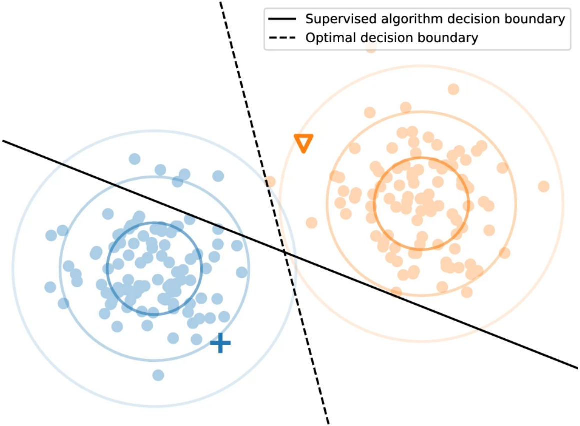
Variance reduction for semi-supervised learning

Hugo Schmutz

INRIA, Equipe MAASAI
TIRO-MATOS, CEA UMR 4320
Laboratoire LJAD, UMR CNRS 7351
Université Côte d'Azur
3iA Côte d'Azur
hugo.schmutz@inria.fr - @HugoSchmutz2

Deep semi-supervised learning ? What for ?

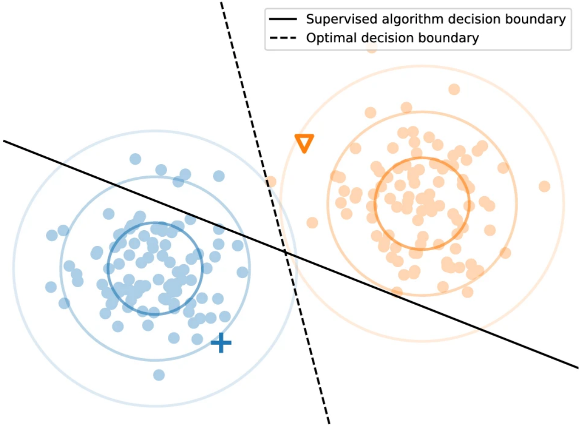
Goal: Using both labelled and unlabelled data to build better learners, than using each one alone



V.Engelen & Hoos [Machine Learning, 2020]

Deep semi-supervised learning ? What for ?

Goal: Using both labelled and unlabelled data to build better learners, than using each one alone



V.Engelen & Hoos [Machine Learning, 2020]

Why bother ?

- unlabelled data are cheap
- labelled data can be hard to get

Learning theory relies on the unbiased estimator of the risk

We have two spaces of objects X and Y and would like to learn a function $f_\theta : X \to Y$.
L is a loss function which measures how different the prediction $f_\theta(x)$ is from the true outcome y.
We define the risk:

$$\mathcal{R}(\theta) = \mathbb{E}[L(\theta; x_i, y_i)]$$

The ultimate goal of a learning algorithm is to find θ^* among a fixed class of functions:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{R}(\theta)$$

Learning theory relies on the unbiased estimator of the risk

We have two spaces of objects X and Y and would like to learn a function $f_\theta : X \to Y$.
L is a loss function which measures how different the prediction $f_\theta(x)$ is from the true outcome y.
We define the risk:

$$\mathcal{R}(\theta) = \mathbb{E}[L(\theta; x_i, y_i)]$$

The ultimate goal of a learning algorithm is to find θ^* among a fixed class of functions:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{R}(\theta)$$

In practice: $P(x, y)$ is unknown. We have access to a training set: $\{(x_1, y_1), ..., (x_n, y_n)\}$

Learning theory relies on the unbiased estimator of the risk

We have two spaces of objects X and Y and would like to learn a function $f_\theta : X \to Y$.
L is a loss function which measures how different the prediction $f_\theta(x)$ is from the true outcome y.
We define the risk:

$$\mathcal{R}(\theta) = \mathbb{E}[L(\theta; x_i, y_i)]$$

The ultimate goal of a learning algorithm is to find θ^* among a fixed class of functions:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{R}(\theta)$$

In practice: $P(x, y)$ is unknown. We have access to a training set: $\{(x_1, y_1), ..., (x_n, y_n)\}$
Empirical risk: Monte Carlo estimator of the risk

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(\theta; x, y)$$

$\longrightarrow$ Unbiased estimator $\Rightarrow$ basic property that is needed for the development of traditional learning theory and asymptotic statistics.

