However, there are differences: PPP only needs very few iterations (< 20) to outperform other PLS methods here. In the experiments without distributional shift, PPP also achieves accuracy gains over other methods after 10-40 iterations in some setups, see results for share of unlabeled = 0.9 in figure 6 in supplement D.3. The extreme speed of this process for data with a distributional shift, however, appears a bit odd. We do not have an explanation for this phenomenon yet. All in all, however, it appears as if indeed our Bayesian approach to the selection problem of pseudo-labeled data robustifies PLS not only towards initial overfitting and confirmation bias but also towards distributional shift. This of course requires more careful empirical evaluation. We leave this to future work.

$n = 500$, share of unlabeled: 0.8										
Method	20	40	60	80	100	120	140	160	180	200
———————	——	——	——	——	——	——	——	——	——	——
Likelihood (max-max)	0.9020	0.9005	0.8965	0.8950	0.8940	0.8915	0.8915	0.8910	0.8840	0.8830
PPP (bayes-optimal)	0.9515	0.9540	0.9530	0.9520	0.9530	0.9560	0.9520	0.9530	0.9550	0.9555
Predictive Variance	0.9065	0.8975	0.8975	0.8990	0.8990	0.8985	0.8980	0.8985	0.8985	0.8990
Probability Score	0.9005	0.8980	0.8950	0.8940	0.8950	0.8940	0.8940	0.8940	0.8885	0.8790
Supervised Learning	0.9035	0.9035	0.9035	0.9035	0.9035	0.9035	0.9035	0.9035	0.9035	0.9035

Table 11: Mean Accuracies after iterations $\{20, 40, 60, \dots\}$ from experiments on simulated binomially distributed data with distribution shift. $n = 500$.

$n = 500$, share of unlabeled: 0.9										
Method	20	40	60	80	100	120	140	160	180	200
———————	——	——	——	——	——	——	——	——	——	——
Likelihood (max-max)	0.8600	0.8600	0.8655	0.8590	0.8530	0.8520	0.8505	0.8500	0.8515	0.8525
PPP (bayes-optimal)	0.9100	0.9160	0.9150	0.9180	0.9110	0.9145	0.9120	0.9105	0.9105	0.9110
Predictive Variance	0.8655	0.8540	0.8550	0.8580	0.8565	0.8555	0.8550	0.8555	0.8555	0.8565
Probability Score	0.8605	0.8600	0.8585	0.8580	0.8530	0.8540	0.8535	0.8505	0.8540	0.8625
Supervised Learning	0.8585	0.8585	0.8585	0.8585	0.8585	0.8585	0.8585	0.8585	0.8585	0.8585

Table 12: Mean Accuracies after iterations $\{20, 40, 60, \dots\}$ from experiments on simulated binomially distributed data with distribution shift. $n = 500$.

$n = 1000$, share of unlabeled: 0.8														
Method	20	40	60	80	100	120	140	160	180	200	220	240	260	280
———————	——	——	——	——	——	——	——	——	——	——	——	——	——	——
Likelihood (max-max)	0.9128	0.9128	0.9128	0.9124	0.9120	0.9112	0.9112	0.9104	0.9104	0.9096	0.9084	0.9084	0.9068	0.9072
PPP (bayes-optimal)	0.9752	0.9752	0.9752	0.9724	0.9744	0.9756	0.9756	0.9760	0.9760	0.9760	0.9764	0.9764	0.9760	0.9760
Predictive Variance	0.9208	0.9188	0.9228	0.9204	0.9208	0.9200	0.9212	0.9200	0.9208	0.9208	0.9212	0.9212	0.9212	0.9212
Probability Score	0.9128	0.9128	0.9128	0.9120	0.9120	0.9112	0.9112	0.9104	0.9104	0.9096	0.9084	0.9080	0.9068	0.9064
Supervised Learning	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128	0.9128

Table 13: Mean Accuracies after iterations $\{20, 40, 60, \dots\}$ from experiments on simulated binomially distributed data with distribution shift. $n = 1000$.

