

CONFIDENT SINKHORN ALLOCATION FOR PSEUDO-LABELING

Anonymous authors

Paper under double-blind review

ABSTRACT

Semi-supervised learning is a critical tool in reducing machine learning’s dependence on labeled data. It has been successfully applied to structure data, such as image and language data, by exploiting the inherent spatial and semantic structure therein with pretrained models or data augmentation. Some of these methods are no longer applicable for the data where domain structures are not available because the pretrained models or data augmentation can not be used.

Due to simplicity, existing pseudo-labeling (PL) methods can be widely used without any domain assumption, but are vulnerable to noise samples and to greedy assignments given a predefined threshold which is typically unknown. This paper addresses this problem by proposing a Confident Sinkhorn Allocation (CSA), which assigns labels to only samples with high confidence scores and learns the best label allocation via optimal transport. CSA outperforms the current state-of-the-art in this practically important area of semi-supervised learning.

1 INTRODUCTION

The impact of machine learning continues to grow in fields as disparate as biology (Libbrecht & Noble, 2015; Tunyasuvunakool et al., 2021), quantum technology (Biamonte et al., 2017; van Esbroeck et al., 2020; Nguyen et al., 2021), brain stimulation (Boutet et al., 2021; van Bueren et al., 2021), and computer vision (Esteva et al., 2021; Yoon et al., 2022). Much of this impact depends on the availability of large numbers of annotated examples for the machine learning models to be trained on. The data annotation task by which such labeled data is created is often expensive, and sometimes impossible, however. Rare genetic diseases, stock market events, and cyber-security threats, for example, are hard to annotate due to the volumes of data involved, the rate at which the significant characteristics change, or both.

Related work. Fortunately, for some classification tasks, we can overcome a scarcity of labeled data using semi-supervised learning (SSL) (Zhu, 2005; Huang et al., 2021; Killamsetty et al., 2021; Olsson et al., 2021). SSL exploits an additional set of unlabeled data with the goal of improving on the performance that might be achieved using labeled data alone (Lee et al., 2019; Carmon et al., 2019; Ren et al., 2020; Islam et al., 2021).

Domain specific: Semi-supervised learning for image and language data has made rapid progress (Oymak & Gulcu, 2021; Zhou, 2021; Sohn et al., 2020) largely by exploiting the inherent spatial and semantic structure of images (Komodakis & Gidaris, 2018) and language (Kenton & Toutanova, 2019). This is achieved typically either using pretext tasks (eg. (Komodakis & Gidaris, 2018; Alexey et al., 2016)) or contrastive learning (eg. (Van den Oord et al., 2018; Chen et al., 2020)). Both approaches assume that specific transformations applied to each data element will not affect the associated label.

Greedy pseudo-labeling: Without domain assumption, a simple but effective way for SSL is pseudo-labeling (PL) (Lee et al., 2013) which generates ‘pseudo-labels’ for unlabeled samples using a model trained on labeled data. A label k is assigned to an unlabeled sample \mathbf{x}_i where a predicted class probability is larger than a predefined threshold γ as

$$y_i^k = \mathbb{1}\left[p(y_i = k \mid \mathbf{x}_i) \geq \gamma\right] \quad (1)$$

Table 1: Comparison with the related approaches in terms of properties and their relative trade-offs.

Algorithms	Not domain specific	Uncertainty consideration	Non-greedy
Pseudo-Labeling (Lee et al., 2013)	✓	✗	✗
FlexMatch (Zhang et al., 2021)	✓	✗	✗
Vime (Yoon et al., 2020)	✗	✗	NA
MixMatch (Berthelot et al., 2019)	✗	✗	NA
FixMatch (Sohn et al., 2020)	✗	✗	NA
UPS (Rizve et al., 2021)	✓	✓	✗
SLA (Tai et al., 2021)	✓	✗	✓
CSA	✓	✓	✓

where $\gamma \in [0, 1]$ is a threshold used to produce hard labels and $p(y_i = k | \mathbf{x}_i)$ is the predictive probability of the i -th data point belonging to the class k -th. A classifier can then be trained using both the original labeled data and the newly pseudo-labeled data. Pseudo labeling is naturally an iterative process, with the next round of pseudo-labels being generated using the most-recently trained classifier. The key advantage of pseudo-labeling is that it does not inherently require any domain assumption and can be generally applied to most domains, including tabular data.

Greedy PL with uncertainty: Rizve et al. (2021) propose an uncertainty-aware pseudo-label selection (UPS) that aims to reduce the noise in the training process by using the uncertainty score – together with the probability score for making assignments:

$$y_i^k = \mathbb{1}\left[p(y_i = k | \mathbf{x}_i) \geq \gamma\right] \mathbb{1}\left[\mathcal{U}(p(y_i = k | \mathbf{x}_i)) \leq \gamma_u\right] \quad (2)$$

where γ_u is an additional threshold on the uncertainty level and $\mathcal{U}(p)$ is the uncertainty of a prediction p . As shown in Rizve et al. (2021), selecting predictions with low uncertainties greatly reduces the effect of poor calibration, thus improving robustness and generalization.

However, the aforementioned works in PL are *greedy* in assigning the labels by simply comparing the prediction value against a predefined threshold γ irrespective of the relative prediction values across samples and classes. Such greedy strategies will be sensitive to the choice of a threshold.

Non-greedy pseudo-labeling: FlexMatch (Zhang et al., 2021) considers adaptively selecting a threshold γ_k for each class based on the level of difficulty. This threshold is adapted using the predictions across classes. However, the selection process is still heuristic in comparing the prediction score with an adjusted threshold. Recently, Tai et al. (2021) provide a novel view in connecting the pseudo-labeling assignment task to optimal transport problem, called SLA, which inspires our work. SLA and FlexMatch are better than existing PL in that their *non-greedy* label assignments not only use the single prediction value but also consider the relative importance of this value across rows and columns in a holistic way. However, both SLA and FlexMatch can overconfidently assign labels to noise samples and have not considered utilizing uncertainty values in making assignments.

Contributions. We propose here a semi-supervised learning method that does not require any domain-specific assumption for the data. We hypothesize that this is by far the most common case for the vast volumes of data that exist. The method we propose is based on pseudo-labeling of a set of unlabeled data using Confident Sinkhorn Allocation (CSA). Our method is theoretically driven by the role of uncertainty in robust label assignment in SSL. CSA utilizes Sinkhorn’s algorithm (Cuturi, 2013) to assign labels to only the data samples with high confidence scores. By learning the label assignment with optimal transport, CSA eliminates the need to predefine the heuristic thresholds used in existing pseudo-labeling methods, which can be greedy. The proposed CSA is applicable to any data domain, and could be used in concert with consistency-based approaches (Sohn et al., 2020), but is particularly useful for data domain where pretext tasks and data augmentation are not applicable, such a tabular data.

2 CONFIDENT SINKHORN ALLOCATION (CSA)

We consider the semi-supervised learning setting whereby we have access to a dataset consisting of labeled examples $\mathcal{D}_l = \{\mathbf{x}_i, y_i\}_{i=1}^{N_l}$, and one of unlabeled examples $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=1}^{N_u}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{1, \dots, K\}$. We define also $\mathcal{X} = \{\mathbf{x}_i\}$, $i = \{1, \dots, N_l + N_u\}$. Our goal is to utilize $\mathcal{D}_l \cup \mathcal{D}_u$ to

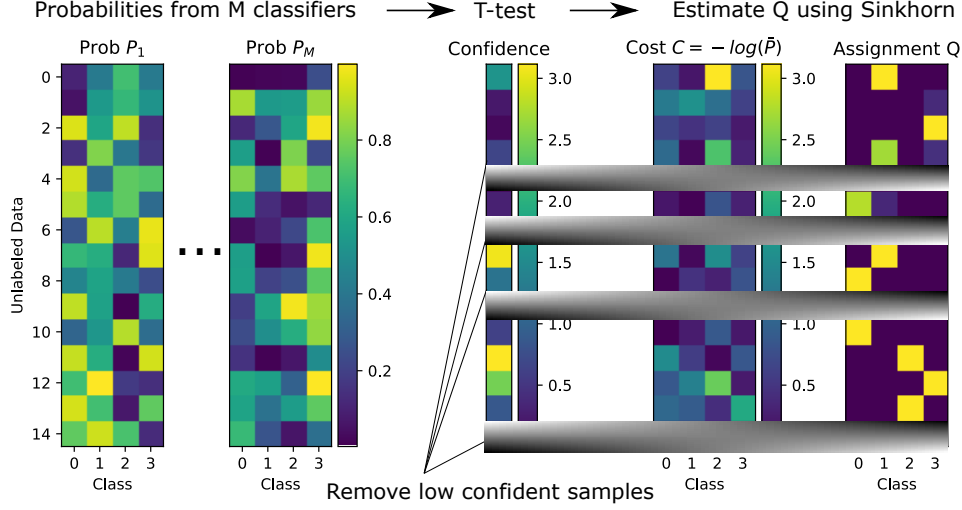


Figure 1: A depiction of CSA in application. We estimate the ensemble of predictions P on unlabeled data using M classifiers which can result in different probabilities. We then identify high-confidence samples by performing a T-test. Next, we estimate the label assignment Q using Sinkhorn’s algorithm. The cost C of the optimal transport problem is the negative of the probability averaging across classifiers, $C = -\log \bar{P}$. We repeat the process on the remaining unlabeled data as required.

learn a predictor $f: \mathcal{X} \rightarrow \mathcal{Y}$ that is more accurate than a predictor trained using labeled data \mathcal{D}_l alone. A notation summary is provided in Appendix Table 3.

Generating high-quality pseudo labels is critical to the final classification performance, as erroneous label assignment can quickly lead the iterative pseudo-labeling process astray. We provide in Sec. 2.1 a theoretical analysis of the role and impact of uncertainty in pseudo-labeling, the first such analysis as far as we are aware. Based on the theoretical result, we propose two approaches to identify the high-confidence samples for assigning labels and use the Sinkhorn’s algorithm to find the best label assignment. We name the proposed algorithm Confident Sinkhorn Allocation (CSA). We provide a diagram demonstrating CSA in Fig 1, and a comparison against related algorithms in Table 1.

2.1 THEORETICAL ANALYSIS ON THE EFFECT OF UNCERTAINTY IN PSEUDO-LABELING

Our theoretical results highlight two properties of PL settings: (i) less noise and uncertainty are beneficial, and (ii) more unlabeled data is beneficial. For the analysis, we make a popular assumption that input features $\mathbf{x}_i \in \mathcal{D}_l$ and \mathcal{D}_u are sampled i.i.d. from a feature distribution P_X , and the labeled data pairs (\mathbf{x}_i, y_i) in \mathcal{D}_l are drawn from a joint distribution $P_{X,Y}$.

We consider the binary classification problem in one-dimensional space as follows: generate the feature $x | y = +1 \sim \mathcal{N}(\mu_+, \sigma^2)$ and similarly $x | y = -1 \sim \mathcal{N}(\mu_-, \sigma^2)$, where μ_+ and μ_- are the means of the positive and negative classes respectively, σ is the standard deviation indicating the level of *data noisiness*. Without loss of generality, let $\mu_+ > \mu_-$. The optimal Bayes’s classifier is $f(x) = \text{sign}(x - \frac{\mu_+ + \mu_-}{2})$, eg. classify $y_i = +1$ if $x_i > (\mu_+ + \mu_-)/2$. Therefore, in the following, we measure the probability bound to learn $(\mu_+ + \mu_-)/2$ as a criterion for achieving good performance.