The complete case, the simplest: get rid of unlabelled data

- labelled data: $\{(x_1, y_1), ..., (x_{n_l}, y_{n_l})\}$
- unlabelled data: $\{x_{n_l+1}, ..., x_{n_l+n_u}\}$
- $n_l + n_u = n$

Semi-supervised is a missing data problem. We consider the case missing completely at random (MCAR) i.e. y being missing is independent x.

The complete case, the simplest: get rid of unlabelled data

- labelled data: $\{(x_1, y_1), \ldots, (x_{n_l}, y_{n_l})\}$
- unlabelled data: $\{x_{n_l+1}, \ldots, x_{n_l+n_u}\}$
- $n_l + n_u = n$

Semi-supervised is a missing data problem. We consider the case missing completely at random (MCAR) i.e. y being missing is independent x.
Complete case. Get rid of unlabelled data:

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

$\longrightarrow$ Under MCAR, also unbiased $\Rightarrow$ basic property that is needed for the development of traditional learning theory and asymptotic statistics.

Using unlabelled data leads to biased estimator of the risk

- labelled data: $\{(x_1, y_1), \ldots, (x_{n_l}, y_{n_l})\}$
- unlabelled data: $\{x_{n_l+1}, \ldots, x_{n_l+n_u}\}$
- $n_l + n_u = n$

Semi-supervised is a missing data problem. We consider the case missing completely at random (MCAR) i.e. y being missing is independent x.

Including unlabelled data in the risk estimator:

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

Using unlabelled data leads to biased estimator of the risk

- labelled data: $\{(x_1, y_1), \dots, (x_{n_l}, y_{n_l})\}$
- unlabelled data: $\{x_{n_l+1}, \dots, x_{n_l+n_u}\}$
- $n_l + n_u = n$

Semi-supervised is a missing data problem. We consider the case missing completely at random (MCAR) i.e. y being missing is independent x.

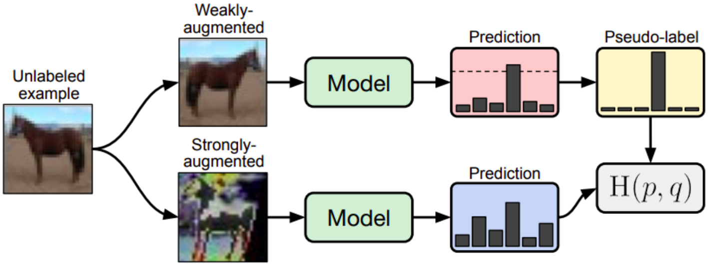
Including unlabelled data in the risk estimator:

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

Examples:

- Entropy minimization: $H(\theta; x_i) = -\sum f_\theta(x)_k \log(p_\theta(x)_k)$ (Grandvalet and Bengio 2005)
- Consistency based: $H(\theta; x_i) = Div(p_\theta(x), p_\theta(pert(x)))$ (Sohn et al. 2020, Xie et al. 2020, Miyato et al. 2018, Laine and Aila 2017, ...)
- Pseudo-label (PL): $H(\theta; x_i) = -\log(\max p_\theta(u|x))\mathbb{1}[\max_y p_{\hat{\theta}}(y|x) > \tau]$ (Scudder 1965, Lee 2013, Rizve 2021)

Fixmatch (Sohn et al., NeurIPS [2020])



$$H(\theta; x) = \mathbb{E}_{x_1 \sim weak(x)} \left[\mathbb{1}[\max_y p_{\hat{\theta}}(y|x_1) > \tau] \mathbb{E}_{x_2 \sim strong(x)} [-\log(p_\theta(\arg\max_y p_{\hat{\theta}}(y|x_1)|x_2))] \right]$$

Using unlabelled data leads to biased estimator of the risk

Including unlabelled data in the risk estimator:

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

Good performance on a various of (deep) learning tasks.

