Information-Theoretic Generalization Bounds for Transductive Learning and its Applications

Huayi Tang HUAYITANG@RUC.EDU.CN
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, 100080, China

Yong Liu* LIUYONGGSAI@RUC.EDU.CN
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, 100080, China

Abstract

We develop data-dependent and algorithm-dependent generalization bounds for transductive learning algorithms based on information theory. We first show that the generalization gap of transductive learning algorithms can be controlled by the mutual information between training labels and the hypothesis. By introducing the concept of transductive supersamples, we go beyond inductive learning setting and establish upper bounds in terms of various information measures. Furthermore, we derive novel PAC-Bayesian bounds and build connection between generalization and loss landscape flatness under the setting of transductive learning. Finally, we present the upper bounds for adaptive optimization algorithms, and demonstrate the applications of results on semi-supervised learning and graph learning. Our theoretic results are validated on both synthetic and real-world datasets.

Keywords: Transductive Learning, Generalization, Information Theory

1 Introduction

In standard supervised learning paradigm (Shai and Shai, 2014; Mohri et al., 2018), we receive a set of instances composed of features and targets, which are assumed to be sampled independently from an unknown distribution. Our task is to build a learner (or model) by specific optimization algorithm that maps features to corresponding targets based on the received finite instances. A modern practice of this learning paradigm is to train a deep neural network for image classification task (Krizhevsky et al., 2012) by SGD. Generalization ability, referring to the prediction performance of a learner on unseen examples, is one of the key quantities we are concerned about. The past decades have witnessed efforts of exploring theory to characterize and understand the generalization ability of machine learning models. In the category of classical learning theory, generalization ability is connected to the complexity of hypothesis space (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2002; Bartlett et al., 2005), the stability of algorithm (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010) or the divergence between two probability measure on hypothesis space (Shawe-Taylor and Williamson, 1997; McAllester, 1999). Recently, mutual information and its variants are shown to serve as an ideal metric for generalization ability,

*. Corresponding author

since they reflect both the impact of datasets and algorithms on generalization. Research of this viewpoint originates from work (Russo and Zou, 2016, 2020; Xu and Raginsky, 2017), and is further developed by subsequent studies (Negrea et al., 2019; Haghifam et al., 2020; Steinke and Zakynthinou, 2020; Harutyunyan et al., 2021; Haghifam et al., 2021; Sefidgaran et al., 2022; Wang and Mao, 2023a). Despite the diverse forms of these results, they possess a common key insight: the less information on training data (or its selection) a hypothesis (or the variables it induced) reveals, the better generalization it will have.

The aforementioned supervised learning paradigm are far not enough to cover all machine learning scenarios. Data collected from real-world scenarios could come from various domains, and some examples could lack of targets due to the expensive cost of annotations. This raises the need of exploring new theory to measure the generalization, where the key challenge is relaxing the identical and independent assumptions of instances. In this paper we move towards this direction by analyzing a classical but important regime termed as transductive learning (Vapnik, 1982). In this learning paradigm, we are provided with a fixed set of instances containing both labeled examples and unlabeled examples, and our task is to build a learner that make prediction for those unlabeled ones. Notably, the examples (except their targets) to be predicted are used by the learner during training. As a comparison, the standard supervised learning paradigm mentioned earlier belongs to the category of inductive learning, since the examples to be predicted are not available to the learner. Practices of transductive learning paradigm are semi-supervised learning (Shahshahani and Landgrebe, 1994; Blum and Mitchell, 1998; Joachims, 1999; Zhu et al., 2003) and transductive graph learning (Gori et al., 2005; Scarselli et al., 2009; Gilmer et al., 2017; Kipf and Welling, 2017), along with their applications on other real-world scenarios.

The research topic of this paper is the generalization ability of transductice learning algorithms. Existing results for transductive generalization bound generally fall into three categories: complexity-based bounds derived by VC dimension (Cortes and Mohri, 2006) or transductive Rademacher complexity (El-Yaniv and Pechyony, 2007; Tolstikhin et al., 2015), stability-based bounds (El-Yaniv and Pechyony, 2006; Cortes et al., 2008) and PAC-Bayesian bounds (Derbeko et al., 2004; Bégin et al., 2014). These findings are sufficient to provide learning guarantee for classical learner or algorithms such as transductive support vector machine and unlabeled-labeled representation. However, they are not sufficient to explain the generalization behaviors of all transducitve models, particularly deep transductive models. The reasons are three folds. First, it has been shown by (Esser et al., 2021) that VC dimension results in trivial generalization error bounds of graph neural networks (GNNs), and transductive Rademacher complexity is algorithm-independent and fail to characterize the impact of optimization algorithms on generalization (Tang and Liu, 2023). Second, stability-based bounds depend on Lipschitz and smoothness constant (Cong et al., 2021), which is difficulty to compute or estimate for deep models (Neu et al., 2021). Third, existing transductive PAC-Bayesian bounds are of slow order and not sufficient to provide non-vacuous bounds. Furthermore, it is unclear to what extent they reflect the impact of optimization algorithms. In a nutshell, effort to establish data-dependent and algorithm-dependent generalization bounds for transductive learning is limited.

