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

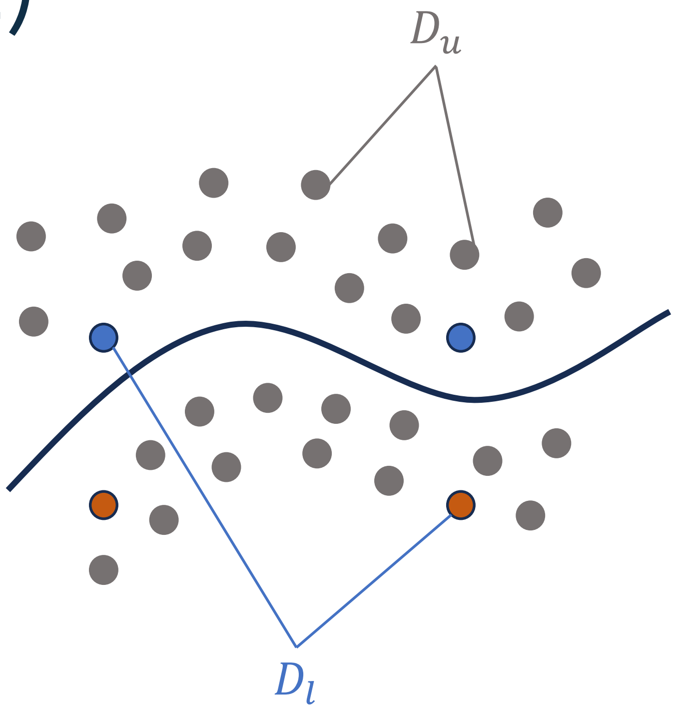
Alexandru Ţifrea*, Gizem Yüce*, Amartya Sanyal, Fanny Yang

ETH Zürich, EPFL




Semi-supervised learning (SSL)

Setting: labeled data $\mathcal{D}_l$; unlabeled data $\mathcal{D}_u$



Two naïve (and wasteful) learning algorithms:

- **Supervised learning (SL):** Use **only** $\mathcal{D}_l$, and ignore $\mathcal{D}_u$
- **Unsupervised learning+ (UL+):**
  1) Use **only** $\mathcal{D}_u$ to learn decision boundary
  2) Assign labels to decision regions using $\mathcal{D}_l$

Empirical evidence: SSL algorithms can use $\mathcal{D}_l$ and $\mathcal{D}_u$ more effectively
$\Rightarrow$ lower prediction error than both (optimal) SL and UL+

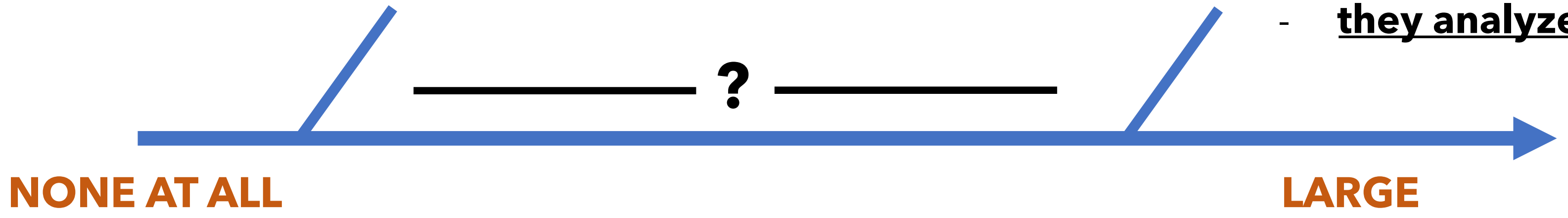
How fundamental is the improvement of SSL over SL and UL+?
e.g. rate improvement?

