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

In this paper, an information-theoretical approach, inspired by Xu and Raginsky (2017) and Russo and Zou (2019), is applied to characterize the generalization error, we propose a new SSL algorithm inspired by our framework and the upper bounds on excess risk and generalization 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-theoretical 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
\[
\text{TV}(P,Q) = \frac{1}{2} \int_{\mathcal{Z}} |dP - dQ|
\]
and the variational representation of total variation distance is as follows (Polyanskiy and Wu, 2014):
\[
\text{TV}(P,Q) = \sup_{g \in \mathcal{G}_L} \{ \mathbb{E}[g(P)] - \mathbb{E}[g(Q)] \}
\]
where \( \mathcal{G}_L = \{ g : \mathcal{Z} \to \mathbb{R}, ||g||_\infty \leq L \} \). Note that the total variation is bounded, \( \text{TV}(P,Q) \leq 1 \).

## 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-labeling 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-labeling – 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 $\alpha$-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, Masihia 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 $(X^L, Y^L) = \{(X^l_i, Y^l_i)\}_{i=1}^n$ 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 $\mu_X \otimes P_{Y|X}$, and $X^U = \{X^u_i\}_{i=1}^m$ a training unlabelled
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:

  $\ell_c(w, x) \triangleq \int_{\mathcal{Y}} \ell(w, x, y)P_{Y|X}(y|x)dy$ \hspace{1cm} (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:

$$L_P(w, P_{X,Y}) \triangleq \int_{\mathcal{X} \times \mathcal{Y}} \ell(w, x, y)P_{X,Y}(x, y)dx dy$$ \hspace{1cm} (4)

$$L_{E}^{SL}(w, x^L, y^L) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(w, x^L_i, y^L_i)$$ \hspace{1cm} (5)

$$L_{E}^{UL}(w, x^L, x^U) \triangleq \frac{1}{n+m} \left( \sum_{i=1}^{n} \ell_u(w, x^L_i) + \sum_{j=1}^{m} \ell_u(w, x^U_j) \right)$$ \hspace{1cm} (6)

$$L_{E}^{SSL}(w, x^L, y^L, x^U) \triangleq \beta L_{E}^{SL}(w, x^L, y^L) + (1 - \beta) L_{E}^{UL}(W, X^L, X^U),$$ \hspace{1cm} (7)

where $\beta \in [0, 1]$ balances between the supervised and unsupervised empirical risk.

**Remark 1. [Choice of $\beta$]** 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:

$$\text{gen}(P_{W|X,Y,L,X,Y}, P_{X,Y}) \triangleq L_P(w, P_{X,Y}) - L_E^{SSL}(w, x^L, y^L, x^U)$$ \hspace{1cm} (8)

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

$$\mathbb{E}_{P_{W,X,L,Y,L,Y,U}}[\text{gen}(P_{W|X,Y,L,Y,U}, P_{X,Y})]$$ \hspace{1cm} (9)

For the covariate-shift scenario, we define the generalization error as $\text{gen}(P_{W|X,Y,L,Y,U}, \mu^X \otimes P_{Y|X})$ where $\mu^X \otimes P_{Y|X}$ 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., $\{P_{Y=y|w,x}\}_{y=1}^{q}$, 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

$$\ell_u(w, x_i) = -\log(\max_j \hat{P}_Y|w,x_i(y_j|w,x_i))$$

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:

$$\ell_u(w, x_i) = \sum_{j=1}^{q} \hat{P}_{Y|w,x_i}(y_j|w,x_i)\ell(w, x_i, y_j).$$ \hspace{1cm} (10)
We begin by offering an upper bound on the expected supervised loss function,
\[ \ell(w, x, y) = -\log(\hat{P}_{Y|w,x}(y|w, x)), \]
then, the unsupervised loss function (10) would be equal to the conditional entropy,
\[ \ell_u(w, x_i) = H(\hat{P}_{Y|w,x_i}) \]
(11)
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., \( \ell(w, x, y) = -\log^2(\hat{P}_{Y|w,x}(y|w, x)) \), as supervised loss function and unsupervised loss function would be as follows:
\[ \ell_u(w, x_i) = -\sum_{j=1}^{q} \hat{P}_{Y|w,x_i}(y_j|w, x_i) \log^2(\hat{P}_{Y|w,x_i}(y_j|w, x_i)) \]