In this paper, we comprehensively study the generalization theory of transductive learning in the context of information theory. First, we derive upper bounds of transductive generalization gap in expectation and high probability. These results reveal that the dependence of output hypothesis on the randomness of training labels serve as a measure to quantify the generalization performance of tranductive learner. Second, we propose the concept of transductive supersamples and establish upper bounds based on various information measure. These bounds are non-vacuous and convenient in estimation. Third, by observing the connection between information theory and PAC-Bayesian theory, we give novel transductive PAC-Bayesian bounds that has weaker assumption and faster rate. With this result, we further show that the flatness of loss landscape affects generalization still holds in transductive learning setting, which is supported by the empirical evidence in recent study (Chen et al., 2023). Forth, we apply these results to analyze adaptive optimization algorithms and derive the corresponding upper bounds. Fifth, we illustrate the applications of the theoretic results on semi-supervised learning and transductive graph learning scenarios. The main contributions of this work are summarized as follows.

- For the first time, we systematically establish information-theoretic generalization bounds for transductive learning and reveal their connection with PAC-Bayesian bounds, which provide new perspective to understanding the generalization of transductive learning.
- We propose the concept of transductive supersamples and use it to bridge the gap between supersample setting in inductive learning and transductive learning setting. This also help us obtain the first non-vacuous bound for deep transductive learner.
- We demonstrate the application of our theoretic results on semi-supervised learning and transductive graph learning and verify them by experiments on both synthetic and real-world datasets.

In the remainder of this paper, we begin with an overview of the literature related to our work in Section 2. After that, we introduce major notations and important concepts in Section 3. The main theoretic results are presented in Section 4, and their applications are given in Section 5. The setting and results of experiments are provided in Section 6. All the proofs are placed in the Appendix.

2 Related Work

Information-theoretic Generalization Theory. (Russo and Zou, 2016, 2020; Xu and Raginsky, 2017) are the pioneering work that associates expected generalization error with the mutual information between training examples and algorithm output. The subsequent studies mainly fall into four categories: (i) deriving sharper upper bounds by introducing new information measure (Harutyunyan et al., 2021; Hellström and Durisi, 2022; Wang and Mao, 2023a), problem setting (Steinke and Zakynthinou, 2020; Rammal et al., 2022; Haghifam et al., 2022) or proof techniques (Asadi et al., 2018; Bu et al., 2020; Hafez-Kolahi et al., 2020a; Gálvez et al., 2021; Zhou et al., 2022; Clerico et al., 2022), (ii) establishing bounds described by various divergences (Lopez and Jog, 2018; Wang et al., 2019a; Esposito et al., 2021; Aminian et al., 2021a,b) (iii) applying existing results to establish upper bounds for optimization algorithms such as SGD (Neu et al., 2021; Wang and Mao, 2022) or SDLD (Pensia et al., 2018; Negrea et al., 2019; Wang et al., 2021), and (iv) extending the theoretical results to diverse scenarios such as meta-learning (Jose and Simeone, 2021a; Rezazadeh et al., 2021; Chen et al., 2021; Jose et al., 2022), transfer learning (Wu et al., 2020; Jose and Simeone, 2021b; Masiha et al., 2021; Bu et al., 2022), semi-supervised learning (Aminian et al., 2022; He et al., 2022), self-supervised learning (Yuan et al., 2022) and domain adaption (Wang and Mao, 2023b). However, the training and test examples are independent in existing studies, which makes them not applicable on transductive learning. Another related topic is information bottleneck theory (Tishby et al., 2000) and its applications to explaining the representation (Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017) and generalization (Hafez-Kolahi et al., 2020b; Wang et al., 2022; Kawaguchi et al., 2023) of deep neural networks, which are parallel to our work. Interested readers are referred to a recent monograph (Hellström et al., 2023) for more illustrations.

PAC-Bayesian Generalization Theory. The classical results in PAC-Bayesian generalization theory include McAllester's bound (McAllester, 1999), Seeger's bound (Seeger, 2002), Catoni's bound (Catoni, 2007) and Maurer's bound (Maurer, 2004). Based on these, there have been numerous studies that applying or extending these results to the analysis of deep neural networks, including computing non-vacuous bounds for deep neural networks (Dziugaite and Roy, 2017; Zhou et al., 2019; Pérez-Ortiz et al., 2021; Lotfi et al., 2022) and establishing upper bounds for optimization algorithms (London, 2017; Neyshabur et al., 2018; Rivasplata et al., 2018; Arora et al., 2018; Mou et al., 2018; Yang et al., 2019; Li et al., 2020; Luo et al., 2022) or specific neural networks (Liao et al., 2021; Mbacke et al., 2023). We refer interested readers to a comprehensive survey (Alquier, 2021) for more detail on this topic. All the above results are derived under the inductive learning setting and could not be applied on transductive learning setting.

