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

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the semi-supervised learning (SSL) setting both labeled and unlabeled datasets are available to the learning algorithm. While it is well-established from prior theoretical and empirical works that the inclusion of unlabeled data can help to improve over the error of supervised learning algorithms, existing theoretical examinations of SSL suggest a limitation: these algorithms might not efficiently leverage labeled data beyond a certain threshold. In this study, we derive a tight lower bound for 2-Gaussian mixture model distributions which exhibits an explicit dependence on the sizes of both the labeled and the unlabeled dataset. Surprisingly, our lower bound indicates that no SSL algorithm can surpass the sample complexities of minimax optimal supervised (SL) or unsupervised learning (UL) algorithms, which exclusively use either the labeled or the unlabelled dataset, respectively. Despite a change in the statistical error rate being unattainable, SSL can still outperform both SL and UL (up to permutation) in terms of absolute error. To this end, we provide evidence that there exist algorithms that can provably achieve lower error than both SL and UL algorithms. We validate our theoretical findings through linear classification experiments on synthetic and real-world data.

1 Introduction

Semi-Supervised Learning (SSL) has recently gained significant attention, often surpassing traditional supervised learning (SL) methods in practical applications [5, 8, 21]. Within this framework, the learning algorithm leverages both labeled and unlabeled datasets sampled from the same distribution. Numerous empirical studies suggest that SSL can effectively harness the joint information from both datasets, outperforming both SL and unsupervised learning (UL) approaches [20, 39, 16, 24]. This observation prompts the question: how fundamental is the improvement of SSL over SL and UL?

From a theoretical standpoint, this inquiry translates to determining if SSL algorithms genuinely showcase enhancements in statistical error rates compared to SL and UL, or if the improvements are simply of a constant factor. Our research focuses on this theoretical aspect in the context of linear classification. Specifically, we contrast lower and upper bounds of the SSL error with established rates for SL and UL for 2-Gaussian mixture models (GMMs) with two symmetrical components. This investigation revolves around the question:

Can semi-supervised classification algorithms simultaneously improve over the minimax rates of both SL and UL for 2-GMMs?

Previous upper bounds for SSL have focused on a regime where SSL improves the labeled sample complexity compared to SL, while matching the unlabeled sample complexity of UL algorithms [29, 30, 17]. In this regime, the unlabeled data (i.e. information about the marginal $P(X)$) contains information about the labeling function $P(Y|X)$. Conversely, prior lower bounds have been restricted to worst-case scenarios where SSL is equivalent to SL, where even oracle knowledge about the

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

marginal $P(X)$ fails to improve the error rates of SSL algorithms. In this regime, the marginal $P(X)$ does not carry any information about the labeling function $P(Y|X)$.

Intuitively, the utility of unlabeled data in SSL improving over SL hinges on the marginal distribution $P(X)$ carrying "any amount of" information about the conditional $P(Y|X)$. However, the above mentioned upper and lower bounds are insufficient for providing general insights into the statistical error rates of SSL since they focus on specific, disjoint, and extreme regimes. Therefore, in order to answer the aforementioned motivating question, we derive the minimax rates for SSL over 2-GMMs. As discussed in Section 3, the error rates are explicitly influenced by a specific measure – termed the Signal-to-Noise Ratio (SNR) – which quantifies the amount of information the marginal distribution $P(X)$ offers about the labeling function $P(Y|X)$. This allows us to analyze the whole spectrum of problem difficulties for 2-GMMs, rather than just the extremes.

Our main contribution is the finding that SSL cannot simultaneously improve over the statistical rates of both SL and UL. However, it is possible to improve upon the errors of SL and UL[1] by a constant factor. Appendix B provides guarantees for an algorithm that achieves lower error than both SL and UL algorithms. Finally, linear classification experiments on both synthetic and real-world datasets confirm our theoretical findings. Furthermore, our empirical analysis reveals that other commonly used SSL algorithms like self-training [38, 7] may also be able to improve over both SL and UL, underscoring the need for further theoretical analyses of these algorithms.

2 Problem setting and motivation

Before providing our main results, in this section, we discuss our problem setting, evaluation metrics, and the types of learning algorithms considered in this paper.

2.1 Linear classification for 2-GMM data

Data distribution. We consider linear binary classification problems where the data is drawn from a Gaussian Mixture Model consisting of two identical spherical gaussians with identity covariance and uniform mixing weights. The means of the two components $\boldsymbol{\theta}^*, -\boldsymbol{\theta}^*$ are symmetric with respect to the origin but can have arbitrary non-zero norm. We denote this family of distributions as $\mathcal{P}_{\text{2-GMM}} := \{P_{XY}^{\boldsymbol{\theta}^*} : \boldsymbol{\theta}^* \in \mathbb{R}^d\}$ where the joint probability is written as $P_{XY}^{\boldsymbol{\theta}^*}(X, Y) = P_{\boldsymbol{\theta}^*}(X|Y)P(Y)$ with

$$P(Y) = \text{Unif}\{-1, 1\} \text{ and } P_{\boldsymbol{\theta}^*}(X|Y) = \mathcal{N}(Y\boldsymbol{\theta}^*, I_d). \tag{1}$$

This family of distributions has often been considered in the context of analysing both SSL [29, 17] and SL/UL [2, 23, 37] algorithms. For $s \in (0, \infty)$, $\mathcal{P}_{\text{2-GMM}}^{(s)} \subset \mathcal{P}_{\text{2-GMM}}$ denotes the set of distributions $P_{XY}^{\boldsymbol{\theta}^*}$ with $\|\boldsymbol{\theta}^*\| = s$. We consider algorithms $\mathcal{A}$ that take as input a labeled dataset $\mathcal{D}_l \sim \left(P_{XY}^{\boldsymbol{\theta}^*}\right)^{n_l}$ of size n_l, an unlabeled dataset $\mathcal{D}_u \sim \left(P_{X}^{\boldsymbol{\theta}^*}\right)^{n_u}$ of size n_u, or both, and output an estimator $\hat{\boldsymbol{\theta}} = \mathcal{A}(\mathcal{D}_l, \mathcal{D}_u) \in \mathbb{R}^d$. The estimator is used to predict the label of a test point x as $\hat{y} = \text{sign}\left(\langle \hat{\boldsymbol{\theta}}, x \rangle\right)$.

Evaluation metrics In this work, we consider two natural error metrics for this class of problems: prediction error and parameter estimation error[2]. For an estimator $\hat{\boldsymbol{\theta}} = \mathcal{A}(\mathcal{D}_l, \mathcal{D}_u)$, we define

$$\textbf{Prediction error: } \mathcal{R}_{\text{pred}}\left(\mathcal{A}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}^{\boldsymbol{\theta}^*}\right) := P_{XY}^{\boldsymbol{\theta}^*}\left(\text{sign}\left(\langle \hat{\boldsymbol{\theta}}, X \rangle\right) \neq Y\right), \tag{2}$$

With a slight abuse of notation, we write $\mathcal{R}_{\text{pred}}\left(\boldsymbol{\theta}^*, P_{XY}^{\boldsymbol{\theta}^*}\right)$ to denote the prediction error of the Bayes optimal linear classifier $\boldsymbol{\theta}^*$. Since the distributions in $\mathcal{P}_{\text{2-GMM}}$ are not linearly separable, and hence suffer non-vanishing Bayes prediction error, we also consider the *excess* prediction error:

$$\textbf{Excess prediction error: } \mathcal{E}\left(\mathcal{A}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}^{\boldsymbol{\theta}^*}\right) := \mathcal{R}_{\text{pred}}\left(\mathcal{A}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}^{\boldsymbol{\theta}^*}\right) - \mathcal{R}_{\text{pred}}\left(\boldsymbol{\theta}^*, P_{XY}^{\boldsymbol{\theta}^*}\right).$$

For the set of all classification algorithms $\mathfrak{A}$, we study the minimax expected error over a family of distributions $\mathcal{P}$. This worst-case error over $\mathcal{P}$ indicates the limits of what is achievable with the algorithm class $\mathfrak{A}$. For instance, the minimax expected excess error of $\mathfrak{A}$ over $\mathcal{P}$ takes the form:

$$\textbf{Minimax excess error: } \epsilon(n_l, n_u, \mathcal{P}) := \inf_{\mathcal{A} \in \mathfrak{A}} \sup_{P_{XY} \in \mathcal{P}} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}\right)\right]. \tag{3}$$

2.2 Supervised, unsupervised, and semi-supervised learning

Based on the kind of available data, we distinguish between three kinds of learning settings and the associated algorithms. Although our discussion is confined to the context of learning under distributions in $\mathcal{P}_{\text{2-GMM}}$, the underlying intuitions are applicable to a broader set of problems.

[1]By referring to error of UL, we refer to prediction error up to sign, we formalise this as UL+

[2]See Appendix C for more details regarding the estimation error bounds.

1) SSL SSL algorithms, $\mathcal{A}_{\text{SSL}}$, utilise both labeled $\mathcal{D}_l$ and unlabeled samples $\mathcal{D}_u$ to produce an estimator $\hat{\boldsymbol{\theta}}_{\text{SSL}} = \mathcal{A}_{\text{SSL}}(\mathcal{D}_l, \mathcal{D}_u)$. The promise of SSL is that by combining labeled and unlabeled data SSL can reduce both the labeled and unlabeled sample complexities compared to algorithms that only use one either dataset. In Appendix A.1 we give an overview of past error bounds for SSL.

2) SL SL algorithms, represented by $\mathcal{A}_{\text{SL}}$, rely exclusively on the labeled dataset $\mathcal{D}_l$ to yield an estimator $\hat{\boldsymbol{\theta}}_{\text{SL}} = \mathcal{A}_{\text{SL}}(\mathcal{D}_l, \emptyset)$. The minimax rate of SL for distributions from $\mathcal{P}^{(s)}_{\text{2-GMM}}$ is known to be given by $\epsilon_{\text{SL}}\left(n_l, 0, \mathcal{P}^{(s)}_{\text{2-GMM}}\right) \asymp e^{-s^2/2}\frac{d}{sn_l}$ for excess risk [23] and $\epsilon_{\text{SL}}\left(n_l, 0, \mathcal{P}^{(s)}_{\text{2-GMM}}\right) \asymp \sqrt{\frac{d}{n_l}}$ for estimation error [3]. Both are achieved by the mean estimator $\hat{\boldsymbol{\theta}}_{\text{SL}} = \frac{1}{n_l}\sum_{i=1}^{n_l} Y_i X_i$.