Using unlabelled data leads to biased estimator of the risk

Including unlabelled data in the risk estimator:

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

Good performance on a various of (deep) learning tasks.
Problems:

- Potential degradation reported in previous works (Schölkopf et al. [ICML, 2012], V.Engelen & Hoos, [Machine Learning, 2020], Zhu et al. [ICLR, 2022])

Using unlabelled data leads to biased estimator of the risk

Including unlabelled data in the risk estimator:

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

Good performance on a various of (deep) learning tasks.
Problems:

- Potential degradation reported in previous works (Schölkopf et al. [ICML, 2012], V.Engelen & Hoos, [Machine Learning, 2020], Zhu et al. [ICLR, 2022])
- Few theoretical guarantees using strong distributional assumptions (Mey & Loog [2019])

Using unlabelled data leads to biased estimator of the risk

Including unlabelled data in the risk estimator:

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

Good performance on a various of (deep) learning tasks.

Problems:

- Potential degradation reported in previous works (Schölkopf et al. [ICML, 2012], V.Engelen & Hoos, [Machine Learning, 2020], Zhu et al. [ICLR, 2022])
- Few theoretical guarantees using strong distributional assumptions (Mey & Loog [2019])
- No assymptotic consistency: may fail even with an infinite number of labelled datapoints

Using unlabelled data leads to biased estimator of the risk

Including unlabelled data in the risk estimator:

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

Good performance on a various of (deep) learning tasks.
Problems:

- Potential degradation reported in previous works (Schölkopf et al. [ICML, 2012], V.Engelen & Hoos, [Machine Learning, 2020], Zhu et al. [ICLR, 2022])
- Few theoretical guarantees using strong distributional assumptions (Mey & Loog [2019])
- No assymptotic consistency: may fail even with an infinite number of labelled datapoints
- Choice of H can be confusing (Corduneanu & Jaakkola [UAI, 003], Krause et al. [NeurIPS, 2010])

Using unlabelled data leads to biased estimator of the risk

Including unlabelled data in the risk estimator:

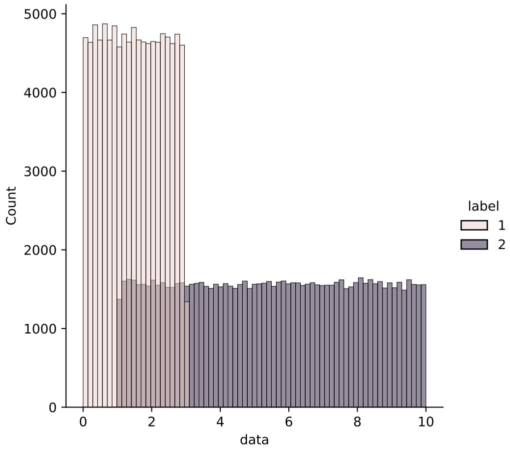
$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l}\sum_{i=1}^{n_l} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i=1}^{n_u} H(\theta; x_i)$$

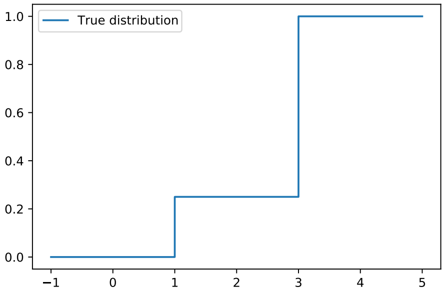
Good performance on a various of (deep) learning tasks.