Transductive Learning Generalization Theory. The concept of transductive learning and the earliest bounds are presented in (Vapnik, 1982). The authors in (El-Yaniv and Pechyony, 2006) study the stability of transductive learning algorithms. They further propose another tool named transductive Rademacher Complexity (El-Yaniv and Pechyony, 2007) as a complexity measure of hypothesis space under transductive setting. Permutational Rademacher Complexity is latter introduced in (Tolstikhin et al., 2015), which is shown to be more suitable for transductive learning setting than transductive Rademacher Complexity. By considering the variance of functions, the authors in (Tolstikhin et al., 2014) establish new concentration inequalities and derive sharper bounds. Different from them, we establish upper bounds based on information theory. (Derbeko et al., 2004) is the first work to analyze the generalization of transductive learning in the context of PAC-Bayesian, and their results are improved latter in (Bégin et al., 2014). We further improve their result and apply it to reveal the impact of loss landscape flatness on generalization. Furthermore, the above theoretic results have been applied on the theoretical analysis in transductive graph learning (Shivanna and Bhattacharyya, 2014; Shivanna et al., 2015; De et al., 2018; Oono and Suzuki, 2020; Esser et al., 2021; Cong et al., 2021; Tang and Liu, 2023), semi-supervised learning (Maximov et al., 2018; Gong et al., 2018; Xu et al., 2023), matrix completion (Giménez-Febrer et al., 2020; Shamir and Shalev-Shwartz, 2014), distributed optimization (Shamir, 2016) and collaborative filtering (Xu et al., 2021; Deng et al., 2022), among other areas. We select semi-supervised learning and transductive graph learning as illustrated examples of our theoretic results, and leave the task of extension to other areas for future work.

3 Preliminaries

3.1 Notations

We stipulate that random variables and their realizations are denoted by uppercase and lowercase letters, respectively. For given random variable X, we denote its distribution measure by P_X. Similarly, the conditional distribution measure of X given Y is given by $P_{X|Y}$. We use $\mathrm{D_{KL}}(P||Q)$ to denote the Kullback–Leibler (KL) divergence between two probability measure P and Q on the same probability space, under the condition that the Radon-Nikodym derivative of P with respect to Q is well defined. With this notation, the mutual information between X and Y is defined as $I(X;Y) \triangleq \mathrm{D_{KL}}(P_{X,Y}||P_X P_Y)$. Furthermore, we use $I^z(X;Y) \triangleq \mathrm{D_{KL}}(P_{X,Y|Z=z}||P_{X|Z=z}P_{Y|Z=z})$ to represent the disintegrated mutual information, whose expectation taking over $Z \sim P_Z$ is the conditional mutual information $I(X;Y|Z) = \mathbb{E}_Z[I^Z(X;Y)]$. Besides, we use $\{\cdot\}$ and $(\cdot)$ to denote set and sequence, respective. Notice that sets are unordered yet sequences are ordered. Therefore, two sequence are equal if and only if the element at each position are equal. For concise, we use $[m]$ to represent the set $\{1, 2, \dots, n\}$. The Hadamard Product and Kronecker Product are denoted by $\odot$ and $\otimes$, respectively.