Let $\{\tilde{X}_i^+\}_{i=1}^{\tilde{n}_+}$ and $\{\tilde{X}_i^-\}_{i=1}^{\tilde{n}_-}$ be the sets of unlabeled data whose pseudo-labels are $+1$ and -1 , respectively. Let $\{I_i^+\}_{i=1}^{\tilde{n}_+}$ be the binary indicators of correct assignments, such as if $I_i^+ = 1$, then $\tilde{X}_i^+ \sim \mathcal{N}(\mu_+, \sigma^2)$ and otherwise $\tilde{X}_i^+ \sim \mathcal{N}(\mu_-, \sigma^2)$. Similarly, we define $\{I_i^-\}_{i=1}^{\tilde{n}_-}$ indicating correct assignment for negative class. Instead of generating from Bernoulli distribution as used in Yang & Xu (2020), we generate the binary indicators I_i^+ and I_i^- from a classifier and denote the expectations $\mathbb{E}(I_i^+), \mathbb{E}(I_i^-)$ and variances $\text{Var}(I_i^+), \text{Var}(I_i^-)$, respectively. In the uncertainty estimation literature (Der Kiureghian & Ditlevsen, 2009), σ^2 is called the *aleatoric uncertainty* (noise from the observations) and $\text{Var}(I_i^+), \text{Var}(I_i^-)$ are the *epistemic uncertainties* (from the model knowledge). As mentioned above, we aim to learn $\frac{\mu_+ + \mu_-}{2}$, via the extra unlabeled data. It is natural from the binary

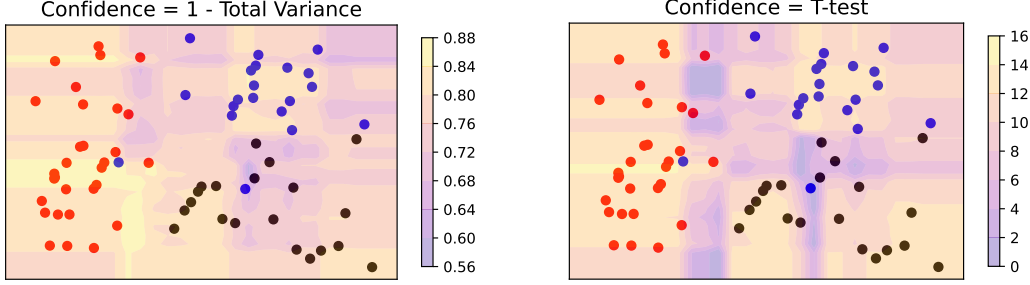


Figure 2: Examples of using total variance and T-test for estimating the confidence level on $K = 3$ classes. The yellow area indicates the highest confidence. We exclude the samples falling into the dark area where the value of T-test less than 2. We have simplified the tabular data (with possibly mixed categorical-continuous-text features) into the $d = 2$ dimensional representation.

classification setting above to construct our estimate as $\hat{\theta} = \frac{1}{2} \left(\sum_{i=1}^{\tilde{n}_+} \tilde{X}_i^+ / \tilde{n}_+ + \sum_{i=1}^{\tilde{n}_-} \tilde{X}_i^- / \tilde{n}_- \right)$. Please see Appendix A.1 for details.

Theorem 1. For $\delta > 0$, with a probability at least $1 - 2\exp\left\{-\frac{2\delta^2[\tilde{n}_+ + \tilde{n}_-]}{9\sigma^2[\tilde{n}_+ + \tilde{n}_-]}\right\} - \frac{9[\text{Var}(I_i^+) + \text{Var}(I_i^-)]}{4\delta^2}$, our estimate $\hat{\theta}$ satisfies $\left|\hat{\theta} - \frac{\Delta(\mu_+ + \mu_-)}{2} - \frac{(\mu_+ + \mu_-)}{2}\right| \leq \delta$ where $\Delta = |\mathbb{E}(I_i^+) - \mathbb{E}(I_i^-)|$.

Proof. Our proof (see Appendix A.1) extends the result from Yang & Xu (2020) taking into account the data noisiness σ^2 and uncertainty $\text{Var}(I_i^+), \text{Var}(I_i^-)$ for pseudo-labeling. \square

Interpretation. The theoretical result reveals several interesting aspects. First, training data imbalance affects the accuracy of our estimation, as discussed in Yang & Xu (2020). The more imbalanced the data is, the larger the gap Δ would be, which influences the closeness between the estimated $\hat{\theta}$ and desired value $\frac{(\mu_+ + \mu_-)}{2}$. Note that data imbalance is not the focus of this paper. In addition, the second term of the bound $\frac{\tilde{n}_+ + \tilde{n}_-}{\tilde{n}_+ \tilde{n}_-}$ shows that more unlabeled data $\tilde{n}_+ \uparrow$ or $\tilde{n}_- \uparrow$ is helpful for a good estimation. We empirically validate this property in Fig. 5. Finally, our analysis takes a step further to show that both aleatoric uncertainty σ^2 and epistemic uncertainty $\text{Var}(I_i^+) + \text{Var}(I_i^-)$ can reduce the probability of obtaining a good estimation. In other words, less uncertainty is more helpful. Note that the interpretation through the uncertainty has not been explored in Yang & Xu (2020). To the best of our knowledge, we are the first to theoretically characterize the role of data noise σ^2 and uncertainties $\text{Var}(I_i^+) + \text{Var}(I_i^-)$ affecting the performance in SSL.

2.2 IDENTIFYING HIGH-CONFIDENT SAMPLES FOR ASSIGNING LABELS

In Theorem 1, we show that the data sample noise and uncertainty can worsen the estimated classifier. Therefore, we propose to identify and ignore the high-uncertain samples from being assigned labels.

It is notorious that machine learning predictions can vary with different choices of hyperparameters. Under these variations, it is unclear to identify the most probable class to assign labels for some data points. Not only the predictive values but also the ranking order of these predictive samples vary with different hyperparameters. The variations in the prediction can be explained due to (i) the uncertainties of the prediction coming from the noise observations (aleatoric uncertainty) and model knowledge (epistemic uncertainty) (Hüllermeier & Waegeman, 2019) and (ii) the gap between the highest and second-highest scores is small. These bring a challenging fact that the highest score class can be changed with a different set of hyperparameters. This leads to the confusion for assigning pseudo labels because (i) the best set of hyperparameter is unknown given limited labeled data and (ii) we consider ensemble learning setting wherein multiple models are used together.

To address the two challenges above, we are motivated by the theoretical result in Theorem 1 to only assign labels to samples in which the most probable class is statistically significant than the second-most probable class. For this purpose, we propose to use two criteria for measuring the confidence level: (i) Welch’s T-test and (ii) total variance across classes.

Algorithm 1 Confident Sinkhorn Label Allocation

Input: lab data $\{\mathbf{X}_l, y_l\}$, unlab data \mathbf{X}_u , Sinkhorn reg $\varepsilon > 0$, fraction of assigned label $\rho \in [0, 1]$
Output: Allocation matrix $Q \in \mathbb{R}^{N \times K}$

```

1 for  $t = 1, \dots, T$  do
2   Train  $M$  models  $\theta_1, \dots, \theta_M$  given the limited labelled data  $\mathbf{X}_l, y_l$ 
3   Obtain a confidence set  $\mathbf{X}'_U \subset \mathbf{X}_U$  using T-value from Eq. (3) where  $N := |\mathbf{X}'_U|$ 
4   Define a cost matrix  $C$  from Eq. (16) using  $\mathbf{X}'_U$ 
5   Set marginal distributions  $\mathbf{r} = [\mathbf{1}_N^T; N(b_+^T \mathbf{1}_K - \rho b_-^T \mathbf{1}_K)]$  and  $\mathbf{c} = [b_+ N; N(1 - \rho b_-^T \mathbf{1}_K)]$ 
6   /* Sinkhorn's algorithm */
7   while  $j = 1, \dots, J$  or until converged do
8     Initialize  $a^{(j)} = \mathbf{1}_K^T$ ; Update  $b^{(j+1)} = \frac{\mathbf{r}}{\exp(-C_{i,k}/\varepsilon)a^{(j)}}$  and  $a^{(j+1)} = \frac{\mathbf{c}}{\exp(-C_{i,k}/\varepsilon)^T b^{(j+1)}}$ 
9     Obtain  $Q = \text{diag}(a^{(j)}) \exp(-C_{i,k}/\varepsilon) \text{diag}(b^{(j)})$  // the assignment matrix
10    Augment the assigned labels to  $\{\mathbf{X}_l, y_l\}$  and remove these points from the unlabeled data  $\mathbf{X}_u$ .
```

Welch's T-test. For each data point i , we define two empirical distributions (see Appendix Fig. 6) of predicting the highest $\mathcal{N}(\mu_{i,\diamond}, \sigma_{i,\diamond}^2)$ and second-highest class $\mathcal{N}(\mu_{i,\oslash}, \sigma_{i,\oslash}^2)$,¹ estimated from the predictive probability across M classifiers, such as $\mu_{i,\diamond} = \frac{1}{M} \sum_{m=1}^M p_m(y = \diamond | \mathbf{x}_i)$ and $\mu_{i,\oslash} = \frac{1}{M} \sum_{m=1}^M p_m(y = \oslash | \mathbf{x}_i)$ are the empirical means; $\sigma_{i,\diamond}^2$ and $\sigma_{i,\oslash}^2$ are the variances. We consider Welch's T-test (Welch, 1947) to compare the statistical significance of two empirical distributions:

$$\text{T-value}(\mathbf{x}_i) = \frac{\mu_{i,\diamond} - \mu_{i,\oslash}}{\sqrt{(\sigma_{i,\diamond}^2 + \sigma_{i,\oslash}^2)/M}}. \quad (3)$$

The degree of freedom for the above statistical significance is $\frac{(M-1)(\sigma_1^2 + \sigma_2^2)^2}{\sigma_1^4 + \sigma_2^4}$. As a standard practice in statistical testing (Neyman & Pearson, 1933; Fisher, 1955), we calculate the degree of freedom and look at the T-value distribution. In this particular setting, we get the following rules if the T-value is less than 2, the two considered classes are from the same distribution – *eg.*, the sample might fall into the dark area in *right* Fig. 2. Thus, we exclude a sample from assigning labels when its T-value is less than 2.

The estimation of the T-test above encourages separation between the highest and second-highest classes relates to entropy minimization (Grandvalet & Bengio, 2004) which encourages a classifier to output low entropy predictions on unlabeled data.

Total variance. In multilabel classification settings (Kapoor et al., 2012), multiple high score classes can be considered together. Thus, the Welch's T-test between the highest and second-highest is no longer applicable. We consider the following total variance across classes as the second criteria. This total variance for uncertainty estimation has also been used in UPS (Rizve et al., 2021) using Monte Carlo dropout (Gal & Ghahramani, 2016) in the context of deep neural network:

$$\mathcal{V}[p(y | \mathbf{x})] \approx \frac{1}{K} \sum_{k=1}^K \underbrace{\left[\frac{1}{M} \sum_{m=1}^M \left(p_m(y = k | \mathbf{x}) - \sum_{m=1}^M \frac{p_m(y = k | \mathbf{x})}{M} \right)^2 \right]}_{\text{variance of assigning } \mathbf{x} \text{ to class } k}. \quad (4)$$

We exclude the data point with high uncertainty measured by the total variance. It makes sense because a consensus of multiple classifiers is generally a good indicator of the labeling quality. In our setting, a high consensus is represented by low variance or high confidence. While the T-test naturally has a threshold of 2 to reject a data point, the total variance does not have such a threshold and thus we need to impose our own value. In the experiment, we reject 50% of the points with a high total variance score. We visualize the confidence estimated by T-test and total variance in Fig. 2.

We further consider the entropy criteria proposed in Malinin et al. (2020) as another way of estimating the confidence score in Appendix B.1.

Discussion. In Appendix Table 4, we empirically compare the performance using different uncertainty choices. We show that the Welch T-test is the best for this pseudo-labeling problem which can

¹ Denote the highest score class by $\diamond := \diamond(i)$ and second-highest score class by $\oslash := \oslash(i)$ for a data point i .

explicitly capture the uncertainty gap in predicting the highest and second-highest classes while total variance and entropy can not.

