Leveraging Ensemble Diversity for Robust Self-Training

Ambroise Odonnat[1,2,3], **Vasilii Feofanov**[1], **Ievgen Redko**[1]

[1]Huawei Noah's Ark Lab, [2]École des Ponts ParisTech, [3]ENS Paris-Saclay

Contributions

- $\mathcal{T}$-similarity, a **calibrated** confidence measure built upon a **diverse** ensemble of **linear** classifiers.
- Analysis of ensemble's **convergence and diversity**.
- **Robust** self-training under **distribution shift**.

Self-Training

Training data: Labeled set $(\mathbf{X}_\ell, \mathbf{y}_\ell)$, unlabeled set $\mathbf{X}_u$.

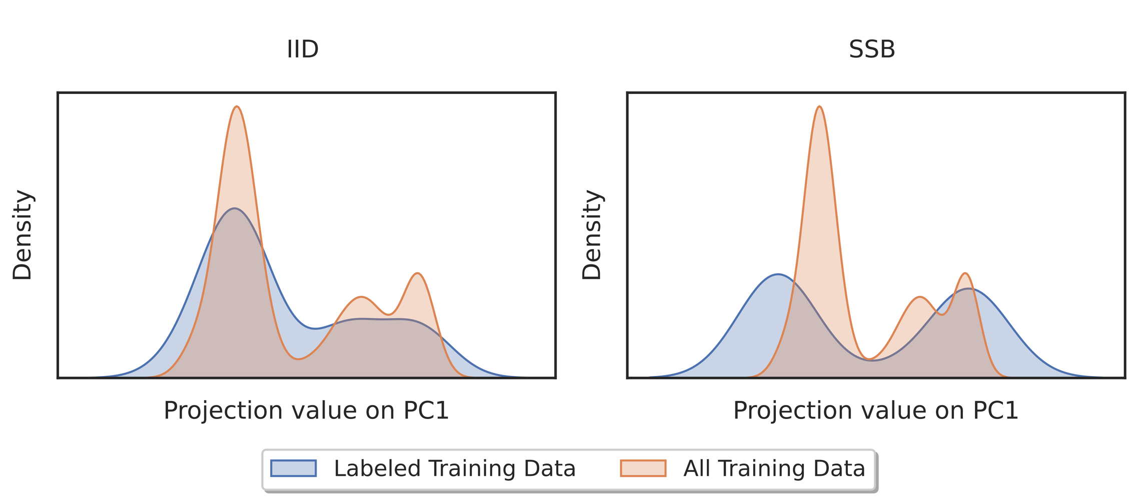
1. Train base classifier h on $(\mathbf{X}_\ell, \mathbf{y}_\ell)$,
2. Predict labels and confidence score on $\mathbf{X}_u$,
3. Pseudo-label most confident data and add them to $\mathbf{X}_\ell$,
4. Repeat until $\mathbf{X}_u = \emptyset$.

Self-training will fail if the confidence measure is biased, which can occur under distribution shifts.

Sample Selection Bias (SSB)

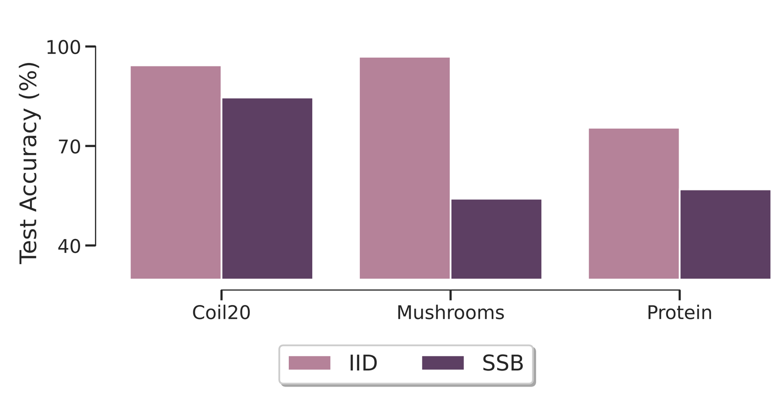
- **IID**: usual labeling that verifies the i.i.d. assumption.
- **SSB**: model shift btw. labeled and unlabeled data.

$$\mathbb{P}(\text{ to label } \mathbf{x} \mid y = c) \propto \exp(r \times |\mathrm{PCA}_1(\mathbf{x})|).$$



Failure of Self-Training with Softmax

- Classifier is biased toward the labeled set under SSB.
- Softmax gives high scores even to wrong predictions.