Another classification loss function is \( \alpha \)-loss (Sypherd et al., 2019), i.e., \( \ell(w, x, y) = \frac{\alpha}{\alpha-1} (1 - \hat{P}_{Y|w,x}^{-1/\alpha}(y|w, x)) \) for \( \alpha \in (0, \infty) \), and the unsupervised loss function based on \( \alpha \)-loss would be as follows:
\[ \ell_u(w, x_i) = \frac{\alpha}{\alpha-1} \sum_{j=1}^{q} \hat{P}_{Y|w,x_i}(y_j|w, x_i) (1 - \hat{P}_{Y|w,x_i}^{-1/\alpha}(y_j|w, x_i)) \]

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 \( \sigma_l \)-sub-Gaussian \footnote{A random variable \( X \) is \( \sigma \)-subgaussian if \( E[e^{\lambda(X-E[X])}] \leq e^{\frac{\sigma^2 \lambda^2}{2}} \) for all \( \lambda \in \mathbb{R} \).} under the \( \mu_X^w \otimes P_{Y|X} \) for all \( w \in \mathcal{W} \) and \( \ell_u(w, x) \) is \( \sigma_e \)-sub-Gaussian under marginal distribution \( \mu_X^w \) for all \( w \in \mathcal{W} \). The following expected generalization error upper bound under covariate-shift holds:
\[
\mathbb{E} \left[ \left| P_{W|X^L, Y^L, U^L, \mu_X^w \otimes P_{Y|X}} \right| \leq \right. \beta \sqrt{ \frac{2\sigma_l^2}{n} I(W; X^L, Y^L) + 2\sigma_l^2 D(\mu_X^w \| \mu_X^L) + n(1-\beta) \sqrt{ \frac{2\sigma_e^2}{n} I(W; X^L) + 2\sigma_e^2 D(\mu_X^w \| \mu_X^L) + \frac{m}{n+m} \sqrt{ \frac{2\sigma_e^2}{m} I(W; X^U) } \right] \]
\]

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

It is interesting to interpret each term in (12). The first term,
\[
\beta \sqrt{ \frac{2\sigma_l^2}{n} I(W; X^L, Y^L) + 2\sigma_l^2 D(\mu_X^w \| \mu_X^L) }
\]
can be interpreted as an upper bound on the supervised learning part of the SSL algorithm. We also have the term \( D(\mu_X^w \| \mu_X^L) \), which can be interpreted as the cost of covariate-shift between training and test feature distributions. The second term,
\[
\beta \sqrt{ \frac{2\sigma_e^2}{n} I(W; X^L) + 2\sigma_e^2 D(\mu_X^w \| \mu_X^L) }
\]
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,
\[
\frac{m}{n+m} \beta \sqrt{ \frac{2\sigma_e^2}{m} I(W; X^U) }
\]
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 \( \sigma_l \)-sub-Gaussian under the \( \mu_X^w \otimes P_{Y|X} \) for all \( w \in \mathcal{W} \) and the unsupervised loss function is \( \alpha \)-loss (Sypherd et al., 2019), i.e., \( \ell_u(w, x) = \frac{\alpha}{\alpha-1} (1 - \hat{P}_{Y|w,x}^{-1/\alpha}(y|w, x)) \) for \( \alpha \in (0, \infty) \).
and \( \ell_u(w, x) \) is \( \sigma_u \)-sub-Gaussian under marginal distribution \( \mu_X^u \) 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 \( \mu_X^u \otimes P_{Y|X} \):

\[
|\mathfrak{g}(P_{W|X^L,Y^L,X^U}, \mu_X^u \otimes P_{Y|X})| \leq \frac{\beta \sqrt{\frac{2 \sigma_u^2}{n} I(W; X^L, Y^L) + 2 \sigma_u^2 D(\mu_X^u \| \mu_X^u)}}{n + m} + \frac{m(1 - \beta)}{n + m} \sqrt{\frac{L_u^2}{2m} I(W; X^U)} + (1 - \beta) \Delta_{SSL}^{SSL},
\]

