Don’t fear the unlabelled: safe deep semi-supervised learning via simple debiasing

Hugo Schmutz[1,2], Olivier Humbert[1] & Pierre-Alexandre Mattei[2]

3iA Côte d'Azur – Interdisciplinary Institute for Artificial Intelligence · Investir l'avenir · Inria · Université Côte d'Azur

Motivation for Semi-supervised learning.

- Unlabelled data are cheap: $\{(x_1, y_1), \ldots, (x_{n_l}, y_{n_l})\}$
- Labelled data can be hard to get: $\{x_{n_l+1}, \ldots, x_{n_l+n_u}\}$

Missing Completely At Random (MCAR) i.e. y being missing is independent of x.

The complete case

$$\hat{\mathcal{R}}_{CC}(\theta) = \frac{1}{n_l}\sum_{i=1}^{n_l} L(\theta; x_i, y_i) \quad (1)$$

$\longrightarrow$ Under MCAR unbiased $\Rightarrow$ learning theory + asymptotic statistics.

Safeness: A SSL algorithm is safe if it has theoretical guarantees that are similar to or stronger than the complete case baseline.

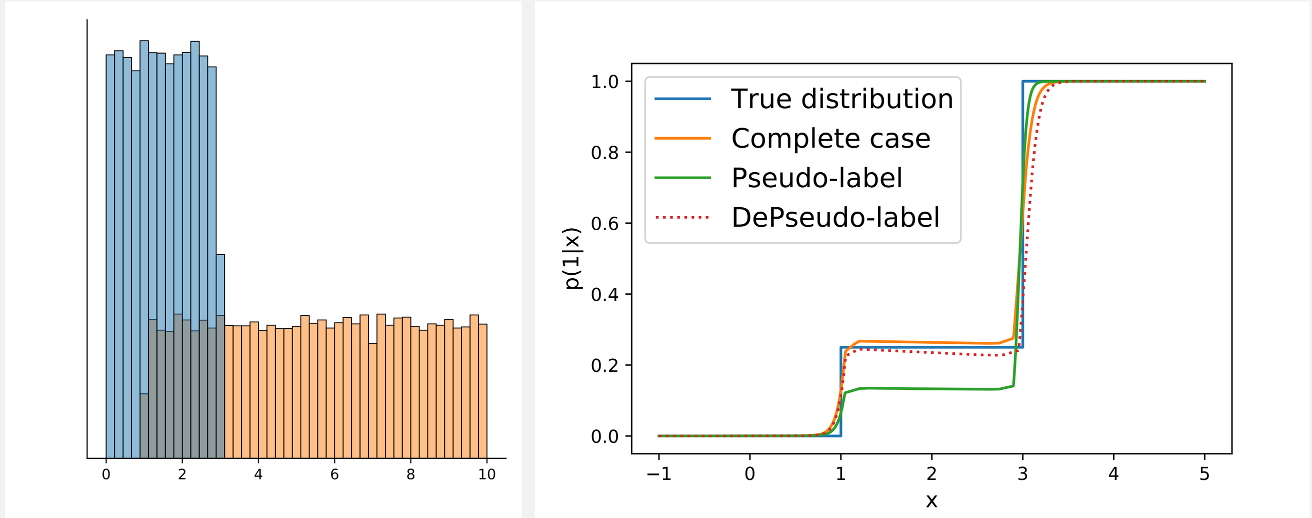
Including unlabelled data

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l}\sum_{i=1}^{n_l} L(\theta; x_i, y_i) + \frac{\lambda}{n}\sum_{i=1}^{n} H(\theta; x_i) \quad (2)$$

- Entropy minimization: $H(\theta; x_i) = -\sum f_\theta(x)_k \log(f_\theta(x)_k)$.
- Consistency based: $H(\theta; x_i) = Div(f_{\hat{\theta}}(x), f_\theta(pert(x)))$.
- Pseudo-label: $H(\theta; x_i) = -\begin{cases} \log(\max f_\theta(x)_k) & \text{if } \max f_\theta(x)_k > \tau \\ 0 & \text{elsewhere.} \end{cases}$