3.2 Transductive Learning

Let $\mathrm{D} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ be a given set with finite cardinality, where $\mathbf{z} = (\mathbf{x}, y)$ is an instance composed of attribute $\mathbf{x} \in \mathcal{X}$ and target $y \in \mathcal{Y}$ from $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. We use $\mathrm{Perm}(\mathrm{D})$ to denote the set containing all bijections $\pi : \mathrm{D} \to \mathrm{D}$. Here each mapping $\pi \in \mathrm{Perm}(\mathrm{D})$ could be regarded as a permutation on D. Notice that sampling without replacement from D is equivalent to firstly sampling a permutation from $\mathrm{Perm}(\mathrm{D})$ with equal probability and then applying it on D. Denote by Π a random variable follows uniform distribution over $\mathrm{Perm}(\mathrm{D})$, namely $\mathbb{P}\{\Pi = \pi\} = \frac{1}{n!}$ holds for any $\pi \in \mathrm{Perm}(\mathrm{D})$. With this notation, we use $Z \triangleq (Z_1, \dots, Z_n)$ to denote the random permutation vector induced by Π, where $Z_j = \Pi(\mathbf{z}_j)$ represents the j-th element of the sequence after permutation. For example, the permutation vector $Z = (\mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_1)$ is induced by the mapping Π with $\Pi(\mathbf{z}_1) = \mathbf{z}_2$, $\Pi(\mathbf{z}_2) = \mathbf{z}_3$ and $\Pi(\mathbf{z}_3) = \mathbf{z}_1$. For determined Z, the training set is defined as $\mathrm{D_{train}} \triangleq \{Z_1, \dots, Z_m, X_{m+1}, \dots, X_{m+u}\}$, where X is the feature of Z. m and u are the number of training and test instances, respectively. Notice that there is a hidden fact that $m, u \in \mathbb{N}_+$ and $m + u = n$. Let $\mathcal{W}$ be the space of parameter, the transductive learning algorithm takes $\mathrm{D_{train}}$ as input and outputs a random element $W \in \mathcal{W}$ as the hypothesis, which is characterized by a Markov kernel $P_{W|Z}$. Let $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ be the objective function, the transductive training and test error of a hypothesis W are defined as $R_m(W, Z) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(W, Z_i)$ and $R_u(W, Z) = \frac{1}{u}\sum_{i=m+1}^{m+u} \ell(W, Z_i)$, respectively. The *transductive generalization error* is then defined as $\mathcal{E}(W, Z) \triangleq R_u(W, Z) - R_m(W, Z)$. Furthermore, we use $\mathbb{E}_{W,Z}[\mathcal{E}(W, Z)]$ to denote the expectation of $\mathcal{E}(W, Z)$ taking over $P_{W,Z} = P_Z \otimes P_{W|Z}$, which represents the the average performance difference of the hypothesis W between testing and training instances over all permutations Z. The objective $\ell(w, z)$ can also be represented as $r(f_w(x), y)$, where $f_w(x)$ is the prediction of the model with parameter w on x, and $r : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is the criterion. For example, we have $r(\hat{y}, y) = \mathbb{1}_{\hat{y} \neq y}$ when the criterion is zero-one loss, where $\mathbb{1}$ is the indicator function.

4 Theoretic Results

4.1 Establishing Upper Bounds by Mutual Information

Different from supervised learning, the randomness of training and testing examples in transductive learning come from the partition determined by permutation rather than from sampling. This also brings another challenge, namely the dependence of training and testing examples, since the testing examples are uniquely determined once training examples are chosen. The most widely adopt technique to tackle the dependence is the martingales method, which enables us to derive similar "sub-Gaussian" property for the transductive generalization error. Together with the Donsker-Varadhan's variational formula, we establish the following transductive generalization bounds.

Theorem 1 *Suppose that $\ell(\mathbf{w}, \mathbf{z}) \in [0, B]$ holds for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{z} \in \mathrm{D}$, where $B > 0$ is a constant. Define $C_{m,u} \triangleq \frac{2B^2(m+u)\max(m,u)}{(m+u-1/2)(2\max(m,u)-1)}$, we have*

$$|\mathbb{E}_{W,Z}\left[R_u(W,Z) - R_m(W,Z)\right]| \leq \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m} + \frac{1}{u}\right) I(W;Z)}, \tag{1}$$

$$\mathbb{E}_{W,Z}\left[(R_u(W,Z) - R_m(W,Z))^2\right] \leq C_{m,u}\left(\frac{1}{m} + \frac{1}{u}\right)(I(W;Z) + \log 3). \tag{2}$$

Theorem 1 shows that the expectation of transductive generalization error is upper bounded by the mutual information between permutation Z and hypothesis W. This result implies that the less dependence the output hypothesis has on the selection of training labels, the better generalization a transductive learning algorithms will have. One can image that if the algorithm only "memorize" the partition of training and testing examples (or heavily depends on the training labels it sees), we could not expect that it will have strong generalization ability. As a comparison, a similar result (Theorem 1 in (Xu and Raginsky, 2017)) under inductive learning setting says that the generalization error is upper bounded by the mutual information between training set S and hypothesis W, which could be regarded as a special case of our result that u is infinite. In this event, the transductive training and test error corresponding to the supervised training error and test error, respectively. Also, m is the number of examples in S, and the term $1/u$ naturally vanishes. Note that since the test error is computed over infinite examples, there is no dependence between it and the training error. Furthermore, the assumption of Theorem 1 is slightly stronger than that in supervised learning setting, where they only requires the loss to be sub-Gaussian while we require the loss to be bounded. However, we believe that our result could be extend to the unbounded loss setting under proper assumptions.