$n = 2000$, share of unlabeled: 0.8											
Method	20	40	60	80	100	120	140	160	180	200	220
Likelihood (max-max)	0.9169444	0.9158333	0.9147222	0.9122222	0.9080556	0.9063889	0.9108333	0.9108333	0.9169444	0.9158333	0.9147222
PPP (bayes-optimal)	0.9655556	0.9688889	0.9691667	0.9691667	0.9688889	0.9688889	0.9675000	0.9686111	0.9655556	0.9688889	0.9691667
Predictive Variance	0.9363889	0.9200000	0.9155556	0.9136111	0.9191667	0.9188889	0.9105556	0.9111111	0.9363889	0.9200000	0.9155556
Probability Score	0.9163889	0.9152778	0.9152778	0.9125000	0.9080556	0.9036111	0.9072222	0.9063889	0.9163889	0.9152778	0.9152778
Supervised Learning	0.9111111	0.9111111	0.9111111	0.9111111	0.9111111	0.9111111	0.9111111	0.9111111	0.9111111	0.9111111	0.9111111

Table 14: Mean Accuracies after iterations $\{20, 40, 60, \dots\}$ from experiments on simulated binomially distributed data with distribution shift. $n = 2000$.

$n = 4000$, share of unlabeled: 0.8											
Method	20	40	60	80	100	120	140	160	180	200	220
Likelihood (max-max)	0.9551852	0.9550926	0.9550926	0.9546296	0.9540741	0.9556481	0.9545370	0.9540741	0.9531481	0.9523148	0.9518519
PPP (bayes-optimal)	0.9665741	0.9676852	0.9673148	0.9679630	0.9665741	0.9657407	0.9656481	0.9654630	0.9661111	0.9673148	0.9673148
Predictive Variance	0.9623148	0.9680556	0.9711111	0.9740741	0.9749074	0.9751852	0.9722222	0.9725926	0.9725000	0.9722222	0.9723148
Probability Score	0.9550926	0.9550000	0.9550000	0.9546296	0.9541667	0.9535185	0.9524074	0.9518519	0.9508333	0.9501852	0.9497222
Supervised Learning	0.9552778	0.9552778	0.9552778	0.9552778	0.9552778	0.9552778	0.9552778	0.9552778	0.9552778	0.9552778	0.9552778

Table 15: Mean Accuracies after iterations $\{20, 40, 60, \dots\}$ from experiments on simulated binomially distributed data with distribution shift. $n = 4000$.

$n = 8000$, share of unlabeled: 0.8											
Method	20	40	60	80	100	120	140	160	180	200	220
Likelihood (max-max)	0.9132237	0.9125000	0.9123026	0.9116447	0.9112500	0.9107895	0.9097368	0.9111184	0.9136184	0.9138816	0.9146053
PPP (bayes-optimal)	0.9673684	0.9676974	0.9586842	0.9586184	0.9536184	0.9536184	0.9551974	0.9581579	0.9584211	0.9584868	0.9596711
Predictive Variance	0.9351316	0.9363816	0.9397368	0.9406579	0.9403289	0.9409868	0.9410526	0.9410526	0.9410526	0.9411842	0.9412500
Probability Score	0.9133553	0.9124342	0.9123684	0.9116447	0.9112500	0.9107237	0.9098684	0.9112500	0.9134868	0.9138816	0.9150000
Supervised Learning	0.9132895	0.9132895	0.9132895	0.9132895	0.9132895	0.9132895	0.9132895	0.9132895	0.9132895	0.9132895	0.9132895

Table 16: Mean Accuracies after iterations $\{20, 40, 60, \dots\}$ from experiments on simulated binomially distributed data with distribution shift. $n = 8000$.

K REPRODUCIBILITY AND OPEN SCIENCE

The implementation of the proposed methods as well as reproducible scripts for the experiments are provided in the following repository named **Bayesian-pls** ("*Bayesian, please!*"): `https://github.com/rodemann/Bayesian-pls`. Please follow the instructions on the Readme-file to reproduce the experiments.

L DATA SETS

The following tables provide details on data sources as well as features and target variables of the eight real-world datasets from the UCI machine learning repository [Dua and Graff, 2017].

