An Information-theoretical Approach to Semi-supervised Learning under Covariate-shift

Gholamali Aminian[⋆†] **Mahed Abroshan**[⋆‡]
Mohammad Mahdi Khalili[††] **Laura Toni**[†] **Miguel R. D. Rodrigues**[†]
[†] University College London [‡] Alan Turing Institute [††] University of Delaware

Abstract

A common assumption in semi-supervised learning is that the labeled, unlabeled, and test data are drawn from the same distribution. However, this assumption is not satisfied in many applications. In many scenarios, the data is collected sequentially (e.g., healthcare) and the distribution of the data may change over time often exhibiting so-called covariate shifts. In this paper, we propose an approach for semi-supervised learning algorithms that is capable of addressing this issue. Our framework also recovers some popular methods, including entropy minimization and pseudo-labeling. We provide new information-theoretical based generalization error upper bounds inspired by our novel framework. Our bounds are applicable to both general semi-supervised learning and the covariate-shift scenario. Finally, we show numerically that our method outperforms previous approaches proposed for semi-supervised learning under the covariate shift.

1 INTRODUCTION

There are many applications, from natural language processing to bio-informatics in which the labeled data is scarce, while plenty of unlabeled data is accessible. Under these circumstances, semi-supervised learning (SSL) algorithms allow us to take advantage of both labeled and unlabeled data. There are different approaches to SSL (Yang et al., 2021), e.g., self-training or input-consistency regularization. Self-training algorithms (Ouali et al., 2020), which are also the focus of this paper, are a class of SSL approaches. These algorithms predict the label of unlabeled data using the model's own confident predictions. However, most of self-training SSL algorithms are designed by assuming labelled training data, and test data admit the same distribution.

There are though many applications in which unlabeled features have a different distribution from labeled features, and the test features distribution could be different or the same as unlabeled features distribution. This situation – commonly known as covariate-shift – arises where data is collected sequentially. For example, in many healthcare applications, the labels will be available with a delay (e.g., studying five years survival analysis or drug discovery) and the distribution of new unlabeled data may change (Ryan and Culp, 2015). Another scenario associated with covariate shift relates to cases where we have a limited number of labeled data on a particular task, and there is plenty of unlabeled data on some other related tasks (Oliver et al., 2018). Studying generalization error bounds is critical to understanding SSL models' performance – and designing suitable SSL approaches – in the presence of the aforementioned distribution shifts.

Various approaches have been developed to characterize the generalization error of SSL algorithms. SSL Generalization error upper bound using Bayes classifiers are provided in (Göpfert et al., 2019) and (Zhu, 2020). The VC-dimension approach is applied by (Göpfert et al., 2019) for the SSL algorithm. (Zhu, 2020) provides an upper bound on the excess risk of SSL algorithm by considering an exponentially concave function[1] based on conditional mutual information. The generalization error of iterative SSL algorithms based on pseudo-labels is studied by He et al. (2021). Generalization error upper bound based on Rademacher complexity for binary classification with a squared-loss mutual information regularization is provided in (Niu et al., 2013). An upper bound on the generalization error of binary classifi-

[⋆] Equal contribution. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

[1] A function $f(x)$ is called β-exponentially concave function if $\exp(-\beta f(x))$ is concave

cation under cluster assumption is derived in (Rigollet, 2007). We refer to the survey paper (Mey and Loog, 2019) and references therein for a thorough review of other theoretical aspects of SSL.

However, these upper bounds on excess risk and generalization error do not entirely capture the role of covariate-shift between labeled data and unlabeled data thereby limiting our ability to characterize the performance of existing SSL methods or to design new ones. In this paper, an information-theoretic approach, inspired by Xu and Raginsky (2017) and Russo and Zou (2019), is applied to characterize the generalization ability of various self-training based SSL algorithms. Such approaches often express the generalization error in terms of certain information measures between the learning algorithm input (the training dataset including labeled and unlabeled data) and output (the hypothesis), thereby incorporating the various ingredients associated with the SSL problem, including the labeled and unlabeled data distribution, the hypothesis space, and the learning algorithm itself. Finally, inspired by our framework and the upper bounds on generalization error, we propose a new SSL algorithm that is able to take advantage of unlabeled data under covariate-shift.

Our main contributions are as follows:

- We propose a novel framework for self-training SSL algorithms that encompasses traditional SSL approaches such as the entropy minimization and the Pseudo-labeling approaches. Our framework is applicable to different loss functions beyond the typically used log-loss function.
- We provide an information-theoretical upper bound on the expected generalization error of the SSL algorithms under covariate-shift in terms of KL divergence and total variation distance. We show that the unlabeled data in our framework can improve the generalization error convergence rate.
- We provide novel information-theoretical upper bounds on *the estimation error of true conditional probabilities* in terms of KL divergence and total variation distance, essential in self-training approaches.
- Inspired by our theoretical results, we then propose a method for SSL algorithm which outperforms traditional SSL algorithms in the presence of covariate shifts.

Notations: We adopt the following notation in the sequel. Upper-case letters denote random variables (e.g., Z), lower-case letters denote random variable realizations (e.g. z), and calligraphic letters denote spaces (e.g. $\mathcal{Z}$). We denote the distribution of the random variable Z by P_Z, the joint distribution of two random variables (Z_1, Z_2) by P_{Z_1,Z_2}.

Information Measures: If P and Q are probability measures over space $\mathcal{Z}$, and P is absolutely continuous with respect to Q, the Kullback-Leibler (KL) divergence between P and Q is given by $D(P\|Q) \triangleq \int_{\mathcal{Z}} \log\left(\frac{dP}{dQ}\right) dP$. If Q is also absolutely continuous with respect to P, then the KL divergence is bounded. The entropy of probability measure, P, is given by $H(P) = \int_{\mathcal{Z}} -dP\log(dP)$.

The mutual information between two random variables Z and T is defined as the KL divergence between the joint distribution and product-of-marginal distributions $I(Z;T) \triangleq D(P_{Z,T}\|P_Z \otimes P_T)$, or equivalently, the conditional KL divergence between $P_{T|Z}$ and P_T averaged over P_Z, $D(P_{T|Z}\|P_T|P_Z) \triangleq \int_{\mathcal{Z}} D(P_{T|Z=z}\|P_T)dP_Z(z)$.

The total variation distance for two probability measures, P and Q, is defined as

$$\mathbb{TV}(P,Q) = \frac{1}{2}\int_{\mathcal{Z}} |dP - dQ| \tag{1}$$

and the variational representation of total variation distance is as follows (Polyanskiy and Wu, 2014):

$$\mathbb{TV}(P,Q) = \frac{1}{2L}\sup_{g\in\mathcal{G}_L}\{\mathbb{E}[g(P)] - \mathbb{E}[g(Q)]\} \tag{2}$$

where $\mathcal{G}_L = \{g : \mathcal{Z} \to \mathbb{R}, ||g||_\infty \leq L\}$. Note that the total variation is bounded, $\mathbb{TV}(P,Q) \leq 1$.

2 RELATED WORK

We now highlight some of the key relevant works in the surrounding fields, including SSL, covariate-shift, domain adaptation, and information-theoretic based generalization error upper bounds.

Semi-Supervised Learning: Entropy minimization and Pseudo-labeling are two fundamental approaches in self-training based SSL. In entropy minimization, an entropy function of the predicted conditional distribution is added to the main empirical risk function, which depends on unlabeled data (Grandvalet et al., 2005). The entropy function can be viewed as a regularization term that penalizes uncertainty in the prediction of the label of the unlabelled data. There are some assumptions for the performance of entropy minimization algorithm, including manifold assumption (Iscen et al., 2019)– where it is assumed that labelled and unlabelled features are drawn from a common data manifold – or cluster assumptions (Chapelle et al., 2003)– where it is assumed that similar data features have a similar label.

In contrast, in Pseudo-labeling, the model is trained using labeled data in a supervised manner and then used to provide a pseudo-label for the unlabeled data with high confidence (Lee et al., 2013). These pseudo labels are then used as inputs in another model, which is trained based on labeled and pseudo-labeled data in a supervised manner. However, pseudo-labelling approaches can underperform because they largely rely on the accuracy of the pseudo-labeling process. To bypass this challenge, an uncertainty-aware Pseudo-labeling approach is proposed in (Rizve et al., 2020). A theoretical framework for using input-consistency regularization combined with self-training algorithms in deep neural networks is proposed in (Wei et al., 2020). Some works also discuss how to combine SSL with causal learning (Schölkopf et al., 2012; Janzing and Schölkopf, 2010, 2015).

Our work departs from existing SSL literature a novel framework – which encompasses existing SSL methods such as entropy minimization or pseudo-labelling – that can extended to other loss function. We also propose to consider the labeled data in the unsupervised loss function in order to improve SSL performance further.

Covariate-shift: Covariate-Shift has been studied in supervised learning (Sugiyama et al., 2007) and (Shimodaira, 2000) and SSL scenarios (Kawakita and Kanamori, 2013). An approximate Bayesian inference scheme by using posterior regularisation for SSL in the presence of covariate-shift is provided in (Chan et al., 2020). The performance of self-training for the SSL algorithms, including entropy minimization and pseudo-labeling in the presence of covariate-shift, with spurious features, is studied in (Chen et al., 2020). Our work differs from this body of research in the sense that we provide an algorithm dependent upper bound on the generalization error of SSL algorithms under covariate-shift.