2.3 OPTIMAL TRANSPORT ASSIGNMENT

We use the confidence scores to filter out the uncertain data points and assign labels to the remaining using Sinkhorn algorithm (Cuturi, 2013). Particularly, we follow the novel view in Tai et al. (2021) to interpret the label assignment process as an optimal transportation problem between examples and classes, wherein the cost of assigning an example to a class is dictated by the predictions of the classifier.

Let us denote $N \leq N_u$ be the number of accepted points, *eg.* by T-test, from the unlabeled set. Let define an assignment matrix $Q \in \mathbb{R}^{N \times K}$ of N unlabeled data points to K classes such that we assign \mathbf{x}_i to a class k if $Q_{i,k} > 0$. We seek an assignment Q that minimizes the total assignment cost $\sum_{ik} Q_{ik} C_{ik}$ where C_{ik} is the cost of assigning an example i to a class k given by the corresponding negative probability as used in Tai et al. (2021), i.e., $C_{ik} := -\log p(y_i = k | \mathbf{x}_i)$ where $0 \leq p(y_i = k | \mathbf{x}_i) \leq 1$

$$\text{minimize}_Q \quad \langle Q, C \rangle \quad (5) \quad \text{minimize}_{Q, \mathbf{u}, \mathbf{v}, \tau} \quad \langle Q, C \rangle \quad (10)$$

$$\text{s.t.} \quad Q_{ik} \geq 0 \quad (6) \quad \text{s.t.} \quad Q_{ik} \geq 0, \mathbf{u} \succeq 0, \mathbf{v} \succeq 0, \tau \geq 0 \quad (11)$$

$$Q \mathbf{1}_K \leq \mathbf{1}_N \quad (7) \quad Q \mathbf{1}_K + \mathbf{u} = \mathbf{1}_N \quad (12)$$

$$Q^T \mathbf{1}_N \leq N \mathbf{w}_+ \quad (8) \quad Q^T \mathbf{1}_N + \mathbf{v} = \mathbf{w}_+ N \quad (13)$$

$$\mathbf{1}_N^T Q \mathbf{1}_K \geq N \rho \mathbf{w}_-^T \mathbf{1}_K \quad (9) \quad \mathbf{u}^T \mathbf{1}_N + \tau = N(1 - \rho \mathbf{w}_-^T \mathbf{1}_K) \quad (14)$$

$$\mathbf{v}^T \mathbf{1}_K + \tau = N(\mathbf{w}_+^T \mathbf{1}_K - \rho \mathbf{w}_-^T \mathbf{1}_K) \quad (15)$$

where $\mathbf{1}_K$ and $\mathbf{1}_N$ are the vectors one in K and N dimensions, respectively; $\rho \in [0, 1]$ is the fraction of assigned label, i.e., $\rho = 1$ is full allocation; $\mathbf{w}_+, \mathbf{w}_- \in \mathbb{R}^K$ are the vectors of upper and lower bound assignment per class which can be estimated empirically from the class label frequency in the training data or from prior knowledge.

Our formulation has been motivated and modified from the original SLA (Tai et al., 2021) to introduce the lower bound \mathbf{w}_- specifying the minimum number of data points to be assigned in each class. We refer to Appendix B.6 for the ablation study on the effect of this lower bound and Appendix A.2 for the derivations on the equivalence between the inequality in Eqs. (6,7,8,9) for linear programming and equality in Eqs. (11,12,13,14,15) for optimal transport.

We define the marginal distributions for row $\mathbf{r} = [\mathbf{1}_N^T; N(\mathbf{w}_+^T \mathbf{1}_K - \rho \mathbf{w}_-^T \mathbf{1}_K)]^T \in \mathbb{R}^{N+1}$ and column $\mathbf{c} = [\mathbf{w}_+ N; N(1 - \rho \mathbf{w}_-^T \mathbf{1}_K)]^T \in \mathbb{R}^{K+1}$. We define the prediction matrix over the unlabeled set, which satisfied the confidence test, $P \in \mathbb{R}^{M \times N \times K}$ such that each element $P_{m,i,k} = p_m(y = k | \mathbf{x}_i)$ and the averaging prediction over M models is $\bar{P} = \frac{1}{M} \sum_{m=1}^M P_{m,*,*} \in \mathbb{R}^{N \times K}$. The cost and the assignment matrices are defined below

$$C := \begin{bmatrix} -\log \bar{P} & \mathbf{0}_N \\ \mathbf{0}_K^T & 0 \end{bmatrix} \in \mathbb{R}^{(N+1) \times (K+1)} \quad (16) \quad \tilde{Q} := \begin{bmatrix} Q & \mathbf{u} \\ \mathbf{v}^T & \tau \end{bmatrix} \in \mathbb{R}^{(N+1) \times (K+1)}. \quad (17)$$

We derive in Appendix A.3 the optimization process to learn Q by initializing $b^{(j)} = \mathbf{1}_K^T$ and iteratively updating for $j = 1 \dots J$ iterations or until convergence:

$$a^{(j+1)} = \frac{\mathbf{r}}{\exp\left(-\frac{C_{i,k}}{\epsilon}\right) b^{(j)}} \quad (18) \quad b^{(j+1)} = \frac{\mathbf{c}}{\exp\left(-\frac{C_{i,k}}{\epsilon}\right)^T a^{(j+1)}}. \quad (19)$$

After estimating \tilde{Q} , we get Q from Eq. (17) and assign labels to unlabeled data, i.e., where $Q_{i,k} > 0$, and repeat the whole pseudo-labeling process for a few iterations as shown in Algorithm 1. The optimization process above can also be estimated using mini-batches (Fattras et al., 2020) if needed.

Discussion. Instead of performing greedy selection using a threshold γ like other PL algorithms, our CSA specifies the frequency of assigned labels including the lower bound \mathbf{w}_- and upper bound \mathbf{w}_+ per class as well as the fraction of data points $\rho \in (0, 1)$ to be assigned. Then, the optimal transport will automatically perform row and column scalings to find the ‘optimal’ assignments such that the selected element, $Q_{i,k} > 0$, is among the highest values in the row (within a data point i -th) and at the

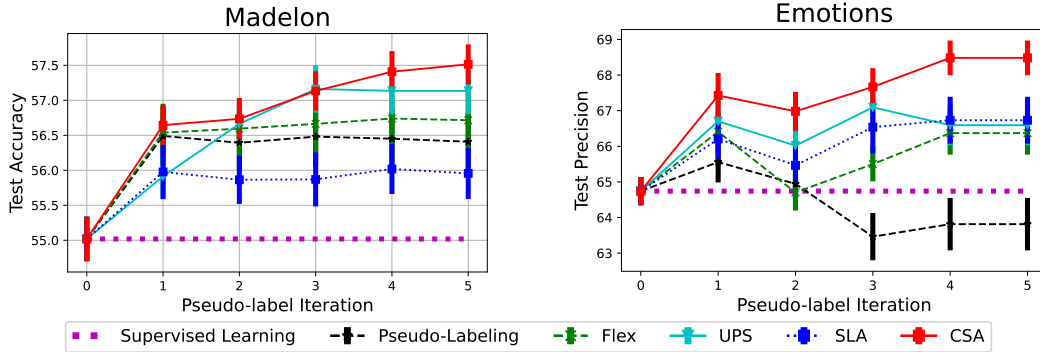


Figure 3: Comparison with the pseudo-labeling methods on tabular data for single-label classification (*left*) and multi-label classification (*right*). x-axis indicates the iteration $t = 1, \dots, T$.

same time receive the highest values in the column (within a class k -th). In contrast, existing PL will either look at the highest values in row or column, separately – but not jointly like ours. Here, the high value refers to high probability P or low-cost C in Eq. (5).

Multi-label. The resulting matrix Q can naturally assign each data point to multiple classes. By that way, we can perform multi-label classification (Read et al., 2011; Nguyen et al., 2016) by assigning a vector of labels k to each data point i s.t. $Q_{i,k} > 0$. This multi-label option is also readily available for other pseudo-labeling methods (Lee et al., 2013). In this multi-label setting, we will use the total variance as the main criteria for confidence estimation as the T-test is not suitable to consider the co-existence of multiple labels.

3 EXPERIMENTS

Baselines. We select to compare our CSA with the following baselines which are suitable for the data setting without data augmentation: Pseudo-labeling (PL) (Lee et al., 2013), FlexMatch (Zhang et al., 2021), UPS (Rizve et al., 2021) and SLA (Tai et al., 2021). We adapt the label assignment mechanisms in Zhang et al. (2021); Rizve et al. (2021) for tabular domains although their original designs are for computer vision tasks. Without explicitly stated, we use a pre-defined threshold of 0.8 for PL, FlexMatch, UPS. We further vary this threshold with different values in Section 3.2. We also compare with the supervised learning method – trained using the labeled data. The multi-layer perceptron (MLP) implemented in Yoon et al. (2020) includes two layers using 100 hidden units. Additional to these pseudo-labeling based approaches, we compare the proposed method with Vime.

Data. We use public datasets for single-label and multi-label classifications from UCI repository (Asuncion & Newman, 2007). Since we do not use any domain assumption, these data are presented as vectorized or tabular data formats. In addition, we obtain three products, renamed as A, B, and C from Amazon catalog. The Amazon data include product numerical and categorical attributes as well as text features of a product review. We summarize all dataset statistics in Appendix Table 7. Our data covers various domains including: image, language, medical, biology, audio, finance and retail.

Classifier and hyperparameters. Given limited labeled data presented in tabular format, we choose XGBoost (Chen & Guestrin, 2016) as the main classifier which typically outperforms state-of-the-art deep learning approaches in this setting (Shwartz-Ziv & Armon, 2022) although our method is more general and not restricted to XGBoost. We use $M = 20$ XGBoost models for ensembling. We refer to Appendix B.3 on the empirical analysis with different choices of M . The ranges for these hyperparameters are summarized in Appendix Table 6.

Setting. Given the setting with limited labels data, we do not use a validation set for tuning hyperparameters of the XGBoost model. We repeat the experiments 30 times with different random seeds, then report the mean and standard error. All Python implementations will be released in the final version.

3.1 COMPARISON OF PSEUDO-LABELING FOR SINGLE AND MULTI-LABEL CLASSIFICATION

We use accuracy as the main metric for evaluating single-label and precision for multi-label classification (Wu & Zhou, 2017). We first compare our CSA with the methods in pseudo-labeling family.

Table 2: Comparison with pseudo-labeling methods.

Datasets	Supervised Learning		Vime	PL	Flex	UPS	SLA	CSA
	XGBoost	MLP						
Segment	95.42 ± 1	94.63 ± 1	92.71 ± 1	95.68 ± 1	95.68 ± 1	95.67 ± 1	95.80 ± 1	95.90 ± 1
Wdbc	89.20 ± 3	91.33 ± 2	91.83 ± 5	91.23 ± 3	91.23 ± 3	91.62 ± 3	90.61 ± 2	91.83 ± 3
Analcatdata	91.63 ± 2	96.17 ± 1	96.13 ± 2	90.95 ± 2	90.62 ± 3	91.33 ± 3	90.98 ± 2	96.60 ± 2
German-credit	70.73 ± 2	71.10 ± 3	70.50 ± 4	70.72 ± 3	70.72 ± 3	71.15 ± 2	70.72 ± 3	71.47 ± 3
Madelon	54.80 ± 3	50.80 ± 2	52.97 ± 2	56.45 ± 4	56.74 ± 4	57.13 ± 3	56.53 ± 4	57.51 ± 3
Dna	88.53 ± 1	76.50 ± 2	79.07 ± 3	88.17 ± 1	88.17 ± 1	88.51 ± 1	88.09 ± 2	89.24 ± 1
Agar Lepiota	57.63 ± 1	63.80 ± 1	63.83 ± 1	58.98 ± 1	59.53 ± 1	58.88 ± 1	58.96 ± 1	59.53 ± 1
Breast cancer	93.20 ± 2	86.40 ± 6	85.87 ± 5	92.89 ± 2	92.89 ± 2	93.38 ± 2	92.76 ± 2	93.55 ± 2
Digits	82.23 ± 3	86.80 ± 1	84.10 ± 1	81.67 ± 3	81.44 ± 3	83.78 ± 3	81.51 ± 3	88.10 ± 2
Category A	72.71 ± 3	57.15 ± 9	50.55 ± 6	72.48 ± 3	72.35 ± 3	72.71 ± 3	72.24 ± 3	73.97 ± 4
Category B	79.48 ± 3	42.85 ± 2	36.45 ± 4	79.41 ± 2	80.49 ± 1	79.45 ± 2	79.82 ± 2	80.62 ± 2
Category C	61.35 ± 1	51.90 ± 1	52.80 ± 1	61.23 ± 1	61.98 ± 1	61.65 ± 1	62.04 ± 1	63.01 ± 1

We show the performance achieved on the held-out test set as we iteratively assign pseudo-label to the unlabeled samples, i.e., varying $t = 1, \dots, T$. As presented in Fig. 3 and Appendix Fig. 11, CSA significantly outperforms the baselines by a wide margin in the datasets of Analcatdata, Synthetic Control and Digits. CSA improves approximately 6% compared with fully supervised learning. CSA gains 2% in Madelon dataset and in the range of $[0.4\%, 0.8\%]$ on other datasets.