The result presented in Theorem 1 is a expectation bound over all possible selection of training data. In real-world application particularly deep learning scenario, only a few partitions (determined by random seed) are adopted to verify the quality of a transductive learning algorithm, and the empirical results show that whose performance could generally be guaranteed. This urges us to establish the high probability bound in order to better describe the generalization behavior of deep transductive learners. Achieving this relies on the monitor technique proposed in (Bassily et al., 2016). Another derivant is the expectation bound on the absolute value of transductive generalization error, which serves as a supplement of Theorem 1. The aforementioned results are summarized in Theorem 2.

Theorem 2 *Suppose that $\ell(\mathbf{w}, \mathbf{z}) \in [0, B]$ holds for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{z} \in \mathrm{D}$, where $B > 0$ is a constant. With probability at least $1 - \delta$ over the randomness of Z and W:*

$$|R_u(W, Z) - R_m(W, Z)| \leq \sqrt{2C_{m,u}\left(\frac{1}{m} + \frac{1}{u}\right)\left(\log\left(\frac{1}{\delta}\right) + \frac{I(W; Z)}{\delta}\right)}, \tag{3}$$

where $C_{m,u}$ follows the definition in Theorem 1. Furthermore, we have

$$\mathbb{E}\,|R_u(W, Z) - R_m(W, Z)| \leq \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m} + \frac{1}{u}\right)(I(W; Z) + \log 2)}. \tag{4}$$

Since $C_{m,u} \approx B^2$ when $m + u$ is large, the high probability bound presented in Eq. 3 is of order $(1/m + 1/u)^{\frac{1}{2}}$. Despite a degenerated constant factor from $\log(1/\delta)$ to $1/\delta$, our bound is sharper than previous result (Oono and Suzuki, 2020; Esser et al., 2021). Although the mutual information term $I(Z; W)$ could not be easily computed, we will show in SubSection 4.4 that it has a unique advantage when the learner is optimized by stochastic algorithms such as stochastic gradient descent and its variants. Besides, result similar to Eq. (4) can be derived from Eq. (2), despite that the constant factor is slightly larger.

4.2 Establishing Upper Bounds by Conditional Mutual Information

So far, all the bounds we have established contain the mutual information term $I(W; Z)$, either in expectation or high probability. One unsatisfied property of mutual information is that it does not have a finite upper bound, which may lead to vacuous bounds under some circumstances. Fortunately, this issue could be addressed by adopting "supersamples" setting proposed in (Steinke and Zakynthinou, 2020) under supervised learning setting. The key insight is that introducing another random variable to control the randomness of training and test examples partition, which is independent of the instances. However, directly applying this technique on transductive learning setting is not feasible. The reason is that the training and test examples given returned by this setting are independent, which are yet dependent in transductive learning setting. To bridge this gap, we propose the following transductive supersamples under a specific condition that the number of training examples is equal to that of test examples, namely $m = u$.

Definition 3 (Transductive Supersamples) *Let $\mathrm{D} = \{\mathbf{z}_i\}_{i=1}^n$ be a given set where n is a finite even number. Denote by $m = \frac{1}{2}n$, the transductive supersamples is a sequence $\widetilde{Z} \triangleq (\widetilde{Z}_1, \ldots, \widetilde{Z}_m)$ generated by sampling without replacement from D, where $\widetilde{Z}_i \triangleq \left\{\widetilde{Z}_{i,0}, \widetilde{Z}_{i,1}\right\}$ for $i \in [m]$ is an unordered set with cardinality 2.*

Definition 3 shows that transductive supersamples are obtained by continuously sampling an unorderd instance pairs from a fixed set until there are no remained instances. Please refer to Appendix E for an illustrated example of this definition. As a comparison, transductive samples Z is obtained by each time sampling an instance from D. A deeper relationship between Z and $\widetilde{Z}$ is given by the following Proposition.

Proposition 4 *Denote by $\widetilde{\mathcal{Z}}$ the set containing all $\widetilde{Z}$. Let $U \triangleq (U_1, \ldots, U_m) \sim \mathrm{Unif}(\{0,1\})^m$ be the sequence of random variables that is independent of $\widetilde{Z}$. Sampling without replacement from D is equivalent to firstly sampling $\widetilde{Z}$ from $\widetilde{\mathcal{Z}}$ and applying U to permute $\widetilde{Z}$.*

Proposition 4 implies that there is another way to obtain the random permutation vector Z based on transductive supersamples. Let $\widetilde{Z}$ and U be the random variables described in Proposition 4, the random permutation vector Z can be expressed by $Z = (\widetilde{Z}_{1,U_1}, \ldots, \widetilde{Z}_{m,U_m}, \widetilde{Z}_{1,1-U_1}, \ldots, \widetilde{Z}_{m,1-U_m})$. Let $\mathcal{E}(W, \widetilde{Z}, U)$ be the transductive generalization error under supersampling setting defined by

$$\mathcal{E}(W, \widetilde{Z}, U) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(W, \widetilde{Z}_{i,U_i}) - \ell(W, \widetilde{Z}_{i,1-U_i}), \tag{5}$$

we have $\mathbb{E}_{W,\widetilde{Z},U}[\mathcal{E}(W, \widetilde{Z}, U)] = \mathbb{E}_{W,Z}[\mathcal{E}(W, Z)]$. This enables us to characterize the generalization bounds using conditional mutual information, as presented in Theorem 5.