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

The term \( \Delta_{SSL}^{SSL} \) 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 \( \Delta_{SSL}^{SSL} \) 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 \( \sigma_u \)-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:

\[
|\Delta_{SSL}^{SSL}| \leq \frac{2 \sigma_u^2 D(\tilde{P}_{Y|W,X} \| P_{Y|X} | P_W \otimes \mu_X^u)}{n + m}. \quad (14)
\]

**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 \( \mu_X^u \) is absolutely continuous with respect to \( \nu_X^u \), then we have \( D(\mu_X^u \| \nu_X^u) \neq \infty \) 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, \( \ell(w, x, y) \) is bounded in \([0, L]\) 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 \( \mu_X^u \otimes P_{Y|X} \):

\[
|\mathfrak{g}(P_{W|X^L,Y^L,X^U}, \mu_X^u \otimes P_{Y|X})| \leq \beta \left( \frac{L_u^2}{2n} I(W; X^L, Y^L) + 2L_u TV(\mu_X^u \| \mu_X^u) \right)
\]

where \( L_u = \beta \sqrt{\frac{2 \sigma_u^2}{n} I(W; X^L, Y^L) + 2 \sigma_u^2 D(\mu_X^u \| \mu_X^u)} \) and \( L_u = \beta \sqrt{\frac{2 \sigma_u^2}{n} I(W; X^L, Y^L) + 2 \sigma_u^2 D(\mu_X^u \| \mu_X^u)} + \frac{m(1 - \beta)}{n + m} \sqrt{\frac{L_u^2}{2m} I(W; X^U)} + (1 - \beta) \Delta_{SSL}^{SSL}, \)

for the SSL algorithm:

\[
|\Delta_{SSL}^{SSL}| \leq 2L_u TV(\tilde{P}_{Y|W,X} | P_{Y|X} | P_W \otimes \mu_X^u), \quad (16)
\]

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

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 \( \mu_X^u = \mu_X \).

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, \( \mu_X = \mu_X^u \). Consider also hypothesis space is countable, \( |\mathcal{H}| = k \), and \( \mathbb{E}_{P_W \otimes \mu_X^u}[\ell_w(W, X) - \ell_u(W, X)] \leq 2 \sigma_u^2 \epsilon \). Then, the following upper bounds holds on the expected generalization error of the SSL algorithm:

\[
|\mathfrak{g}(P_{W|X^L,Y^L,X^U}, P_{X,Y})| \leq \beta \left( \frac{2 \sigma_u^2 log(k)}{n} + \frac{m(1 - \beta)}{n + m} \right) + (1 - \beta) \sqrt{\frac{2 \sigma_u^2}{n} I(W; X^L, Y^L) + 2 \sigma_u^2 D(\mu_X^u \| \mu_X^u)} + \frac{m(1 - \beta)}{n + m} \sqrt{\frac{L_u^2}{2m} I(W; X^U)} + (1 - \beta) \Delta_{SSL}^{SSL}.
\]

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:

\[
O\left( \frac{\beta}{\sqrt{n}} + \frac{m(1 - \beta)}{n + m} \right). \quad (18)
\]

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

\[
O\left( \max\left( \frac{\sqrt{n}}{n + m}, \frac{m}{(n + m)^{3/2}} \right) \right), \quad (19)
\]
we consider one shared block of parameters (hypothesis), $W_\gamma$, 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 $L^{UL}_E$ 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,

$$L^{SL}_E(W_L, W_{LU}, W_\gamma, x^L, y^L) = \frac{1}{n} \sum_{i=1}^{n} \ell(W_L, W_\gamma, x^L_i, y^L_i) + \frac{1}{n} \sum_{i=1}^{n} \ell(W_{LU}, W_\gamma, x^L_i, y^L_i)$$

$$L^{UL}_E(W_L, W_{LU}, W_\gamma, x^L, x^U) = \frac{1}{n + m} \left( \sum_{i=1}^{n} \ell_u(W_L, W_\gamma, x^L_i) + \sum_{j=1}^{m} \ell_u(W_{LU}, W_\gamma, x^U_j) \right)$$