3) UL UL algorithms, symbolised by $\mathcal{A}_{\text{UL}}$, employ only unlabeled data to identify underlying structures in the distribution. For distributions in $\mathcal{P}_{\text{2-GMM}}$, $\mathcal{A}_{\text{UL}}$ can identify the Gaussian components in the distribution, but without labeled data, it is unable to determine the class labels of the individual components. Formally, UL algorithms output a set of estimators $\left\{\hat{\boldsymbol{\theta}}_{\text{UL}}, -\hat{\boldsymbol{\theta}}_{\text{UL}}\right\} = \mathcal{A}_{\text{UL}}(\emptyset, \mathcal{D}_u)$ one of which is guaranteed to be close to the true $\boldsymbol{\theta}^*$. The minimax rate (up to permutation) of UL algorithms over $\mathcal{P}^{(s)}_{\text{2-GMM}}$ is given by $\epsilon_{\text{UL}}\left(0, n_u, \mathcal{P}^{(s)}_{\text{2-GMM}}\right) \asymp e^{-s^2/2}\frac{d}{s^3 n_u}$ for excess risk and $\epsilon_{\text{UL}}\left(0, n_u, \mathcal{P}^{(s)}_{\text{2-GMM}}\right) \asymp \sqrt{\frac{d}{s^2 n_u}}$ for estimation error [23, 37]. These rates are achieved by the unsupervised estimator $\hat{\boldsymbol{\theta}}_{\text{UL}} = \sqrt{(\hat{\lambda}-1)_+}\hat{v}$, where $(\hat{\lambda}, \hat{v})$ is the leading eigenpair of the sample covariance matrix $\hat{\Sigma} = \frac{1}{n_u}\sum_{j=0}^{n_u} X_j X_j^T$ and we use the notation $(x)_+ := \max(0, x)$.

To choose from the set $\left\{\hat{\boldsymbol{\theta}}_{\text{UL}}, -\hat{\boldsymbol{\theta}}_{\text{UL}}\right\}$, one can use a two-stage approach: i) run a UL algorithm $\mathcal{A}_{\text{UL}}$ to estimate $\boldsymbol{\theta}^*$ up to sign; then ii) use labeled data to select the best sign, e.g. via majority voting. We refer to this class of two-stage algorithms as **UL+** , and denote it by $\mathcal{A}_{\text{UL+}}$. These algorithms operate essentially in the same setting as SSL. Both $\mathcal{D}_l$ and $\mathcal{D}_u$ are available; however, labeled data is exclusively used to ascertain the sign (or permutation of labels) of the estimator obtained using unlabeled data. Several early analyses of semi-supervised learning focus, in fact, on algorithms that fit the description of UL+ [29, 30].

UL+ algorithms are "wasteful" SSL algorithms. As described above, UL+ algorithms follow a precise structure where labeled data is used solely to select from the set of estimators output by a UL algorithm. This approach, however, may not always achieve optimal error. Consider a scenario where n_u is finite, but $n_l \to \infty$. The error of a UL+ algorithm will, at best, mirror the error of a UL algorithm with the correct sign (e.g. $\Theta(d/n_u)$ for the excess risk). However, a more effective use of the labeled dataset would be to employ a consistent SL or SSL algorithm, like self-training [38, 9, 17], to obtain vanishing excess risk. Thus, despite using both labeled and unlabeled data, UL+ algorithms bear a close resemblance to UL algorithms that only use unlabeled data.

2.3 Improvement rates for SSL

To understand whether an SSL algorithm is using the labeled and unlabeled data effectively, we compare the error rate of SSL algorithms to the minimax rates for SL and UL+ algorithms.

Definition 1 (SSL improvement rates)**.** *For a family of distributions $\mathcal{P}$, we define the improvement rates of SSL over SL and UL+ as h_l and h_u, respectively, where*

$$h_l(n_l, n_u, \mathcal{P}) := \frac{\inf_{\mathcal{A}_{SSL}} \sup_{P_{XY}\in\mathcal{P}} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{SSL}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}\right)\right]}{\inf_{\mathcal{A}_{SL}} \sup_{P_{XY}\in\mathcal{P}} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{SL}(\mathcal{D}_l, \emptyset), P_{XY}\right)\right]}, \tag{4}$$

$$h_u(n_l, n_u, \mathcal{P}) := \frac{\inf_{\mathcal{A}_{SSL}} \sup_{P_{XY}\in\mathcal{P}} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{SSL}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}\right)\right]}{\inf_{\mathcal{A}_{UL+}} \sup_{P_{XY}\in\mathcal{P}} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{UL+}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}\right)\right]}, \tag{5}$$

where the expectations are over $\mathcal{D}_l \sim P_{XY}^{n_l}$ and $\mathcal{D}_u \sim P_X^{n_u}$.

To simplify notation, we denote the improvement rates of SL and UL+ over $\mathcal{P}^{(s)}_{\text{2-GMM}}$ as $h_l(n_l, n_u, s)$ and $h_u(n_l, n_u, s)$, respectively. For SSL to demonstrate an enhanced error rate over SL and UL+, the conditions $\lim_{n_l, n_u\to\infty} h_l(n_l, n_u, \mathcal{P}) = 0$ and $\lim_{n_l, n_u\to\infty} h_u(n_l, n_u, \mathcal{P}) = 0$ must be satisfied.

[3]The notation $f(x) \asymp g(x)$ is equivalent to $f = \Theta(g)$.

SNR Regime	Rate of growth of n_u vs n_l	$h_l(n_l, n_u, s)$	$h_u(n_l, n_u, s)$
$s = o\left(\sqrt{1/n_u}\right)$	Any	c_{SL}	0
fixed $s > 0$	$n_u = o(n_l)$	c_{SL}	0
	$n_u = \omega(n_l)$	0	c_{UL}
	$\lim_{n_l, n_u \to \infty} \frac{n_u}{n_l} = c$	$\left(\frac{1}{1+cs^2}\right) c_{\text{SL}}$	$\left(\frac{s^2 c}{1+s^2 c}\right) c_{\text{UL}}$

Table 1: SSL improvement rates over SL and UL+ for different regimes of s and n_u, where h_l, h_u are evaluated for $\lim_{n_l, n_u \to \infty}$. c_{SL} and c_{UL} denote constants.

3 Minimax rates for SSL

In this section we provide tight minimax lower bounds for SSL algorithms and 2-GMM distributions in $\mathcal{P}^{(s)}_{\text{2-GMM}}$. Our results indicate that it is, in fact, not possible for SSL algorithms to simultaneously achieve faster minimax rates than both SL and UL+.

3.1 Excess risk minimax rate

We present a tight lower bound on the excess risk of a linear estimator obtained using both labeled and unlabeled data. The formal conditions required by the theorem as well as the proofs of the lower and upper bounds can be found in Appendix E.

Theorem 1 (SSL Minimax Rate for Excess Risk)**.** *Let $P^{\boldsymbol{\theta}^*}_{XY}$ be a distribution from $\mathcal{P}^{(s)}_{\text{2-GMM}}$. For any $s \in (0, 1]$, sufficiently large d and $d < n_l < n_u$, we have*

$$\inf_{\mathcal{A}_{SSL}} \sup_{\|\boldsymbol{\theta}^*\| = s} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{SSL}\left(\mathcal{D}_l, \mathcal{D}_u\right), P^{\boldsymbol{\theta}^*}_{XY}\right)\right] \asymp e^{-s^2/2} \min\left\{s, \frac{d}{sn_l + s^3 n_u}\right\}, \tag{6}$$

where the infimum is over all the possible SSL algorithms that have access to both unlabeled and labeled data and the expectation is over $\mathcal{D}_l \sim \left(P^{\boldsymbol{\theta}^*}_{XY}\right)^{n_l}$ and $\mathcal{D}_u \sim \left(P^{\boldsymbol{\theta}^*}_{X}\right)^{n_u}$.

A direct implication of the theorem is that $\epsilon_{\text{SSL}}\left(n_l, n_u, \mathcal{P}^{(s)}_{\text{2-GMM}}\right) \asymp \min\left(\epsilon_{\text{SL}}\left(n_l, 0, \mathcal{P}^{(s)}_{\text{2-GMM}}\right), \epsilon_{\text{UL+}}\left(n_l, n_u, \mathcal{P}^{(s)}_{\text{2-GMM}}\right)\right)$, i.e. the minimax rate of SSL is the same as either that of SL or UL+, depending on the values of s, n_u and n_l. We can conclude the following.

Remark 1. *No SSL algorithm can improve the rates* ***of both*** *SL and UL+ for $P_{XY} \in \mathcal{P}^{(s)}_{\text{2-GMM}}$.*

In order to prove the theorem, we derive both a minimax lower bound for SSL, and a matching upper bound. The proof of the upper bound is constructive. The algorithm that achieves the upper bound simply chooses between using a (minimax optimal) SL or UL+ algorithm based on the values of s, n_l, and n_u, as shown in Algorithm 2. We call this the **SSL Switching Algorithm (SSL-S)**.

While the *rates* of either SL or UL+ cannot be improved further using SSL algorithms, it is nonetheless possible to improve the error by a constant factor, independent of n_l and n_u. To see this, in Appendix B we describe an algorithm that uses both $\mathcal{D}_l$ and $\mathcal{D}_u$ effectively and can hence achieve a provable improvement in error over both SL and UL+.

3.1.1 Fine-grained analysis of different improvement regimes for SSL

The observation in Remark 1 can be made formal using the improvement rates from Definition 1.

Corollary 1. *Assuming the setting of Theorem 1, the improvement rates of SSL can be written as:*

$$\textbf{\textit{Improvement rate over SL:}}\ h_l\left(n_l, n_u, s\right) \asymp \frac{n_l}{n_l + s^2 n_u}. \tag{7}$$

$$\textbf{\textit{Improvement rate over UL+:}}\ h_u\left(n_l, n_u, s\right) \asymp \frac{s^2 n_u}{n_l + s^2 n_u}. \tag{8}$$

We distinguish between the different scenario summarized in Table 1, based on the nature of the rate improvement over SL and UL+. Noticeably, SSL cannot achieve better rates than both UL+ and SL at the same time since there is no regime for which h_l and h_u are simultaneously 0.

4 Conclusions and limitations

In this study, we demonstrate that SSL cannot simultaneously improve the error rates of both SL and UL across all signal-to-noise ratios. Our theoretical analysis focuses exclusively on isotropic and symmetric GMMs due to limitations in the technical tools used for the proofs. Similar constraints can be observed in recent examinations of SL or UL algorithms [23, 37].

References

[1] Martin Azizyan, Aarti Singh, and Larry Wasserman. Density-sensitive semisupervised inference. *The Annals of Statistics*, 2013.

[2] Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, 2013.

[3] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 2017.

[4] Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 2010.

[5] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv:2304.12210*, 2023.

[6] Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In *Annual Conference on Learning Theory (COLT)*, 2008.

[7] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Annual Conference on Learning Theory (COLT)*, 1998.

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems 33*, 2020.

[9] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems 33*, 2020.

[10] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2006.

[11] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two gaussians. In *Annual Conference on Learning Theory (COLT)*, 2017.

[12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012.

[13] Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Michael I. Jordan, Martin J. Wainwright, and Bin Yu. Singularity, misspecification and the convergence rate of em. *The Annals of Statistics*, 2018.

[14] Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J Wainwright, and Michael I Jordan. Theoretical guarantees for EM under misspecified gaussian mixture models. In *Advances in Neural Information Processing Systems 31*, 2018.