Problems:

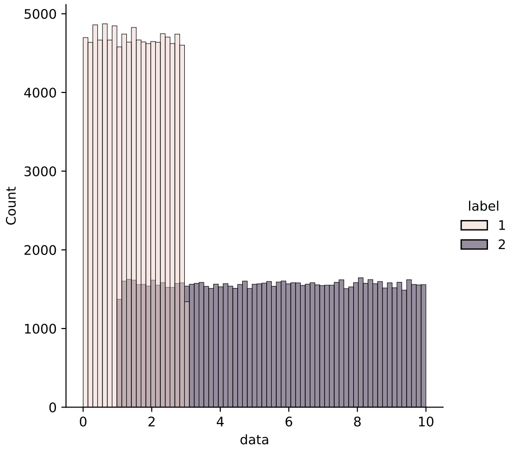
- Potential degradation reported in previous works (Schölkopf et al. [ICML, 2012], V.Engelen & Hoos, [Machine Learning, 2020], Zhu et al. [ICLR, 2022])
- Few theoretical guarantees using strong distributional assumptions (Mey & Loog [2019])
- No assymptotic consistency: may fail even with an infinite number of labelled datapoints
- Choice of H can be confusing (Corduneanu & Jaakkola [UAI, 003], Krause et al. [NeurIPS, 2010])
- An additional hyperparameter λ and no realistic validation (Oliver et al. [NeurIPS, 2018])

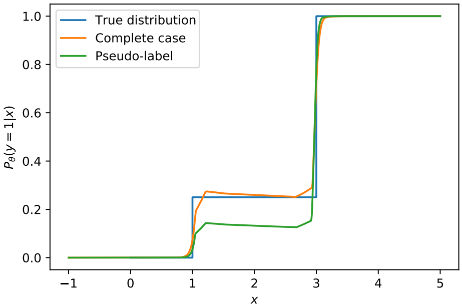
FEAR THE UNLABELLED !



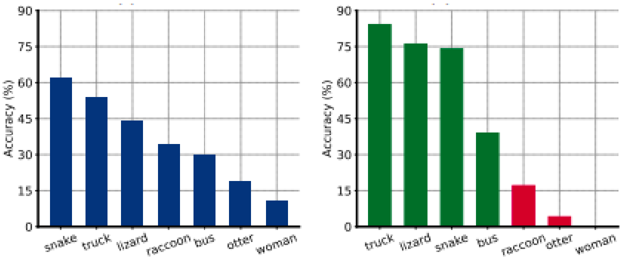


FEAR THE UNLABELLED !





The rich get richer ! (Zhu et al. [ICLR, 2022])



Top-1 accuracy of 7 randomly selected categories with different training methods on CIFAR-100. (Left) Complete case. (Right) FixMatch

FixMatch largely increases the bias of poor-behaved categories. (Chen et al., [arXiv:2202.07136])

DeSSL: Debiased version of SSL

Don't fear the unlabelled: Safe semi-supervised learning via debiasing, Schmutz et al., ICLR 2023

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

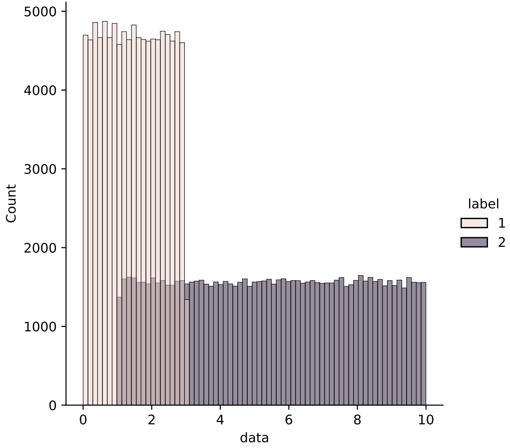
DeSSL: Debiased version of SSL

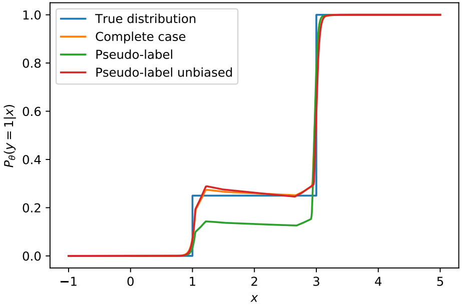
Don't fear the unlabelled: Safe semi-supervised learning via debiasing, Schmutz et al., ICLR 2023

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

- Under the MCAR assumption, estimator is unbiased estimator of the true risk.
- Close relationship with control variates (Owen [2013]).
- Other motivations:
  - Penalising the confidence of a model on labelled data (Pereyra et al., [2017])
  - Maximising the plausibility (Barndorff-Nielsen [1976]).
  - The risk estimate is a Lagrangian!