Table 17: Breast Cancer Data, Details: [Street et al., 1993]

Name	Class	Values
target	factor	'0' '1'
radius_mean	numeric	Num: 6.981 to 28.11
texture_mean	numeric	Num: 9.71 to 33.81
perimeter_mean	numeric	Num: 43.79 to 188.5
area_mean	numeric	Num: 143.5 to 2501
smoothness_mean	numeric	Num: 0.053 to 0.145
compactness_mean	numeric	Num: 0.019 to 0.311
concavity_mean	numeric	Num: 0 to 0.427
concave_points_mean	numeric	Num: 0 to 0.201
symmetry_mean	numeric	Num: 0.117 to 0.304
fractal_dimension_mean	numeric	Num: 0.05 to 0.097
radius_se	numeric	Num: 0.112 to 2.873
texture_se	numeric	Num: 0.36 to 4.885
perimeter_se	numeric	Num: 0.757 to 21.98
area_se	numeric	Num: 6.802 to 542.2
smoothness_se	numeric	Num: 0.002 to 0.031
compactness_se	numeric	Num: 0.002 to 0.106
concavity_se	numeric	Num: 0 to 0.396
concave_points_se	numeric	Num: 0 to 0.053
symmetry_se	numeric	Num: 0.008 to 0.061
fractal_dimension_se	numeric	Num: 0.001 to 0.03
radius_worst	numeric	Num: 7.93 to 36.04
texture_worst	numeric	Num: 12.02 to 49.54
perimeter_worst	numeric	Num: 50.41 to 251.2
area_worst	numeric	Num: 185.2 to 4254
smoothness_worst	numeric	Num: 0.071 to 0.223
compactness_worst	numeric	Num: 0.027 to 1.058
concavity_worst	numeric	Num: 0 to 1.252
concave_points_worst	numeric	Num: 0 to 0.287
symmetry_worst	numeric	Num: 0.156 to 0.664
fractal_dimension_worst	numeric	Num: 0.055 to 0.208

Table 18: Sonar Data Set, Details: [Gorman and Sejnowski, 1988]

Name	Class	Values
V1	numeric	Num: 0.002 to 0.137
V2	numeric	Num: 0.001 to 0.234
V3	numeric	Num: 0.002 to 0.306
V4	numeric	Num: 0.006 to 0.426
V5	numeric	Num: 0.007 to 0.401
V6	numeric	Num: 0.01 to 0.382
V7	numeric	Num: 0.003 to 0.373
V8	numeric	Num: 0.005 to 0.459
V9	numeric	Num: 0.007 to 0.683
V10	numeric	Num: 0.011 to 0.711
V11	numeric	Num: 0.029 to 0.734
V12	numeric	Num: 0.024 to 0.706
V13	numeric	Num: 0.018 to 0.713
V14	numeric	Num: 0.027 to 0.997
V15	numeric	Num: 0.003 to 1
V16	numeric	Num: 0.016 to 0.999
V17	numeric	Num: 0.035 to 1
V18	numeric	Num: 0.038 to 1
V19	numeric	Num: 0.049 to 1
V20	numeric	Num: 0.066 to 1
V21	numeric	Num: 0.051 to 1
V22	numeric	Num: 0.022 to 1
V23	numeric	Num: 0.056 to 1
V24	numeric	Num: 0.024 to 1
V25	numeric	Num: 0.024 to 1
V26	numeric	Num: 0.092 to 1
V27	numeric	Num: 0.048 to 1
V28	numeric	Num: 0.028 to 1
V29	numeric	Num: 0.014 to 1
V30	numeric	Num: 0.061 to 1
V31	numeric	Num: 0.048 to 0.966
V32	numeric	Num: 0.04 to 0.931
V33	numeric	Num: 0.048 to 1
V34	numeric	Num: 0.021 to 0.965
V35	numeric	Num: 0.022 to 1
V36	numeric	Num: 0.008 to 1
V37	numeric	Num: 0.035 to 0.95
V38	numeric	Num: 0.038 to 1
V39	numeric	Num: 0.037 to 0.986
V40	numeric	Num: 0.012 to 0.93
V41	numeric	Num: 0.036 to 0.899
V42	numeric	Num: 0.006 to 0.825
V43	numeric	Num: 0 to 0.773
V44	numeric	Num: 0 to 0.776
V45	numeric	Num: 0 to 0.703
V46	numeric	Num: 0 to 0.729
V47	numeric	Num: 0 to 0.552
V48	numeric	Num: 0 to 0.334
V49	numeric	Num: 0 to 0.198
V50	numeric	Num: 0 to 0.082
V51	numeric	Num: 0 to 0.1
V52	numeric	Num: 0.001 to 0.071
V53	numeric	Num: 0 to 0.039
V54	numeric	Num: 0.001 to 0.035
V55	numeric	Num: 0.001 to 0.045
V56	numeric	Num: 0 to 0.039
V57	numeric	Num: 0 to 0.035
V58	numeric	Num: 0 to 0.044
V59	numeric	Num: 0 to 0.036
V60	numeric	Num: 0.001 to 0.044
V61	matrix	Num: 1 to 2