Domain Adaptation: Domain adaptation involves training a model based on labeled data from the source domain and unlabeled data from the target domain. The covariate-shift reduces to the domain adaptation scenario by considering the same conditional distribution of label given feature but different marginal distributions for the features and unlabeled data. The work (Ben-David et al., 2010), proposed $\mathcal{H}\Delta\mathcal{H}$-divergence as a similarity metric, and generalization bound based on VC-dimension approach is provided. The authors in (Mansour et al., 2009) proposed the discrepancy distance for the general loss function, and a generalization bound based on the Rademacher complexity is derived. Inspired by Ben-David et al. (2010), a domain adversarial algorithm which minimizes $\mathcal{H}\Delta\mathcal{H}$-divergence between source and target domains, is provided by Ganin et al. (2016). The application of entropy minimization and the combination of domain adversarial and entropy minimization in domain adaptation are proposed in (Wang et al., 2020) and (Shu et al., 2018). Our work differs from this area of research as our theoretical results are derived based on the covariate-shift assumption, i.e., the same conditional distribution of features given data under labeled and unlabeled data. In addition, the estimation of conditional distributions in the domain adversarial approach is induced by mostly labeled data. However, in our algorithm, this estimation is induced by both labeled and unlabeled data.

Information-theoretic upper bounds: Recently, Russo and Zou (2019); Xu and Raginsky (2017) proposed to use the mutual information between the input training set and the output hypothesis to upper bound the expected generalization error. Bu et al. (2020b) provides tighter bounds by considering the individual sample mutual information, (Asadi et al., 2018) proposes using chaining mutual information, and some works advocate the conditioning and processing techniques (Steinke and Zakynthinou, 2020; Hafez-Kolahi et al., 2020; Haghifam et al., 2020). Information-theoretic generalization error bounds using other information quantities are also studied, such as α-Rényi divergence and maximal leakage (Esposito et al., 2021), Jensen-Shannon divergence (Aminian et al., 2021b), power divergence (Aminian et al., 2021c), and Wasserstein distance (Lopez and Jog, 2018; Wang et al., 2019). An exact characterization of the generalization error for the Gibbs algorithm is provided in (Aminian et al., 2021a). Using rate-distortion theory, Masiha et al. (2021) and Bu et al. (2020a) provide information-theoretic generalization error upper bounds for model misspecification and model compression. Information theoretical approaches are applied mostly to the supervised learning scenario. However, our work offers an information-theoretical upper bound for the generalization error of SSL in the presence of covariate shift.

3 SSL FRAMEWORK

We consider a SSL setting we wish to learn a hypothesis given a set of labeled and unlabeled features. We also wish to use this hypothesis to predict new labels given new features.

We model the features (also known as inputs) using a random variable $X \in \mathcal{X}$ where $\mathcal{X}$ represents the input space; we model the labels (also known as outputs) using a random variable $Y \in \mathcal{Y}$ where $\mathcal{Y}$ represents the output set. We also let $(X^L, Y^L) = \{(X_i^l, Y_i^l)\}_{i=1}^n$ be a training labelled set consisting of a number of input-output data points drawn i.i.d. from $\mathcal{X}\times\mathcal{Y}$ according to $\mu_X^l \otimes P_{Y|X}$, and $X^U = \{X_i^u\}_{i=1}^m$ a training unlabelled set consisting of a number of inputs data point drawn i.i,d. from $\mathcal{X}$ according to the marginal distribution μ_X^u. Note that in the traditional SSL scenario without covariate shift we consider $\mu_X^l = \mu_X^u$.

Under the covariate-shift scenario, we assume that the test and unlabeled feature distribution, μ_X^u, are shifted with respect to labeled inputs distribution, μ_X^l, but the conditional distribution of labels given inputs, $P_{Y|X}$, is the same for test and training dataset.

We represent hypotheses using a random variable $W \in \mathcal{W}$ where $\mathcal{W}$ is a hypothesis space. We also represent an SSL algorithm via a Markov kernel that maps a given training set (Y^L, X^L, X^U) onto a hypothesis W of the hypothesis class $\mathcal{W}$ according to the probability law $P_{W|X^L,Y^L,X^U}$.

Let us define the following loss functions:

- **Supervised loss function:** A (non-negative) loss function $\ell : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ that measures how well a hypothesis predicts a label (output) given a feature (input).
- **Conditional expectation of supervised loss function:** The expectation of supervised loss function with respect to true conditional distribution, $P_{Y|X}(y|x)$, is defined as follows:

  $$\ell_c(w,x) \triangleq \int_{\mathcal{Y}} \ell(w,x,y) P_{Y|X}(y|x)dy \tag{3}$$

  Note that the conditional distribution, $P_{Y|X}$, is unknown.
- **Unsupervised loss function:** A (non-negative) loss function $\ell_u : \mathcal{X} \times \mathcal{W} \to \mathbb{R}^+$ that measures the loss related to inputs including unlabeled and labeled features.

We can now define the population risk, the supervised empirical risk, the unsupervised empirical risk and the semi-supervised empirical risk as follows:

$$L_P(w, P_{X,Y}) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} \ell(w,x,y) P_{X,Y}(x,y)dxdy \tag{4}$$

$$L_E^{SL}(w, x^L, y^L) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, x_i^L, y_i^L) \tag{5}$$

$$L_E^{UL}(w, x^L, x^U) \triangleq \frac{1}{n+m}\left(\sum_{i=1}^{n} \ell_u(w, x_i^L) + \sum_{j=1}^{m} \ell_u(w, x_j^U)\right) \tag{6}$$

$$L_E^{SSL}(w, x^L, y^L, x^U) \triangleq \beta L_E^{SL}(w, x^L, y^L) + (1-\beta) L_E^{UL}(W, X^L, X^U), \quad 0 \le \beta \le 1. \tag{7}$$

Quantify the performance of a hypothesis w delivered by the SSL algorithm on a testing set (population) and the training set, respectively. The hyper-parameter β balances between the supervised and unsupervised empirical risk.

Remark 1. *[Choice of β] Choosing $\beta = 0$ reduces our problem to an unsupervised learning scenario by considering the unsupervised empirical risk function, $L_E^{UL}(W, X^L, X^U)$, and if we choose $\beta = 1$ our problem reduces to a supervised learning setting by considering the supervised empirical risk, $L_E^{SL}(w, x^L, y^L)$.*

Remark 2. *[Ideal SSL Empirical Risk] If we substitute the $\ell_u(w,x)$ with $\ell_c(w,x)$ in (6), then the unsupervised empirical risk is an unbiased estimation of population risk (4) (See appendix A).*

We can also define the generalization error as follows:

$$\mathrm{gen}(P_{W|X^L,Y^L,X^U}, P_{X,Y}) \triangleq L_P(w, P_{X,Y}) - L_E^{\mathrm{SSL}}(w, x^L, y^L, x^U) \tag{8}$$

which quantifies how much the population risk deviates from the SSL empirical risk. We can also define the expected generalization error as follows:

$$\overline{\mathrm{gen}}(P_{W|X^L,Y^L,X^U}, P_{X,Y}) = \mathbb{E}_{P_{W,X^L,Y^L,X^U}}[\mathrm{gen}(P_{W|X^L,Y^L,X^U}, P_{X,Y})] \tag{9}$$

For the covariate-shift scenario, we define the generalization error as $\overline{\mathrm{gen}}(P_{W|X^L,Y^L,X^U}, \mu_X^u \otimes P_{Y|X})$ where $\mu_X^u \otimes P_{Y|X}$ is the distribution of test data.

Now, we will show how our framework reduces to the Pseudo-labeling and entropy minimization. Let us consider a classification task with q labels, i.e., $|\mathcal{Y}| = q$. Suppose that the estimation of true underlying conditional distributions of labels given features and hypothesis, i.e., $\{\widehat{P}_{y=i|w,x}\}_{i=1}^q$, are available. For example, the output of the Softmax layer in a deep neural network could be considered as the estimation of true underlying conditional distributions of labels given features and hypotheses.

Pseudo-labeling: Consider the log-loss function as supervised loss function and consider the following function

$$\ell_u(w, x_i) = -\log(\max_j(\widehat{P}_{Y|w,x_i}(y_j|w,x_i)))$$

as unsupervised loss function, then our framework reduces to the Pseudo-labeling approach in (Lee et al., 2013).

Entropy Minimization: Let us consider the following unsupervised loss function in classification problem:

$$\ell_u(w, x_i) = \sum_{j=1}^{q} \widehat{P}_{Y|w,x_i}(y_j|w,x_i)\ell(w,x_i,y_j). \tag{10}$$

Now, if we choose negative log-loss function as supervised loss function,

$$\ell(w,x,y) = -\log(\widehat{P}_{Y|w,x}(y|w,x)),$$

then, the unsupervised loss function (10) would be equal to the conditional entropy,

$$\begin{aligned}\ell_u(w,x_i) &= H(\widehat{P}_{Y|w,x_i}) \qquad (11)\\ &= -\sum_{j=1}^{q} \widehat{P}_{Y|w,x_i}(y_j|w,x_i)\log(\widehat{P}_{Y|w,x_i}(y_j|w,x_i)),\end{aligned}$$

and our framework reduces to the entropy minimization (Grandvalet et al., 2005).

