Can semi-supervised learning use all the data effectively? A lower bound perspective

Alexandru Ţifrea*, Gizem Yüce*, Amartya Sanyal, Fanny Yang

ETH Zürich, EPFL, MPI-IS Tübingen

Semi-supervised classification

Setting: Access to labeled data $\mathcal{D}_l$ and unlabeled data $\mathcal{D}_u$.

"Wasteful" strategies:

- Supervised learning (SL) – **ignore** $\mathcal{D}_u$
- Unsupervised learning up to permutation (UL+)
  1. learn decision boundary using $\mathcal{D}_u$ – **ignore** $\mathcal{D}_l$
  2. label prediction regions using $\mathcal{D}_l$

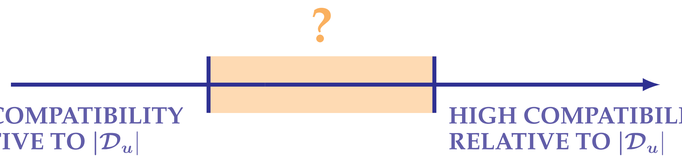
Alternative: SSL algorithms that use both $\mathcal{D}_l$ and $\mathcal{D}_u$ effectively

- e.g. experiments suggest SimCLRv2, {Mix,Fix,Free}Match can improve over both SL and UL+

How fundamental is the improvement over SL and UL+

Prior theoretical works:

- focus on specific regimes of **Compatibility relative to $\mathcal{D}_u$** i.e. information about $Y \mid X$ captured in $\mathcal{D}_u$



Information-theoretic lower bounds for SSL:

SSL cannot improve SL rates

Examples:
Ben-David et al, 2008
Tolstikhin et al, 2016
Göpfert et al, 2019

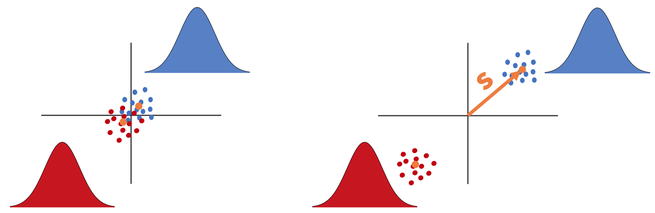
Upper bounds for specific SSL algorithms:

SSL improves SL rates

Examples:
Ratsaby et al, 1995
Rigollet, 2006
Frei et al, 2022

Can SSL **simultaneously** improve over the minimax rates of **both** SL and UL+?

Our setting: 2-GMM family



Low compatibility High compatibility

Suitable setting to study SSL improvement because:

- can vary explicitly the compatibility of the SSL task
- there exist known minimax rates for SL and UL+.

Main result: Adaptive SSL lower bound

Goal: SSL minimax rate for **excess risk** that depends on SNR s

Definition

The minimax rate of algorithms $\mathcal{A}$ over distributions $\mathcal{P}$ is

$$\inf_{\mathcal{A}} \sup_{\mathcal{P}} \mathbb{E}[\mathcal{E}(\mathcal{A}(\mathcal{D}_l, \mathcal{D}_u), \mathcal{P})]$$

Denoting $n_l = |\mathcal{D}_l|$ and $n_u = \mathcal{D}_u$:

SL	UL+	**SSL (this paper)**
$\Theta\left(e^{-s^2/2}\frac{d}{sn_l}\right)$	$\tilde{\Theta}\left(e^{-s^2/2}\frac{d}{s^3 n_u}\right)$	$\tilde{\Theta}\left(e^{s^2/2}\min\left\{s, \frac{d}{sn_l}, \frac{d}{s^3 n_u}\right\}\right)$

- Proof employs prior techniques developed for SL/UL [Azizyan et al, 2013; Li et al, 2017; Wu et al, 2021].
- Upper bound achieved by using either SL or UL+ depending on (s, n_l, n_u).
- SSL minimax rate of parameter estimation error in the paper.

No simultaneous rate improvement of SSL over both SL and UL+

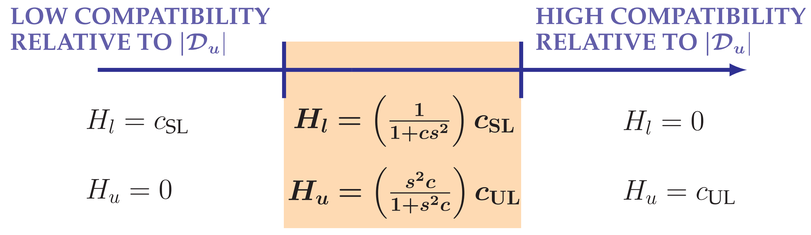
Definition: Rate improvement of SSL over SL and UL+

$$h_l(n_l, n_u, s) := \frac{\text{SSL rate}}{\text{SL rate}} \quad \text{and} \quad h_u(n_l, n_u, s) := \frac{\text{SSL rate}}{\text{UL+ rate}}$$

Ideally: SSL improves upon the rates of both SL and UL+ simultaneously if

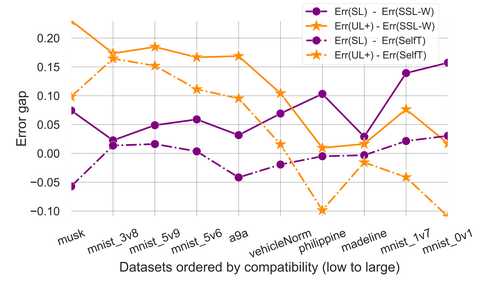
- $H_l := \lim_{n_l, n_u \to \infty} h_l(n_l, n_u, s) = 0$, and
- $H_u := \lim_{n_l, n_u \to \infty} h_u(n_l, n_u, s) = 0$

Corollary: Rate improvement over both SL and UL+ is not possible with any SSL algorithm for 2-GMMs.



Future work

Empirically, SSL algorithms can **simultaneously** improve over both SL and UL+, i.e. use all the data more effectively



- need for constant-tracking in bounds for SSL algorithms e.g. self-training
- benchmarking SSL algorithms should also consider the intermediate regime of moderate n_u