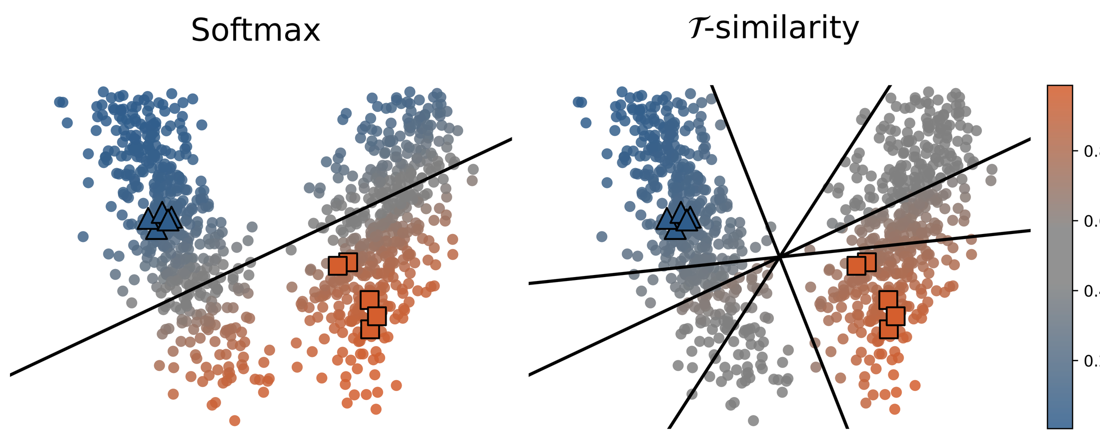


Challenges

1. **Reliable** confidence estimation is fundamental,
2. The widely-used `softmax` **cannot be trusted**,
3. The solution must have a **lightspeed** computation.

Learning with the $\mathcal{T}$-similarity

Diverse classifiers **disagree a lot** on samples in ***unsafe*** regions and have a **strong agreement** inside ***safe*** regions.



We train an ensemble $\mathcal{T}$ to **fit the labeled set** while being **diverse on the unlabeled set** by minimizing