Pseudo-label unbiased success under the cluster assumption





$\hat{\mathcal{R}}_{DeSSL}(\theta)$ is unbiased but is it an accurate risk estimator ?

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

$$\lambda_{opt} = \frac{n_u}{n} \frac{\mathrm{Cov}(L(\theta; x, y), H(\theta; x))}{\mathbb{V}(H(\theta; x))} \tag{1}$$

and

$$\begin{aligned} \mathbb{V}(\hat{\mathcal{R}}_{DeSSL}(\theta))|_{\lambda_{opt}} &= (1 - \frac{n_u}{n}\rho_{L,H}^2)\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)) \\ &\leq \mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)) \end{aligned} \tag{2}$$

where $\rho_{L,H} = \mathrm{Corr}(L(\theta; x, y), H(\theta; x))$.

Justification on the heuristic idea that H should be a surrogate of L.

$\hat{\mathcal{R}}_{DeSSL}(\theta)$ is an accurate risk estimator but $\nabla\hat{\mathcal{R}}_{DeSSL}(\theta)$ is even better

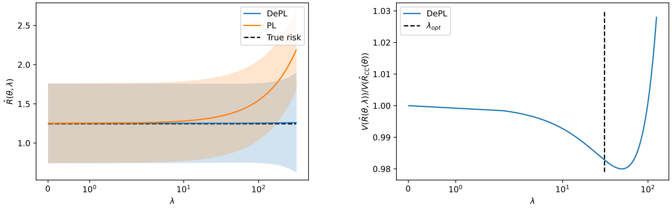


Figure: (Left) Risk estimate value for PseudoLabel (PL) and DePseudoLabel (DePL) compared to the true value of the risk. (Right) The influence of λ on the ratio $\mathbb{V}(\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)|r)$. (Down) The influence of λ on the ratio $\mathbb{V}(\nabla\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\nabla\hat{\mathcal{R}}_{CC}(\theta)|r)$.

$\hat{\mathcal{R}}_{DeSSL}(\theta)$ is an accurate risk estimator but $\nabla\hat{\mathcal{R}}_{DeSSL}(\theta)$ is even better

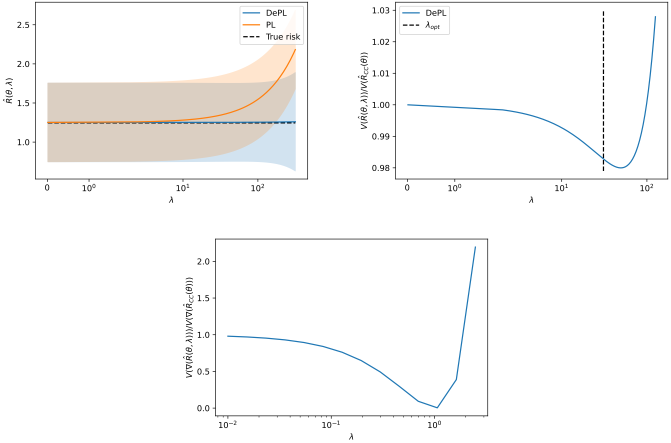


Figure: (Left) Risk estimate value for PseudoLabel (PL) and DePseudoLabel (DePL) compared to the true value of the risk. (Right) The influence of λ on the ratio $\mathbb{V}(\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\hat{\mathcal{R}}_{CC}(\theta)|r)$. (Down) The influence of λ on the ratio $\mathbb{V}(\nabla\hat{\mathcal{R}}_{DePL}(\theta)|r)/\mathbb{V}(\nabla\hat{\mathcal{R}}_{CC}(\theta)|r)$.