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

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 the 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 $(X \rightarrow Y)$ does not hold. This is because, if $(X \rightarrow Y)$ 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:

$$X_C \rightarrow Y \rightarrow X_E.$$

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 $\mathcal{N}((a_1, \cdots, a_1), s_I)$ and the second one is $\mathcal{N}((-a_1, \cdots, -a_1), s_I)$. The mixing probability $\pi$ is $(0.5, 0.5)$. The binary label $Y$ is defined as follows

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

Figure 1: Structure of The CSSL Method
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

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

(24)

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

(26)

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

(27)

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

(28)

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

(29)

where (25) and (27) hold from Bayes rule, and we used $(X_C \rightarrow Y \rightarrow X_E)$ 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 $W_L$. 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 $W_L$ 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 $W_L$. The value of $\beta$ (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 handwritten 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
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 $W_\gamma$ 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 $W_\gamma$ 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 $W_\gamma$.

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


Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities: A nonasymptotic


Mohammad Saeed Masiha, Amin Gohari, Mohammad Hossein Yassaee, and Mohammad Reza Aref. Learning under distribution mismatch and model


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:

\[
L_{E}(w, x^L, y^L, x^U) \triangleq \frac{\beta}{n} \sum_{i=1}^{n} \ell(w, x^L_i, y^L_i) + \frac{(1 - \beta)}{n + m} \left( \sum_{i=1}^{n} \ell_{c}(w, x^L_i) + \sum_{j=1}^{m} \ell_{c}(w, x^U_j) \right), \quad 0 \leq \beta \leq 1
\] (30)

We have:

\[
\mathbb{E}_{P_X^{Y}}[\ell_{c}(w, x)] = \mathbb{E}_{P_X^{Y}}[\ell(w, x, y)]
\] (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:

\[
\mathbb{E}_{P_X^{Y}}[L_{E}(w, x^L, y^L, x^U)] = \beta \mathbb{E}_{P_X^{Y}} \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(w, x^L_i, y^L_i) \right] + (1 - \beta) \mathbb{E}_{P_X^{Y}} \left[ \frac{1}{n + m} \left( \sum_{i=1}^{n} \ell_{c}(w, x^L_i) + \sum_{j=1}^{m} \ell_{c}(w, x^U_j) \right) \right]
\]

\[
= \beta \mathbb{E}_{P_X^{Y}}[\ell(w, x, y)] + (1 - \beta) \mathbb{E}_{P_X^{Y}}[\ell(w, x, y)]
\]

\[
= L_{P}(w, P_{XY}), \quad 0 \leq \beta \leq 1
\] (32)

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:

\[
\mathbb{E}[P_{W|X,Y,X^L,Y^L,X^U}\mu_X^W \otimes P_Y|X] = \mathbb{E}_{P_{W,X^L,Y^L,X^U}}[L_{P}(W, \mu_X^W \otimes P_Y|X) - L_{E}^{SSL}(W, X^L, Y^L, X^U)]
\] (33)

\[
= \mathbb{E}_{P_{W,X^L,Y^L,X^U}}[L_{P}(W, \mu_X^W \otimes P_Y|X) - \beta L_{E}^{SL}(W, X^L, Y^L) - (1 - \beta) L_{E}^{UL}(W, X^L, X^U)] =
\] (34)

\[
\beta(\mathbb{E}[P_{W,X^L,Y^L}|L_{P}(W, \mu_X^W \otimes P_Y|X) - L_{E}^{SSL}(W, X^L, Y^L)])
\]

\[
+ (1 - \beta) \mathbb{E}_{P_{W,X^L,Y^L,X^U}}[L_{P}(W, \mu_X^W \otimes P_Y|X) - L_{E}^{UL}(W, X^L, X^U)]
\]

\[
= \beta \left( \mathbb{E}_{P_{W} \otimes \mu_X^W \otimes P_Y|X \otimes n}[L_{E}^{SSL}(W, X^L, Y^L)] - \mathbb{E}_{P_{W,X^L,Y^L}}[L_{E}^{UL}(W, X^L)] \right)
\]