Unlike other pseudo-labeling methods (PL, FlexMatch, UPS), CSA can get rid of the requirement to predefine the suitable threshold $\gamma \in [0, 1]$, which is unknown in advance and should vary with different datasets. In addition, the selection in CSA is non-greedy that considers the relative importance within and across rows and columns by the Sinkhorn’s algorithm while the existing PL methods can be greedy when only compared against a predefined threshold.

Furthermore, our CSA outperforms SLA (Tai et al., 2021), which does not take into consideration the confidence score. Thus, SLA may assign labels to ‘uncertain’ samples which can degrade the performance. From the comparison with SLA, we have shown that uncertainty estimation provides us a tool to identify susceptible points to ignore from the data and thus can be beneficial for SSL.

We then compare our CSA with Vime (Yoon et al., 2020), the current state-of-the-art methods in semi-supervised learning for tabular data in Table 2. CSA performs better than Vime in most cases, except the Agaricus Lepiota dataset which has an adequately large size (6500 samples, see Table 7) – thus the deep network component in Vime can successfully learn and perform well.

Multi-label classification. The advantage of CSA is further demonstrated to assign labels in multi-label classification settings. We present the comparison in the bottom row of Fig. 3 where we show that CSA performs competitively and better than the baselines on the three datasets considered.

3.2 COMPARISON OF CSA AGAINST PSEUDO-LABELING WITH DIFFERENT THRESHOLDS

One of the key advantages of using optimal transport is that it can learn the non-greedy assignment by performing row-scaling and column-scaling to get the best assignment such that the cost C is minimized, or the likelihood is maximized with respect to the given constraints from the row \mathbf{r} and column \mathbf{c} marginal distributions. By doing this, our CSA can get rid of the necessity to define the threshold γ and perform greedy selection in most of the existing pseudo-labeling methods.

We may be wondering if the optimal transport assignment Q can be achieved by simply varying the thresholds γ in pseudo-labeling methods? The answer is no. To back up our claim, we perform two analyses. First, we visualize and compare the assignment matrix achieved by CSA and by varying a threshold $\gamma \in \{0.5, 0.7, 0.9\}$ in PL. As shown in *left* Fig. 4, the outputs are different and there are some assignments which could not be selected by PL with varying γ , *eg.*, the following $\{\text{row index, column index}\}$: $\{1, 3\}$, $\{9, 3\}$ and $\{14, 3\}$ annotated in red square in *left* Fig. 4.

Second, we empirically compare the performance of CSA against varying $\gamma \in \{0.7, 0.8, 0.9, 0.95\}$ in PL. We present the results in *right* Fig. 4, which again validates our claim that the optimal assignment in CSA will lead to better performance consistently than PL with changing values of γ .

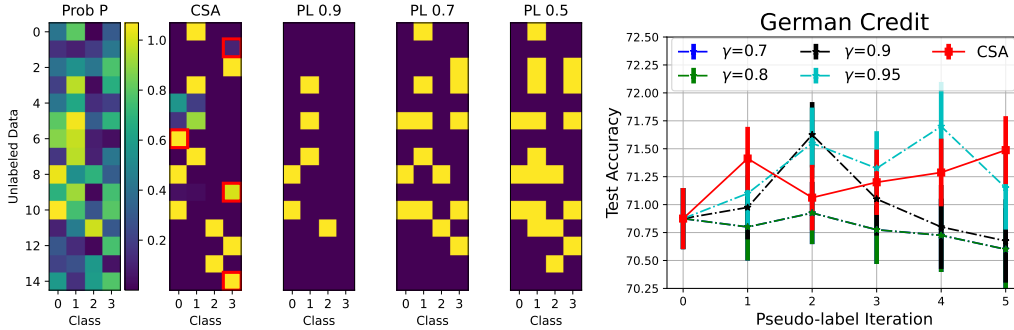


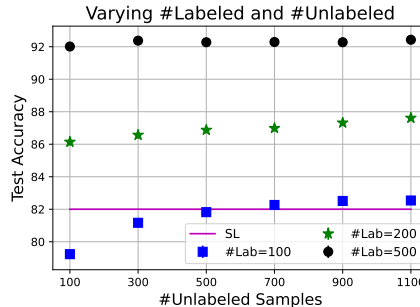
Figure 4: Comparison between CSA versus pseudo-labeling with different thresholds γ . *Left*: the assignments by CSA can not be simply achieved by varying the threshold in PL, *eg.* see the locations highlighted in red square. *Right*: Comparison when varying thresholds in PL against CSA.

3.3 ABLATION STUDIES

Varying the number of labels and unlabeled. Different approaches exhibit substantially different levels of sensitivity to the amount of labeled and unlabeled data, as mentioned in [Oliver et al. \(2018\)](#). We, therefore, validate our claim in Theorem 1 by empirically demonstrating the model performance with increasing the number of unlabeled examples. In *left* Fig. 5, we show that the model will be beneficial with increasing the number of unlabeled data points.

Limitation. In some situations, such as when labeled data is too small in size or contains outliers, our CSA will likely assign incorrect labels to the unlabeled points at the earlier iteration. This erroneous will be accumulated further and lead to poor performance, as shown in *left* Fig. 5 when the number of labels is 100 and the number of unlabeled is less than 500. We note, however, that other pseudo-labeling methods will share the same limitation.

Further analysis. We refer to Appendix B for other empirical analysis which can be sensitive to the performance, including (i) different choices of pseudo-labeling iterations T and (ii) different number of choices for XGB models M , (iii) computational time for each component and (iv) statistics on the number of points selected by each component per iteration.



4 CONCLUSION AND FUTURE WORK

Although there has been significant progress in SSL for images and text domains, limited attention has been invested for unstructured domains in which data augmentation or pretraining is not available, such as data presenting in tabular format. We propose CSA, a new method for pseudo-labeling particularly suitable for tabular domains but not restricted to.

Our CSA has two key ingredients. First, it estimates and takes into account the confidence levels in assigning labels. Second, it learns the optimal assignment using Sinkhorn algorithm which appropriately scales rows and columns probability to get assignment without the greedy selection using a threshold as used in most of PL methods.

The proposed CSA maintains the benefits of PL in its simplicity, generality, and ease of implementation while CSA can significantly improve PL performance by addressing the overconfidence issue and better assignment with optimal transport.

Future work can extend the ensembling process by not only using XGBoost, but also other classifiers, such as AdaBoost ([Wyner et al., 2017](#)) or CatBoost ([Prokhorenkova et al., 2018](#)). Another extension is to use CSA as the main label assignment method for SSL and integrate it into training deep learning model to build CSAMatch, analogous to FixMatch ([Sohn et al., 2020](#)), MixMatch ([Berthelot et al., 2019](#)) and FlexMatch ([Zhang et al., 2021](#)).

REFERENCES

- Dosovitskiy Alexey, Philipp Fischer, Jost Tobias, Martin Riedmiller Springenberg, and Thomas Brox. Discriminative, unsupervised feature learning with exemplar convolutional, neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1734–1747, 2016. 1
- Arthur Asuncion and David Newman. Uci machine learning repository, 2007. 7
- David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 9
- Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017. 1
- Alexandre Boutet, Radhika Madhavan, Gavin JB Elias, Suresh E Joel, Robert Gramer, Manish Ranjan, Vijayashankar Paramanandam, David Xu, Jurgen Germann, Aaron Loh, et al. Predicting optimal deep brain stimulation parameters for parkinson’s disease using functional mri and machine learning. *Nature communications*, 12(1):1–13, 2021. 1
- Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019. 1
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SigKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. 7
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 1
- Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013. 2, 6, 15, 16
- Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 3, 13
- Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn’s algorithm. In *International conference on machine learning*, pp. 1367–1376. PMLR, 2018. 15
- Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021. 1
- Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *AISTATS 2020-23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1–20, 2020. 6
- Ronald Fisher. Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(1):69–78, 1955. 5
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016. 5
- Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004. 5
- Zhuo Huang, Chao Xue, Bo Han, Jian Yang, and Chen Gong. Universal semi-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021. 1
- Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. 2019. 4, 13

- Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *Advances in Neural Information Processing Systems*, 34:3584–3595, 2021. 1
- Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. Multilabel classification using bayesian compressed sensing. In *Advances in Neural Information Processing Systems*, pp. 2645–2653, 2012. 5
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. 1
- Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. Retrieve: Coreset selection for efficient and robust semi-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021. 1
- Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 1
- Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013. 1, 2, 7
- Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 312–321, 2019. 1
- Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015. 1
- Andrey Malinin. *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, University of Cambridge, 2019. 18
- Andrey Malinin, Liudmila Prokhorenkova, and Aleksei Ustimenko. Uncertainty in gradient boosting via ensembles. In *International Conference on Learning Representations*, 2020. 5, 18
- Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. 5
- V Nguyen, SB Orbell, Dominic T Lennon, Hyungil Moon, Florian Vigneau, Leon C Camenzind, Liuqi Yu, Dominik M Zumbühl, G Andrew D Briggs, Michael A Osborne, et al. Deep reinforcement learning for efficient measurement of quantum devices. *npj Quantum Information*, 7(1):1–9, 2021. 1
- Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. A Bayesian nonparametric approach for multi-label classification. In *Proceedings of The 8th Asian Conference on Machine Learning (ACML)*, pp. 254–269, 2016. 7
- Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018. 9
- Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1369–1378, 2021. 1
- Samet Oymak and Talha Cihad Gulcu. A theoretical characterization of semi-supervised learning with self-training for gaussian mixture models. In *International Conference on Artificial Intelligence and Statistics*, pp. 3601–3609. PMLR, 2021. 1
- Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018. 9

- Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011. 7
- Zhongzheng Ren, Raymond Yeh, and Alexander Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. *Advances in Neural Information Processing Systems*, 33:21786–21797, 2020. 1
- 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*, 2021. 2, 5, 7, 18
- Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022. 7
- 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. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 9
- Kai Sheng Tai, Peter D Bailis, and Gregory Valiant. Sinkhorn label allocation: Semi-supervised classification via annealed self-training. In *International Conference on Machine Learning*, pp. 10065–10075. PMLR, 2021. 2, 6, 7, 8, 15, 21, 22
- Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Židek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021. 1
- Nienke ER van Bueren, Thomas L Reed, Vu Nguyen, James G Sheffield, Sanne HG van der Ven, Michael A Osborne, Evelyn H Kroesbergen, and Roi Cohen Kadosh. Personalized brain stimulation for effective neurointervention across participants. *PLoS computational biology*, 17(9):e1008886, 2021. 1
- Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pp. arXiv–1807, 2018. 1
- Nina M van Esbroeck, Dominic T Lennon, Hyungil Moon, Vu Nguyen, Florian Vigneau, Leon C Camenzind, Liuqi Yu, Dominik M Zumbühl, G Andrew D Briggs, Dino Sejdinovic, et al. Quantum device fine-tuning using unsupervised embedding learning. *New Journal of Physics*, 22(9):095003, 2020. 1
- Bernard L Welch. The generalization of student problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947. 5
- Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. In *International Conference on Machine Learning*, pp. 3780–3788. PMLR, 2017. 7
- Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017. 9
- Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *NeurIPS*, 2020. 3, 4, 13, 14
- Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020. 2, 7, 8
- Sora Yoon, Aditi Chandra, and Golnaz Vahedi. Stripenn detects architectural stripes from chromatin conformation data using computer vision. *Nature Communications*, 13(1):1–14, 2022. 1
- Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 7, 9
- Zhi-Hua Zhou. Semi-supervised learning. In *Machine Learning*, pp. 315–341. Springer, 2021. 1
- Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005. 1

Table 3: Notations. We use column vectors in all notations.