Our framework could be extended by choosing different supervised loss function in (10). For example, we could consider the squared log loss (Janocha and Czarnecki, 2016), i.e., $\ell(w,x,y) = -\log^2(\widehat{P}_{Y|w,x}(y|w,x))$, as supervised loss function and unsupervised loss function would be as follows:

$$\ell_u(w,x_i) = -\sum_{j=1}^{q} \widehat{P}_{Y|w,x_i}(y_j|w,x_i)\log^2(\widehat{P}_{Y|w,x_i}(y_j|w,x_i))$$

Another classification loss function is α-loss (Sypherd et al., 2019), i.e., $\ell(w,x,y) = \frac{\alpha}{\alpha-1}(1-\widehat{P}_{Y|w,x}^{1-1/\alpha}(y|w,x))$ for $\alpha \in (0,\infty)$, and the unsupervised loss function based on α-loss would be as follows:

$$\ell_u(w,x_i) = \frac{\alpha}{\alpha-1}\sum_{j=1}^{q} \widehat{P}_{Y|w,x_i}(y_j|w,x_i)(1-\widehat{P}_{Y|w,x_i}^{1-1/\alpha}(y_j|w,x_i)).$$

In entropy minimization (Grandvalet et al., 2005), the authors consider solely unlabeled features for conditional entropy. However, in our framework, we also consider the labeled features in the unsupervised empirical risk inspired by Remark 2. Actually, the labeled features can also help to improve the unsupervised performance of the SSL algorithm. We will show in Section 5, this helps us to have better performance in comparison to the case considering solely unlabeled features in entropy minimization method.

4 BOUNDING THE EXPECTED GENERALIZATION ERROR

We begin by offering an upper bound on the expected generalization error of the SSL scenario under covariate-shift by considering the conditional expectation of supervised loss function instead of unsupervised loss function in (7).

Theorem 1 (Proved in Appendix B). *Assume that the supervised loss functions, $l(w,x,y)$ is σ_l-sub-Gaussian* [2] *under the $\mu_X^u \otimes P_{Y|X}$ for all $w \in \mathcal{W}$ and $\ell_c(w,x)$ is σ_c-sub-Gaussian under marginal distribution μ_X^u for all $w \in \mathcal{W}$. The following expected generalization error upper bound under covariate-shift holds:*

$$\begin{aligned}&|\overline{gen}(P_{W|X^L,Y^L,X^U}, \mu_X^u \otimes P_{Y|X})| \leq \qquad (12)\\ &\beta\sqrt{\frac{2\sigma_l^2}{n}I(W;X^L,Y^L) + 2\sigma_l^2 D(\mu_X^l\|\mu_X^u)}\\ &\quad+\frac{n(1-\beta)}{n+m}\sqrt{\frac{2\sigma_c^2}{n}I(W;X^L) + 2\sigma_c^2 D(\mu_X^l\|\mu_X^u)}\\ &\quad+\frac{m(1-\beta)}{n+m}\sqrt{\frac{2\sigma_c^2}{m}I(W;X^U)}.\end{aligned}$$

If the supervised loss function is bounded in $[a,b]$, then the conditional expectation of supervised loss function, $\ell_c(w,x)$, is also bounded in $[a,b]$ and is $\frac{(b-a)}{2}$-sub-Gaussian under all distributions over $\mathcal{X}$ and all $w \in \mathcal{W}$ and we have $\sigma_l = \sigma_c = \frac{b-a}{2}$.

It is interesting to interpret each term in (12). The first term,

$$\sqrt{\frac{2\sigma_l^2}{n}I(W;X^L,Y^L) + 2\sigma_l^2 D(\mu_X^l\|\mu_X^u)},$$

can be interpreted as an upper bound on the supervised learning part of the SSL algorithm. We also have the term $D(\mu_X^l\|\mu_X^u)$, which can be interpreted as the cost of covariate-shift between training and test feature distributions. The second term,

$$\frac{n}{n+m}\sqrt{\frac{2\sigma_c^2}{n}I(W;X^L) + 2\sigma_c^2 D(\mu_X^l\|\mu_X^u)},$$

could be interpreted as an upper bound on the unsupervised performance of the SSL algorithm by considering conditional expectation of supervised loss function and labeled features. And finally,

$$\frac{m}{n+m}\sqrt{\frac{2\sigma_c^2}{m}I(W;X^U)},$$

could be interpreted as an upper bound on unsupervised performance of the SSL algorithm by considering the unlabeled data.

Now, we provide another expected generalization error upper bound by substituting the conditional expectation of supervised loss function with the unsupervised loss function.

Proposition 1 (Proved in Appendix C). *Assume that the supervised loss functions, $l(w,x,y)$ is σ_l-sub-Gaussian under the $\mu_X^u \otimes P_{Y|X}$ for all $w \in \mathcal{W}$*

[2] A random variable X is σ-subgaussian if $E[e^{\lambda(X-E[X])}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for all $\lambda \in \mathbb{R}$.

and $\ell_u(w,x)$ is σ_u-sub-Gaussian under marginal distribution μ_X^u for all $w \in \mathcal{W}$. The following upper bound holds on the expected generalization error under covariate-shift by considering the test data distribution as $\mu_X^u \otimes P_{Y|X}$:

$$\begin{aligned}
&|\overline{gen}(P_{W|X^L,Y^L,X^U}, \mu_X^u \otimes P_{Y|X})| \le \qquad (13)\\
&\beta\sqrt{\frac{2\sigma_l^2}{n}I(W;X^L,Y^L) + 2\sigma_l^2 D(\mu_X^l \| \mu_X^u)}\\
&+ \frac{n(1-\beta)}{n+m}\sqrt{\frac{2\sigma_u^2}{n}I(W;X^L) + 2\sigma_u^2 D(\mu_X^l\|\mu_X^u)}\\
&+ \frac{m(1-\beta)}{n+m}\sqrt{\frac{2\sigma_u^2}{m}I(W;X^U)}\\
&+ (1-\beta)\Delta^{SSL},\\
&\textit{where} \quad \Delta^{SSL} = \mathbb{E}_{P_W \otimes \mu_X^u}[\ell_c(W,X) - \ell_u(W,X)].
\end{aligned}$$

The term Δ^{SSL} in (13) can be interpreted as the estimation error of conditional distributions of labels given features (prediction uncertainty), (Guo et al., 2017), under the learning algorithm. Now we provide an upper bound on the absolute value of Δ^{SSL} for the classification task.

Corollary 1 (Proved in Appendix D). *Consider the same assumption as in Proposition 1. We suppose that the supervised loss function $\ell(w,x,y)$ is also σ_u-sub-Gaussian under distribution $P_{Y|X=x}$ for all $x \in \mathcal{X}$ and $w \in \mathcal{W}$. The following upper bound holds on estimation error of conditional distributions in classification task:*

$$|\Delta^{SSL}| \le \sqrt{2{\sigma_u}^2 D(\widehat{P}_{Y|W,X}\|P_{Y|X}|P_W \otimes \mu_X^u)}. \qquad (14)$$

Remark 3 (Calibration). *Based on Corollary 1, a poor network calibration would result the looser generalization error upper bound in compare to a calibrated network. The same fact is also discussed by Rizve et al. (2020).*

In Proposition 1, if the distribution of μ_X^u is not absolutely continuous with respect to μ_X^l, then we have $D(\mu_X^l||\mu_X^u) = \infty$ leading up to a vacuous upper bound. In the following, we therefore propose an alternative upper bound based on total variation that bypasses this issue.

Corollary 2 (Proved in Appendix E). *Assume that the supervised loss functions, $l(w,x,y)$ is bounded in $[0,L_l]$ and $\ell_u(w,x)$ is bounded in $[0,L_u]$. The following upper bound holds on the expected generalization error under covariate-shift by considering the test data distribution as $\mu_X^u \otimes P_{Y|X}$:*

$$\begin{aligned}
&|\overline{gen}(P_{W|X^L,Y^L,X^U}, \mu_X^u \otimes P_{Y|X})| \le \qquad (15)\\
&\beta\left(\sqrt{\frac{L_l^2}{2n}I(W;X^L,Y^L)} + 2L_l\mathbb{TV}(\mu_X^l,\mu_X^u)\right)\\
&+ \frac{n(1-\beta)}{n+m}\left(\sqrt{\frac{L_u^2}{2n}I(W;X^L)} + 2L_u\mathbb{TV}(\mu_X^l,\mu_X^u)\right)\\
&+ \frac{m(1-\beta)}{n+m}\sqrt{\frac{L_u^2}{2m}I(W;X^U)}\\
&+ (1-\beta)\Delta^{SSL},
\end{aligned}$$

where $\Delta^{SSL} = \mathbb{E}_{P_W\otimes\mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]$.

In Corollary 1, if the estimation of conditional probabilities, i.e., $\widehat{P}_{Y|W,X}$, is not absolutely continuous with respect to true conditional probability, i.e., $P_{Y|X}$, for all $x \in \mathcal{X}$ and $w \in \mathcal{W}$, then we have $D(\widehat{P}_{Y|W,X}\|P_{Y|X}|P_W \otimes \mu_X^u) = \infty$. In the following Corollary, we derive an upper bound for estimation error of conditional distributions of labels given features, in terms of total variation distance which is bounded.