[15] Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin Wainwright, Michael Jordan, and Bin Yu. Sharp analysis of expectation-maximization for weakly identifiable models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[16] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv:2112.10740*, 2021.

[17] Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[18] Christophe Giraud. *Introduction to high-dimensional statistics*. CRC Press, 2021.

[19] Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya O. Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate? In *Annual Conference Computational Learning Theory*, 2019.

[20] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *International Conference on Computer Vision*, 2019.

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems 33*, 2020.

[22] Yann LeCun and Corinna Cortes. MNIST handwritten digit database, 2010. URL `http://yann.lecun.com/exdb/mnist/`.

[23] Tianyang Li, Xinyang Yi, Constantine Carmanis, and Pradeep Ravikumar. Minimax Gaussian Classification & Clustering. In *20th International Conference on Artificial Intelligence and Statistics*, 2017.

[24] Thomas Lucas, Philippe Weinzaepfel, and Gregory Rogez. Barely-supervised learning: semi-supervised learning with very few labeled images. In *AAAI Conference on Artificial Intelligence*, 2022.

[25] Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[26] Alexander Mey and Marco Loog. Improvability through semi-supervised learning: A survey of theoretical results. *arXiv:1908.09574*, 2020.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[28] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 1999.

[29] Joel Ratsaby and Santosh S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Annual Conference on Learning Theory (COLT)*, 1995.

[30] Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 2006.

[31] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning*, 2012.

[32] Aarti Singh, Robert D. Nowak, and Xiaojin Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems 21*, 2008.

[33] Erixhen Sula and Lizhong Zheng. On the semi-supervised expectation maximization. *arXiv:2211.00537*, 2022.

[34] Ilya O. Tolstikhin and David Lopez-Paz. Minimax lower bounds for realizable transductive classification. *arXiv:1602.03027*, 2016.

[35] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *SIGKDD Explorations*, 2013.

[36] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

[37] Yihong Wu and Harrison H. Zhou. Randomly initialized EM algorithm for two-component gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 2021.

[38] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting on Association for Computational Linguistics*, 1995.

[39] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In *Advances in Neural Information Processing Systems 33*, 2020.

A Related work

Other theoretical analyses of SSL algorithms. Beyond the theoretical studies highlighted in Section 2, there are a few others pertinent to our research. Specifically, Azizyan et al. [1], Singh et al. [32] present upper bounds for semi-supervised regression, which are contingent on the degree to which the marginal P_X informs the labeling function. This is akin to the results we derive in this work. However, obtaining a minimax lower bound for semi-supervised regression remains an exciting direction for future work. We refer to [26] for an overview of prior theoretical results for SSL.

Balcan and Blum [4] introduced a compatibility score, denoted as $\chi(f, P_X) \in [0, 1]$, which connects the space of marginal distributions to the space of labeling functions. While their findings hint that SSL may surpass the SL minimax rates, they offer no comparisons with UL/UL+. Moreover, the paper does not discuss minimax optimality of the proposed SSL algorithms.

On another note, even though SSL does not enhance the rates of UL, Sula and Zheng [33] demonstrate that labeled samples can bolster the convergence speed of Expectation-Maximization within the context of our study.

To conclude, Schölkopf et al. [31] leveraged a causality framework to pinpoint scenarios where SSL does not offer any advantage over SL. In essence, when the covariates, represented by X, act as causal ancestors to the labels Y, the independent causal mechanism assumption dictates that the marginal P_X offers no insights about the labeling function.

Minimax rates for SL and UL. The proofs in this work rely on techniques used to derive minimax rates for SL and UL algorithms. Most of these prior results consider the same distributional assumptions as our paper. Wu and Zhou [37] show a tight minimax lower bound for estimation error for spherical 2-GMMs from $\mathcal{P}_{\text{2-GMM}}$. Moreover, Azizyan et al. [2], Li et al. [23] derive minimax rates over $\mathcal{P}_{\text{2-GMM}}$ for classification and clustering (up to permutation).

In addition to the SL and UL algorithms considered in Section 3, Expectation-Maximization (EM) is another family of algorithms that is commonly analyzed for the same distributional setting considered in our paper. For instance, Wu and Zhou [37] rely on techniques from several previous seminal papers [11, 3, 13–15] to obtain upper bounds for EM-style algorithms.

A.1 Brief overview of prior error bounds for SSL

Upper bounds. The optimal condition for SSL is when both h_l and h_u approach zero as $n_l \to \infty$. There are numerous known upper bounds on the excess risk of SSL algorithms for $\mathcal{P}_{\text{2-GMM}}$ distributions. Nevertheless, existing results fall short of establishing that SSL algorithms can consistently outperform both SL and UL+. Earlier bounds primarily match the UL+ minimax rates [29, 30] or exhibit slower rates than UL+ [17]. In this work, we aim to discern if SSL can ever excel over the minimax rates of both SL and UL+ within the $\mathcal{P}_{\text{2-GMM}}$ distribution family.

Lower bounds. To our knowledge, three distinct minimax lower bounds for SSL have been proposed. Each suggests that there exists a distribution P_{XY} where SSL cannot outperform the SL minimax rate. Ben-David et al. [6] substantiate this claim for learning thresholds from univariate data sourced from a uniform distribution on $[0, 1]$. Göpfert et al. [19] expand upon this by considering arbitrary marginal distributions P_X and a "rich" set of realizable labeling functions, such that no volume of unlabeled data can differentiate between possible hypotheses. Lastly, Tolstikhin and Lopez-Paz [34] set a lower bound for scenarios with no implied association between the labeling function and the marginal distribution, a condition recognized as being unfavorable for SSL improvements [31].

Each of the aforementioned results contends that a particular worst-case distribution P_{XY} exists, where the labeled sample complexity for SSL matches that of SL, even with limitless unlabeled data. Within the spherical 2-GMM distributions $\mathcal{P}_{\text{2-GMM}}^{(s)}$ with $\|\boldsymbol{\theta}^*\| = s$, this "hard" setting (where SSL and SL rates are equivalent) emerges for extremely low SNR s. Further insights on this topic are available in Section 3.1.1. Prior lower bounds do not capture other levels of the SNR s, and hence, cannot predict the best achievable error rate with SSL algorithms for moderate or large s.

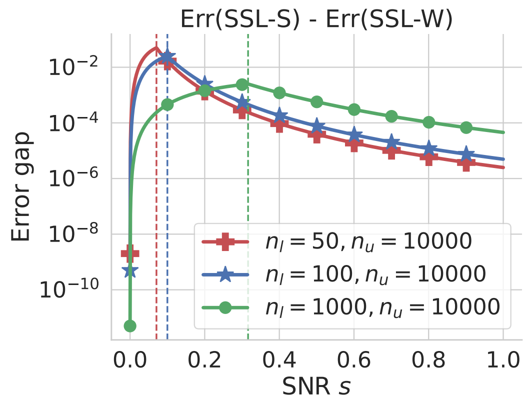


Figure 1: Estimation error gap between SSL-S and SSL-W as revealed by Theorem 2 for varying SNR and n_l ($n_u = 10000$). The maximum gap is reached at the switching point, indicated by the vertical dashed lines.

B Finding better SSL algorithms

Section 3 shows that a simple algorithm that switches between the optimal SL and the optimal UL+ algorithm achieves the minimax SSL rates discussed in Theorem 2. However, the SSL Switching algorithm, albeit optimal in terms of rates, does not take full advantage of all the available data – it either uses only the labeled data for SL, or the unlabeled data and a small fraction of labeled samples for UL+ .

In this section we describe a simple algorithm that has the desirable property that it utilises all the data at its disposal. We argue that this algorithm can lead to strictly lower error than the SSL-S algorithm. Unsurprisingly, this improvement is only in the constants and not in the actual learning rate for which Algorithm 2 is already minimax optimal. We show experimentally that the proposed algorithm, as well as other SSL algorithms such as self-training [38], can improve over the error of SSL-S on synthetic and real-world data. It remains an exciting direction for future work to characterize the exact improvement of self-training algorithms over SL and UL+.

B.1 A weighted ensemble of $\hat{\boldsymbol{\theta}}_{\text{UL+}}$ and $\hat{\boldsymbol{\theta}}_{\text{SL}}$

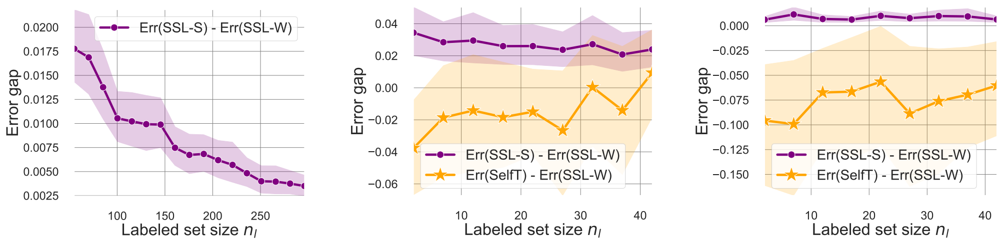
A natural means to use both the labeled and unlabeled datasets in an SSL algorithm is to construct an ensemble of an SL and a UL+ estimator, trained on $\mathcal{D}_l$ and $\mathcal{D}_u$, respectively, where the influence of each estimator on the final prediction is controlled by a hyperparameter t. We call this the **SSL Weighted algorithm (SSL-W)** shown in Algorithm 1. With an appropriate choice of the weight t, it is possible to show that the performance of the SSL-W algorithm is better (up to sign permutation) than SSL-S. In practice, one can fix the sign permutation of the $\hat{\boldsymbol{\theta}}_{\text{SSL-W}}$ estimator using a small amount of labeled data. The formal statement of this result together with the proof are deferred to Appendix F. The intuition for this improvement is that the ensemble estimator $\hat{\boldsymbol{\theta}}_{\text{SSL-W}}$ achieves better error than the individual estimators that are part of the ensemble (i.e. $\hat{\boldsymbol{\theta}}_{\text{SL}}$ and $\hat{\boldsymbol{\theta}}_{\text{UL+}}$), which, in turn, determine the error of the SSL-S algorithm.

Algorithm 1: SSL-W algorithm

Input : $\mathcal{D}_l, \mathcal{D}_u, t$
Result: $\hat{\boldsymbol{\theta}}_{\text{SSL-W}}$
$\hat{\boldsymbol{\theta}}_{\text{SL}} \leftarrow \mathcal{A}_{\text{SL}}(\mathcal{D}_l)$
$\hat{\boldsymbol{\theta}}_{\text{UL+}} \leftarrow \mathcal{A}_{\text{UL+}}(\mathcal{D}_l, \mathcal{D}_u)$
$\hat{\boldsymbol{\theta}}_{\text{SSL-W}}(t) = t\hat{\boldsymbol{\theta}}_{\text{SL}} + (1-t)\hat{\boldsymbol{\theta}}_{\text{UL+}}$
return $\hat{\boldsymbol{\theta}}_{\text{SSL-W}}(t)$

B.2 Empirical improvements over SSL Switching Algorithm

In this section we present linear classification experiments on synthetic and real-world data to show that there indeed exist SSL algorithms that can improve over the error of the SSL Switching Algorithm. For both synthetic and real-world data, we use $\hat{\boldsymbol{\theta}}_{\text{SL}} = \frac{1}{n_l}\sum_{i=1}^{n_l} Y_i X_i$ as the SL estimator and an Expectation-Maximization (EM) algorithm for the UL method (see Appendix G for implementation details). The optimal switching point for SSL-S and the optimal weight for SSL-W, as well as the optimal ℓ_2 penalty for logistic regression are chosen using a holdout validation set.