Variable	Domain	Meaning
K	\mathcal{N}	number of classes
M	\mathcal{N}	number of XGBoost models for ensembling
d	\mathcal{N}	feature dimension
\mathbf{x}_i	\mathbb{R}^d	a data sample and a collection of them $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$
y_i	$\{1, 2, \dots, K\}$	a data point label and a collection of labels $Y = [y_1, \dots, y_N] \in \mathbb{R}^N$
$\mathcal{D}_l, \mathcal{D}_u$	sets	a collection of labeled $\{\mathbf{x}_i, y_i\}_{i=1}^{N_l}$ or unlabeled data $\{\mathbf{x}_i\}_{i=N_l+1}^{N_u+N_l}$
$\mathbf{1}_K, \mathbf{1}_N$	$[1, \dots, 1]^T$	a vector of ones in K (or N) dimensions
ρ	$[0, 1]$	allocation fraction, $\rho = 0$ is no allocation or $\rho = 1$ is full allocation
ε	\mathbb{R}^+	Sinkhorn regularization parameter
Q	$\mathbb{R}^{N \times K}$	assignment matrix, i.e., $Q_{i,k} > 0$ assigns a label k to a data point i
C	$\mathbb{R}^{N \times K}$	a cost matrix estimated from the probability matrix i.e., $C = 1 - P$
$\mathbf{w}, \mathbf{w}_+, \mathbf{w}_-$	vector $[0, 1]^K$	label frequency, upper bound and lower bound of label frequency
γ	$[0, 1]$	a threshold for assigning labels used in pseudo-labeling methods

A APPENDIX: THEORETICAL RESULTS

We derive the proof for Theorem 1 in Sec. A.1, interpret the linear programming inequality to optimal transport in Sec. A.2, present the derivations for the Sinkhorn algorithm used in our method in Sec. A.3 and consider another criterion for confident estimation in Sec. B.1.

A.1 PROOF OF THEOREM 1

Setting. To highlight the role of variances and data noises toward SSL, we consider a simple binary classification problem on one-dimensional space. We generate the feature $x | y = +1 \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_+, \sigma^2)$ and similarly $x | y = -1 \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_-, \sigma^2)$ where μ_+ and μ_- are the means of the positive and negative classes respectively, σ is the standard deviation indicating the level of data noisiness, and $\mathcal{N}(\cdot, \cdot)$ the suitably parameterized normal distribution. Without loss of generality, let $\mu_+ > \mu_-$. The optimal Bayes’s classifier is $f(x) = \text{sign}(x - \frac{\mu_+ + \mu_-}{2})$, i.e. classify $y_i = +1$ if $x_i > (\mu_+ + \mu_-)/2$. Therefore, in the following, we measure the probability bound to learn $(\mu_+ + \mu_-)/2$ as a criterion for achieving good performance.

Let $\{\tilde{X}_i^+\}_{i=1}^{\tilde{n}_+}$ and $\{\tilde{X}_i^-\}_{i=1}^{\tilde{n}_-}$ be the set of unlabeled data whose pseudo-label is $+1$ and -1 , respectively. Let $\{I_i^+\}_{i=1}^{\tilde{n}_+}$ be the binary indicator of correct assignments, such as if $I_i^+ = 1$, then $\tilde{X}_i^+ \sim \mathcal{N}(\mu_+, \sigma^2)$ and otherwise $\tilde{X}_i^+ \sim \mathcal{N}(\mu_-, \sigma^2)$. Similarly, we define $\{I_i^-\}_{i=1}^{\tilde{n}_-}$ indicating the correct assignment for negative class.

Instead of generating from Bernoulli distribution as used in Yang & Xu (2020), we assume the binary indicators I_i^+ and I_i^- to be generated from a classifier. We denote the expectations $\mathbb{E}(I_i^+), \mathbb{E}(I_i^-)$ and variances of $\text{Var}(I_i^+), \text{Var}(I_i^-)$ for the generating processes of I_i^+ and I_i^- , respectively. We denote the gap in the expectations as $\Delta = |\mathbb{E}(I_i^+) - \mathbb{E}(I_i^-)|$ which can be used to characterize the imbalance property of the data, as shown in Yang & Xu (2020).

In uncertainty estimation literature (Der Kiureghian & Ditlevsen, 2009; Hüllermeier & Waegeman, 2019), σ^2 is called the *aleatoric uncertainty* (from the observations) and $\text{Var}(I_i^+), \text{Var}(I_i^-)$ are the *epistemic uncertainty* (from the model knowledge).

As mentioned, we aim to learn $\frac{\mu_+ + \mu_-}{2}$ with the above setup via the extra unlabeled data. It is natural to construct our estimate as $\hat{\theta} = \frac{1}{2} \left(\sum_{i=1}^{\tilde{n}_+} \tilde{X}_i^+ / \tilde{n}_+ + \sum_{i=1}^{\tilde{n}_-} \tilde{X}_i^- / \tilde{n}_- \right)$.

Theorem 2. For $\delta > 0$, with a probability at least $1 - 2\exp\left\{-\frac{2\delta^2[\tilde{n}_+ + \tilde{n}_-]}{9\sigma^2[\tilde{n}_+ + \tilde{n}_-]}\right\} - \frac{9[\text{Var}(I_i^+) + \text{Var}(I_i^-)]}{4\delta^2}$, our estimate $\hat{\theta}$ for the above binary classification satisfies

$$\left| \hat{\theta} - \frac{\Delta(\mu_+ - \mu_-)}{2} - \frac{(\mu_+ + \mu_-)}{2} \right| \leq \delta \quad (20)$$

where $\Delta = |\mathbb{E}(I_i^+) - \mathbb{E}(I_i^-)|$.

Proof. Our proof extends the result from Yang & Xu (2020) taking into account the uncertainty for pseudo-labeling. We rewrite $\tilde{X}_i^+ = (1 - I_i^+)(\mu_- - \mu_+) + Z_i^+$ where $Z_i^+ \sim \mathcal{N}(\mu_+, \sigma^2)$ and I_i^+ is generated from a classifier with the expectation $\mathbb{E}(I_i^+)$ and variance $\text{Var}(I_i^+)$. This means if the pseudo-label is correct, $\tilde{X}_i^+ \sim Z_i^+$ and otherwise $\tilde{X}_i^+ \sim (\mu_- - \mu_+) + Z_i^+$. Similarly, $\tilde{X}_i^- = (1 - I_i^-)(\mu_+ - \mu_-) + Z_i^-$ where $Z_i^- \sim \mathcal{N}(\mu_-, \sigma^2)$

$$\hat{\theta} = \frac{1}{2} \left(\frac{\sum_{i=1}^{\tilde{n}_+} \tilde{X}_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} \tilde{X}_i^-}{\tilde{n}_-} \right) \quad (21)$$

$$= \frac{1}{2} \left(\frac{\sum_{i=1}^{\tilde{n}_+} (1 - I_i^+)(\mu_- - \mu_+) + Z_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} (1 - I_i^-)(\mu_+ - \mu_-) + Z_i^-}{\tilde{n}_-} \right) \quad (22)$$

$$= \frac{1}{2} \left(\frac{\sum_{i=1}^{\tilde{n}_+} I_i^+}{\tilde{n}_+} (\mu_+ - \mu_-) + \frac{\sum_{i=1}^{\tilde{n}_-} I_i^-}{\tilde{n}_-} (\mu_- - \mu_+) + \frac{\sum_{i=1}^{\tilde{n}_+} Z_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} Z_i^-}{\tilde{n}_-} \right). \quad (23)$$

We bound each term separately. First, we bound $\frac{\sum_{i=1}^{\tilde{n}_+} I_i^+}{\tilde{n}_+}$ using Chebyshev's inequality as follows

$$P \left(\left| \frac{\sum_{i=1}^{\tilde{n}_+} I_i^+}{\tilde{n}_+} - \mathbb{E}(I_i^+) \right| > t \right) \leq \frac{\text{Var}(I_i^+)}{t^2}. \quad (24)$$

Similarly, we obtain the probability bound for the negative class

$$P \left(\left| \frac{\sum_{i=1}^{\tilde{n}_-} I_i^-}{\tilde{n}_-} - \mathbb{E}(I_i^-) \right| > t \right) \leq \frac{\text{Var}(I_i^-)}{t^2}. \quad (25)$$

Next, we bound the term

$$\left(\frac{\sum_{i=1}^{\tilde{n}_+} Z_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} Z_i^-}{\tilde{n}_-} \right) \sim \mathcal{N} \left(\mu_+ + \mu_-, \sigma^2 \left(\frac{1}{\tilde{n}_+} + \frac{1}{\tilde{n}_-} \right) \right). \quad (26)$$

By using Gaussian concentration, we have that

$$P \left(\left| \frac{\sum_{i=1}^{\tilde{n}_+} Z_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} Z_i^-}{\tilde{n}_-} - (\mu_+ + \mu_-) \right| > t \right) \leq 2\exp \left\{ -\frac{t^2}{2\sigma^2[1/\tilde{n}_+ + 1/\tilde{n}_-]} \right\}. \quad (27)$$

By triangle inequality,

$$\left| \hat{\theta} - \frac{\Delta(\mu_+ - \mu_-)}{2} - \frac{(\mu_+ + \mu_-)}{2} \right| \quad (28)$$

$$= \left| \frac{1}{2} \left(\frac{\sum_{i=1}^{\tilde{n}_+} \tilde{X}_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} \tilde{X}_i^-}{\tilde{n}_-} \right) - \frac{\mathbb{E}(I_i^+)(\mu_+ - \mu_-)}{2} - \frac{\mathbb{E}(I_i^-)(\mu_- - \mu_+)}{2} - \frac{\mu_+ + \mu_-}{2} \right| \quad (29)$$

$$\leq \left| \frac{1}{2} \frac{\sum_{i=1}^{\tilde{n}_+} I_i^+}{\tilde{n}_+} - \frac{\mathbb{E}(I_i^+)}{2} \right| |\mu_+ - \mu_-| + \left| \frac{1}{2} \frac{\sum_{i=1}^{\tilde{n}_-} I_i^-}{\tilde{n}_-} - \frac{\mathbb{E}(I_i^-)}{2} \right| |\mu_+ - \mu_-| \quad (30)$$

$$+ \left| \frac{1}{2} \left(\frac{\sum_{i=1}^{\tilde{n}_+} Z_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} Z_i^-}{\tilde{n}_-} \right) - \frac{\mu_+ + \mu_-}{2} \right|. \quad (31)$$