Table 19: Mushrooms Data Set, Details: [Schlimmer, 1987]

Name	Class	Values
cap.diameter	numeric	Num: 0.71 to 54.6
stem.height	numeric	Num: 0 to 28.33
stem.width	numeric	Num: 0 to 52.22
target	factor	’0’ ’1’

Table 20: Banknote Data Set, Details: archive.ics.uci.edu/ml/datasets/banknote+authentication

Name	Class	Values
target	factor	'0' '1'
Length	numeric	Num: 213.8 to 216.3
Left	numeric	Num: 129 to 131
Right	numeric	Num: 129 to 131.1
Bottom	numeric	Num: 7.2 to 12.7
Top	numeric	Num: 7.7 to 12.3
Diagonal	numeric	Num: 137.8 to 142.4

Table 21: Abalone Data Set, Details: [Waugh, 1995]

Name	Class	Values
target	factor	'0' '1'
rings	numeric	Num: 4 to 29
length	numeric	Num: 0.165 to 0.775
weight	numeric	Num: 0.024 to 2.493
height	numeric	Num: 0.04 to 0.24
diameter	numeric	Num: 0.125 to 0.605
shell_weight	numeric	Num: 0.008 to 0.885

Table 22: Cars Data Set, Details: [Ezekiel, 1930]

Name	Class	Values
wt	numeric	Num: 1.513 to 5.424
qsec	numeric	Num: 14.5 to 22.9
vs	factor	'0' '1'

Table 23: EEG Data Set, Details: [Zhang et al., 1995]

Name	Class	Values
V1	numeric	Num: -2.035 to 1
V2	numeric	Num: -1.005 to 1
V3	numeric	Num: -0.912 to 1
V4	numeric	Num: -1.107 to 1
V5	numeric	Num: -1.078 to 1
V6	numeric	Num: -1.073 to 1
V7	numeric	Num: -1.651 to 1
V8	numeric	Num: -1.024 to 1
V9	numeric	Num: -1.864 to 1
V10	numeric	Num: -1.604 to 1
V11	numeric	Num: -0.883 to 1
V12	numeric	Num: -1.087 to 1
target	factor	'0' '1'

Table 24: Ionosphere Data, Details: [Sigillito et al., 1989b]