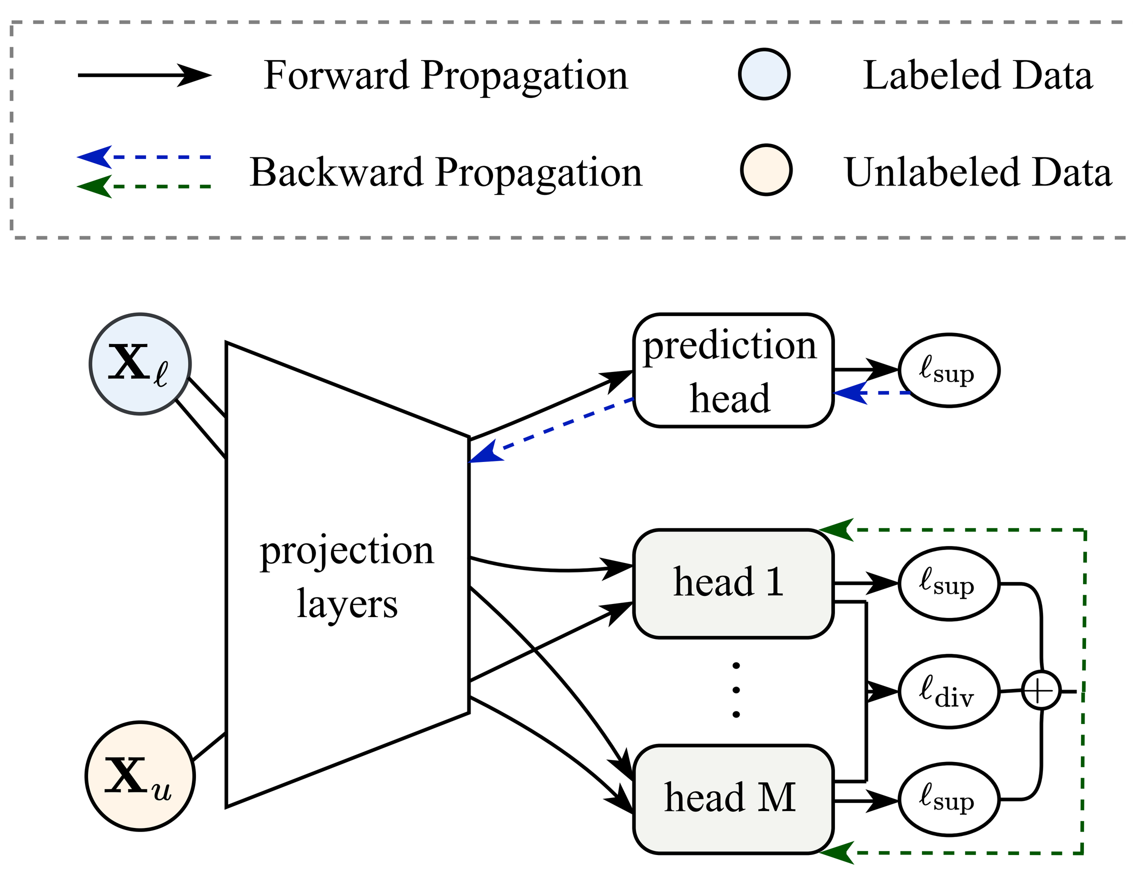
$$\mathcal{L}(\mathcal{T}) = \underbrace{\frac{1}{M}\sum_{m=1}^{M} \ell_{\text{sup}}(h_m, \mathbf{X}_\ell, \mathbf{y}_\ell)}_{\text{label fidelity term}} + \underbrace{\frac{\gamma}{n_u}\sum_{\mathbf{x}\in\mathbf{X}_u} s_{\mathcal{T}}(\mathbf{x})}_{\text{agreement term}},$$

where the agreement is quantified by the $\mathcal{T}$-similarity

$$s_{\mathcal{T}}(\mathbf{x}) = \frac{1}{M(M-1)}\sum_{m\neq k} h_m(\mathbf{x})^\top h_k(\mathbf{x}).$$

Practical Implementation

- Projection layers learned via the prediction head.
- Learning $\mathcal{T}$ without influencing the representation.
- Ensemble $\mathcal{T}$ of 5 **linear** heads.



We obtain a lightweight implementation suitable to any SSL method with neural networks as backbones.

Theoretical Analysis

- Binary classification with an ensemble of linear heads.
- ℓ_{sup} is the least-square loss with Tikhonov regularization.
- Gradient descent finds stationary points of $\mathcal{L}$, i.e., $\mathcal{T}$ s.t.

$$\nabla\mathcal{L}(\mathcal{T}) = 0.$$

Findings

- Finding stationary points is a **linear** problem in $\mathcal{T}$.
- Under mild assumptions, $\mathcal{L}$ has a **unique minimizer**.
- High diversity when classifiers cover the directions of **large variance** in the labeled data.
- High diversity when labeled data cover the input space evenly $\rightarrow$ motivation for **contrastive learning**.