Let $\delta > 0$, we consider an event

$$E = \left\{ \left| \frac{\sum_{i=1}^{\tilde{n}_+} I_i^+}{\tilde{n}_+} - \mathbb{E}(I_i^+) \right| |\mu_+ - \mu_-| \leq \frac{2\delta}{3} \quad \text{and} \quad \left| \frac{\sum_{i=1}^{\tilde{n}_-} I_i^-}{\tilde{n}_-} - \mathbb{E}(I_i^-) \right| |\mu_+ - \mu_-| \leq \frac{2\delta}{3} \quad \text{and} \right. \quad (32)$$

$$\left. \left| \left(\frac{\sum_{i=1}^{\tilde{n}_+} Z_i^+}{\tilde{n}_+} + \frac{\sum_{i=1}^{\tilde{n}_-} Z_i^-}{\tilde{n}_-} \right) - (\mu_+ + \mu_-) \right| \leq \frac{2\delta}{3} \right\}. \quad (33)$$

Using union bound and the concentration inequalities

$$P(E) \geq 1 - 2 \exp \left\{ -\frac{2\delta^2}{9\sigma^2[1/\tilde{n}_+ + 1/\tilde{n}_-]} \right\} - \frac{9\text{Var}(I_i^+)}{4\delta^2} - \frac{9\text{Var}(I_i^-)}{4\delta^2}. \quad (34)$$

Equivalently, we write

$$P \left(\left| \hat{\theta} - \frac{\Delta(\mu_+ - \mu_-)}{2} - \frac{(\mu_+ + \mu_-)}{2} \right| \leq \delta \right) \geq 1 - 2 \exp \left\{ -\frac{2\delta^2}{9\sigma^2[\frac{1}{\tilde{n}_+} + \frac{1}{\tilde{n}_-}]} \right\} - \frac{9[\text{Var}(I_i^+) + \text{Var}(I_i^-)]}{4\delta^2} \quad (35)$$

where $\Delta = |\mathbb{E}(I_i^+) - \mathbb{E}(I_i^-)|$. \square

A.2 DERIVATION OF THE OPTIMAL TRANSPORT LINEAR PROGRAMING

We transform the original linear programming inequality in Eqs. (5,6,7,8) to optimal transport setup. As used in [Tai et al. \(2021\)](#), we introduce non-negative slack variables u , v and τ to replace the inequality constraints on the marginal distributions. We yield the following optimization problem:

$$\text{minimize}_{Q, \mathbf{u}, \mathbf{v}, \tau} \quad \langle Q, C \rangle \quad (36)$$

$$\text{s.t.} \quad Q_{ik} \geq 0, \mathbf{u} \succeq 0, \mathbf{v} \succeq 0, \tau \geq 0 \quad (37)$$

$$Q\mathbf{1}_K + \mathbf{u} = \mathbf{1}_N \quad (38)$$

$$Q^T\mathbf{1}_N + \mathbf{v} = \mathbf{w}_+ N \quad (39)$$

$$\mathbf{1}_N^T Q \mathbf{1}_K = \tau + N\rho \mathbf{w}_-^T \mathbf{1}_K \quad (40)$$

We substitute Eq. (40) into Eq. (38) to have:

$$\mathbf{u}^T \mathbf{1}_N + \tau = N(1 - \rho \mathbf{w}_-^T \mathbf{1}_K) \quad (41)$$

Similarly, we substitute Eq. (40) into Eq. (39) to obtain:

$$\mathbf{v}^T \mathbf{1}_K + \tau = N(\mathbf{w}_+^T \mathbf{1}_K - \rho \mathbf{w}_-^T \mathbf{1}_K) \quad (42)$$

Thus, we yield Eqs. (11,12,13,14,15) and recognize them as an optimal transport problem ([Cuturi, 2013](#); [Tai et al., 2021](#)) with row marginal distribution $\mathbf{r} = [\mathbf{1}_N^T; N(\mathbf{w}_+^T \mathbf{1}_K - \rho \mathbf{w}_-^T \mathbf{1}_K)]^T \in \mathbb{R}^{N+1}$ and column marginal distribution $\mathbf{c} = [\mathbf{w}_+ N; N(1 - \rho \mathbf{w}_-^T \mathbf{1}_K)]^T \in \mathbb{R}^{K+1}$.

A.3 SINKHORN OPTIMIZATION STEP

Given the cost matrix $C \in \mathbb{R}^{N \times K}$, our original objective function is

$$\text{minimize}_{Q \in \mathbb{R}^{N \times K}} \langle Q, C \rangle \quad (43)$$

where $\langle Q, C \rangle := \sum_{ik} Q_{ik} C_{ik}$. For a general matrix C , the worst case complexity of computing that optimum scales in $\mathcal{O}(N^3 \log N)$ ([Cuturi, 2013](#)). To overcome this expensive computation, we can utilize the entropic regularization to reduce the complexity to $\mathcal{O}\left(\frac{N^2}{\epsilon^2}\right)$ ([Dvurechensky et al., 2018](#)) by writing the objective function as:

$$\mathcal{L}(Q) = \text{minimize}_{Q \in \mathbb{R}^{N \times K}} \langle Q, C \rangle - \epsilon \mathcal{H}(Q) \quad (44)$$

where $\mathbb{H}(Q)$ is the entropy of Q . This is a strictly convex optimization problem (Cuturi, 2013) and has a unique optimal solution.

Since the solution Q^* of the above problem is surely nonnegative, i.e., $Q_{i,k}^* > 0$, we can thus ignore the positivity constraints when introducing dual variables $\lambda^{(1)} \in \mathbb{R}^N, \lambda^{(2)} \in \mathbb{R}^K$ for each marginal constraint as follows:

$$\mathcal{L}(Q, \lambda^{(1)}, \lambda^{(2)}) = \langle Q, C \rangle + \varepsilon \mathbb{H}(Q) + \langle \lambda^{(1)}, \mathbf{r} - Q \mathbf{1}_K \rangle + \langle \lambda^{(2)}, \mathbf{c} - Q^T \mathbf{1}_N \rangle \quad (45)$$

where \mathbf{r} and \mathbf{c} are the row and column marginal distributions. We take partial derivatives w.r.t. each variable to solve the above objective. We have,

$$\frac{\partial \mathcal{L}(Q, \lambda^{(1)}, \lambda^{(2)})}{\partial Q_{i,k}} = C_{i,k} + \varepsilon (\log Q_{i,k} + 1) - \lambda_i^{(1)} - \lambda_k^{(2)} = 0 \quad (46)$$

$$Q_{i,k}^* = \exp \left\{ \frac{\lambda_i^{(1)} + \lambda_k^{(2)} - C_{i,k}}{\varepsilon} + 1 \right\} \quad (47)$$

$$= \exp \left(\frac{\lambda_i^{(1)}}{\varepsilon} - 1/2 \right) \exp \left(-\frac{C_{i,k}}{\varepsilon} \right) \exp \left(\frac{\lambda_k^{(2)}}{\varepsilon} - 1/2 \right). \quad (48)$$

The factorization of the optimal solution in the above equation can be conveniently rewritten in matrix form as $\text{diag}(a) \exp \left(-\frac{C_{i,k}}{\varepsilon} \right) \text{diag}(b)$ where a and b satisfy the following non-linear equations corresponding to the constraints

$$\text{diag}(a) \exp \left(-\frac{C_{i,k}}{\varepsilon} \right) \text{diag}(b) \mathbf{1}_K = \mathbf{r} \quad (49) \quad \text{diag}(b) \exp \left(-\frac{C_{i,k}}{\varepsilon} \right)^T \text{diag}(a) \mathbf{1}_N = \mathbf{c}. \quad (50)$$

Now, instead of optimizing $\lambda^{(1)}$ and $\lambda^{(2)}$, we alternatively solve for the vectors

$$a = \exp \left(\frac{\lambda^{(1)}}{\varepsilon} - 1/2 \right) \quad (51) \quad b = \exp \left(\frac{\lambda^{(2)}}{\varepsilon} - 1/2 \right). \quad (52)$$

The above forms the system of equations with two equations and two unknowns. We solve this system by iteratively updating for multiple iterations j :

- Initialize $b^{(j)} = \mathbf{1}_K^T$
- Update $a^{(j+1)} = \frac{\mathbf{r}}{\exp \left(-\frac{C_{i,k}}{\varepsilon} \right) b^{(j)}}$
- Update $b^{(j+1)} = \frac{\mathbf{c}}{\exp \left(-\frac{C_{i,k}}{\varepsilon} \right)^T a^{(j+1)}}$

The division operator used above between two vectors is to be understood entry-wise. After J iterations, we estimate the final assignment matrix for pseudo-labeling

$$Q^* = \text{diag}(a^{(J)}) \exp \left(-\frac{C_{i,k}}{\varepsilon} \right) \text{diag}(b^{(J)}). \quad (53)$$

A.4 REPRESENTING THE TWO EMPIRICAL DISTRIBUTIONS FOR T-TEST

We interpret the confidence estimation using T-test in Fig. 6. We train XGB as the main classifier using different sets of hyperparameters. Let denote the predictive probability using M classifiers across N unlabeled data points over K classes as $P \in \mathbb{R}^{M \times N \times K}$ in top Fig. 6. Given a data point i , we define the highest score and second-highest score classes as follows: $\diamond = \arg \max_{k=\{1 \dots K\}} \frac{1}{M} \sum_{m=0}^M P_{m,i,k}$ and $\oslash = \arg \max_{k=\{1 \dots K\} \setminus \diamond} \frac{1}{M} \sum_{m=0}^M P_{m,i,k}$ where $\{1 \dots K\} \setminus \diamond$ means that we exclude an index \diamond from a set $\{1 \dots K\}$. Note that these indices of the highest \diamond and second-highest score class \oslash vary with different data points and thus can also be defined as the function $\diamond(i)$ and $\oslash(i)$.

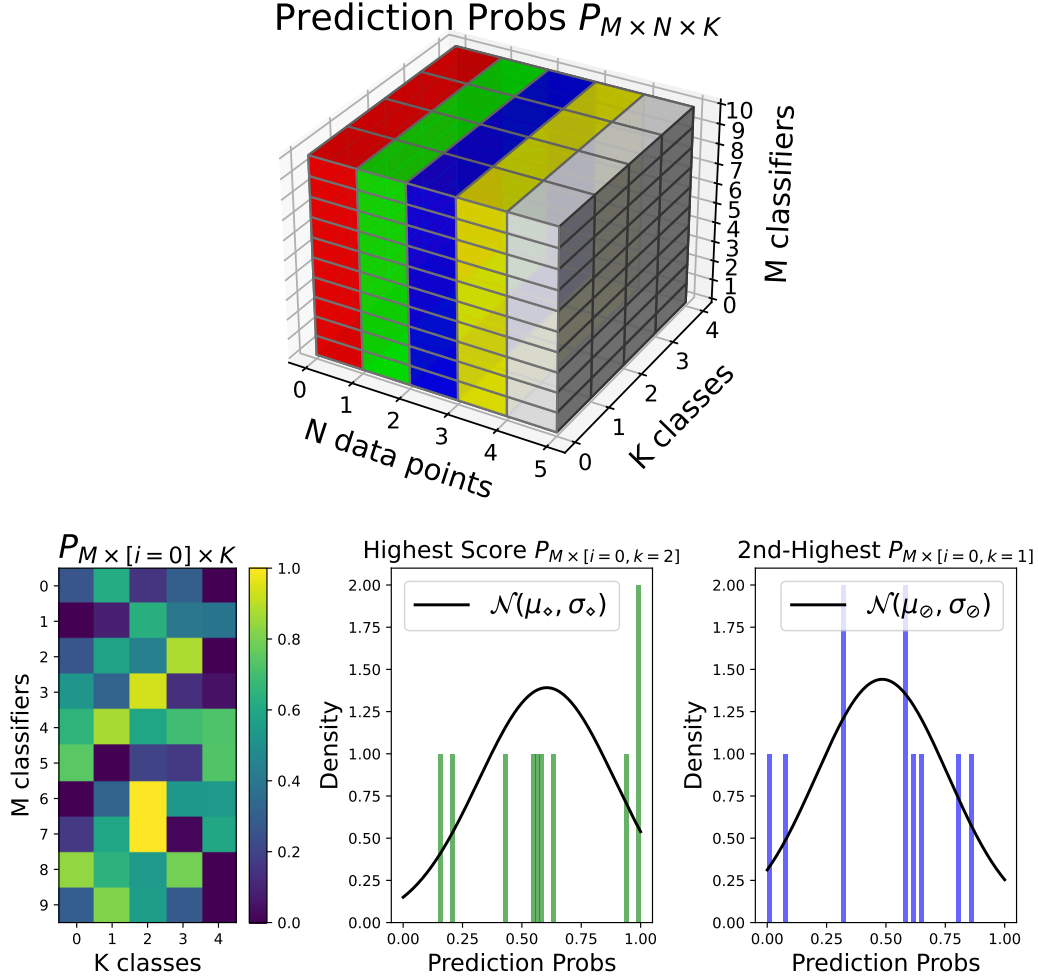


Figure 6: The empirical distributions used for the T-test. We estimate a predictive probability for a data point $i = 0$ using M classifiers across K classes, denoted as $P_{M, i=0, K}$ (bottom left). The highest and second-highest score classes are denoted as \diamond and \circ , respectively, and their empirical distributions $\mathcal{N}(\mu_{i, \diamond}, \sigma_{i, \diamond}^2)$ in bottom middle and $\mathcal{N}(\mu_{i, \circ}, \sigma_{i, \circ}^2)$ in bottom right. See Sec. A.4 for further details.