DeSSL is calibrated and benefits from generalisation error bounds

Calibration:

- We can expect DeSSL to be as **well-calibrated** as the complete case, while SSL will generally overfit.

Generalisation error bound:

- Under classical assumptions on L and H, DeSSL benefits of generalisation error bounds derived from the **Rademacher complexity**.

DeSSL is consistent and improves the supervised baseline

Consistency:

- Under classical assumptions on L and H, DeSSL provides **asymptotically consistent** models.
- SSL may fail with an infinite number of labelled data when DeSSL will not

Asymptotic normality:

- Under classical assumptions on L and H, the parameter estimated using DeSSL is asymptotically normal.
- It exists λ such as the asymptotic variance of the estimated parameters is lower than the one of the complete case.
- **DeSSL outperforms the complete case baseline in term of parameters estimation**

DeFixmatch improves Fixmatch accuracy and calibration

Table: Test accuracy, worst class accuracy and cross-entropy of Complete Case, Fixmatch and DeFixmatch on 5 folds of CIFAR-10 and one fold of CIFAR-100.

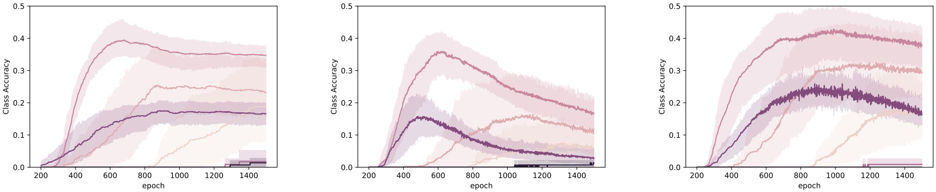
	CIFAR-10 ($n_l = 4000$)			CIFAR-100 ($n_l = 10000$)		
	Complete Case	Fixmatch	DeFixmatch	Complete Case	Fixmatch	DeFixmatch
Accuracy	87.27 $\pm$ 0.25	93.87 $\pm$ 0.13	**95.44** $\pm$ **0.10**	62.62	69.28	**71.22**
Worst class accuracy	70.08 $\pm$ 0.93	82.25 $\pm$ 2.27	**87.16** $\pm$ **0.46**	28.00	23.00	**31.00**
Cross entropy	0.60 $\pm$ 0.01	0.27 $\pm$ 0.01	**0.20** $\pm$ **0.01**	1.87	1.52	**1.42**
Brier score	0.214 $\pm$ 0.005	0.101 $\pm$0.003	**0.076** $\pm$ **0.001**	0.56	0.47	**0.44**

DeSSL mitigates the disparact effect of SSL

Table: Mean accuracy per class and mean benefit ratio ($\mathcal{BR}$, Zhu et al. [2022]) on 5 splits.

	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

DeSSL mitigates the disparact effect of SSL



Class accuracies (without the majority class) on DermaMNIST trained with $n_l = 1000$ labelled data on five folds. (Left) CompleteCase (B-Acc: $26.88 \pm 2.26\%$); (Middle) PseudoLabel (B-Acc: $22.03 \pm 1.45\%$); (Right) DePseudoLabel (B-Acc: $\mathbf{28.84} \pm 1.02\%$), with 95% CI.

Conclusion

More results on CIFAR-100, SVHN, STL10, UCI datasets and MedMNIST datasets in the paper.

- DeSSL comes with theoretical guarantees using only the MCAR assumption
- Estimator unbiased, reduction of variance, asymptotically consistent, well calibrated
- Formula for the hyperparameter
- Mitigates the disparact effect of SSL
- Performs better than the biased estimator on various datasets (see our paper)

On going works:

- Computation of λ_{opt}
- Extend to non MCAR settings
- Extend to segmentation
- Stratified sampling to select automatically the number of labelled and unlabelled data per batch

Don't fear the unlabelled: Safe semi-supervised learning via debiasing, Schmutz et al., ICLR 2023
Monte Carlo theory, methods and examples , Art B. Owen, 2013