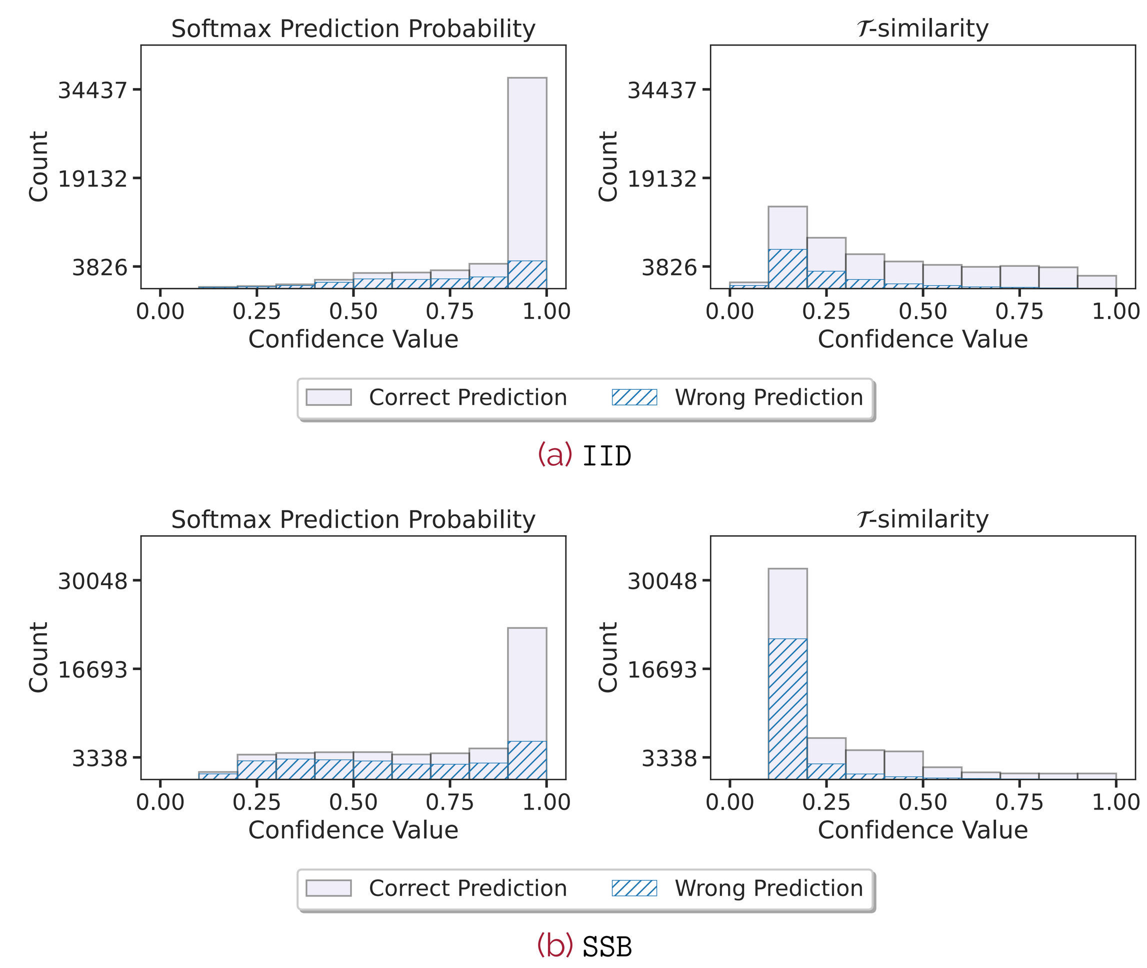
Main References

- **Quionero-Candela et al.** - MIT Press 2009
  Dataset Shift in Machine Learning
- **Zhang and Zhou** - DMKD 2013
  Exploiting unlabeled data to enhance ensemble diversity
- **Odonnat et al.** - **AISTATS 2024 (this work)**
  Leveraging Ensemble Diversity for Robust Self-Training

Experiments

- `ERM` is supervised learning with the labeled set.
- $\mathrm{PL}_{\theta=0.8}$ is self-training with a fixed threshold. $\theta = 0.8$