SSL failure on a toy example

,



- No **safeness** guarantees even under strong assumptions (domain-specific data augmentations; manifold, low-density or cluster assumption).
- **Biased** risk estimator $\Rightarrow$ learning theory does not hold.
- No asymptotic **consistency** $\Rightarrow$ fail even with an infinite amount of labelled data points

DeSSL: unbiased under MCAR

$$\hat{\mathcal{R}}_{DeSSL}(\theta) = \frac{1}{n_l}\sum_{i=1}^{n_l} L(\theta; x_i, y_i) + \frac{\lambda}{n}\sum_{i=1}^{n} H(\theta; x_i) - \frac{\lambda}{n_l}\sum_{i=1}^{n_l} H(\theta; x_i) \quad (3)$$

Is $\hat{\mathcal{R}}_{DeSSL}(\theta)$ an accurate risk estimate?

Theorem 1 *The function $\lambda \mapsto \mathbb{V}(\hat{\mathcal{R}}_{DeSSL}(\theta)|r)$ reaches its minimum for:*

$$\lambda_{opt} = \frac{\mathrm{Cov}(L(\theta; x, y), H(\theta; x))}{\mathbb{V}(H(\theta; x))} \quad (4)$$

and

$$\mathbb{V}(\hat{\mathcal{R}}_{DeSSL}(\theta))|_{\lambda_{opt}} = (1 - \frac{n_u}{n}\rho^2_{L,H})\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)) \leq \mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)). \quad (5)$$

But wait! There is more:

Calibration *If the original loss is a proper scoring rule, then DeSSL is also a proper scoring rule.*

Consistency *Under usual regularity conditions of M-estimators for both L and H, $\hat{\theta} = \arg\min \hat{\mathcal{R}}_{DeSSL}$ is asymptotically consistent with respect to n.*

Generalisation error bounds *If L and H are bounded, DeSSL benefits of generalisation error bounds based on the Rademacher complexity.*

Asymptotic normality *Under usual conditions on L, H and the risk $\mathcal{R}(\theta)$, DeSSL is asymptotically normal. Its asymptotic variance can be optimised in λ and is smaller than the complete case's variance at its optimum.*