Comparing SSL to SL and UL+ for classification

Goal: Compare SSL, SL, UL+ via **error lower and upper bounds** for classification

Existing lower bounds for SSL:
- SSL achieves same rates as SL

Existing upper bounds for SSL:
- SSL improves over rates SL
- **they analyze UL+**



Amount of information about Y | X in unlabeled data $\mathcal{D}_u$

e.g. clusterability

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

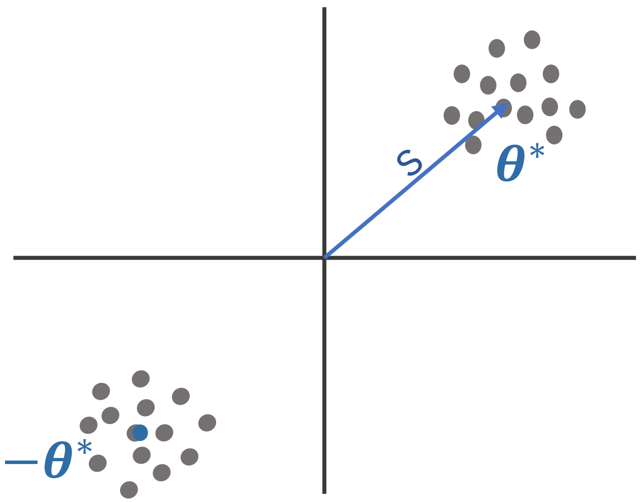
To answer, we need a tight minimax lower bound for SSL

Choosing a problem setting

Setting = class of distributions P_{XY}, class of algorithms

We want a setting such that:

1) Can vary the difficulty of the SSL task
   - *i.e. amount of information about Y | X captured in* $\mathcal{D}_u$
2) There exist known minimax rates for SL and UL+



Distributions: symmetric 2-GMM with isotropic components

$P_Y = Unif\{-1, 1\}$ and $P_{X|Y}^{\theta^*} = \mathcal{N}(Y\theta^*, I_d)$, with $||\theta^*||_2 = s$

Algorithms: $\mathcal{A}$ that outputs a linear classifier $\hat{\theta} = \mathcal{A}(\mathcal{D}_l, \mathcal{D}_u)$ i.e. $\hat{y} = sign(\langle\hat{\theta}, x\rangle)$

where $\mathcal{D}_l \sim \left(P_{XY}^{\theta^*}\right)^{n_l}$ and $\mathcal{D}_u \sim \left(P_X^{\theta^*}\right)^{n_u}$

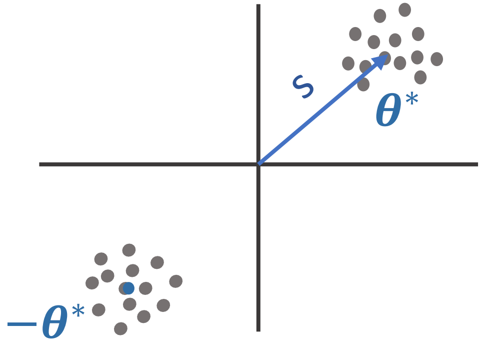
Minimax rate of SSL for 2-GMMs

Evaluation metric: $\mathcal{R}_{estim}(\theta, \theta^*) \coloneqq \|\theta - \theta^*\|_2$ (see paper for excess risk)

Theorem (Informal)

For large enough n_l, n_u and $s \in (0,1]$

$$\inf_{\mathcal{A}_{SSL}} \sup_{\|\theta^*\|=s} \mathbb{E}[\mathcal{R}_{estim}(\mathcal{A}_{SSL}(\mathcal{D}_l, \mathcal{D}_u), \theta^*)] \asymp \min\left\{s, \sqrt{\frac{d}{n_l + s^2 n_u}}\right\}$$



- $\asymp$ hides constant factors
- **dependence on s** allows to study the intermediate regime