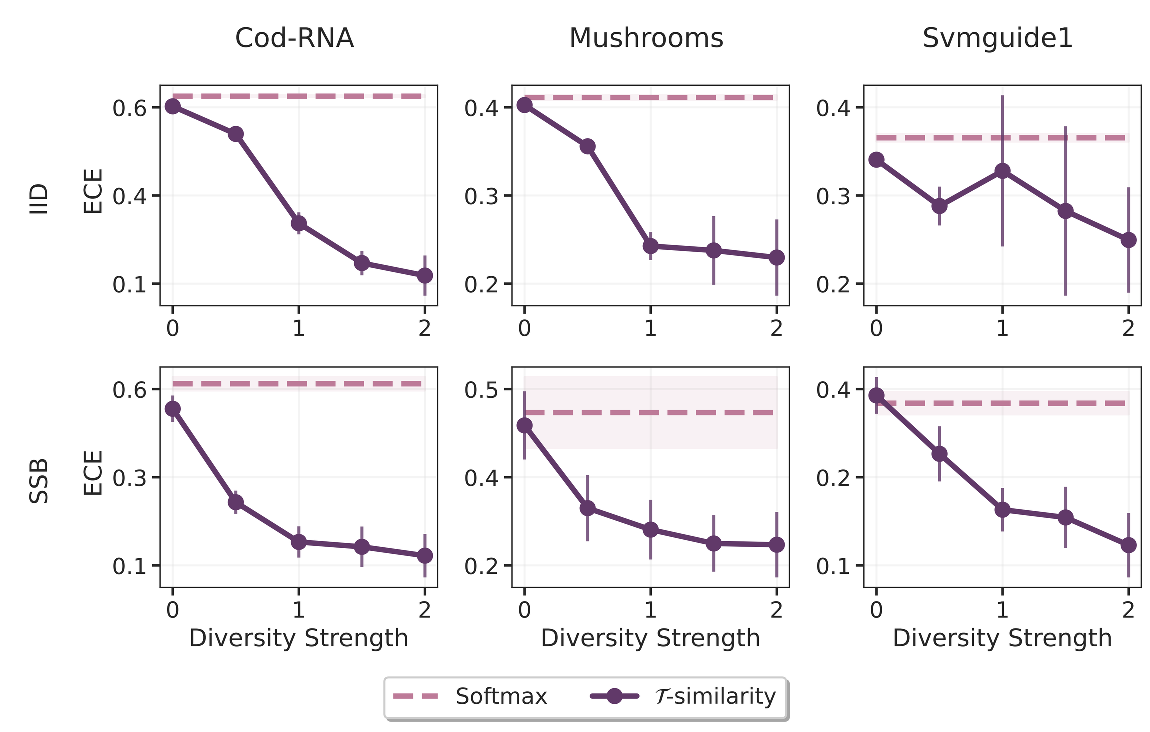
Diversity and Calibration



(a) IID

(b) SSB

$\mathcal{T}$-similarity corrects the softmax overconfidence and gives high confidence only to accurate predictions.



Increasing the diversity of the ensemble of classifiers improves the calibration of predicted probabilities.

Robust Self-Training under SSB

Dataset	ERM	$\mathrm{PL}_{\theta=0.8}$ softmax	$\mathrm{PL}_{\theta=0.8}$ $\mathcal{T}$-similarity
Cod-RNA	74.51 ± 8.86	74.75 ± 8.14	$\mathbf{80.06 \pm 3.55}$
HAR	82.57 ± 1.96	82.87 ± 3.02	$\mathbf{83.12 \pm 2.27}$
Mnist	50.74 ± 2.25	51.08 ± 2.55	$\mathbf{52.69 \pm 2.42}$
Mushrooms	69.45 ± 7.29	59.53 ± 10.46	$\mathbf{71.36 \pm 6.63}$
Phishing	67.42 ± 3.55	66.08 ± 5.66	$\mathbf{77.41 \pm 3.93}$
Protein	57.57 ± 6.33	57.45 ± 6.36	$\mathbf{57.61 \pm 6.23}$
Rice	79.19 ± 5.12	80.54 ± 4.31	$\mathbf{81.1 \pm 4.28}$
Splice	66.13 ± 4.47	67.14 ± 2.62	$\mathbf{67.45 \pm 2.53}$
Svmguide1	70.89 ± 10.98	70.35 ± 11.74	$\mathbf{81.07 \pm 5.39}$

The $\mathcal{T}$-similarity is better than softmax and can enable self-training to go from degradation to improvement.

Take Home Message

Confidence estimation should be made with care in semi-supervised settings under distribution shifts.
$\rightarrow$ **Start using our $\mathcal{T}$-similarity to avoid trouble!**

Want to Know More?

paper

code

ambroiseodonnattechnologie@gmail.com