Table 4: Comparison with different choices for estimating the confidence scores. The performance with total variance reflects the uncertainty choice used in UPS (Rizve et al., 2021). CSA using T-test achieves the best performance. We also highlight that using CSA (either with Total variance of T-test) will consistently outperform SLA.

Datasets	SLA	CSA Entropy	CSA Total variance	CSA T-test
Segment	95.82 \pm 1	95.74 (1.0)	95.70 \pm 1	95.90 \pm 1
Wdbc	90.64 \pm 3	89.50(4.3)	91.62 \pm 3	91.83 \pm 3
Analcatdata	90.43 \pm 3	92.00 \pm 3	92.40 \pm 2	94.79 \pm 2
German-credit	70.79 \pm 3	70.88(3.0)	71.75(2.8)	71.47 \pm 3
Madelon	55.96 \pm 3	57.50 (2.5)	55.56 (3.8)	57.51 \pm 2
Dna	87.86 \pm 2	88.23 (1.6)	87.12(2.2)	89.24 \pm 1
Agaricus Lepiota	59.01 \pm 1	57.67(1.0)	59.59(1.0)	59.53 \pm 1
Breast cancer	92.65 \pm 2	93.33(2.8)	93.25(2.6)	93.55 \pm 2
Digits	81.03 \pm 3	84.65(3.1)	84.37(3.1)	86.31 \pm 3

In bottom left Fig. 6, we consider a predictive probability of a data point $i = 0$, i.e., $P_{M,i=0,K} \in \mathbb{R}^{M \times K}$. We then define two empirical distributions of predicting the highest $\mathcal{N}(\mu_{i,\diamond}, \sigma_{i,\diamond}^2)$ and second-highest class $\mathcal{N}(\mu_{i,\oslash}, \sigma_{i,\oslash}^2)$, such as the empirical means are:

$$\mu_{i,\diamond} = \frac{1}{M} \sum_{m=1}^M p_m(y = \diamond | \mathbf{x}_i) \quad (54)$$

$$\mu_{i,\oslash} = \frac{1}{M} \sum_{m=1}^M p_m(y = \oslash | \mathbf{x}_i) \quad (55)$$

and the variances are defined respectively as

$$\sigma_{i,\diamond}^2 = \frac{1}{M} \sum_{m=1}^M \left(p_m(y = \diamond | \mathbf{x}_i) - \mu_{i,\diamond} \right)^2 \quad (56) \quad \sigma_{i,\oslash}^2 = \frac{1}{M} \sum_{m=1}^M \left(p_m(y = \oslash | \mathbf{x}_i) - \mu_{i,\oslash} \right)^2. \quad (57)$$

We illustrate the estimated distributions $\mathcal{N}(\mu_{i,\diamond}, \sigma_{i,\diamond}^2)$ in bottom middle Fig. 6 and $\mathcal{N}(\mu_{i,\oslash}, \sigma_{i,\oslash}^2)$ in bottom right Fig. 6. We also plot the empirical data samples used to estimate these distributions, i.e., a collection of $[P_{1,i=0,k=\diamond}, P_{2,i=0,k=\diamond}, \dots, P_{M,i=0,k=\diamond}]$ and $[P_{1,i=0,k=\oslash}, P_{2,i=0,k=\oslash}, \dots, P_{M,i=0,k=\oslash}]$.

B APPENDIX: ADDITIONAL ABLATION STUDIES

B.1 EXPERIMENTS WITH DIFFERENT UNCERTAINTY ESTIMATION CHOICES

We consider the *entropy* as another metric for estimating the confidence score for each data point. The entropy across multiple ensembles is as known as the total uncertainty (Malinin, 2019; Malinin et al., 2020), which is the sum of data and knowledge uncertainty. By considering the mutual information, we can estimate the *total uncertainty* as follows:

$$\mathbb{H}[p(y | \mathbf{x}, D)] \sim \mathbb{H} \left[\frac{1}{M} \sum_{m=1}^M \underbrace{p_m(y = 1, \dots, K | \mathbf{x})}_{\in \mathbb{R}^K} \right] \quad (58)$$

where $p_m(y | \mathbf{x}) := p(y | \mathbf{x}, \theta^{(m)})$ for brevity. We compare this total uncertainty against the proposed T-test and the total variance in Table 4. However, this entropy does not perform as well as the T-test and total variance. The possible explanation is that this total uncertainty can not well capture the variation in these class predictions, such as can not explicitly measure the gap between the highest and second-highest classes. Additionally in Table 4, we also show that the proposed T-test outperforms the total variance which is used in the existing UPS (Rizve et al., 2021). In addition, we highlight that using CSA (either with Total variance of T-test) will consistently outperform SLA.

B.2 EXPERIMENTS WITH DIFFERENT NUMBER OF ITERATIONS T FOR PSEUDO-LABELING

Pseudo-labeling is an iterative process by repeatedly augmenting the labeled data with selected samples from the unlabeled set. Under this repetitive mechanism, the error can also be accumulated

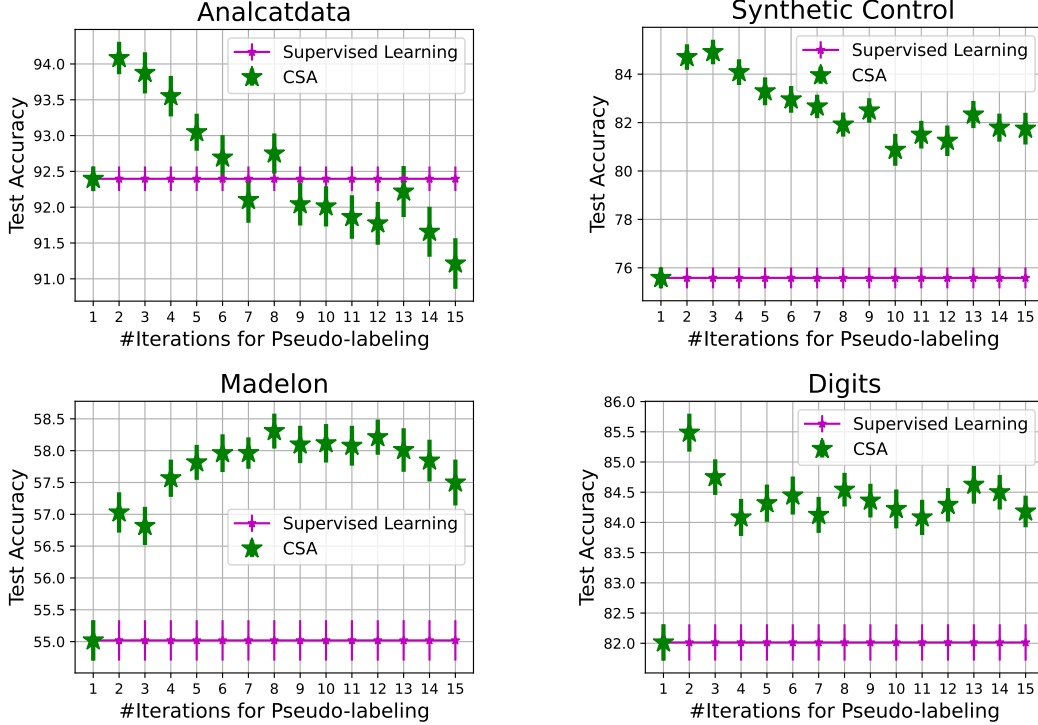


Figure 7: Performance with different choices of the iterations $T \in [1, \dots, 15]$ for pseudo-labeling.

during the PL process if the wrong assignment is made at an earlier stage. Therefore, not only the choice of pseudo-labeling threshold is important, but also the number of iterations T used for the pseudo-labeling process can affect the final performance.

Therefore, we analyze the performance with respect to the choices of the number of iterations $T \in [1, \dots, 15]$ in Fig. 7. We show an interesting fact that the best number of iterations is unknown in advance and varies with the datasets. For example, the best $T = 2$ is for Analcatdata while $T = 3$ is for Synthetic control, $T = 8$ for Madelon and $T = 2$ for Digits. Our CSA is robust with different choices of T for Synthetic control, Madelon and Digits that our method surpasses the Supervised Learning (SL) by a wide margin. Having said that, CSA with $T > 6$ on Analcatdata can degrade the performance below SL. Therefore, we recommend using $T \leq 5$, and in the main experiment, we have set $T = 5$ for all datasets.

B.3 EXPERIMENTS WITH DIFFERENT NUMBER OF XGB MODELS M FOR ENSEMBLING

To estimate the confidence score for each data point, we have ensembled M XGBoost models with different sets of hyperparameters. Since we are interested in knowing how many models M should we use for the experiments. We below perform the experiments with varying the number of models $M = [1, \dots, 30]$ in Fig. 8. We have shown on two datasets that using $M \geq 5$ will be enough to estimate the confidence score to achieve the best performance.

Additionally, we have demonstrated that making use of the confidence score (the case of CSA) will consistently improve the performance against without using it (the case of SLA). This is because we can ignore and do not assign labels to data points which are in low confidence scores, such as high variance or overlapping between the highest and second-highest class. These uncertain samples can be seen at the dark area in Fig. 2.

B.4 COMPUTATIONAL TIME

We analyze the computational time used in our CSA using Madelon dataset. As part of the ensembling task, we need to train M XGBoost models. As shown previously in Appendix B.3, the recommended

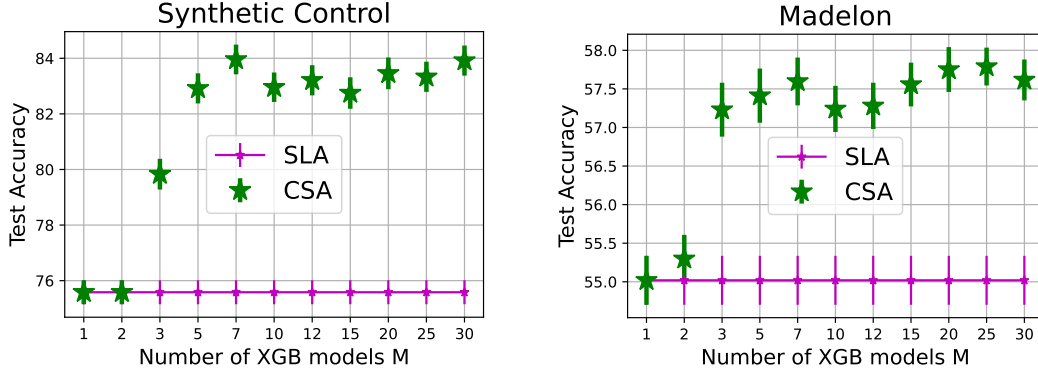


Figure 8: Experiments with different numbers of XGB models for ensembling. The results suggest that using $M \geq 5$ XGBoost models will sufficiently provide a good estimation for the confidence level. Overall, we show that using the confidence score (CSA) will improve the performance than without using it (SLA).