(a) 2-GMM data with $s = 0.1$ (b) VehicleNorm dataset (c) MNIST classes 3 vs 8

Figure 2: Error gap between SSL-S/self-training and SSL-W on synthetic and real-world datasets. The positive gap indicates that SSL-W and self-training outperform SSL-S (and hence, also SL and UL+) for a broad range of n_l values. See Appendix H for more datasets.

Synthetic data. We consider data drawn from symmetric and isotropic 2-GMM distributions $P_{XY}^{\boldsymbol{\theta}^*}$ over $\mathbb{R}^2$. The unlabeled set size is set to 5000 and we vary the SNR s and the labeled set size n_l. Figures 2a and 3a show the gap between the SSL algorithms (i.e. SSL-W, SSL-S) and SL or UL+ as a function of the SNR s and the labeled set size n_l, respectively. There are two main takeaways. First, for varying s and n_l, SSL-W always outperforms SL and UL+, and hence, also SSL-S, as suggested in Appendix B.1. Second, as argued in Section 3.1.1, SSL-S improves more over UL+ for small values of the SNR s, and it improves more over SL for large values of the SNR.

Real-world data. We consider 10 binary classification real-world datasets: five from the OpenML repository [35] and five 2-class subsets of the MNIST dataset [12]. For the MNIST subsets, we choose class pairs that have a linear Bayes error varying between 0.1% and 2.5%.[4] We choose from OpenML datasets that have a large enough number of samples compared to dimensionality (see Appendix G for details on how we choose the datasets). The OpenML datasets span a range of Bayes errors that varies between 3% and 34%.

In the absence of the exact data generating process, we quantify the SNR of the real-world datasets using the fraction of the Bayes error that is captured by UL using the spherical and symmetrical 2-GMM parametric assumption for the distribution. More specifically, we use SNR $= \frac{\mathcal{R}_{\text{pred}}(\boldsymbol{\theta}^*_{UL}) - \mathcal{R}_{\text{pred}}(\boldsymbol{\theta}^*_{\text{Bayes}})}{\mathcal{R}_{\text{pred}}(\boldsymbol{\theta}^*_{\text{Bayes}})\sqrt{d}}$, where d is the dimension of the data, $\boldsymbol{\theta}^*_{\text{Bayes}}$ is obtained via SL on the entire dataset and $\boldsymbol{\theta}^*_{\text{UL}}$ determines the predictor with optimal sign obtained via UL on the entire dataset.

In addition to SSL-S (Algorithm 2) and SSL-W (Algorithm 1) we also evaluate the performance of self-training, using a procedure similar to the one analyzed in Frei et al. [17]. We use a logistic regression estimator for the pseudolabeling, and train logistic regression with a ridge penalty in the second stage of the self-training procedure. Note that an ℓ_2 penalty corresponds to input consistency regularization [36] with respect to ℓ_2 perturbations.

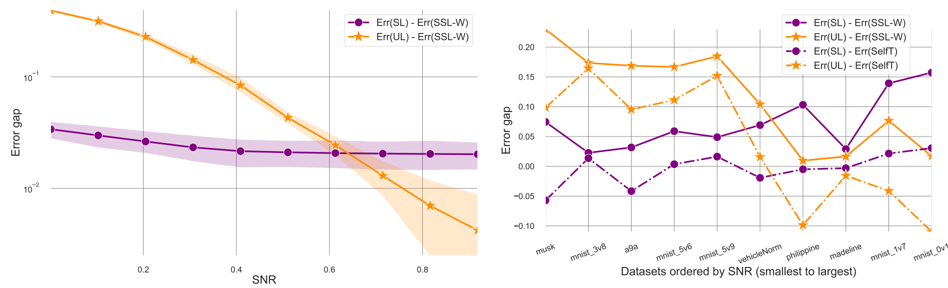
Figure 3 shows the improvement in classification error of SSL algorithms (i.e. SSL-W and self-training) compared to SL and UL+ . Figure 2 shows the gap between SSL-W (or self-training) and SSL-S as the size of the labeled set varies. There is a broad spectrum of n_l values for which the gap is positive indicating that it is indeed possible to improve over the SSL Switching algorithm even for data that does not follow the 2-GMM distribution that we consider in the theoretical analysis.

Furthermore, Figure 3 shows that the gap between SSL-W (or self-training) and SL or UL follows the same trends as the synthetic experiments in Figure 3a. This finding suggests that the intuition presented in Appendix B.1 carries over to more generic distributions, beyond just 2-GMMs.

C Parameter estimation error minimax rate

Beyond the tight lower bound on the excess risk we detailed in Section 3.1, we also formulate a lower bound on the estimation error for the means of class-conditional distributions. This is especially relevant when addressing linear classification of symmetric and spherical GMMs. In this setting, a

[4] We estimate the Bayes error of a dataset by training a linear classifier on the entire labeled dataset.



(a) 2-GMM synthetic data.

(b) Real-world data.

Figure 3: Error gap between SL or UL and SSL-W for varying SNR. We see the same trends for both synthetic and real-world data. Moreover, self-training also exhibits the same trend as $\hat{\boldsymbol{\theta}}_{\text{SSL-W}}$.

reduced estimation error points to not only a low excess risk but also suggests a small calibration error under the assumption of a logistic noise model [28]. The trend suggested by this result mirrors that of Theorem 1, and the arguments presented in Section 3.1.1 also remain applicable to the estimation error minimax rates. Similar to Theorem 1, an optimal algorithm that matches the minimax error rate is the SSL Switching algorithm presented in Algorithm 2. The formal conditions required for the theorem to hold as well as the proofs can be found in Appendix D.

Let us define the estimation error as follows:

$$\textbf{Estimation error: } \mathcal{R}_{\text{estim}}\left(\mathcal{A}\left(\mathcal{D}_l, \mathcal{D}_u\right), P_{XY}^{\boldsymbol{\theta}^*}\right) := \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2^2. \tag{9}$$

Theorem 2 (SSL Minimax Rate for Parameter Estimation)**.** *Let $P_{XY}^{\boldsymbol{\theta}^*}$ be a distribution from $\mathcal{P}_{\text{2-GMM}}^{(s)}$. For any $s \in (0, 1]$, $d \geq 2$, and sufficiently large n_l and n_u, we have*

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

where the infimum is over all the possible SSL algorithms and the expectation is over $\mathcal{D}_l \sim \left(P_{XY}^{\boldsymbol{\theta}^*}\right)^{n_l}$ and $\mathcal{D}_u \sim \left(P_X^{\boldsymbol{\theta}^*}\right)^{n_u}$.

C.1 Proof sketch

For the estimation error lower bound, we use Fano's method with the packing construction in Wu and Zhou [37], who have employed this method to derive lower bounds in the context of unsupervised learning. Similarly, for the excess risk we adopt the packing construction in Li et al. [23]. Directly applying Fano's method to derive the lower bound for the excess risk poses a challenge, given that the excess risk does not conform to the traditional framework of a (distribution-independent) metric. To overcome this challenge, we use techniques introduced in Azizyan et al. [2]. These mathematical tools make it possible to reduce the estimation problem to hypothesis testing by only using a property reminiscent of the triangle inequality instead of metric axioms.

Since the algorithms have access to both labeled and unlabeled datasets in the semi-supervised setting, KL-divergences between the marignal and the joint distributions show up together in the lower bound after the application of Fano's method, which is the key difference from its SL and UL counterparts.

The lower bounds reveal that the SSL rate is either determined by the SL rate or the UL+ rate depending on s and the ratio of the sizes of the labeled and unlabeled samples. Hence, it follows that an algorithm that chooses between an SL and an UL+ algorithm can match the minimax error rate for SSL, for an appropriate choice of the switching point, that depends on s, n_l and n_u. We further show that selecting the optimal sign for the estimator returned by running UL using labeled samples only adds an exponential term to the UL upper bound.

D Proof of Theorem 2

In this section we provide the proofs for the lower and upper bounds on the estimation error presented in Theorem 2. We formalize the conditions under which Theorem 2 holds in the following assumption: $d \geq 2$, $n_u > O(\frac{d}{s^2})$ and $n_l > O(\frac{\log n_u}{s^2})$.

D.1 Proof of lower bound

We first prove the estimation error lower bound in Theorem 2. As discussed in Section 2, consider the 2-GMM distributions from $\mathcal{P}_{\text{2-GMM}}^{(s)}$, with isotropic components and identical covariance matrices.

Consider an arbitrary set of predictors $\mathcal{M} = \{\boldsymbol{\theta}_i\}_{i=0}^{M}$ and . We can apply Fano's method [10] to obtain that the following holds:

Algorithm 2: SSL-S algorithm

Input : $\mathcal{D}_l$, $\mathcal{D}_u$, s, $\mathcal{A}_{\text{SL}}$, $\mathcal{A}_{\text{UL+}}$
Result: $\hat{\boldsymbol{\theta}}_{\text{SSL-S}}$
$\hat{\boldsymbol{\theta}}_{\text{SL}} \leftarrow \mathcal{A}_{\text{SL}}(\mathcal{D}_l)$
$\hat{\boldsymbol{\theta}}_{\text{UL+}} \leftarrow \mathcal{A}_{\text{UL+}}(\mathcal{D}_u, \mathcal{D}_l)$
if $s \leq \min\left\{\sqrt{\frac{d}{n_l}}, \left(\frac{d}{n_u}\right)^{1/4}\right\}$
  $\hat{\boldsymbol{\theta}}_{\text{SSL-S}} = 0$
else if $\min\left\{\sqrt{\frac{d}{n_l}}, \left(\frac{d}{n_u}\right)^{1/4}\right\} < s \leq \sqrt{\frac{n_l}{n_u}}$
  $\hat{\boldsymbol{\theta}}_{\text{SSL-S}} = \hat{\boldsymbol{\theta}}_{\text{SL}}$
else
  $\hat{\boldsymbol{\theta}}_{\text{SSL-S}} = \hat{\boldsymbol{\theta}}_{\text{UL+}}$
return $\hat{\boldsymbol{\theta}}_{\text{SSL-S}}$