Theorem 5 *Suppose that $\ell(\mathbf{w}, \mathbf{z}) \in [0, B]$ holds for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{z} \in \mathrm{D}$, where $B > 0$ is a constant. We have*

$$|\mathbb{E}_{Z,W}\left[R_u(W,Z) - R_m(W,Z)\right]| \le \mathbb{E}_{\widetilde{Z}}\sqrt{\frac{2B^2}{m} I^{\widetilde{Z}}(W;U)}, \tag{6}$$

$$\mathbb{E}_{S,W}\left[(R_u(W,Z) - R_m(W,Z))^2\right] \le \frac{4B^2}{m}(I(W;U|\widetilde{Z}) + \log 3). \tag{7}$$

By (Steinke and Zakynthinou, 2020), we have $I(W;U|\widetilde{Z}) \le I(W;U) \le m\log 2$ holds, suggesting that the conditional mutual information has a finite upper bound and thus provides a non-vacuous generalization bound. Eq. (6) is consistent with the results in supervised learning setting (Steinke and Zakynthinou, 2020) in formulation, and the only difference is that $\widetilde{Z}$ should be interpreted as the transductive supersamples. Also, we can recover the result in supervised learning (Theorem 1.2 in (Steinke and Zakynthinou, 2020)) whereas the full sample set D has a infinite cardinality. In this event, D is exactly the space containing all instances, and the entries in the sequence $\widetilde{Z}$ are independent to each other.

Although the mutual information term $I(W;U|\widetilde{Z})$ in Theorem 5 is bounded, computing its numerical value is still difficult, as W is commonly a high-dimensional random variable in deep learning scenarios. Thanks to the transductive supersampling setting, various information-theoretical measures (Harutyunyan et al., 2021; Hellström and Durisi, 2022; Wang and Mao, 2023a) adopted in supervised learning setting can be extended to transductive learning setting, as shown in the following Corollary.

Corollary 6 *Suppose that $r(\hat{y}, y) \in [0, B]$ holds for any $\hat{y} \in \widehat{\mathcal{Y}}$ and $y \in \mathcal{Y}$, where $B > 0$ is a constant. Let $f_w(\mathbf{x}) \in \mathbb{R}^K$ be the prediction given by the learner. Denote by $F \in \mathbb{R}^{m \times 2K}$ the prediction matrix where the i-th row is given by $F_{i,:} \triangleq (f_W(\widetilde{X}_{i,0}), f_W(\widetilde{X}_{i,1}))$, where $\widetilde{X}$ is the feature of $\widetilde{Z} = (\widetilde{X}, \widetilde{Y})$. We have*

$$|\mathbb{E}_{Z,W}\left[R_u(W,Z) - R_m(W,Z)\right]| \le \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(F_i;U_i)}. \tag{8}$$

Denote by $L \in \{0,1\}^{m\times 2}$ *the loss value matrix, where the* i*-th row is* $L_{i,:} \triangleq (\ell(W, \widetilde{Z}_{i,0}), \ell(W, \widetilde{Z}_{i,1}))$*. Let* $\Delta_i \triangleq \ell(W, \widetilde{Z}_{i,1}) - \ell(W, \widetilde{Z}_{i,0})$ *be the difference of loss value. We have*

$$|\mathbb{E}_{S,W}\left[R_u(W,Z) - R_m(W,Z)\right]| \le \frac{B}{m}\sum_{i=1}^{m} \mathbb{E}_{\widetilde{Z}} \sqrt{2I^{\widetilde{Z}}(L_i; U_i)}, \tag{9}$$

$$|\mathbb{E}_{S,W}\left[R_u(W,Z) - R_m(W,Z)\right]| \le \frac{B}{m}\sum_{i=1}^{m} \mathbb{E}_{\widetilde{Z}} \sqrt{2I^{\widetilde{Z}}(\Delta_i; U_i)}. \tag{10}$$

According to the type of conditional mutual information they contained, the bounds in Eqs. (8,9,10) are termed as f-CMI (Harutyunyan et al., 2021), e-CMI (Hellström and Durisi, 2022) and Id-CMI bounds (Wang and Mao, 2023a), respectively. The only difference between these results and previous one is that here $\widetilde{Z}$ is the *transductive supersamples*. In applications, the prediction of learner is a low-dimension vector and thus reduce the difficulty of computing the conditional mutual information $I(W; U|\widetilde{Z})$. Note that L_i in Eq. (9) and Δ_i in Eq. (10) are two-dimensional and one-dimensional random variable, yielding more convenient computation and sharper bounds. Here we point out that each result has its advantage. The vanilla bound in Eq. (6) is more informative to understanding generalization, and the expense is is difficulty of calculating numerical value. In contrast, the other bounds in Corollary 6 has computation convenience, yet they are inferior in reflecting factors that affect generalization. Despite the existence of this trade-off, these results are sufficient for us to understand the generalization behavior of transductive learner or establish non-vacuous bounds for them.