\[
+ (1 - \beta) \left( \mathbb{E}_{P_{W}} \otimes \mu_X^W \otimes n[L_{E}^{UL}(W, X^L)] - \mathbb{E}_{P_{W,X^L,Y^L}}[L_{E}^{UL}(W, X^L)] \right)
\]

\[
+ (1 - \beta) \left( \mathbb{E}_{P_{W}} \otimes \mu_X^W \otimes n[L_{E}^{UL}(W, X^U)] - \mathbb{E}_{P_{W,X^L,Y^L}}[L_{E}^{UL}(W, X^U)] \right)
\]

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

\[
\beta \left( E_{P_W \otimes (\mu_X^{\otimes n} \otimes P_{Y|x})}^{L_{UL}^{SL}(W, X^L), Y^L} - E_{P_{W,X^L,Y^L}}^{L_{UL}^{SL}(W, X^L, Y^L)} \right) \\
+ (1 - \beta) \frac{n}{m + n} \left( E_{P_W \otimes \mu_X^{\otimes n}}^{L_{UL}^{UL}(W, X^L)} - E_{P_{W,X^L}}^{L_{UL}^{UL}(W, X^L)} \right) \\
+ (1 - \beta) \frac{m}{m + n} \left( E_{P_{W,X^U}}^{L_{UL}^{UL}(W, X^U)} - E_{P_{W,X^U}}^{L_{UL}^{UL}(W, X^U)} \right)
\]

\[
\leq \beta \left| E_{P_{W,X^L,Y^L}}^{L_{UL}^{UL}(W, X^L, Y^L)} - E_{P_{W,X^L}}^{L_{UL}^{UL}(W, X^L)} \right| \\
+ (1 - \beta) \frac{n}{m + n} \left| E_{P_{W,X^L}}^{L_{UL}^{UL}(W, X^L)} - E_{P_{W,X^U}}^{L_{UL}^{UL}(W, X^U)} \right| \\
+ (1 - \beta) \frac{m}{m + n} \left| E_{P_{W,X^U}}^{L_{UL}^{UL}(W, X^U)} - E_{P_{W,X^U}}^{L_{UL}^{UL}(W, X^U)} \right|
\]

\[
\leq \beta \sqrt{2\sigma^2 D(P_{W,X^L,Y^L} \| P_W \otimes (\mu_X^{\otimes n} \otimes \mu_X^{\otimes m}))} \\
+ (1 - \beta) \frac{n}{m + n} \sqrt{2\sigma^2 D(P_{W,X^L} \| P_W \otimes \mu_X^{\otimes n})} \\
+ (1 - \beta) \frac{m}{m + n} \sqrt{2\sigma^2 D(P_{W,X^U} \| P_W \otimes \mu_X^{\otimes m})}
\]

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

\[
D(P_{W,X^L} \| \mu_X^{\otimes n} \otimes P_W) = D(P_{X^L,W} \| \mu_X^{\otimes n} \otimes P_W) + n D(\mu_X^{\otimes n} \| \mu_X^{\otimes n}) = I(W; X^L) + n D(\mu_X^{\otimes n} \| \mu_X^{\otimes n})
\]

\section{Proof of Proposition 1}

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

\[
m(P_{W,X|X^L,Y^L,X^U}) = L_P(W, \mu_X^{\otimes n} \otimes P_{Y|x}) - L_{UL}^{SL}(W, X^L, Y^L, X^U)
\]

\[
= E_{P_{W,X^L,Y^L,X^U}}[L_P(W, \mu_X^{\otimes n} \otimes P_{Y|x}) - \beta L_{UL}^{SL}(W, X^L, Y^L) - (1 - \beta) L_{UL}^{UL}(W, X^L, X^U)]
\]

\[
= \beta E_{P_{W,X^L,Y^L}}[L_P(W, \mu_X^{\otimes n} \otimes P_{Y|x}) - L_{UL}^{SL}(W, X^L, Y^L)] \\
+ (1 - \beta) E_{P_{W,X^U}}[L_P(W, \mu_X^{\otimes n} \otimes P_{Y|x}) - L_{UL}^{UL}(W, X^L, X^U)]
\]