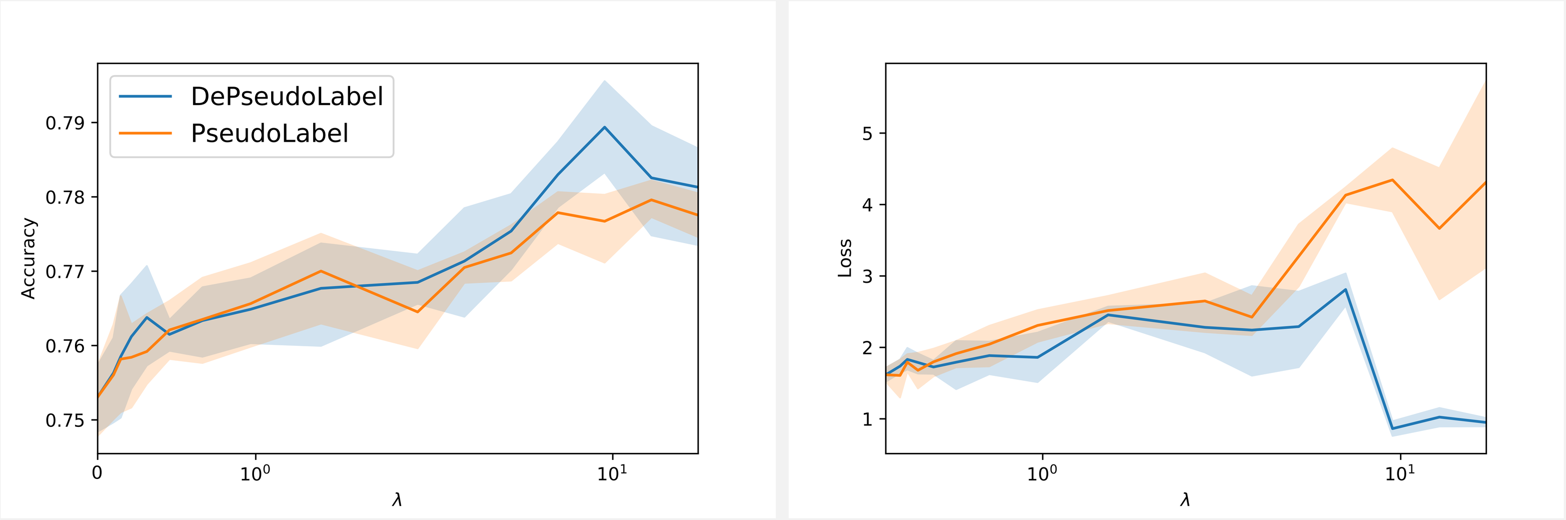
Interpretation of DeSSL

- Debiaising of the SSL risk loss
- Control variates (variance reduction techniques)
- Lagrangian of the following optimization problem:

$$\min_\theta \ \hat{\mathcal{R}}_{CC}(\theta) \quad \text{s.t.} \ \frac{1}{n}\sum_{i=1}^{n} H(\theta; x_i) = \frac{1}{n_l}\sum_{i=1}^{n_l} H(\theta; x_i) \quad (6)$$

- Regularization of the model's confidence on the labelled
- Debiasing with the labelled is optimal regarding the variance

Pseudo-label on CIFAR-10 ($n_l = 4000$)



Fixmatch on CIFAR-10 ($n_l = 4000$)

	Complete Case	Fixmatch	DeFixmatch
Accuracy	87.27 ± 0.25	93.87 ± 0.13	**95.44 ± 0.10**
Worst class accuracy	70.08 ± 0.93	82.25 ± 2.27	**87.16 ± 0.46**
Cross-entropy	0.60 ± 0.01	0.27 ± 0.01	**0.20 ± 0.01**

Disparate effect of SSL ($n_l = 4000$)

	Complete Case	Fixmatch		DeFixmatch	
	Accuracy	Accuracy	$\mathcal{BR}$	Accuracy	$\mathcal{BR}$
airplane	86.94	95.94	0.88	96.62	0.94
automobile	95.26	97.54	0.68	98.22	0.89
bird	**80.46**	**90.80**	**0.68**	**92.64**	**0.80**
cat	**70.08**	**82.50**	**0.56**	**87.16**	**0.78**
deer	88.88	95.86	0.78	97.26	0.94
dog	**79.66**	**87.16**	**0.53**	**90.98**	**0.81**
frog	93.12	97.84	0.80	98.62	0.94
horse	90.96	96.94	0.83	97.64	0.92
ship	94.12	97.26	0.67	98.06	0.84
truck	93.18	96.82	0.84	97.20	0.93

Benefit ratio, $\mathcal{BR}$ = the impact of SSL on a class

References

A. Oliver et al., **Realistic Evaluation of Deep Semi-Supervised Learning Algorithms** *NeurIPS*, 2018.
Y. Grandvalet. & Y. Bengio **Semi-supervised Learning by Entropy Minimization**, *NeurIPS*, 2005.
K. Sohn et al., **FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence** *NeurIPS* 2020.
T. Miyato et al., **Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning** *IEEE transactions on pattern analysis and machine intelligence*, 2017.
A. W. van der Vaart, **Asymptotic Statistics** *Cambridge University Press*, 1992.
G. Pereyra, **Regularizing Neural Networks by Penalizing Confident Output Distributions**, 2017

[1] UMR E4320 TIRO-MATOs, Université Côte d'Azur, Centre Antoine Lacassagne
[2] Inria (Maasai team), Laboratoire J.A. Dieudonné, UMR CNRS 7351