$$\inf_{\mathcal{A}_{\text{SSL}}} \sup_{\|\boldsymbol{\theta}^*\|=s} \mathbb{E}_{\mathcal{D}_l,\mathcal{D}_u}\left[\mathcal{R}_{\text{estim}}(\mathcal{A}_{\text{SSL}}(\mathcal{D}_l,\mathcal{D}_u), P_{XY}^{\boldsymbol{\theta}^*})\right]$$
$$\geq \frac{1}{2}\min_{\substack{i,j\in[M]\\ i\neq j}}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_j\|\left(1-\frac{1+\frac{1}{M}\sum_{i=1}^{M} D\left({P_{XY}^{\boldsymbol{\theta}_i}}^{n_l}\, {P_X^{\boldsymbol{\theta}_i}}^{n_u}\|{P_{XY}^{\boldsymbol{\theta}_0}}^{n_l}\, {P_X^{\boldsymbol{\theta}_0}}^{n_u}\right)}{log(M)}\right)$$
$$= \frac{1}{2}\min_{\substack{i,j\in[M]\\ i\neq j}}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_j\|\left(1-\frac{1+\frac{1}{M}\sum_{i=1}^{M} n_l D\left(P_{XY}^{\boldsymbol{\theta}_i}\|P_{XY}^{\boldsymbol{\theta}_0}\right)+n_u D\left(P_X^{\boldsymbol{\theta}_i}\|P_X^{\boldsymbol{\theta}_0}\right)}{log(M)}\right) \tag{10}$$
$$\geq \frac{1}{2}\min_{\substack{i,j\in[M]\\ i\neq j}}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_j\|\left(1-\frac{1+n_l \max_{i\in[M]} D\left(P_{XY}^{\boldsymbol{\theta}_i}\|P_{XY}^{\boldsymbol{\theta}_0}\right)+n_u \max_{i\in[M]} D\left(P_X^{\boldsymbol{\theta}_i}\|P_X^{\boldsymbol{\theta}_0}\right)}{log(M)}\right), \tag{11}$$

where $D(\cdot||\cdot)$ denotes the KL divergence. In Equation (10), we use the fact that the labeled and unlabeled samples are drawn i.i.d. from P_X and P_{XY} and in Equation (11) we upper bound the average with the maximum. The next step of the proof consists in choosing an appropriate packing $\{\boldsymbol{\theta}_i\}_{i=1}^{M}$ and $\boldsymbol{\theta}_0$ on the sphere of radius s, i.e. $\frac{1}{s}\boldsymbol{\theta}_i \in S^{d-1}$, that optimizes the trade-off between zhe minimum and the maxima in Equation (11).

For the packing, we use the same construction that was employed by Wu and Zhou [37] for deriving adaptive bounds for unsupervised learning. This construction has the advantage that it also leads to a tight lower bound for the supervised setting. Let c_0 and C_0 be positive absolute constants and let $\tilde{\mathcal{M}} = \{\psi_1, ..., \psi_M\}$ be a c_0-net on the unit sphere S^{d-2} such that we have $|\tilde{\mathcal{M}}| = M \geq e^{C_0 d}$. For an absolute constant $\alpha \in [0,1]$, we construct the following packing of the sphere of radius s in $\mathbb{R}^d$:

$$\mathcal{M} = \left\{\boldsymbol{\theta}_i = s\begin{bmatrix}\sqrt{1-\alpha^2}\\ \alpha\psi_i\end{bmatrix}\middle|\psi_i \in \tilde{\mathcal{M}}\right\},$$

and define $\boldsymbol{\theta}_0 = [s, 0, ..., 0]$. Note that, by definition, $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \geq c_0 s\alpha$, for any distinct $i, j \in [M]$, which lower bounds the first term in (11). Furthermore, $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_0\| \leq \sqrt{2}\alpha s$, for all $i \in [M]$.

In the next step, we upper bound the maxima in Equation (11). First, we write the KL divergence between two GMMs with identitiy covariance matrices: we have that

$$D\left(P_{XY}^{\boldsymbol{\theta}_i}||P_{XY}^{\boldsymbol{\theta}_0}\right) = \frac{1}{2}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_0\|_2^2 \leq \alpha^2 s^2, \text{ for all } i = [M]. \tag{12}$$

Second, we can upper bound the KL divergence between marginal distributions, namely $D\left(P_X^{\boldsymbol{\theta}_i}||P_X^{\boldsymbol{\theta}_0}\right)$, using Lemma 27 in Wu and Zhou [37], which implies that:

$$\max_{i\in[M]} D\left(P_X^{\boldsymbol{\theta}_i}||P_X^{\boldsymbol{\theta}_0}\right) \leq C \max_{i\in[M]} \|\frac{1}{s}\boldsymbol{\theta}_i - \frac{1}{s}\boldsymbol{\theta}_0\|^2 s^4 \leq 2C\alpha^2 s^4. \tag{13}$$

Plugging Equations (12) and (13) into Equation (11) we obtain the following lower bound for the minimax error, which holds for any $\alpha \leq 1$:

$$\inf_{\mathcal{A}_{\text{SSL}}} \sup_{\|\boldsymbol{\theta}^*\|=s} \mathbb{E}_{\mathcal{D}_l,\mathcal{D}_u}\left[\mathcal{R}_{\text{estim}}(\mathcal{A}_{\text{SSL}}(\mathcal{D}_l,\mathcal{D}_u), P_{XY}^{\boldsymbol{\theta}^*})\right] \geq \frac{1}{2}c_o\alpha s\left(1 - \frac{1 + n_l s^2\alpha^2 + n_u C_1 s^4\alpha^2}{C_0 d}\right).$$

Minimizing over α yields the optimum value $\alpha = \min\left\{1, \sqrt{\frac{C_0 d - 1}{3s^2 n_l + 3C_1 s^4 n_u}}\right\}$, where the minimum comes from how we have constructed the packing, which requires that $\alpha \leq 1$. Using this value for α concludes the proof.

D.2 Proof of upper bound

We now prove the tightness of our lower bound by establishing the upper bound for the estimation error of the SSL Switching algorithm presented in Algorithm 2. We choose the following minimax optimal SL and UL+ estimators

$$\hat{\boldsymbol{\theta}}_{\text{SL}} = \frac{1}{n_l}\sum_{i=1}^{n_l} Y_i X_i \tag{14}$$

$$\hat{\boldsymbol{\theta}}_{\text{UL+}} = \text{sign}\left(\hat{\boldsymbol{\theta}}_{\text{SL}}^\top \hat{\boldsymbol{\theta}}_{\text{UL}}\right)\hat{\boldsymbol{\theta}}_{\text{UL}}, \text{ with } \hat{\boldsymbol{\theta}}_{\text{UL}} = \sqrt{(\hat{\lambda}-1)_+}\hat{v}, \tag{15}$$

where $(\hat{\lambda}, \hat{v})$ is the leading eigenpair of the sample covariance matrix $\hat{\Sigma} = \frac{1}{n_u}\sum_{j=0}^{n_u} X_j X_j^T$ and we use the notation $(x)_+ := \max(0, x)$. By [37], this UL estimator is known to match the minimax rate. As the vanilla UL estimation problem is agnostic to the sign as discussed in section 2.2, in order to classify, the UL+ estimator needs to choose a sign, which it does in a way that aligns better with the SL estimator.

We first bound the expected error incurred by the UL+ estimator:

Proposition 1 (Fixing the sign of $\hat{\boldsymbol{\theta}}_{\text{UL}}$)**.** *Consider the UL+ estimator $\hat{\boldsymbol{\theta}}_{UL+}$ defined in Equation* (15)*. There exist universal constants $C, C' > 0$ such that for $n_u \geq (160/s)^2 d$*

$$\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_{UL+} - \boldsymbol{\theta}^*\|\right] \leq C\sqrt{\frac{d}{s^2 n_u}} + C' s e^{-\frac{1}{2}n_l s^2\left(1 - c_0\sqrt{\frac{d\log(n_u)}{s^2 n_u}}\right)^2}.$$

The proof, given in Appendix D.3 uses prior results for upper bounds for the UL estimator and additionally characterizes the price that needs to be paid for selecting the best sign for $\hat{\boldsymbol{\theta}}_{\text{UL}}$.

For the SL estimator $\hat{\boldsymbol{\theta}}_{\text{SL}}$, we apply standard results for Gaussian distributions, to upper bound the estimation error that holds for any regime of n, d.

$$\mathbb{E}_{\mathcal{D}_l\sim\left(P_{XY}^{\boldsymbol{\theta}^*}\right)^{n_l}}\left[\|\hat{\boldsymbol{\theta}}_{\text{SL}} - \boldsymbol{\theta}^*\|\right] \leq \sqrt{\frac{d}{n_l}}. \tag{16}$$

Using Equation (16) and Proposition 1 and switching between $\hat{\boldsymbol{\theta}}_{\text{SL}}$ and $\hat{\boldsymbol{\theta}}_{\text{UL+}}$ according to the conditions in Algorithm 2, picking the better performing of the two depending on the regime, we can show that there exist universal constants $C, c_0 > 0$ such that for $0 \leq s \leq 1$, $d \geq 2$ and $n_u \geq (160/s)^2 d$, we have

$$\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_{\text{SSL-S}} - \boldsymbol{\theta}^*\|\right] \leq C \min\left\{s, \sqrt{\frac{d}{n_l}}, \sqrt{\frac{d}{s^2 n_u}} + s e^{-\frac{1}{2} n_l s^2 \left(1 - c_0 \sqrt{\frac{d \log(n_u)}{s^2 n_u}}\right)^2}\right\}, \tag{17}$$

where the expectation is over $\mathcal{D}_l \sim \left(P_{XY}^{\boldsymbol{\theta}^*}\right)^{n_l}$ and $\mathcal{D}_u \sim \left(P_X^{\boldsymbol{\theta}^*}\right)^{n_u}$.

Matching lower and upper bound When $n_l > O(\frac{\log(n_u)}{s^2})$, the first additive component dominates in the last term in the right-hand side of Equation (17). Basic calculations then yield that the expected error of the switching algorithm is upper bounded by $C' \min\left\{s, \sqrt{\frac{d}{n_l + s^2 n_u}}\right\}$ for some constant C', which concludes the proof of the theorem.

D.3 Proof of Proposition 1

Recall that we consider the UL+ estimator $\hat{\boldsymbol{\theta}}_{\text{UL+}} = \text{sign}\left(\hat{\boldsymbol{\theta}}_{\text{SL}}^\top \hat{\boldsymbol{\theta}}_{\text{UL}}\right) \hat{\boldsymbol{\theta}}_{\text{UL}}$ and denote $\hat{\beta} := \text{sign}\left(\hat{\boldsymbol{\theta}}_{\text{SL}}^\top \hat{\boldsymbol{\theta}}_{\text{UL}}\right)$. Now let $\beta := \text{sign}(\boldsymbol{\theta}^{*\top} \hat{\boldsymbol{\theta}}_{\text{UL}}) = \arg\min_{\tilde{\beta} \in \{-1,+1\}} \|\tilde{\beta} \hat{\boldsymbol{\theta}}_{\text{UL}} - \boldsymbol{\theta}^*\|^2$.