We close this part by briefly discuss how to extend the aforementioned results to more ordinary cases. According to Definition 3, Theorem 5 and Corollary 6 only apply to the case that $m = u$. Here we point out that they can be extended to cases that $m = ku$ or $u = km$ where $k \in \mathbb{N}_+$. Considering the symmetric, it is sufficient to discuss the case that $u = km$. To this end, the transductive supersamples are extended to the following k-transductive supersamples.

Definition 7 (k-Transductive Supersamples) *Let* $\mathrm{D} = \{\mathbf{z}_i\}_{i=1}^n$ *be a given set where* n *is a finite even number. Let* $k \in [n-1]$ *be a given integer. Denote by* $m = \frac{n}{k+1} \in \mathbb{N}_+$*, the* k*-transductive supersamples is a sequence* $\widetilde{Z} \triangleq (\widetilde{Z}_1, \ldots, \widetilde{Z}_m)$ *generated by sampling without replacement from* D*, where* $\widetilde{Z}_i \triangleq \{\widetilde{Z}_{i,0}, \ldots, \widetilde{Z}_{i,k}\}$ *is an unordered set with cardinality* $k+1$*.*

Note that Definition 3 is a special case of Definition 7 where $k = 1$. Similarly we need to extend the definition of the indicator variable U. Let $U \triangleq (U_1, \ldots, U_m) \sim \mathrm{Unif}(\{0, \ldots, k\})^m$ be the sequence of random variables that is independent of $\widetilde{Z}^m$. Let $V = (V_1, \ldots, V_m) \sim \mathrm{Unif}(\mathrm{Perm}([k]))^m$ be the sequence of random variables that is independent of $\widetilde{Z}^m$ and U. For a fixed set $\mathcal{S}$, we use $\Pi(\mathcal{S}, V)$ to represent the permutation of $\mathcal{S}$ induced by random variable V. Here each entry in V corresponding to a bijection $\mathcal{S} \to \mathcal{S}$. For example, $\Pi(\widetilde{Z}_i \setminus \widetilde{Z}_{i,0}, V_1)$ denotes the result of drawing a permutation of $\widetilde{Z}_i \setminus \widetilde{Z}_{i,0}$ according to V_1, which is essentially equivalent to sampling without replacement from $\widetilde{Z}_i \setminus \widetilde{Z}_{i,0}$. With this definition, the permutation vector Z can be expressed by $Z^n = (\widetilde{Z}_{1,U_1}, \ldots, \widetilde{Z}_{m,U_m}, \Pi(\widetilde{Z}_1 \setminus \widetilde{Z}_{1,U_1}, V_1), \ldots, \Pi(\widetilde{Z}_m \setminus \widetilde{Z}_{m,U_m}, V_m))$. By this way, results in Theorem 5 and Corollary 6 can be extended to the case that $u = km$, and we place the details in Appendix E. One main differences between the case that $k = 1$ and $k > 1$ is the increased computation cost of estimating the conditional mutual information, since each entry of S has more possible values to take. In other words, with the increase of k, we need to accordingly increase the samples of S to reduce the estimated error of the conditional mutual information.

4.3 Connection with Transductive PAC-Bayesian Bounds

PAC-Bayesian methods and Information-theoretic methods are closely related, since both of them are based on Donsker-Varadhan's variational formulation. Borrowing the proof of Theorem 1, we obtain the following new transductive PAC-Bayesian bounds.

Theorem 8 *Suppose that $\ell(\mathbf{w}, \mathbf{z}) \in [0, B]$ holds for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{z} \in D$, where $B > 0$ is a constant. Let P be a prior distribution P on $\mathcal{W}$. With probability at least $1 - \delta$ over the randomness of Z, for any distribution Q on $\mathcal{W}$ we have*

$$|\mathbb{E}_{W\sim Q}\left[R_u(W,Z) - R_m(W,Z)\right]| \leq \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\mathrm{D}_{\mathrm{KL}}(Q||P) + \log\left(\frac{1}{\delta}\right)\right)}, \tag{11}$$

where $C_{m,u}$ follows the definition in Theorem 1.