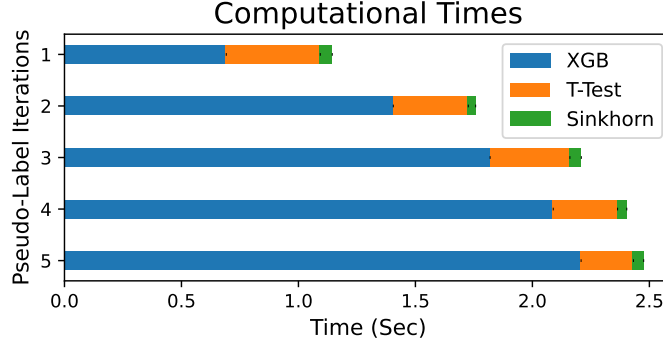


Figure 9: The computational time per each component in CSA for Madelon dataset. We show that the time used for computing both T-test and Sinkhorn is negligible as approximately within a second while the time for XGBoost training takes longer.

number of XGBoost models is $M \geq 5$ and we have used $M = 10$ in all experiments. Note that training M XGBoost models can be taken in parallel to speed up the time.

We present in Fig. 9 a relative comparison of the computational times taken by each component in CSA, including: time for training a single model of XGBoost, time for calculating T-test, and time used by Sinkhorn’s algorithm. We show that the XGB training takes the most time, T-test estimation and Sinkhorn algorithm will consume less time. Especially, Sinkhorn’s algorithm takes an unnoticeable time, such as 0.1 sec for Madelon dataset.

We also observe that the XGBoost computational time used in each iteration will increase with iterations because more data samples are augmented into the labeled set at each pseudo-label iteration. On the other hand, the T-test will take less time with iteration because the number of unlabeled samples reduces over time.

B.5 STATISTICS ON THE NUMBER OF POINTS REJECTED BY T-TEST AND ACCEPTED BY CSA

We plot the analysis on the number of samples accepted by CSA as well as rejected by the confidence score w.r.t. iterations $t = 1, \dots, T$ in Fig. 10. We can see that the time used for training XGBoost model is increasing over iterations because more samples are added into the labeled set which enlarges the training set. Overall, the computational time for XGB and T-test is within a second.

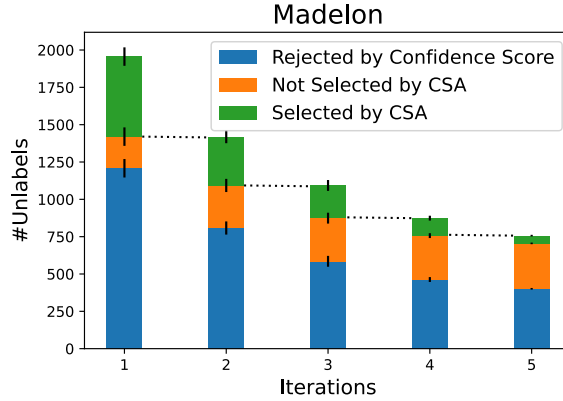


Figure 10: The statistic of the number of samples rejected by the confidence score (t-test), selected by CSA and not selected. The remaining unlabeled data in the next iteration excludes the samples being selected by CSA. Thus, the number of unlabeled samples is reduced across pseudo-label iterations.

Table 5: Comparison on the effect of introducing the lower bound constraint w_- in the optimal transport formulation. We show that adding the w_- into the OT formulation will help the performance slightly and adding the confident score will boost the performance significantly.

Datasets	OT without w_- SLA (Tai et al., 2021)	OT with w_- CSA without confidence	CSA
Segment	95.80 ± 1	95.96 ± 1	95.90 ± 1
Wdbc	90.61 ± 2	91.27 ± 3	91.83 ± 3
AnaCatdata	90.98 ± 2	91.38 ± 3	96.60 ± 2
German-credit	70.72 ± 3	70.93 ± 3	71.47 ± 3
Madelon	56.53 ± 4	56.87 ± 4	57.51 ± 3
Dna	88.09 ± 2	89.14 ± 2	89.24 ± 1
Agar Lepiota	58.96 ± 1	59.12 ± 1	59.53 ± 1
Breast cancer	92.76 ± 2	92.81 ± 2	93.55 ± 2
Digits	81.51 ± 3	81.68 ± 3	88.10 ± 2

Table 6: XGBoost hyperparameters range for ensembling.

Name	Min	Max
Learning rate	0.01	0.3
Max depth	3	20
Subsample	0.5	1
Colsample_bytree	0.4	1
Colsample_bylevel	0.4	1
n_estimators	100	1000

Table 7: Dataset statistics including the number of classes K , number of feature d , number of test samples, number of labeled, unlabeled and the ratio between labeled versus unlabeled (last column).

Datasets	Domains	#Classes K	#Feat d	#Test	Train		Ratio Lab/Unlab
					#Labeled	#Unlabeled	
Single-label classification datasets							
Segment	image	7	19	462	739	1109	/
Wdbc	medical	2	30	114	45	410	/
Analcatdata	economic	4	70	169	67	605	/
German-credit	finance	2	24	200	160	640	/
Madelon	artificial	2	500	520	124	1956	/
Dna	biology	3	180	152	638	2,396	/
Agaricus lepiota	biology	7	22	1625	3,249	3,249	/
Breast cancer	image	2	30	114	91	364	/
Digits	image	10	64	360	287	1150	/
Multi-label classification datasets							
Yeast	biology	14	103	726	845	846	/
Emotions	audio	6	72	178	207	208	/
Category A	retail	6	5,007	166	331	717	/
Category B	retail	6	5,010	213	263	1,207	/
Category C	retail	7	5,006	637	642	19,968	/

B.6 EFFECT OF THE LOWER BOUND \mathbf{w}_- FOR THE OPTIMAL TRANSPORT ASSIGNMENT

In the proposed CSA, we have modified the original optimal transport formulation in [Tai et al. \(2021\)](#) to introduce the lower bound \mathbf{w}_- to ensure that the proportion of the assign labels in each class should be above this threshold. This design is useful in practice to prevent from being dominated by the high-frequent class while low-frequent class receives low or zero allocation. We summarize the comparison in Table 5 which shows that introducing the lower bound \mathbf{w}_- will improve marginally the performance (middle column). Then, introducing the confident estimation in CSA will significantly boost the performance (right column).

C APPENDIX: EXPERIMENTAL SETTINGS

We provide additional information about our experiments. We empirically calculate the label frequency vector $\mathbf{w} \in \mathbb{R}^K$ from the training data, then set the upper $\mathbf{w}_+ = 1.1 \times \mathbf{w}$ and lower $\mathbf{w}_- = 0.9 \times \mathbf{w}$ while we note that these values can vary with the datasets. We set the Sinkhorn regularizer $\varepsilon = 0.01$. We use $T = 5$ pseudo-label iterations for the main experiments while we provide analysis with different choices of the iterations in Appendix B.2.

We present the hyperparameters ranging for XGBoost in Table 6. Then, we summarize statistics for all of the datasets used in our experiments in Table 7.

We design to allocate more data points at the earlier iterations and less at later iterations by setting a decreasing vector over time $\rho_t = (T - t + 1) / (T + 1)$, where T is the maximum number of pseudo iterations, which is then normalized $\rho_t = \frac{\rho_t}{\sum_{t=1}^T \rho_t}$.

D APPENDIX: ADDITIONAL EXPERIMENTS

We present a further comparison with the pseudo-labeling baselines in Fig. 11 using 6 additional datasets.

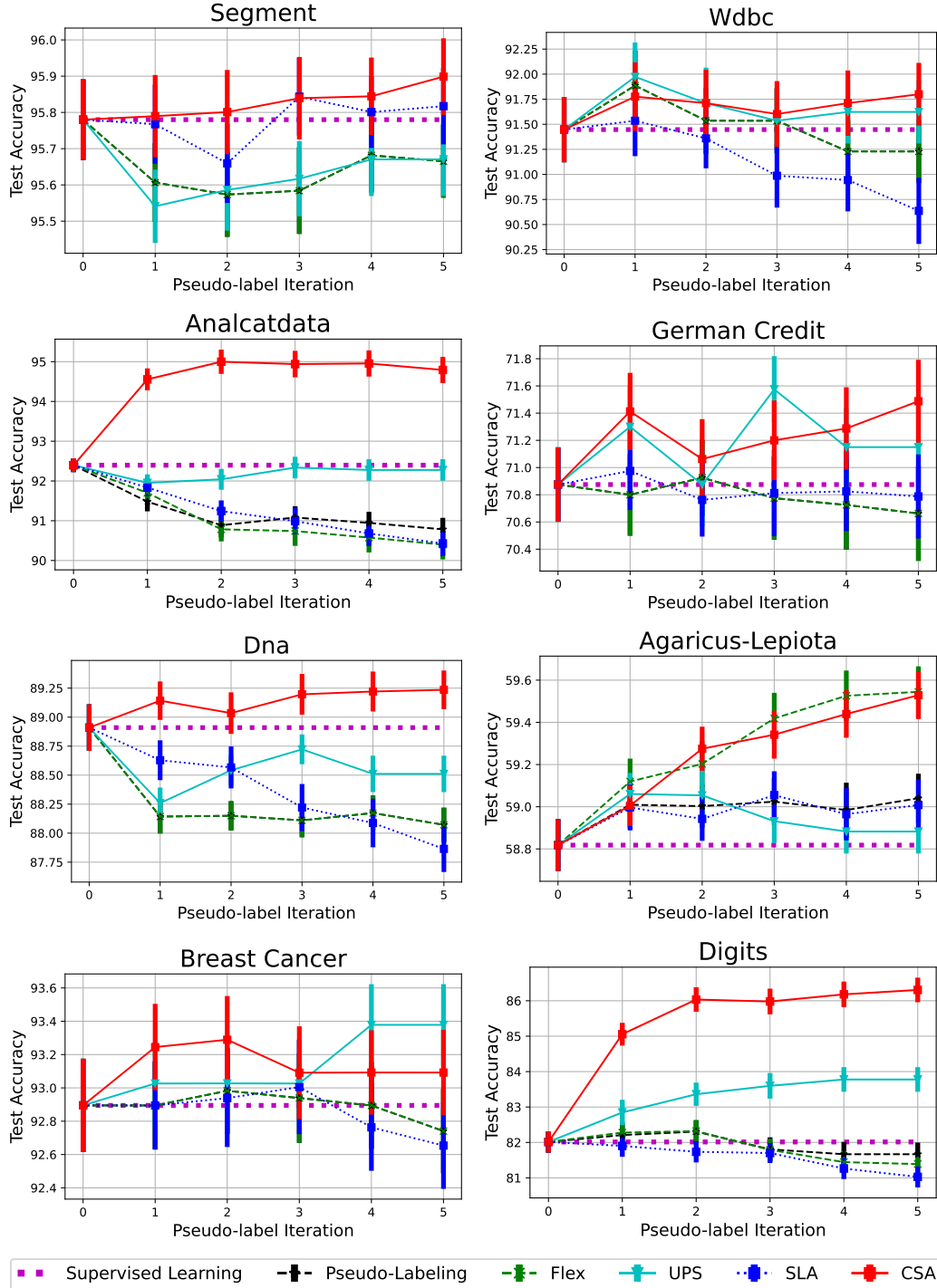


Figure 11: Additional experimental results in comparison with the pseudo-labeling methods on tabular data. The results clearly demonstrate the efficiency of the proposed CSA against the baselines.