Note that we can write the expected squared estimation error of $\hat{\boldsymbol{\theta}}_{\text{UL+}}$ as

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_{\text{UL+}} - \boldsymbol{\theta}^*\|\right] &= \mathbb{E}\left[\|\hat{\beta}\hat{\boldsymbol{\theta}}_{\text{UL}} - \boldsymbol{\theta}^*\|\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{\hat{\beta}=\beta\}} \|\beta\hat{\boldsymbol{\theta}}_{\text{UL}} - \boldsymbol{\theta}^*\| + \mathbb{1}_{\{\hat{\beta}\neq\beta\}} \|\beta\hat{\boldsymbol{\theta}}_{\text{UL}} + \boldsymbol{\theta}^*\|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{\{\hat{\beta}=\beta\}} \|\beta\hat{\boldsymbol{\theta}}_{\text{UL}} - \boldsymbol{\theta}^*\|\right] + \mathbb{E}\left[\mathbb{1}_{\{\hat{\beta}\neq\beta\}} (\|\beta\hat{\boldsymbol{\theta}}_{\text{UL}} - \boldsymbol{\theta}^*\| + 2\|\boldsymbol{\theta}^*\|)\right] \\
&\leq \mathbb{E}\left[\|\beta\hat{\boldsymbol{\theta}}_{\text{UL}} - \boldsymbol{\theta}^*\|\right] + 2s\mathbb{P}(\hat{\beta} \neq \beta). \qquad (18)
\end{aligned}$$

First, Wu and Zhou [37] established for this particular UL estimator that $\mathbb{E}\left[\|\beta\hat{\boldsymbol{\theta}}_{\text{UL}} - \boldsymbol{\theta}^*\|^2\right] \leq C\frac{d}{s^2 n_u}$. Moreover, the probability of incorrectly estimating the sign (permutation) can be written as

$$\begin{aligned}
\mathbb{P}(\hat{\beta} \neq \beta) &= \mathbb{P}\Big(\text{sign}\left(\hat{\boldsymbol{\theta}}_{\text{SL}}^\top \hat{\boldsymbol{\theta}}_{\text{UL}}\right) \neq \text{sign}\left(\boldsymbol{\theta}^{*\top} \hat{\boldsymbol{\theta}}_{\text{UL}}\right)\Big), \text{ where } \hat{\boldsymbol{\theta}}_{\text{SL}} \sim \mathcal{N}(\boldsymbol{\theta}^*, \frac{1}{n_l} I_d) \\
&\leq \mathbb{P}\left(\text{sign}(\tilde{Z}) \neq \text{sign}\left(\boldsymbol{\theta}^{*\top} \hat{\boldsymbol{\theta}}_{\text{UL}}\right)\right), \text{ where } \tilde{Z} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_{\text{UL}}^\top \boldsymbol{\theta}^*, \frac{1}{n_l}(\hat{\boldsymbol{\theta}}_{\text{UL}}^\top \hat{\boldsymbol{\theta}}_{\text{UL}})) \\
&\leq \mathbb{P}\left(Z' \geq |\hat{\boldsymbol{\theta}}_{\text{UL}}^\top \boldsymbol{\theta}^*|\right), \text{ where } Z' \sim \mathcal{N}(0, \frac{1}{n_l}(\hat{\boldsymbol{\theta}}_{\text{UL}}^\top \hat{\boldsymbol{\theta}}_{\text{UL}})) \\
&= \mathbb{P}\Big(Z \geq \sqrt{n_l s^2} S_C(\hat{\boldsymbol{\theta}}_{\text{UL}}, \boldsymbol{\theta}^*)\Big) \qquad \text{where } Z \sim \mathcal{N}(0,1),
\end{aligned}$$

where $S_C(\hat{\boldsymbol{\theta}}_{\text{UL}}, \boldsymbol{\theta}^*) = \frac{|\hat{\boldsymbol{\theta}}_{\text{UL}}^\top \boldsymbol{\theta}^*|}{\|\hat{\boldsymbol{\theta}}_{\text{UL}}\|\|\boldsymbol{\theta}^*\|}$ herefore, for any A we have:

$$\begin{aligned}
\mathbb{P}(\hat{\beta} \neq \beta) &\leq \mathbb{P}(Z \geq \sqrt{n_l s^2}(1-A)) + \mathbb{P}\Big(S_C(\hat{\boldsymbol{\theta}}_{\text{UL}}, \boldsymbol{\theta}^*) \leq 1 - A\Big) \\
&\leq e^{-\frac{1}{2} n_l s^2 (1-A)^2} + \mathbb{P}\Big(S_C(\hat{\boldsymbol{\theta}}_{\text{UL}}, \boldsymbol{\theta}^*) \leq 1 - A\Big),
\end{aligned}$$

where we used the Chernoff bound in the last step. Finally, setting $A = c_0\sqrt{\frac{d\log(n_u)}{s^2 n_u}}$ as a corollary of Proposition 6 in Azizyan et al. [2] for $n_u \geq (160/s)^2 d$ we have $\mathbb{P}\Big(S_C(\hat{\boldsymbol{\theta}}_{\text{UL}}, \boldsymbol{\theta}^*) \leq 1 - A\Big) \leq \frac{d}{n_u}$. Therefore, for big enouhg n_u, we have the following upper bound on estimating the sign wrong

$$\mathbb{P}(\hat{\beta} \neq \beta) \leq e^{-\frac{1}{2} n_l s^2 \left(1 - c_0 \sqrt{\frac{d\log(n_u)}{s^2 n_u}}\right)^2} + \frac{d}{n_u}.$$

Combining this result with Equation (18) finishes the proof of the proposition, as we obtain

$$\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_{\text{UL+}} - \boldsymbol{\theta}^*\|\right] \leq C\sqrt{\frac{d}{s^2 n_u}} + C' s e^{-\frac{1}{4} n_l s^2 (1 - c_0 \sqrt{\frac{d \log(n_u)}{s^2 n_u}})^2}.$$

E Proof of Theorem 1

In this section, we prove the minimax lower bound on excess risk for an algorithm that uses both labelled and unlabelled data and a matching (up to logarithmic factors) upper bound.

E.1 Proof of lower bound

We first prove the excess error minimax lower bound in Theorem 1: there exist a constant $C_0 > 0$ such that for any $s > 0$, $n_u, n_l \geq 0$ and $d \geq 4$, we have

$$\inf_{\mathcal{A}_{\text{SSL}}} \sup_{\|\theta_*\|=s} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{\text{SSL}}\left(\mathcal{D}_l, \mathcal{D}_u\right), P_{XY}^{\boldsymbol{\theta}^*}\right)\right] \geq C_0 e^{-s^2/2} \min\left\{\frac{d}{s n_l + s^3 n_u}, s\right\}, \tag{19}$$

where the expectation is over $\mathcal{D}_l \sim \left(P_{XY}^{\boldsymbol{\theta}^*}\right)^{n_l}$ and $\mathcal{D}_u \sim \left(P_X^{\boldsymbol{\theta}^*}\right)^{n_u}$. Our approach to proving this lower bound is again to apply Fano's method [18] using the excess risk as the evaluation method. The reduction from estimation to testing usually hinges on the triangle inequality in metric space. As the excess risk does not satisfy the metric axioms, as previously used in Azizyan et al. [2], we can use Markov's inequality to obtain the same reduction and then use Fano's inequality:

Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M \in \Theta$, $M \geq 2$, and $\gamma > 0$. If for all $1 \leq i \neq j \leq M$ and $\hat{\boldsymbol{\theta}}$,

$$\mathcal{E}\left(\hat{\boldsymbol{\theta}}, P_{XY}^{\boldsymbol{\theta}_i}\right) < \gamma \;\text{ implies }\; \mathcal{E}\left(\hat{\boldsymbol{\theta}}, P_{XY}^{\boldsymbol{\theta}_j}\right) \geq \gamma, \tag{20}$$

then

$$\begin{aligned}\inf_{\mathcal{A}_{\text{SSL}}} \max_{i \in [0..M]} &\mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{\text{SSL}}(\mathcal{D}_l, \mathcal{D}_u), P_{XY}^{\boldsymbol{\theta}_i}\right)\right] \qquad (21)\\ &\geq \gamma\left(1 - \frac{1 + n_l \max_{i \neq j} D\left(P_{XY}^{\boldsymbol{\theta}_i} || P_{XY}^{\boldsymbol{\theta}_j}\right) + n_u \max_{i \neq j} D\left(P_X^{\boldsymbol{\theta}_i} || P_X^{\boldsymbol{\theta}_j}\right)}{\log(M)}\right),\end{aligned}$$

where the expectation is over $\mathcal{D}_l \sim \left(P_{XY}^{\boldsymbol{\theta}_i}\right)^{n_l}$ and $\mathcal{D}_u \sim \left(P_X^{\boldsymbol{\theta}_i}\right)^{n_u}$.

In order to then lower bound the testing problem, we again pick $\boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_M$ to be an appropriate packing, so that Condition (20) can be satisfied. For that purpose, we can simply use the construction from Li et al. [23], which results in tight bounds for supervised and unsupervised settings. Let $p = (d-1)/6$. By Lemma 4.10 in Massart [25], there exists a set $\tilde{\mathcal{M}} = \{\psi_1, \ldots, \psi_M\}$, such that $\|\psi_i\|_0 = p$, $\psi_i \in \{0,1\}^{d-1}$, the Hamming distance $\delta(\psi_i, \psi_j) > p/2$ for all $1 \leq i < j \leq M = |\tilde{\mathcal{M}}|$, and $\log M \geq \frac{p}{5} \log \frac{d}{p} \geq d \log(6)/60 = c_1 d$.