\[
= \beta E_{P_{W,X^L,Y^L}}[L_P(W, \mu_X^{\otimes n} \otimes P_{Y|x}) - L_{UL}^{SL}(W, X^L, Y^L)] \\
+ (1 - \beta) E_{P_{W,X^L,Y^L}}[L_P(W, \mu_X^{\otimes n} \otimes P_{Y|x}) - L_{UL}^{UL}(W, X^L, X^U)]
\]

\[
\geq \beta \left( E_{P_W \otimes (\mu_X^{\otimes n} \otimes P_{Y|x})}^{L_{UL}^{SL}(W, X^L, Y^L)} - E_{P_{W,X^L,Y^L}}^{L_{UL}^{UL}(W, X^L, Y^L)} \right) \\
+ (1 - \beta) \frac{n}{m + n} \left( E_{P_W \otimes (\mu_X^{\otimes n} \otimes P_{Y|x})}^{L_{UL}^{UL}(W, X^L)} - E_{P_{W,X^L}}^{L_{UL}^{UL}(W, X^L)} \right)
\]
An Information-theoretical Approach to Semi-supervised Learning under Covariate-shift

\begin{align}
+ (1 - \beta) \frac{m}{m + n} \left( \mathbb{E}_{P_W \otimes \mu_X^m} [L_E^{UL}(W, X_U)] - \mathbb{E}_{P_{W,X,U}} [L_E^{UL}(W, X_U)] \right) \\
+ (1 - \beta) \frac{n}{m + n} \left( \mathbb{E}_{P_W \otimes \mu_X^n} [\ell_u(W, X) - \ell_u(W, X)] \right) \\
\leq \beta \mathbb{E}_{P_{W,L,Y,L},Y,L} \left[ L_E^{SL}(W, X^L, Y^L) - \mathbb{E}_{P_{W,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^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^m} [L_E^{UL}(W, X^U)] - \mathbb{E}_{P_{W,X,U}} [L_E^{UL}(W, X^U)] \right) \\
+ (1 - \beta) \left[ \mathbb{E}_{P_W \otimes \mu_X^n} [\ell_c(W, X) - \ell_u(W, X)] \right] \\
\leq \beta \sqrt{2\sigma_a^2 D(P_{W,X,L,Y,L} \| P_{W} \otimes \mu_X^{m+n})} \\
+ (1 - \beta) \frac{n}{m + n} \sqrt{2\sigma_a^2 D(P_{W,L,Y,L} \| P_{W} \otimes \mu_X^n)} \\
+ (1 - \beta) \frac{m}{m + n} \sqrt{2\sigma_a^2 D(P_{W,U} \| P_{W} \otimes \mu_X^m)} \\
+ (1 - \beta) \left[ \mathbb{E}_{P_W \otimes \mu_X^n} [\ell_c(W, X) - \ell_u(W, X)] \right],
\end{align}

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

\begin{align}
D(P_{X,W} \| \mu_X^n \otimes P_W) = D(P_{X,W} \| P_X \otimes P_W) + D(\mu_X^n \| \mu_X^n) = I(W; X) + D(\mu_X^n \| \mu_X^n)
\end{align}

D Proof of Corollary 1

For the classification task (10), we have:

\begin{align}
|\Delta^{SSL}| = \left| \mathbb{E}_{P_{W} \otimes \mu_X^n} [\ell_c(W, X) - \ell_u(W, X)] \right| \\
\left| \mathbb{E}_{P_{W} \otimes \mu_X^n} \left[ \sum_{j=1}^{q} \hat{P}_{Y=y_j|W,X} \ell(W, X, y_j) - \sum_{j=1}^{q} \hat{P}_{Y=y_j|X} \ell(W, X, y_j) \right] \right| \\
\leq \sqrt{2\sigma_a^2 D(P_W \otimes \mu_X^n \otimes P_{Y|X} \| P_W \otimes \mu_X^n \otimes \hat{P}_{Y|W,X})} \\
= \sqrt{2\sigma_a^2 D(P_{Y|X} \| \hat{P}_{Y|W,X} | P_W \otimes \mu_X^n),}
\end{align}