Corollary 3 (Proved in Appendix F). *Consider the same assumptions as in Corollary 2, The following upper bound holds on estimation error of conditional distributions in classification task:*

$$|\Delta^{SSL}| \le 2L_u\mathbb{TV}(\widehat{P}_{Y|W,X}, P_{Y|X}|P_W \otimes \mu_X^u), \qquad (16)$$

where $\mathbb{TV}(\widehat{P}_{Y|W,X}\|P_{Y|X}|P_W \otimes \mu_X^u) = \mathbb{E}_{P_W\otimes\mu_X^u}[\mathbb{TV}(\widehat{P}_{Y|W,X}, P_{Y|X})]$.

It is worthwhile to mention that the results in Theorem 1 and Proposition 1 could also be applied to the SSL algorithms for traditional SSL scenario (no covariate-shift), where $\mu_X^u = \mu_X^l$.

In the following, we provide a convergence rate for the expected generalization error of SSL algorithms.

Corollary 4 (Proved in Appendix G). *Consider the same assumptions as in Proposition 1 for traditional SSL scenario, $\mu_X^l = \mu_X^u$. Consider also hypothesis space is countable, $|\mathcal{W}| = k$, and $\left|\mathbb{E}_{P_W\otimes\mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right| \le \sqrt{2\sigma_u^2}\epsilon$. Then, the following upper bounds holds on the expected generalization error of the SSL algorithm:*

$$\begin{aligned}
&\overline{gen}(P_{W|X^L,Y^L,X^U}, P_{X,Y}) \le \beta\sqrt{\frac{2\sigma_l^2\log(k)}{n}} \qquad (17)\\
&\quad + (1-\beta)\sqrt{2\sigma_u^2}\left(\sqrt{\frac{\log(k)}{(n+m)}} + \epsilon\right).
\end{aligned}$$

If the estimation error of unsupervised loss function in Corollary 4 is negligible ($\epsilon \to 0$), the convergence rate of the upper bound in (17), would be as follows:

$$\mathcal{O}\left(\frac{\beta}{\sqrt{n}} + \frac{(1-\beta)}{\sqrt{n+m}}\right). \qquad (18)$$

The convergence rate in (18) depends on the choice of β. If we consider $\beta = \frac{n}{n+m}$, we have

$$\mathcal{O}\left(\max\left(\frac{\sqrt{n}}{n+m}, \frac{m}{(n+m)^{3/2}}\right)\right), \qquad (19)$$

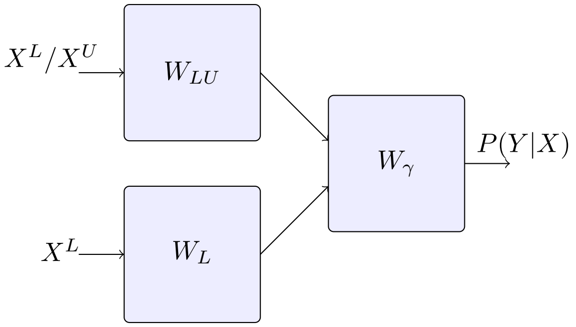


Figure 1: Structure of The CSSL Method

where shows if m is sufficiently large and n is relatively small, then SSL algorithm's generalization error upper bound's convergence rate would be better than the generalization error upper bound's convergence rate for the supervised learning algorithm, $\mathcal{O}(\frac{1}{\sqrt{n}})$, (Xu and Raginsky, 2017).

5 CSSL METHOD AND EXPERIMENTS

There are two inspirations from our theoretical results. First, as shown in Proposition 1 and Corollary 1, the estimation of conditional distributions for both labeled and unlabeled data plays an important role in the performance of SSL algorithm. Second, the unsupervised empirical risk L_E^{UL}, given in (6), is a function of both labeled and unlabeled data and it can help to have the convergence rate as shown in (18).

Considering these two inspirations, we now propose the Covariate-shift SSL (CSSL) method (the structure of our model is shown in Figure 1). The unsupervised empirical risk, L_E^{UL}, is expressed using the unsupervised loss function (11), which itself is dependent on the conditional distribution estimation. And, we have this assumption that the conditional distributions remain invariant for labeled and unlabeled data under covariate-shift. Based on this assumption, in our model, we consider one shared block of parameters (hypothesis), W_γ, to produce the estimation of conditional probability for both labeled and unlabeled data, i.e., X^L and X^U, respectively. As the distribution of X^L and X^U is different, we consider two disjoint blocks of parameters W_L and W_{LU}. The input of W_L is only the labeled data, while the inputs to the W_{LU} are both labeled and unlabeled data. If we only use X^U using the loss L_E^{UL} for the training of W_{LU}, then the model converges to extreme points (producing only zeros and ones at the output). Hence, feeding X_L to W_{LU} is important to avoid converging to degenerated cases. Based on Figure 1, the empirical loss defined in (7) can be written as follows,

$$L_E^{SL}(W_L, W_{LU}, W_\gamma, x^L, y^L) = \frac{1}{n}\sum_{i=1}^{n} \ell(W_L, W_\gamma, x_i^L, y_i^L) + \frac{1}{n}\sum_{i=1}^{n} \ell(W_{LU}, W_\gamma, x_i^L, y_i^L) \tag{20}$$

$$L_E^{UL}(W_L, W_{LU}, W_\gamma, x^L, x^U) = \frac{1}{n+m}\left(\sum_{i=1}^{n} \ell_u(W_L, W_\gamma, x_i^L) + \sum_{j=1}^{m} \ell_u(W_{LU}, W_\gamma, x_j^U)\right) \tag{21}$$

$$L_E^{SSL}(W_L, W_{LU}, W_\gamma, x^L, y^L, x^U) = \beta L_E^{SL}(W_L, W_{LU}, W_\gamma, x^L, y^L) + (1-\beta) L_E^{UL}(W_L, W_{LU}, W_\gamma, x^L, x^U), 0 \leq \beta \leq 1 \tag{22}$$

Now, we show the performance of our CSSL method using two experiments. In the first experiment, we use synthetic data, and in the second experiment, we use in MNIST dataset (LeCun and Cortes, 2010).

5.1 Synthetic data

In the first experiment, we use the synthetic data generated inspired by the first experiments of (Grandvalet et al., 2005) and (Kügelgen et al., 2019). We need to create the dataset and impose covariate-shift while satisfying two conditions. First, $p(Y|X)$ should remain constant with the covariate shift. Secondly, we need to make sure that unlabeled data are indeed useful in a semi-supervised learning setup. As discussed in (Janzing and Schölkopf, 2015) and (Kügelgen et al., 2019), this can be achieved by ensuring that $(X \to Y)$ does not hold. This is because, if $(X \to Y)$ holds then $p(X)$ and $p(Y|X)$ are independent mechanisms (Kügelgen et al., 2019). Thus, we consider a scenario where we have the following causal learning setting:

$$X_C \to Y \to X_E. \tag{23}$$

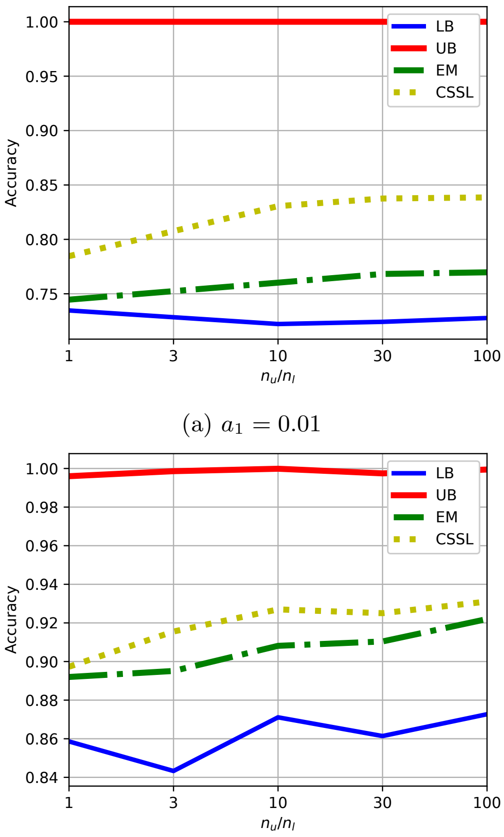
Here, X_C denotes the cause features, and X_E denotes the effect features. This scenario frequently arises in practice. For example, in healthcare, X_C can be genetic characteristics, living conditions, etc., Y could represent the illness, and X_E can represent symptoms of the illness like coughing, fever, etc.