Define

$$\mathcal{M} = \left\{\boldsymbol{\theta}_i = \begin{bmatrix}\sqrt{s^2 - p\alpha^2} \\ \alpha\psi_i\end{bmatrix} \middle| \psi_i \in \tilde{\mathcal{M}}\right\}$$

for some absolute constant α. Note that since $\|\boldsymbol{\theta}_i\| = s$ and $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2 = \alpha^2 \delta(\psi_i, \psi_j)$, we have

$$\frac{p\alpha^2}{2} \leq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|^2 \leq 2p\alpha^2 \tag{22}$$

and

$$s^2 - p\alpha^2 \leq \boldsymbol{\theta}_i^\top \boldsymbol{\theta}_j \leq s^2 - p\alpha^2/4. \tag{23}$$

First, we show that the excess risk satisfies Condition (20). As in the proof of Theorem 1 in Li et al. [23], we have that for any θ,

$$\mathcal{E}_{\boldsymbol{\theta}_i}(\theta) + \mathcal{E}_{\boldsymbol{\theta}_j}(\theta) \geq 2c_0 e^{-s^2/2} \frac{p\alpha^2}{s}.$$

and thus for all i and $j \neq i$, it holds that

$$\mathcal{E}_{\boldsymbol{\theta}_i}(\theta) \leq c_0 e^{-s^2/2} \frac{p\alpha^2}{s} \implies \mathcal{E}_{\boldsymbol{\theta}_j}(\theta) \geq c_0 e^{-s^2/2} \frac{p\alpha^2}{s}. \tag{24}$$

Then since the condition in (20) is satisfied, we obtain

$$\begin{aligned}
&\inf_{\mathcal{A}_{\text{SSL}}} \sup_{\|\theta_*\|=s} \mathbb{E}_{\mathcal{D}_l,\mathcal{D}_u}\left[\mathcal{E}\left(\mathcal{A}_{\text{SSL}}\left(\mathcal{D}_l,\mathcal{D}_u\right), P_{XY}^{\boldsymbol{\theta}^*}\right)\right] \\
&\quad \geq \inf_{\mathcal{A}_{\text{SSL}}} \max_{i\in[0..M]} \mathbb{E}\left[\mathcal{E}\left(\mathcal{A}_{\text{SSL}}(\mathcal{D}_l,\mathcal{D}_u), P_{XY}^{\boldsymbol{\theta}_i}\right)\right] \\
&\quad \geq c_0 e^{-s^2/2} \frac{p\alpha^2}{s}\left(1 - \frac{1 + n_l \max_{i\neq j} D\left(P_{XY}^{\boldsymbol{\theta}_i}||P_{XY}^{\boldsymbol{\theta}_j}\right) + n_u \max_{i\neq j} D\left(P_{X}^{\boldsymbol{\theta}_i}||P_{X}^{\boldsymbol{\theta}_j}\right)}{\log(M)}\right).
\end{aligned} \tag{25}$$

Next, we bound the KL divergence between the two joint distributions and between the two marginals respectively in Equation (25).

$$D\left(P_{XY}^{\boldsymbol{\theta}_i}||P_{XY}^{\boldsymbol{\theta}_j}\right) = \frac{1}{2}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2 \leq p\alpha^2. \tag{26}$$

where the inequality follows from (22). Using Proposition 24 in Azizyan et al. [2], we bound the KL divergence between the two marginals

$$D\left(P_{X}^{\boldsymbol{\theta}_i}||P_{X}^{\boldsymbol{\theta}_j}\right) \lesssim s^4\left(1 - \frac{\boldsymbol{\theta}_i^\top \boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_i\|\|\boldsymbol{\theta}_j\|}\right) \leq ps^2\alpha^2. \tag{27}$$

where the inequality follows from (23). Plugging (26) and (27) into (25) and setting

$$\alpha^2 = c_3 \min\left\{\frac{c_1 d - \log 2}{8(pn_l + s^2 pn_u)}, \frac{s^2}{p}\right\},$$

gives the desired result

$$\inf_{\mathcal{A}_{\text{SSL}}} \sup_{\|\theta_*\|=s} \mathbb{E}_{\mathcal{D}_l,\mathcal{D}_u}\left[\mathcal{E}\left(\mathcal{A}_{\text{SSL}}\left(\mathcal{D}_l,\mathcal{D}_u\right), P_{XY}^{\boldsymbol{\theta}^*}\right)\right] \gtrsim e^{-s^2/2} \min\left\{\frac{d}{sn_l + s^3 n_u}, s\right\}.$$

E.2 Proof of upper bound

Next, we prove the upper bound on the excess risk of the SSL switching estimator $\hat{\boldsymbol{\theta}}_{\text{SSL-S}}$ output by Algorithm 2 with the supervised and unsupervised estimators defined in Appendix D.2 to show the tightness of Theorem 1. In particular, we show that there exist universal constants $C, c_0 > 0$ such that for $0 \leq s \leq 1$, $d \geq 2$ and for sufficiently large n_u and n_l,

$$\mathbb{E}\left[\mathcal{E}(\hat{\boldsymbol{\theta}}_{\text{SSL-S}})\right] \leq Ce^{-\frac{1}{2}s^2} \min\left\{s, \frac{d\log(n_l)}{sn_l}, \frac{d\log(dn_u)}{s^3 n_u} + se^{-\frac{1}{2}s^2\left(n_l\left(1 - c_0\sqrt{\frac{d\log(n_u)}{s^2 n_u}}\right)^2 - 1\right)}\right\},$$

where the expectation is over $\mathcal{D}_l \sim \left(P_{XY}^{\boldsymbol{\theta}^*}\right)^{n_l}$ and $\mathcal{D}_u \sim \left(P_{X}^{\boldsymbol{\theta}^*}\right)^{n_u}$.

The proof follows the same arguments as the proof of in Appendix D.2 where we instead use excess risk upper bounds for SL and UL from Li et al. [23].

In addition, we also use a result that follows from Proposition 1 to choose the sign of the UL+ estimator.

Note that the upper bound on the excess risk of $\hat{\boldsymbol{\theta}}_{\text{SSL-S}}$ is matching the lower bound in (19), up to logarithmic factors. We conjecture that the logarithmic factors are an artifact of the analysis and can be removed. For instance, it may be possible to extend results in Ratsaby and Venkatesh [29] that bound the excess risk using the estimation error upper bound without incurring logarithmic factors. However, their results are not directly applicable here.

F Theoretical guarantees for the SSL Weighted Algorithm

In this section, we show theoretically that the SSL-W procedure introduced in Appendix B.1 can achieve lower squared estimation error (up to sign permutation) compared to SSL-S. This result shows that it is possible to improve the error of the naïve SSL-S algorithm by utilizing *all* the data that is available.

For the purpose of the theoretical analysis, we consider a slightly different SSL-W estimator compared to the one introduced in Section B.1. First, recall that for the classification problem we consider, unsupervised learning produces a set of two feasible predictors $\{\hat{\boldsymbol{\theta}}_{\text{UL}}, -\hat{\boldsymbol{\theta}}_{\text{UL}}\}$ and cannot discern between them without access to a (small) labeled dataset. We denote by $\boldsymbol{\theta}^*_{\text{UL}}$ the UL estimator with correct sign, namely $\boldsymbol{\theta}^*_{\text{UL}} := \arg\min_{\boldsymbol{\theta}\in\{\hat{\boldsymbol{\theta}}_{\text{UL}}, -\hat{\boldsymbol{\theta}}_{\text{UL}}\}} \mathbb{E}\left[\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2\right]$.

In what follows, we study theoretically the error of the SSL-W estimator constructed using $\boldsymbol{\theta}^*_{\text{UL}}$, i.e. $\boldsymbol{\theta}^*_{\text{SSL-W}}(t) := t\hat{\boldsymbol{\theta}}_{\text{SL}} + (1-t)\boldsymbol{\theta}^*_{\text{UL}}$. Therefore, our result characterizes the error of the SSL-W estimator up to a sign permutation. To choose the correct sign, one needs only a small labeled dataset, similar in size to what is prescribed by Proposition 1. While this step is not captured by Proposition 2, SSL-S is unlikely to close the gap to SSL-W when provided with this small amount of additional labeled data.

We can now state Proposition 2, which shows that there exists an optimal weight for which the SSL-W predictor achieves lower estimation error than the SSL Switching predictor, $\hat{\boldsymbol{\theta}}_{\text{SSL-S}}$.

Proposition 2. *Consider a distribution $P_{XY}^{\boldsymbol{\theta}^*} \in \mathcal{P}^{(s)}_{\text{2-GMM}}$ and let $d \geq 2$, and $n_l, n_u > 0$. Let $\boldsymbol{\theta}^*_{\text{SSL-W}}(t^*)$ be the SSL-W estimator introduced above. Then there exists a $t^* \in (0,1)$ for which*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\theta}}_{\text{SSL-S}} - \boldsymbol{\theta}^*\right\|^2\right] - \mathbb{E}\left[\|\boldsymbol{\theta}^*_{\text{SSL-W}}(t^*) - \boldsymbol{\theta}^*\|^2\right] = \min\left\{r, \frac{1}{r}\right\} \mathbb{E}\left[\|\boldsymbol{\theta}^*_{\text{SSL-W}}(t^*) - \boldsymbol{\theta}^*\|^2\right], \quad (28)$$

where $r = \frac{\mathbb{E}\left[\|\boldsymbol{\theta}^*_{UL} - \boldsymbol{\theta}^*\|^2\right]}{\mathbb{E}\left[\left\|\hat{\boldsymbol{\theta}}_{SL} - \boldsymbol{\theta}^*\right\|^2\right]}$, and the expectations are over $\mathcal{D}_l \sim \left(P_{XY}^{\boldsymbol{\theta}^*}\right)^{n_l}, \mathcal{D}_u \sim \left(P_X^{\boldsymbol{\theta}^*}\right)^{n_u}$.

Since the RHS of Equation (28) is always positive, $\boldsymbol{\theta}^*_{\text{SSL-W}}(t^*)$ always outperforms $\hat{\boldsymbol{\theta}}_{\text{SSL-S}}$ as long as the conditions of Proposition 2 are satisfied. The magnitude of the error gap between SSL-S and SSL-W depends on the gap between SL and UL+ (see Figure 1). The maximum gap is reached for $\mathbb{E}\left[\|\boldsymbol{\theta}^*_{\text{UL}} - \boldsymbol{\theta}^*\|^2\right] \approx \mathbb{E}\left[\left\|\hat{\boldsymbol{\theta}}_{\text{SL}} - \boldsymbol{\theta}^*\right\|^2\right]$ when SSL-W obtains half the error of SSL-S.

F.1 Proof of Proposition 2

The first step in proving Proposition 2 is to express the estimation error of $\hat{\boldsymbol{\theta}}_{\text{SSL-W}}(t^*)$ in terms of the estimation errors of $\hat{\boldsymbol{\theta}}_{\text{SL}}$ and $\hat{\boldsymbol{\theta}}_{\text{UL+}}$ which is captured by Lemma 1.

Lemma 1. *Let $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ be two statistically independent estimators of $\boldsymbol{\theta}^* \in \mathbb{R}^d$ and let $\hat{\boldsymbol{\theta}}_1$ be unbiased, i.e. $\mathbb{E}\left[\hat{\boldsymbol{\theta}}_1\right] = \boldsymbol{\theta}^*$. Then, the expected squared error of the weighted estimator $\hat{\boldsymbol{\theta}}_{t^*} = t^*\hat{\boldsymbol{\theta}}_1 + (1-t^*)\hat{\boldsymbol{\theta}}_2$ with $t^* = \frac{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2\right]}{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2\right]}$ is given by*

$$\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_{t^*} - \boldsymbol{\theta}^*\|^2\right] = \left(\frac{1}{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2\right]} + \frac{1}{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2\right]}\right)^{-1}.$$

We can apply Lemma 1, since $\hat{\boldsymbol{\theta}}_{\text{SL}}$ is unbiased and $\hat{\boldsymbol{\theta}}_{\text{SL}}$ and $\hat{\boldsymbol{\theta}}_{\text{UL+}}$ are trained on $\mathcal{D}_l$ and $\mathcal{D}_u$ respectively, and hence, are independent. The proof then follows from calculating the difference between the harmonic mean and the minimum of estimation errors of $\hat{\boldsymbol{\theta}}_{\text{SL}}$ and $\hat{\boldsymbol{\theta}}_{\text{UL+}}$. Let $x, y \in \mathbb{R}_+$ and w.l.o.g. assume $x \leq y$. Then we have:

$$x - \left(\frac{1}{x} + \frac{1}{y}\right)^{-1} = x - \frac{xy}{x+y} = \frac{x^2}{x+y} = \frac{x}{y}\frac{xy}{x+y}.$$

Choosing $x = \min\left\{\mathbb{E}[\|\hat{\boldsymbol{\theta}}_{\text{UL+}} - \boldsymbol{\theta}^*\|^2], \mathbb{E}[\|\hat{\boldsymbol{\theta}}_{\text{SL}} - \boldsymbol{\theta}^*\|^2]\right\}$ and $y = \max\left\{\mathbb{E}[\|\hat{\boldsymbol{\theta}}_{\text{UL+}} - \boldsymbol{\theta}^*\|^2], \mathbb{E}[\|\hat{\boldsymbol{\theta}}_{\text{SL}} - \boldsymbol{\theta}^*\|^2]\right\}$ finishes the proof and yields the desired result for $t^* = \frac{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2\right]}{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2\right] + \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2\right]}$.