Name	Class	Values
V1	integer	Num: 0 to 1
V3	numeric	Num: -1 to 1
V4	numeric	Num: -1 to 1
V5	numeric	Num: -1 to 1
V6	numeric	Num: -1 to 1
V7	numeric	Num: -1 to 1
V8	numeric	Num: -1 to 1
V9	numeric	Num: -1 to 1
V10	numeric	Num: -1 to 1
V11	numeric	Num: -1 to 1
V12	numeric	Num: -1 to 1
V13	numeric	Num: -1 to 1
V14	numeric	Num: -1 to 1
V15	numeric	Num: -1 to 1
V16	numeric	Num: -1 to 1
V17	numeric	Num: -1 to 1
V18	numeric	Num: -1 to 1
V19	numeric	Num: -1 to 1
V20	numeric	Num: -1 to 1
V21	numeric	Num: -1 to 1
V22	numeric	Num: -1 to 1
V23	numeric	Num: -1 to 1
V24	numeric	Num: -1 to 1
V25	numeric	Num: -1 to 1
V26	numeric	Num: -1 to 1
V27	numeric	Num: -1 to 1
V28	numeric	Num: -1 to 1
V29	numeric	Num: -1 to 1
V30	numeric	Num: -1 to 1
V31	numeric	Num: -1 to 1
V32	numeric	Num: -1 to 1
V33	numeric	Num: -1 to 1
V34	numeric	Num: -1 to 1
target	factor	’0’ ’1’

M REFERENCES OF SUPPLEMENTARY MATERIAL

REFERENCES

[Augustin et al., 2014] Augustin, T., Coolen, F. P., de Cooman, G., and Troffaes, M. C. M., editors (2014). *Introduction to Imprecise Probabilities*. John Wiley, Chichester.

[Benzing, 2022] Benzing, F. (2022). Gradient descent on neurons and its link to approximate second-order optimization. In *International Conference on Machine Learning*, pages 1817–1853. PMLR.

[Chang et al., 2022] Chang, V., Bailey, J., Xu, Q. A., and Sun, Z. (2022). Pima indians diabetes mellitus classification based on machine learning (ml) algorithms. *Neural Computing and Applications*, pages 1–17.

[Daxberger et al., 2021] Daxberger, E., Nalisnick, E., Allingham, J. U., Antorán, J., and Hernández-Lobato, J. M. (2021). Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pages 2510–2521. PMLR.

[Dempster, 1968] Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232.

[Demšar, 2006] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

[Dua and Graff, 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

[Ezekiel, 1930] Ezekiel, M. (1930). Methods of correlation analysis.

[Fahrmeir et al., 2013] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Regression: Models, methods and applications.

[Friedman, 1937] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

[Friedman, 1940] Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

[Gorman and Sejnowski, 1988] Gorman, R. P. and Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1(1):75–89.

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

[Hastie, 2017] Hastie, T. (2017). Generalized additive models. In Chambers, J. M. and Hastie, T., editors, *Statistical models in S*, pages 249–307. Routledge.

[Hurwicz, 1951] Hurwicz, L. (1951). The generalized Bayes minimax principle: a criterion for decision making under uncertainty. *Cowles Commission Discussion Paper Statistics*, 335:1950.

[Izmailov et al., 2021] Izmailov, P., Nicholson, P., Lotfi, S., and Wilson, A. G. (2021). Dangers of bayesian model averaging under covariate shift. *Advances in Neural Information Processing Systems*, 34:3309–3322.

[Li et al., 2020] Li, S., Wei, Z., Zhang, J., and Xiao, L. (2020). Pseudo-label selection for deep semi-supervised learning. In *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 1–5. IEEE.

[Nelder and Wedderburn, 1972] Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

[Nemenyi, 1963] Nemenyi, P. B. (1963). *Distribution-free multiple comparisons.* Princeton University.

[Ruggeri et al., 2005] Ruggeri, F., Insua, D. R., and Martín, J. (2005). Robust Bayesian analysis. volume 25, pages 623–667. Elsevier.

[Schlimmer, 1987] Schlimmer, J. C. (1987). *Concept acquisition through representational adjustment*. University of California, Irvine.

[Sigillito et al., 1989a] Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989a). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266.

[Sigillito et al., 1989b] Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989b). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins Applied Physics Laboratory Technical Digest*, 10(3):262–266.

[Street et al., 1993] Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE.

[Walley, 1991] Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman & Hall.

[Waugh, 1995] Waugh, S. G. (1995). *Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks*. PhD thesis, University of Tasmania.

[Wenzel et al., 2020] Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the Bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*.

[Wilcoxon, 1992] Wilcoxon, F. (1992). *Individual comparisons by ranking methods*. Springer.

[Zhang et al., 1995] Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538.