In our synthetic dataset, features have dimension of 50. The first 30 features, are the cause features X_C, drawn from a mixture of two multivariate Gaussian distributions. Similar to (Grandvalet et al., 2005), the first Gaussian distribution is $\mathcal{N}((a_1, \cdots, a_1), s_1 I)$ and the second one is $\mathcal{N}((-a_1, \cdots, -a_1), s_1 I)$. The mixing probability π is $(0.5, 0.5)$. The binary label Y is defined as follows

$$Y = \begin{cases} 1 & \text{if} \quad \epsilon_Y < \sigma(\sum_{i=1}^{30} x_i) \\ 0 & \text{if} \quad \epsilon_Y > \sigma(\sum_{i=1}^{30} x_i) \end{cases} \quad , \quad \epsilon_Y \sim U(0,1).$$

Here $\sigma(x) = (1+e^{-x})^{-1}$ is the logistic sigmoid function. The effect features X_E is of dimension 20, and is defined as

$$X_E = \begin{cases} a_2 + \epsilon_E & \text{if} \quad Y = 1 \\ -a_2 + \epsilon_E & \text{if} \quad Y = 0 \end{cases} \qquad \epsilon_Y \sim \mathcal{N}(0, s_2 I). \tag{24}$$



(a) $a_1 = 0.01$

(b) $a_1 = 0.03$

Figure 2: Comparison of Accuracy of Methods For Two Settings of Synthetic Data

Note that variable a_1 determines how far apart are the two Gaussian mixtures. As a_1 increases the expected value of $|\sum_{i=1}^{30} x_i|$ increases. This means that X_C will become a better predictor of Y. Similarly, a_2 determines how good Y can be predicted using X_E. The covariate shift will be applied by changing a_1. We start with a small value for a_1, which means the predictor will rely on X_E for predicting Y, and then for the unlabeled features, we increase a_1. Now it is easy to see that both of our conditions are satisfied with this method of data generation. The causal learning setting holds $(X_C \to Y \to X_E)$ immediately as a consequence of the way we generate data. For the other condition we have

$$p(Y|X_C, X_E) = \frac{p(X_C, X_E|Y)p(Y)}{p(X_C, X_E)} \tag{25}$$

$$= \frac{p(X_C|Y)p(X_E|Y)p(Y)}{p(X_C)p(X_E|X_C)} \tag{26}$$

$$= \frac{\frac{p(Y|X_C)p(X_C)}{p(Y)}p(X_E|Y)p(Y)}{p(X_C)p(X_E|X_C)} \tag{27}$$

$$= \frac{p(Y|X_C)p(X_E|Y)}{\sum_y p(X_E|y, X_C)p(y|X_C)} \tag{28}$$

$$= \frac{p(Y|X_C)p(X_E|Y)}{\sum_y p(X_E|y)p(y|X_C)}, \tag{29}$$

where (25) and (27) hold from Bayes rule, and we used $(X_C \to Y \to X_E)$ in (26) and (29). This shows that $p(Y|X_C, X_E)$ remains invariant if we only change $p(X_C)$.

We have used a single layer fully connected network to implement each of W_L, W_{LU}, and W_γ. In particular. W_L and W_{LU} are neural networks with an input dimension of 50 and output dimension of 10, and with a ReLU activation function. Whereas W_γ gets ten inputs and has two outputs, a softmax function is used at the end to produce the required conditional distributions. The result is reported in Figure 2. The performance of the entropy minimization method (Grandvalet et al., 2005) is also presented. The lower bound is obtained by using only labeled data, and the upper bound is when we used true labels of unlabeled data to train in a supervised manner. We used a three-layer neural network for these two methods, a concatenation of W_L and W_γ. The value of β (and regularization term in EM) can be tuned using ten-fold cross-validation (we have $\beta = 0.02$). In Figure 2, the first figure is corresponding to a scenario where $a_1 = 0.01$, small a_1 means that X_C is not informative and the supervised model relies on X_E for predicting Y, we increase a_1 significantly for unlabeled data $a_1 = 0.8$ (thus the upper bound model always predict correctly). In the second figure, we have a more subtle change in a_1 and also X_E is more noisy forcing models to consider both X_E and X_C (more details about the experiments is reported in Appendix H). In both cases, our proposed method outperforms EM.

5.2 MNIST

In this experiment, we use a hand-written digits dataset, MNIST (LeCun and Cortes, 2010). In order to create covariate shift we impose a selection bias in labeled and unlabeled data. In the labeled data, we choose the majority of images (90 percent of the labeled dataset) from numbers with labels 0 to 4; the remaining 10 percent of the labeled dataset are drawn from images with labels 5 to 9. We reverse this ratio for the unlabeled dataset, with 90 percent of data having labels 5 to 9 and 10 percent with labels 0 to 4. Note that our two conditions in Section 5.1 are satisfied for this

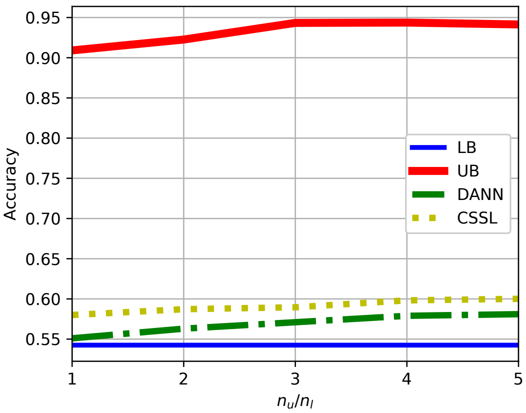


Figure 3: Comparison of Accuracy of CSSL, Domain Adversarial Method (DANN), Lower And Upper Bounds With Varying Ratio of Unlabeled Data For MNIST Dataset.

experiment. First, it is clear that the conditional distribution of $p(Y|X)$ will not change with the selection bias we imposed. Secondly, no direct causal link exists between X and Y. This dataset has been widely used in SSL settings (e.g., in (Ganin et al., 2016)), and it is shown that unlabeled data can indeed improve the performance of the model.

In Figure 3, we present the results for this experiment. Similar to the previous experiment, the lower and upper bounds are derived by training a supervised model using only labeled data, and both labeled and unlabeled data (with true labels), respectively. We also report the performance of the domain adversarial method (Ganin et al., 2016) for comparison. Note that in comparison to domain adversarial approach (Ganin et al., 2016), the final block W_γ in CSSL is trained by both labeled and unlabeled data. However, in domain adversarial approach, the final block is trained solely based on labeled data. Here, we have 1000 labeled images, and we vary the number of unlabeled images (note that because of the limited number of images, we cannot use arbitrarily large numbers of unlabeled data). The networks W_L and W_{LU} have three convolutional layers, and W_γ has only one fully connected layer. We used a similar structure for DANN, the feature extractor network is the same as W_L (or W_{LU}), and the classifier is similar to W_γ.

6 CONCLUSION

We provide a framework for SSL algorithms that can be reduced to other popular SSL algorithms, including entropy minimization and Pseudo-labeling. Inspired by our framework, we propose new expected generalization error upper bounds based on some information measures distance under the covariate-shift assumption, which illuminates the importance of estimating conditional distributions of labels given features. We also provide an upper bound on the estimation error of conditional distributions. Finally, we propose a method for SSL algorithms under covariate-shift, which outperforms entropy minimization under covariate-shift. This work motivates further investigation of other supervised loss functions in SSL algorithms. For example, using our framework, we can extend the support vector machine approach based on the Hinge loss function to include unlabeled data. The calibration algorithms (Guo et al., 2017) can be applied in our method to see if they will reduce the estimation error of conditional distributions. Our theoretical results and our method are based on covariate-shift assumption, and as a feature work could be extended to other scenarios, e.g., concept drift.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback, which helped us to improve the paper greatly. Gholamali Aminian is supported by the Royal Society Newton International Fellowship, grant no. NIF\R1 \192656.

References

Gholamali Aminian, Yuheng Bu, Laura Toni, Miguel Rodrigues, and Gregory Wornell. An exact characterization of the generalization error for the gibbs algorithm. *Advances in Neural Information Processing Systems*, 34, 2021a.

Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Jensen-shannon information based characterization of the generalization error of learning algorithms. In *2020 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2021b.

Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Information-theoretic bounds on the moments of the generalization error of learning algorithms. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 682–687. IEEE, 2021c.

Amir R Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *NeurIPS*, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic* *theory of independence*. Oxford university press, 2013.

Yuheng Bu, Weihao Gao, Shaofeng Zou, and Venugopal Veeravalli. Information-theoretic understanding of population risk improvement with model compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3300–3307, 2020a.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 2020b.

Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela Van Der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International Conference on Machine Learning*, pages 1392–1402. PMLR, 2020.

Olivier Chapelle, Jason Weston, and Bernhard Scholkopf. Cluster kernels for semi-supervised learning. *Advances in neural information processing systems*, pages 601–608, 2003.

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

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate? In *Conference on Learning Theory*, pages 1500–1518. PMLR, 2019.

Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33, 2020.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 2020.

Haiyun He, Hanshu Yan, and Vincent YF Tan. Information-theoretic generalization bounds for iterative semi-supervised learning. *arXiv preprint arXiv:2110.00926*, 2021.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019.

Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *Schedae Informaticae*, 25:49–59, 2016.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10): 5168–5194, 2010.

Dominik Janzing and Bernhard Schölkopf. Semi-supervised interpolation in an anticausal learning scenario. *The Journal of Machine Learning Research*, 16(1):1923–1948, 2015.

Masanori Kawakita and Takafumi Kanamori. Semi-supervised learning with density-ratio estimation. *Machine learning*, 91(2):189–209, 2013.

Julius Kügelgen, Alexander Mey, and Marco Loog. Semi-generative modelling: Covariate-shift adaptation with cause and effect features. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1361–1369. PMLR, 2019.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. *public*, 2010. URL http://yann.lecun.com/exdb/mnist/.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.

Adrian Tovar Lopez and Varun Jog. Generalization error bounds using wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2018.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Mohammad Saeed Masiha, Amin Gohari, Mohammad Hossein Yassaee, and Mohammad Reza Aref. Learning under distribution mismatch and model misspecification. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.