Remark. Note that this lemma holds for arbitrary distributions and estimators as long as they are independent and one of them is unbiased. Therefore, future results that derive upper bounds for SL and UL+ for other distributional assumptions and estimators can seamlessly be plugged into Lemma 1. By the same argument, $\hat{\boldsymbol{\theta}}_{\text{SSL-W}}$ obtained by other SL and UL+ estimators can also be expected to improve over the respective SL and UL+ estimators, given that one of them is unbiased.

F.2 Proof of Lemma 1

By definition of $\hat{\boldsymbol{\theta}}_{t^*}$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_{t^*} - \boldsymbol{\theta}^*\|^2\right] &= \mathbb{E}\left[\|t^*\hat{\boldsymbol{\theta}}_1 + (1-t^*)\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2\right] \\
&= \mathbb{E}\left[{t^*}^2\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 + (1-t^*)^2\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2 + 2t^*(1-t^*)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*)^\top(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*)\right] \\
&= \mathbb{E}\left[{t^*}^2\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 + (1-t^*)^2\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2\right],
\end{aligned}$$

where the last equality holds due to the independence of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ and the unbiasedness of $\hat{\boldsymbol{\theta}}_1$.

Plugging in $t^* = \frac{\mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2}{\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 + \mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2}$, we get

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{\theta}}_{t^*} - \boldsymbol{\theta}^*\|^2 &= \left(\frac{\mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2}{\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 + \mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2}\right)^2 \mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 \\
&\quad + \left(\frac{\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2}{\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 + \mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2}\right)^2 \mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2 \\
&= \frac{\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 \mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2}{\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2 + \mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2} \\
&= \frac{1}{\frac{1}{\mathbb{E}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2} + \frac{1}{\mathbb{E}\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^*\|^2}}.
\end{aligned}$$

G Simulation details

G.1 Methodology

We split each dataset in a test set, a validation set and a training set. The unlabeled set size is fixed to 5000 for the synthetic experiments and 4000 for the real-world datasets. The size of the labeled set n_l is varied in each experiment. For each dataset, we draw a different labeled subset 20 times and report the average and the standard deviation of the error gap (or the error) over these runs. The validation and the test set have 1000 labeled samples each.

We use logistic regression from Scikit-Learn [27] as the supervised algorithm. We use the validation set to select the ridge penalty for SL. For the unsupervised algorithm, we use an implementation of Expectation-Maximization from the Scikit-Learn library. We also use the self-training algorithm from Scikit-Learn with a logistic regression estimator. The best confidence threshold for the pseudolabels is selected using the validation set. Moreover, the optimal weight for SSL-W is also chosen with the help of the validation set. We give SSL-S the benefit of choosing the optimal switching point between SL and UL+ by using the test set. Even with this important advantage, SSL-W (and sometimes self-training) still manage to outperform SSL-S.

G.2 Details about the real-world datasets

Tabular data. We select tabular datasets from the OpenML repository [35] according to a number of criteria. We focus on high-dimensional data with $100 \leq d < 1000$, where the two classes are not suffering from extreme class imbalance, i.e. the imbalance ratio between the majority and the minority class is at most 5. Moreover, we only consider datasets that have substantially more samples than the number of features, i.e. $\frac{n}{d} > 10$. In the end, we are left with 5 datasets, that span a broad range of application domains, from ecology to chemistry and finance.

To assess the difficulty of the datasets, we train logistic regression on the entire data that is available, and report the training error. Since we train on substantially more samples than the number of dimensions, the predictor that we obtain is a good estimate of the linear Bayes classifier for each dataset.

Furthermore, we measure the extent to which the data follows a GMM distribution with spherical components. We fit a spherical Gaussian to data coming from each class and use linear discriminant analysis (LDA) for prediction. We record the training error (of the best permutation). Intuitively, this is a score of how much our assumption about the connection between the marginal P_X and the labeling function $P(Y|X)$ is satisfied. In Figure 3 we rank datasets by SNR using the following formula to estimate SNR: $\text{SNR} = \frac{\mathcal{R}_{\text{pred}}(\boldsymbol{\theta}^*_{\text{UL}}) - \mathcal{R}_{\text{pred}}(\boldsymbol{\theta}^*_{\text{Bayes}})}{\mathcal{R}_{\text{pred}}(\boldsymbol{\theta}^*_{\text{Bayes}})\sqrt{d}}$, where $\boldsymbol{\theta}^*_{\text{Bayes}}$ is the linear Bayes classifier and $\boldsymbol{\theta}^*_{UL}$ the LDA classifier described above.[5] If the data distribution is very similar to an isotropic GMM (i.e. $\mathcal{R}_{\text{pred}}(\boldsymbol{\theta}^*_{\text{UL}}) \leq 0.1$), then we simply take the linear Bayes error as the estimate of the SNR.

Image data. In addition to the tabular data, we also consider a number of datasets that are subsets of the MNIST dataset [22]. More specifically, we create binary classification problems by selecting class pairs from MNIST. We choose 5 classification problems which vary in difficulty, as measured by the Bayes error, from easier (e.g. digit 0 vs digit 1) to more difficult (e.g. digit 5 vs digit 9). Table 2 presents the exact class pairs that we selected. To make the problem more amenable for linear classification, we consider as covariates the 20 principle components of the MNIST images.

Dataset name	d	Linear classif. training error	LDA w/ spherical GMM training error
mnist_0v1	784	0.001	0.009
mnist_1v7	784	0.006	0.036
madeline	259	0.344	0.381
philippine	308	0.240	0.318
vehicleNorm	100	0.141	0.177
mnist_5v9	784	0.024	0.045
mnist_5v6	784	0.024	0.042
a9a	123	0.150	0.216
mnist_3v8	784	0.042	0.105
musk	166	0.037	0.270

Table 2: Some characteristics of the datasets considered in our experimental study.

H More experiments

In this section we present further experiments that complement Figure 2 and indicate that the SSL Weighted algorithm (SSL-W) can indeed outperform the naive baseline of the Switching algorithm (SSL-S) on other real-world datasets. The extent of the error gap is determined by the $\frac{n_u}{n_l}$ ratio as well as the signal-to-noise ratio that is specific to each dataset. In addition, we also show that self-training can outperform SSL-W in some scenarios. While in this work we provide guarantees only for SSL-W, it remains an exciting direction for future work to provide an analysis of self-training that can indicate when it performs best.

[5] Note that we refer to the LDA estimator as *UL* since we use it as a proxy to assess how well unsupervised learning can perform on each dataset.

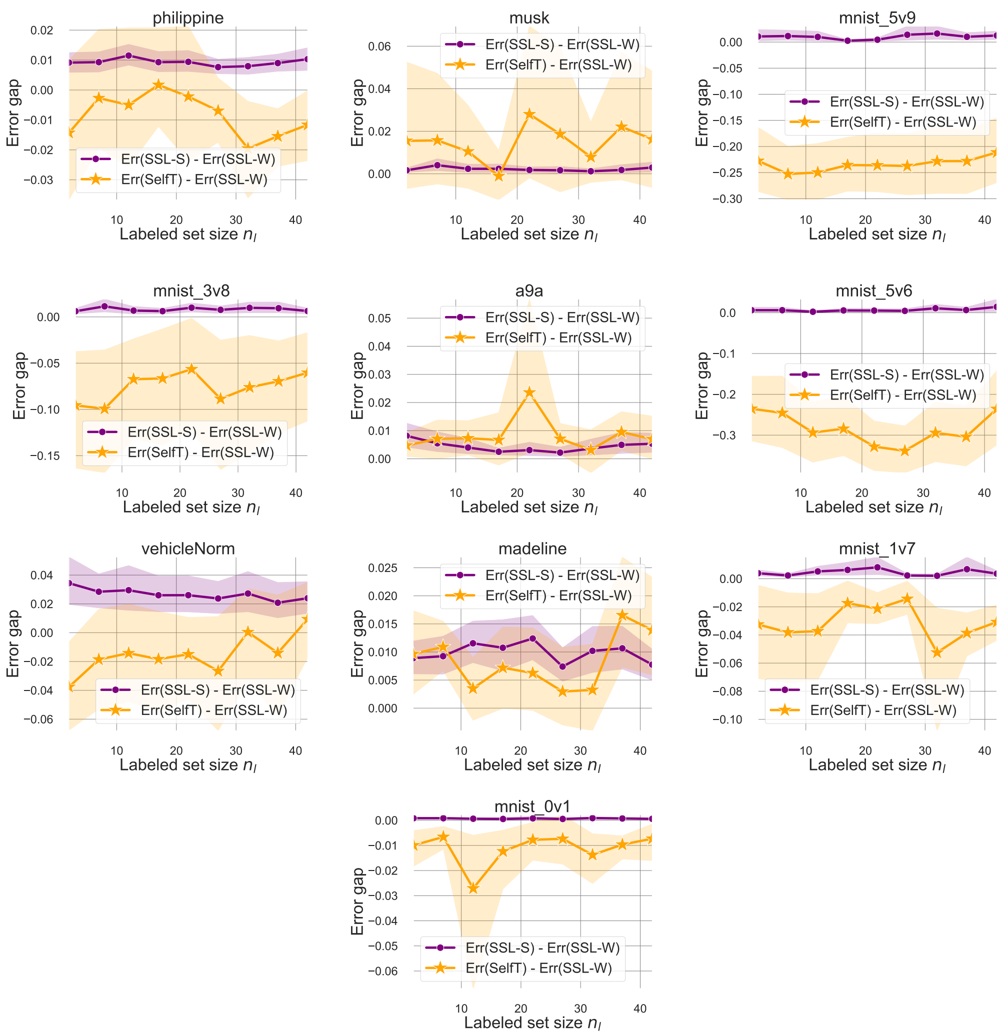


Figure 4: Error gap between SSL-S/self-training and SSL-W on real-world datasets. The positive gap indicates that SSL-W (and, in turn, self-training) outperforms SSL-S (and hence, also SL and UL+) for a broad range of n_l values.