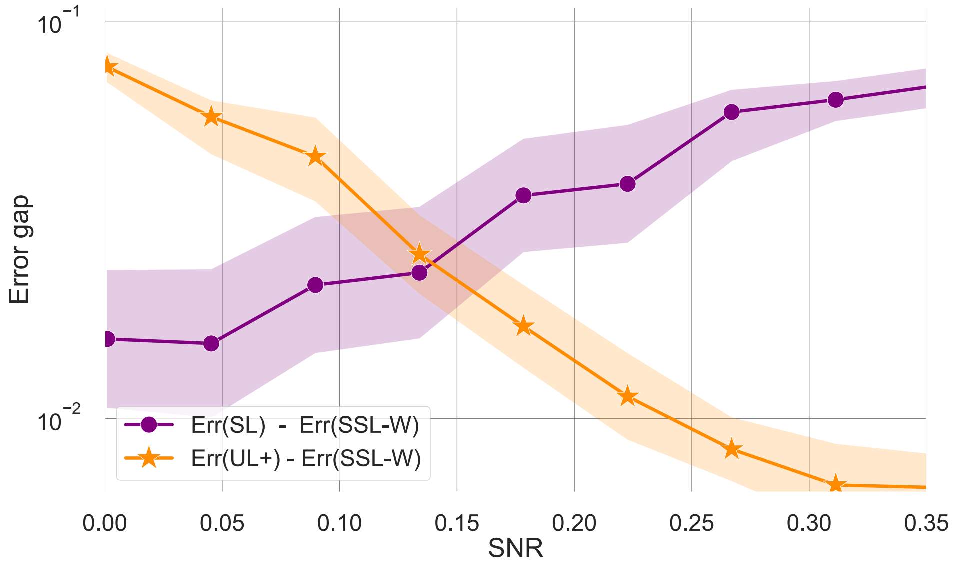
SL	UL+
$\sqrt{\frac{d}{n_l}}$	$\sqrt{\frac{d}{s^2 n_u}}$

Consequence: no SSL algorithm can achieve better rates than both SL and UL+ simultaneously

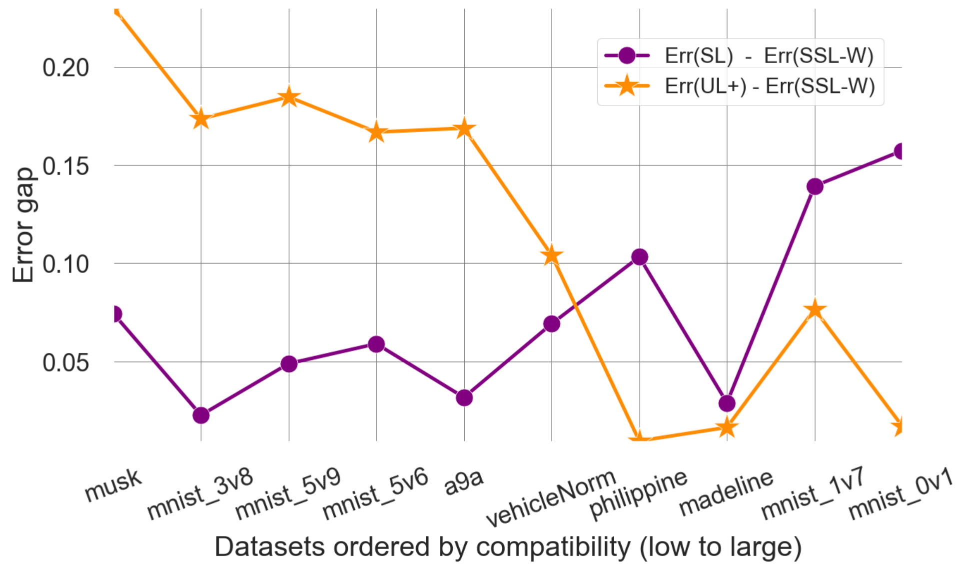
Conclusion & Future work

- No rate improvement possible with SSL over SL and UL+
- But is it possible to prove that certain SSL algorithms improve over both SL and UL+?

e.g. SSL-W is an SSL algorithm that uses both $\mathcal{D}_l$ and $\mathcal{D}_u$



2-GMM synthetic data



Real-world data

Thank you!

Overview of the exercises

Question 1: understand the setting of the question, apply Fano's method to the GMM setting with some modification to arrive at the final expression.

Question 2: understand the packing construction, develop intuition why specific packings are often better than uniform packings. Derive an upper bound on the KL divergence between the joint distributions of GMMs.

Question 3: analyze and understand results from the two previous exercise. Figure out how to put them together to obtain the final result. Make conclusions on the performance of semi-supervised learning algorithms in the GMM setting.