Alexander Mey and Marco Loog. Improvability through semi-supervised learning: A survey of theoretical results. *arXiv preprint arXiv:1908.09574*, 2019.

Gang Niu, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, and Masashi Sugiyama. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3239–3250, 2018.

Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(7), 2007.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Kenneth Joseph Ryan and Mark Vere Culp. On semi-supervised linear regression in covariate shift problems. *The Journal of Machine Learning Research*, 16(1):3183–3217, 2015.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M Mooij. On causal and anticausal learning. In *ICML*, 2012.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Tyler Sypherd, Mario Diaz, John Kevin Cava, Gautam Dasarathy, Peter Kairouz, and Lalitha Sankar. A tunable loss function for robust classification: Calibration, landscape, and generalization. *arXiv preprint arXiv:1906.02314*, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2020.

Hao Wang, Mario Diaz, José Cândido S Santos Filho, and Flavio P Calmon. An information-theoretic view of generalization via wasserstein distance. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 577–581. IEEE, 2019.

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*, 2020.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021.

Jingge Zhu. Semi-supervised learning: the case when unlabeled data is equally useful. In *Conference on Uncertainty in Artificial Intelligence*, pages 709–718. PMLR, 2020.

Supplementary Material:
An Information-theoretical Approach to Semi-supervised Learning under Covariate-shift

A SSL Empirical Risk Discussion

Let's consider SSL empirical risk based on conditional expectation of supervised loss function as follows:

$$L_E(w,x^L,y^L,x^U) \triangleq \frac{\beta}{n}\sum_{i=1}^{n}\ell(w,x_i^L,y_i^L)+\frac{(1-\beta)}{n+m}\left(\sum_{i=1}^{n}\ell_c(w,x_i^L)+\sum_{j=1}^{m}\ell_c(w,x_j^U)\right),0\leq\beta\leq 1 \tag{30}$$

We have:

$$\mathbb{E}_{P_{XY}}[\ell_c(w,x)]=\mathbb{E}_{P_{XY}}[\ell(w,x,y)] \tag{31}$$

Using (31), it could be shown that empirical risk based on conditional expectation of supervised loss function (30) is an unbiased estimator of population risk:

$$\begin{aligned}\mathbb{E}_{P_{XY}}[L_E(w,x^L,y^L,x^U)] &= \beta\mathbb{E}_{P_{XY}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(w,x_i^L,y_i^L)\right]+(1-\beta)\mathbb{E}_{P_{XY}}\left[\frac{1}{n+m}\left(\sum_{i=1}^{n}\ell_c(w,x_i^L)+\sum_{j=1}^{m}\ell_c(w,x_j^U)\right)\right] \qquad (32)\\ &= \beta\mathbb{E}_{P_{XY}}[\ell(w,x,y)]+(1-\beta)\mathbb{E}_{P_{XY}}[\ell(w,x,y)]\\ &= L_P(w,P_{XY}),\quad 0\leq\beta\leq 1\end{aligned}$$

B Proof of Theorem 1

We consider the unsupervised empirical risk functions based on the conditional expectation of supervised loss function, $\ell_c(w,x)$, in the following:

$$\overline{\text{gen}}(P_{W|X^L,Y^L,X^U},\mu_X^u\otimes P_{Y|X})=\mathbb{E}_{P_{W,X^L,Y^L,X^U}}[L_P(W,\mu_X^u\otimes P_{Y|X})-L_E^{\text{SSL}}(W,X^L,Y^L,X^U)] \tag{33}$$

$$=\mathbb{E}_{P_{W,X^L,Y^L,X^U}}[L_P(W,\mu_X^u\otimes P_{Y|X})-\beta L_E^{SL}(W,X^L,Y^L)-(1-\beta)L_E^{UL}(W,X^L,X^U)]= \tag{34}$$

$$\beta(\mathbb{E}_{P_{W,X^L,Y^L}}[L_P(W,\mu_X^u\otimes P_{Y|X})-L_E^{SL}(W,X^L,Y^L)] \tag{35}$$

$$+(1-\beta)\mathbb{E}_{P_{W,X^U,X^L}}[L_P(W,\mu_X^u\otimes P_{Y|X})-L_E^{UL}(W,X^L,X^U)]=$$

$$\beta\mathbb{E}_{P_{W,X^L,Y^L}}[L_P(W,\mu_X^u\otimes P_{Y|X})-L_E^{SL}(W,X^L,Y^L)] \tag{36}$$

$$+(1-\beta)\frac{n}{m+n}\mathbb{E}_{P_{W,X^L}}[L_P(W,\mu_X^u\otimes P_{Y|X})-L_E^{UL}(W,X^L)]$$

$$+(1-\beta)\frac{m}{m+n}\mathbb{E}_{P_{W,X^U}}[L_P(W,\mu_X^u\otimes P_{Y|X})-L_E^{UL}(W,X^U)]$$

$$=\beta\left(\mathbb{E}_{P_W\otimes(\mu_X^u\otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)]-\mathbb{E}_{P_{W,X^L,Y^L}}[L_E^{SL}(W,X^L,Y^L)]\right) \tag{37}$$

$$+(1-\beta)\frac{n}{m+n}\left(\mathbb{E}_{P_W\otimes\mu_X^{u\ \otimes n}}[L_E^{UL}(W,X^L)]-\mathbb{E}_{P_{W,X^L}}[L_E^{UL}(W,X^L)]\right)$$

$$+(1-\beta)\frac{m}{m+n}\left(\mathbb{E}_{P_W\otimes\mu_X^{u\ \otimes m}}[L_E^{UL}(W,X^U)]-\mathbb{E}_{P_{W,X^U}}[L_E^{UL}(W,X^U)]\right),$$

where $L_E^{UL}(W,X^L) = \frac{1}{n}\sum_{i=1}^{n}\ell_c(w,x_i^L)$, $L_E^{UL}(W,X^U) = \frac{1}{m}\sum_{j=1}^{m}\ell_c(w,x_j^U)$, $P_{X^L} = {\mu_X^l}^{\otimes n}$ and $P_{X^U} = {\mu_X^u}^{\otimes m}$. Now we have:

$$\begin{align}
&\beta\left(\mathbb{E}_{P_W\otimes(\mu_X^u\otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_{W,X^L,Y^L}}[L_E^{SL}(W,X^L,Y^L)]\right) \tag{38}\\
&+(1-\beta)\frac{n}{m+n}\left(\mathbb{E}_{P_W\otimes{\mu_X^u}^{\otimes n}}[L_E^{UL}(W,X^L)] - \mathbb{E}_{P_{W,X^L}}[L_E^{UL}(W,X^L)]\right) \nonumber\\
&+(1-\beta)\frac{m}{m+n}\left(\mathbb{E}_{P_W\otimes{\mu_X^u}^{\otimes m}}[L_E^{UL}(W,X^U)] - \mathbb{E}_{P_{W,X^U}}[L_E^{UL}(W,X^U)]\right) \nonumber\\
&\leq \beta\left|\mathbb{E}_{P_{W,X^L,Y^L}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_W\otimes(\mu_X^u\otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)]\right| \tag{39}\\
&+(1-\beta)\frac{n}{m+n}\left|\mathbb{E}_{P_{W,X^L}}[L_E^{UL}(W,X^L)] - \mathbb{E}_{P_W\otimes{\mu_X^u}^{\otimes n}}[L_E^{UL}(W,X^L)]\right| \nonumber\\
&+(1-\beta)\frac{m}{m+n}\left|\mathbb{E}_{P_{W,X^U}}[L_E^{UL}(W,X^U)] - \mathbb{E}_{P_W\otimes{\mu_X^u}^{\otimes m}}[L_E^{UL}(W,X^U)]\right| \nonumber\\
&\leq \beta\sqrt{\frac{2\sigma_l^2 D(P_{W,X^L,Y^L}\|P_W\otimes(\mu_X^u\otimes P_{Y|X})^{\otimes n})}{n}} \tag{40}\\
&+(1-\beta)\frac{n}{m+n}\sqrt{\frac{2\sigma_c^2 D(P_{W,X^L}\|P_W\otimes{\mu_X^u}^{\otimes n})}{n}} \nonumber\\
&+(1-\beta)\frac{m}{m+n}\sqrt{\frac{2\sigma_c^2 D(P_{W,X^U}\|P_W\otimes{\mu_X^u}^{\otimes m})}{m}} \nonumber\\
&\leq \beta\sqrt{\frac{2\sigma_l^2}{n}I(W;X^L,Y^L) + 2\sigma_l^2 D(\mu_X^l\|\mu_X^u)} \tag{41}\\
&+\frac{n(1-\beta)}{n+m}\sqrt{\frac{2\sigma_c^2}{n}I(W;X^L) + 2\sigma_c^2 D(\mu_X^l\|\mu_X^u)} \nonumber\\
&+\frac{m(1-\beta)}{n+m}\sqrt{\frac{2\sigma_c^2}{m}I(W;X^U)} \nonumber
\end{align}$$

The result (40) follows from Donsker-Varadhan representation of KL divergence (Xu and Raginsky, 2017) and the result (41) follows from the fact that

$$D(P_{W,X^L}\|{\mu_X^u}^{\otimes n}\otimes P_W) = D(P_{X^L,W}\|{\mu_X^l}^{\otimes n}\otimes P_W) + nD(\mu_X^l\|\mu_X^u) = I(W;X^L) + nD(\mu_X^l\|\mu_X^u) \tag{42}$$