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{align}
\mathbb{E}_{P_W|X,Y,Y,U} \{ L_P(W, \mu_X^n \otimes P_{Y|X}) - L_E^{SSL}(W, X^L, Y^L, X^U) \} \\
= \mathbb{E}_{P_{W,X,Y,Y,U}} \left[ L_P(W, \mu_X^n \otimes P_{Y|X}) - \beta L_E^{SL}(W, X^L, Y^L) - (1 - \beta) L_E^{UL}(W, X^L, X^U) \right]
\end{align}
For the classification task (10), we have:

\[|\Delta^{SSL}| = |E_{P_W \otimes \mu^u_X} [\ell_c(W, X) - \ell_u(W, X)]|\]  

(64)
We have used pytorch package for implementation of the code. For the first experiment (synthetic data), a single
will become a very good predictor of \( Y \) where (54) is based on the variational representation of total variation (2).

\[
\mathbb{E}_{P_{W} \otimes \mu_{X}} \left[ \sum_{j=1}^{q} \tilde{P}_{Y=y_j|W,X} \ell(W, X, y_j) - \sum_{j=1}^{q} P_{Y=y_j|X} \ell(W, X, y_j) \right] \quad (65)
\]
\[
\leq 2L_u \mathbb{E}_{P_{W} \otimes \mu_{X}} \left[ \text{TV}(\tilde{P}_{Y|W,X}, P_{Y|X}) \right], \quad (66)
\]
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:

\[
|\mathbb{E}_{P_{W}} (P_{W|X^L,Y^L,X^U}) - P_{Y|X}| \leq \beta \sqrt{\frac{2\sigma^2}{n} I(W; X^L, Y^L) + \frac{n(1-\beta)}{n+m} \frac{2\sigma^2}{n} I(W; X^L) + \frac{m(1-\beta)}{n+m} \frac{2\sigma^2}{m} I(W; X^U) + (1-\beta) \Delta^\text{SSL}} \quad (67)
\]
\[
\leq \beta \sqrt{\frac{2\sigma^2}{n} I(W; X^L, Y^L) + (1-\beta) \frac{2\sigma^2}{n+m} I(W; X^L, X^U) + (1-\beta) \sqrt{2\sigma^2 \epsilon}} \quad (68)
\]
\[
\leq \beta \sqrt{\frac{2\sigma^2}{n} H(W) + (1-\beta) \frac{2\sigma^2}{n+m} H(W) + (1-\beta) \sqrt{2\sigma^2 \epsilon}} \quad (69)
\]
\[
\leq \beta \sqrt{\frac{2\sigma^2}{n} \log(k) + (1-\beta) \sqrt{2\sigma^2 \epsilon}}. \quad (70)
\]

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 \( W_u \) 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. \( W_u \) has ten inputs and two output nodes, a softmax function is used at the end to produce the required conditional distributions. The value of \( \beta \) (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

<table>
<thead>
<tr>
<th>Value of $a_1$ for test dataset</th>
<th>$a_1 = 0.9$</th>
<th>$a_1 = 0.6$</th>
<th>$a_1 = 0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower bound</td>
<td>0.761 ± 0.011</td>
<td>0.681 ± 0.010</td>
<td>0.621 ± 0.010</td>
</tr>
<tr>
<td>EM</td>
<td>0.874 ± 0.032</td>
<td>0.781 ± 0.036</td>
<td>0.705 ± 0.030</td>
</tr>
<tr>
<td>CSSL</td>
<td><strong>0.876 ± 0.070</strong></td>
<td><strong>0.792 ± 0.065</strong></td>
<td><strong>0.712 ± 0.053</strong></td>
</tr>
<tr>
<td>Upper bound</td>
<td>0.999 ± 0.0001</td>
<td>0.997 ± 0.004</td>
<td>0.977 ± 0.024</td>
</tr>
</tbody>
</table>