C Proof of Proposition 1

We consider the unsupervised empirical risk functions based on the unsupervised loss function in the following.

$$\begin{align}
&\overline{\text{gen}}(P_{W|X^L,Y^L,X^U},\mu_X^u\otimes P_{Y|X}) = L_P(W,\mu_X^u\otimes P_{Y|X}) - L_E^{\text{SSL}}(W,X^L,Y^L,X^U) \tag{43}\\
&= \mathbb{E}_{P_{W,X^L,Y^L,X^U}}[L_P(W,\mu_X^u\otimes P_{Y|X}) - \beta L_E^{SL}(W,X^L,Y^L) - (1-\beta)L_E^{UL}(W,X^L,X^U)] \tag{44}\\
&= \beta\mathbb{E}_{P_{W,X^L,Y^L}}[L_P(W,\mu_X^u\otimes P_{Y|X}) - L_E^{SL}(W,X^L,Y^L)] \tag{45}\\
&+(1-\beta)\mathbb{E}_{P_{W,X^U,X^L}}[L_P(W,\mu_X^u\otimes P_{Y|X}) - L_E^{UL}(W,X^L,X^U)] \nonumber\\
&= \beta\mathbb{E}_{P_{W,X^L,Y^L}}[L_P(W,\mu_X^u\otimes P_{Y|X}) - L_E^{SL}(W,X^L,Y^L)] \tag{46}\\
&+(1-\beta)\frac{n}{m+n}\mathbb{E}_{P_{W,X^L}}[L_P(W,\mu_X^u\otimes P_{Y|X}) - L_E^{UL}(W,X^L)] \nonumber\\
&+(1-\beta)\frac{m}{m+n}\mathbb{E}_{P_{W,X^U}}[L_P(W,\mu_X^u\otimes P_{Y|X}) - L_E^{UL}(W,X^U)] \nonumber\\
&= \beta\left(\mathbb{E}_{P_W\otimes(\mu_X^u\otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_{W,X^L,Y^L}}[L_E^{SL}(W,X^L,Y^L)]\right) \tag{47}\\
&+(1-\beta)\frac{n}{m+n}\left(\mathbb{E}_{P_W\otimes{\mu_X^u}^{\otimes n}}[L_E^{UL}(W,X^L)] - \mathbb{E}_{P_{W,X^L}}[L_E^{UL}(W,X^L)]\right) \nonumber
\end{align}$$

$$
\begin{aligned}
&+ (1-\beta)\frac{m}{m+n}\left(\mathbb{E}_{P_W\otimes\mu_X^u{}^{\otimes m}}[L_E^{UL}(W,X^U)] - \mathbb{E}_{P_{W,X^U}}[L_E^{UL}(W,X^U)]\right)\\
&+ (1-\beta)(\mathbb{E}_{P_W\otimes\mu_X^u}[\ell_c(W,X) - \ell_u(W,X)])\\
&\leq \beta\left|\mathbb{E}_{P_{W,X^L,Y^L}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_W\otimes(\mu_X^u\otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)]\right| \qquad (48)\\
&+ (1-\beta)\frac{n}{m+n}\left|\mathbb{E}_{P_{W,X^L}}[L_E^{UL}(W,X^L)] - \mathbb{E}_{P_W\otimes\mu_X^u{}^{\otimes n}}[L_E^{UL}(W,X^L)]\right|\\
&+ (1-\beta)\frac{m}{m+n}\left|\mathbb{E}_{P_{W,X^U}}[L_E^{UL}(W,X^U)] - \mathbb{E}_{P_W\otimes\mu_X^u{}^{\otimes m}}[L_E^{UL}(W,X^U)]\right|\\
&+ (1-\beta)\left|\mathbb{E}_{P_W\otimes\mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right|\\
&\leq \beta\sqrt{\frac{2\sigma_l^2 D(P_{W,X^L,Y^L}\|P_W\otimes(\mu_X^u\otimes P_{Y|X})^{\otimes n})}{n}} \qquad (49)\\
&+ (1-\beta)\frac{n}{m+n}\sqrt{\frac{2\sigma_u^2 D(P_{W,X^L}\|P_W\otimes\mu_X^u{}^{\otimes n})}{n}}\\
&+ (1-\beta)\frac{m}{m+n}\sqrt{\frac{2\sigma_u^2 D(P_{W,X^U}\|P_W\otimes\mu_X^u{}^{\otimes m})}{m}}\\
&+ (1-\beta)\left|\mathbb{E}_{P_W\otimes\mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right|\\
&\leq \beta\sqrt{\frac{2\sigma_l^2}{n}I(W;X^L,Y^L) + 2\sigma_l^2 D(\mu_X^l\|\mu_X^u)} \qquad (50)\\
&+ \frac{n(1-\beta)}{n+m}\sqrt{\frac{2\sigma_u^2}{n}I(W;X^L) + 2\sigma_c^2 D(\mu_X^l\|\mu_X^u)}\\
&+ \frac{m(1-\beta)}{n+m}\sqrt{\frac{2\sigma_u^2}{m}I(W;X^U)}\\
&+ (1-\beta)\left|\mathbb{E}_{P_W\otimes\mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right|,
\end{aligned}
$$

where $L_E^{UL}(W,X^L) = \frac{1}{n}\sum_{i=1}^n \ell_u(w,x_i^L)$, $L_E^{UL}(W,X^U) = \frac{1}{m}\sum_{j=1}^m \ell_u(w,x_j^U)$, $P_{X^L} = \mu_X^l{}^{\otimes n}$ and $P_{X^U} = \mu_X^u{}^{\otimes m}$. (49) follows from Donsker-Varadhan representation of KL divergence. (50) follows from the fact that

$$D(P_{X,W}\|\mu_X^u\otimes P_W) = D(P_{X,W}\|P_X\otimes P_W) + D(\mu_X^l\|\mu_X^u) = I(W;X) + D(\mu_X^l\|\mu_X^u) \qquad (51)$$

D Proof of Corollary 1

For the classification task (10), we have:

$$
\begin{aligned}
&\left|\Delta^{SSL}\right| = \left|\mathbb{E}_{P_W\otimes\mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right| = \qquad (52)\\
&\left|\mathbb{E}_{P_W\otimes\mu_X^u}\left[\sum_{j=1}^q \widehat{P}_{Y=y_j|W,X}\ell(W,X,y_j) - \sum_{j=1}^q P_{Y=y_j|X}\ell(W,X,y_j)\right]\right| \qquad (53)\\
&\leq \sqrt{2\sigma_u^2 D(P_W\otimes\mu_X^u\otimes P_{Y|X}\|P_W\otimes\mu_X^u\otimes\widehat{P}_{Y|W,X})} \qquad (54)\\
&= \sqrt{2\sigma_u^2 D(P_{Y|X}\|\widehat{P}_{Y|W,X}|P_W\otimes\mu_X^u)}, \qquad (55)
\end{aligned}
$$

where (54) is based on Donsker-Varadhan representation of KL divergence, (Boucheron et al., 2013).

E Proof of Corollary 2

We consider the unsupervised empirical risk functions based on the unsupervised loss function in the following.

$$
\begin{aligned}
&\overline{\text{gen}}(P_{W|X^L,Y^L,X^U},\mu_X^u\otimes P_{Y|X}) = L_P(W,\mu_X^u\otimes P_{Y|X}) - L_E^{\text{SSL}}(W,X^L,Y^L,X^U) \qquad (56)\\
&= \mathbb{E}_{P_{W,X^L,Y^L,X^U}}[L_P(W,\mu_X^u\otimes P_{Y|X}) - \beta L_E^{SL}(W,X^L,Y^L) - (1-\beta)L_E^{UL}(W,X^L,X^U)] \qquad (57)
\end{aligned}
$$

$$\begin{aligned}
&= \beta \mathbb{E}_{P_{W,X^L,Y^L}}[L_P(W,\mu_X^u \otimes P_{Y|X}) - L_E^{SL}(W,X^L,Y^L)] \\
&+ (1-\beta)\mathbb{E}_{P_{W,X^U,X^L}}[L_P(W,\mu_X^u \otimes P_{Y|X}) - L_E^{UL}(W,X^L,X^U)]
\end{aligned} \tag{58}$$

$$\begin{aligned}
&= \beta \mathbb{E}_{P_{W,X^L,Y^L}}[L_P(W,\mu_X^u \otimes P_{Y|X}) - L_E^{SL}(W,X^L,Y^L)] \\
&+ (1-\beta)\frac{n}{m+n}\mathbb{E}_{P_{W,X^L}}[L_P(W,\mu_X^u \otimes P_{Y|X}) - L_E^{UL}(W,X^L)] \\
&+ (1-\beta)\frac{m}{m+n}\mathbb{E}_{P_{W,X^U}}[L_P(W,\mu_X^u \otimes P_{Y|X}) - L_E^{UL}(W,X^U)]
\end{aligned} \tag{59}$$

$$\begin{aligned}
&= \beta \left(\mathbb{E}_{P_W \otimes (\mu_X^u \otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_{W,X^L,Y^L}}[L_E^{SL}(W,X^L,Y^L)]\right) \\
&+ (1-\beta)\frac{n}{m+n}\left(\mathbb{E}_{P_W \otimes {\mu_X^u}^{\otimes n}}[L_E^{UL}(W,X^L)] - \mathbb{E}_{P_{W,X^L}}[L_E^{UL}(W,X^L)]\right) \\
&+ (1-\beta)\frac{m}{m+n}\left(\mathbb{E}_{P_W \otimes {\mu_X^u}^{\otimes m}}[L_E^{UL}(W,X^U)] - \mathbb{E}_{P_{W,X^U}}[L_E^{UL}(W,X^U)]\right) \\
&+ (1-\beta)(\mathbb{E}_{P_W \otimes \mu_X^u}[\ell_c(W,X) - \ell_u(W,X)])
\end{aligned} \tag{60}$$

$$\begin{aligned}
&\leq \beta \left|\mathbb{E}_{P_{W,X^L,Y^L}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_W \otimes (\mu_X^l \otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)]\right| \\
&+ \beta \left|\mathbb{E}_{P_W \otimes (\mu_X^l \otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_W \otimes (\mu_X^u \otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)]\right| \\
&+ (1-\beta)\frac{n}{m+n}\left|\mathbb{E}_{P_{W,X^L}}[L_E^{UL}(W,X^L)] - \mathbb{E}_{P_W \otimes {\mu_X^u}^{\otimes n}}[L_E^{UL}(W,X^L)]\right| \\
&+ (1-\beta)\frac{n}{m+n}\left|\mathbb{E}_{P_W \otimes {\mu_X^l}^{\otimes n}}[L_E^{UL}(W,X^L)] - \mathbb{E}_{P_W \otimes {\mu_X^u}^{\otimes n}}[L_E^{UL}(W,X^L)]\right| \\
&+ (1-\beta)\frac{m}{m+n}\left|\mathbb{E}_{P_{W,X^U}}[L_E^{UL}(W,X^U)] - \mathbb{E}_{P_W \otimes {\mu_X^u}^{\otimes m}}[L_E^{UL}(W,X^U)]\right| \\
&+ (1-\beta)\left|\mathbb{E}_{P_W \otimes \mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right|
\end{aligned} \tag{61}$$

$$\begin{aligned}
&\leq \beta \left(\sqrt{\frac{L_l^2 I(W;X^L,Y^L)}{2n}} + 2L_l \mathbb{TV}(\mu_X^l,\mu_X^u)\right) \\
&+ \frac{n(1-\beta)}{m+n}\left(\sqrt{\frac{L_u^2 I(W;X^L)}{2n}} + 2L_u \mathbb{TV}(\mu_X^l,\mu_X^u)\right) \\
&+ (1-\beta)\frac{m}{m+n}\sqrt{\frac{L_u^2 I(W;X^U)}{2m}} \\
&+ (1-\beta)\left|\mathbb{E}_{P_W \otimes \mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right|,
\end{aligned} \tag{62}$$

where $L_E^{UL}(W,X^L) = \frac{1}{n}\sum_{i=1}^n \ell_u(w,x_i^L)$, $L_E^{UL}(W,X^U) = \frac{1}{m}\sum_{j=1}^m \ell_u(w,x_j^U)$, $P_{X^L} = {\mu_X^l}^{\otimes n}$ and $P_{X^U} = {\mu_X^u}^{\otimes m}$. (62) follows from Donsker-Varadhan representation of KL divergence, (Boucheron et al., 2013), and the variational representation of total variation (2),

$$\begin{aligned}
&\left|\mathbb{E}_{P_W \otimes (\mu_X^l \otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)] - \mathbb{E}_{P_W \otimes (\mu_X^u \otimes P_{Y|X})^{\otimes n}}[L_E^{SL}(W,X^L,Y^L)]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n |\mathbb{E}_{P_W \otimes \mu_X^l \otimes P_{Y_i|X_i}}[\ell(W,X_i,Y_i)] - \mathbb{E}_{P_W \otimes \mu_X^u \otimes P_{Y_i|X_i}}[\ell(W,X_i,Y_i)]| \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{TV}(P_W \otimes \mu_X^l \otimes P_{Y_i|X_i}, P_W \otimes \mu_X^u \otimes P_{Y_i|X_i}) \\
&= \mathbb{TV}(\mu_X^l,\mu_X^u).
\end{aligned} \tag{63}$$

F Proof of Corollary 3

For the classification task (10), we have:

$$\left|\Delta^{SSL}\right| = \left|\mathbb{E}_{P_W \otimes \mu_X^u}[\ell_c(W,X) - \ell_u(W,X)]\right| \tag{64}$$

$$= \left| \mathbb{E}_{P_W \otimes \mu_X^u} \left[\sum_{j=1}^{q} \widehat{P}_{Y=y_j|W,X} \ell(W,X,y_j) - \sum_{j=1}^{q} P_{Y=y_j|X} \ell(W,X,y_j) \right] \right| \tag{65}$$

$$\leq 2L_u \mathbb{E}_{P_W \otimes \mu_X^u} \left[\mathbb{TV}(\widehat{P}_{Y|W,X}, P_{Y|X}) \right], \tag{66}$$

where (54) is based on the variational representation of total variation (2).

G Proof of Corollary 4

Considering $H(W) \leq \log(k)$ and the fact that the $H(W)$ is the upper bound on the mutual information between W and any other random variable, we have:

$$\begin{aligned}
&|\overline{\text{gen}}(P_{W|X^L,Y^L,X^U}, \mu_X^u \otimes P_{Y|X})| && (67)\\
&\leq \beta\sqrt{\frac{2\sigma_l^2}{n} I(W;X^L,Y^L)} + \frac{n(1-\beta)}{n+m}\sqrt{\frac{2\sigma_u^2}{n} I(W;X^L)} + \frac{m(1-\beta)}{n+m}\sqrt{\frac{2\sigma_u^2}{m} I(W;X^U)} + (1-\beta)\Delta^{SSL} \\
&\leq \beta\sqrt{\frac{2\sigma_l^2}{n} I(W;X^L,Y^L)} + (1-\beta)\sqrt{\frac{2\sigma_u^2}{n+m} I(W;X^L,X^U)} + (1-\beta)\sqrt{2\sigma_u^2}\epsilon && (68)\\
&\leq \beta\sqrt{\frac{2\sigma_l^2}{n} H(W)} + (1-\beta)\sqrt{\frac{2\sigma_u^2}{n+m} H(W)} + (1-\beta)\sqrt{2\sigma_u^2}\epsilon && (69)\\
&\leq \beta\sqrt{\frac{2\sigma_l^2}{n}} \log(k) + (1-\beta)\sqrt{2\sigma_u^2}\left(\sqrt{\frac{\log(k)}{n+m}} + \epsilon\right). && (70)
\end{aligned}$$

H Experiment Details

We have used pytorch package for implementation of the code. For the first experiment (synthetic data), a single layer fully connected network is used to implement each of W_L, W_{LU}, and W_γ networks. In particular. W_L and W_{LU} are neural networks with an input dimension of 50 and output dimension of 10, and with a ReLU activation function. W_γ has ten inputs and two output nodes, a softmax function is used at the end to produce the required conditional distributions. The value of β (and regularization term in EM) can be tuned using ten-fold cross-validation, we have $\beta = 0.02$. Number of labeled data points for both figures is 300.

For Figure 2a, we have used the following parameters for producing the dataset: $a_1 = 0.01$, $s_1 = 0.05$, $a_2 = 0.01$, and $s_2 = 2$. For the unlabeled data and test data we chose $a_1 = 0.8$. Since, a_1 is large for unlabeled data, X_C will become a very good predictor of Y, thus the upper bound model always predict the output correctly.

For Figure 2b, we used these parameters for synthetic data: $a_1 = 0.03$, $s_1 = 0.05$, $a_2 = 0.01$, and $s_2 = 0.2$. Note, that here X_C is more informative in labeled data. Also, the variance of X_E is decreased, hence it becomes more useful and hence the lower bound has improved. We have made a more subtle change in a_1 for unlabeled data, we chose $a_1 = 0.3$ for unlabeled and test data. As a result upper bound is not always 1 anymore. ‘

H.1 Different distribution for unlabeled and test data

Here, we evaluate the performance of different methods when the distribution of test data is different from unlabeled data. When the data is collected sequentially it is possible that the distribution constantly changes, hence it is possible that the distribution of test data does not match with unlabeled data. In fact, it is an interesting direction for future work to consider non-stationary data stream, where the distribution of data constantly changes.

In Table 1, we consider the setup of first experiment (Figure 2a), while changing a_1 for the test data (recall that $a_1 = 0.8$ for unlabeled data). We repeat the experiment five times and report the mean and standard deviation of the accuracy of each method. We have 300 labeled data and 3000 unlabeled data. It can be seen that CSSL outperforms EM.

Table 1: Comparison of different methods when the distribution of unlabeled and test data is different

Value of a_1 for test dataset	$a_1 = 0.9$	$a_1 = 0.6$	$a_1 = 0.4$
Lower bound	0.761 ± 0.011	0.681 ± 0.010	0.621 ± 0.010
EM	0.874 ± 0.032	0.781 ± 0.036	0.705 ± 0.030
CSSL	$\mathbf{0.876 \pm 0.070}$	$\mathbf{0.792 \pm 0.065}$	$\mathbf{0.712 \pm 0.053}$
Upper bound	0.999 ± 0.0001	0.997 ± 0.004	0.977 ± 0.024