Compared with previous transductive PAC-Bayesian bound (Corollary 7(b) in (Bégin et al., 2014)), the advantages of Theorem 8 are as follows. First, the assumptions of Theorem 8 is weaker. Previous result only apply to zero-one loss, and the value of m and n are required to satisfy $n \geq 40$ and $20 \leq m \leq n - 20$. In contrast, our result apply to any bounded loss and there are no constraints on the value of m and n. Second, our result is strictly sharper than previous result by removing the term $\log\left(\log(m)\sqrt{\frac{mu}{m+u}}\right)$. Furthermore, incorporating the technique in (Neyshabur et al., 2018), the results in (Liao et al., 2021) could be directly extend to transductive learning setting and thus providing generalization guarantee for many GNNs on node classification and link prediction task.

One of the most important insights provided by PAC-Bayesian bounds in that the generalization performance is closely related with the sharpness of the loss landscape, and a flat minimum is beneficial for generalization (Foret et al., 2021). With the help of Theorem 8, this result can be extended to transductive learning setting when ℓ is zero-one loss.

Corollary 9 *Suppose that $R_u(\mathbf{w}, Z) \leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[R_u(\mathbf{w}+\boldsymbol{\epsilon}, Z)]$ holds for any random permutation vector Z, where $\mathbf{w} \in \mathbb{R}^d$ is the parameter returned by given transductive learning algorithm and $\boldsymbol{\epsilon} \in \mathbb{R}^d$ is a random Gaussian noise. With probability at least $1 - \delta$ over the randomness of Z,*

$$\begin{aligned} &R_u(\mathbf{w}, Z) \\ &\leq \max_{\|\boldsymbol{\epsilon}\|_2\leq\rho} R_m(\mathbf{w}+\boldsymbol{\epsilon}, Z) \\ &+ \sqrt{\frac{C_{m,u}(m+u)\left(1+\frac{d}{2}\log\left(1+\frac{\left(1+\widetilde{C}_{m,u}\right)^2\|\mathbf{w}\|_2^2}{\rho^2}\right)+\log\left(\frac{1}{6\delta}\right)+2\log\left(\frac{4\pi mu}{m+u}\right)\right)}{mu}}, \end{aligned} \tag{12}$$

where $\widetilde{C}_{m,u} \triangleq \sqrt{\log(4mu/(m+u))/d}$ and $C_{m,u}$ follows the definition in Theorem 1.

The term $\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_m(\mathbf{w}+\boldsymbol{\epsilon}, Z)$ characterizes the change of loss landscape within a ball with $\mathbf{w}$ as the center and ρ as the radius. Formally, we call $\mathbf{w}$ as sharp minima if the loss values around it differ significantly from itself, namely $R_m(\mathbf{w}+\boldsymbol{\epsilon}, Z) \gg R_m(\mathbf{w}, Z)$. Therefore, Corollary 9 suggests that a flat optima could have better transductive generalization performance. This theoretical result has been verify by recent work (Chen et al., 2023), where the authors apply sharpness-aware minimization (Foret et al., 2021) on recommendation task to achieve better generalization performance for GNNs. This result could shed light in understanding the correlation between sharpness and generalization for transductive learning models. Particularly, recent work (Tang and Liu, 2023) reveals that the initial residual and identity mapping adopted in GCNII (Chen et al., 2020) can help the model maintain the generalization gap when the number of layers increases. Investigating how these techniques affect the flatness of loss landscape and ultimately affect the generalization of the model is worth exploring.

4.4 Upper Bounds for Adaptive Optimization Algorithms

As we have mentioned, one advantages of our theoretical results against previous one is that the effect of optimization algorithm on generalization can be fully considered. Now we illustrate this in more detail by analyzing AdaGrad (Duchi et al., 2011), which is one of the most widely adopted optimization algorithms in practice. Different from SGD, the learning rate in AdaGrad is adaptively adjusted during training. Denote by $\{W_t\}_{t\in[T]}$ the weights along the training trajectory of AdaGrad. Following (Wang and Mao, 2022), we consider the setting that mini-batches examples are fixed. Denote by $(B_1, \dots, B_T)$ the sequence of mini-batches where B_t is the examples used in the t-th epoch. For concise we assume that the learner only minimizes the loss on labeled examples, and the number of each mini-batch examples are equal to b. Then the average gradient on the B_t is defined as

$$g(w, B_t(Z)) \triangleq \frac{1}{b}\sum_{\mathbf{z}\in B_t} \nabla_w \ell(w, \mathbf{z}), \tag{13}$$

where $B_t(Z) \subseteq \{\mathbf{z}|\mathbf{z} \in Z^m\}$. Notice that here ℓ means the objective function. For $t \in [T]$, the update rule of AdaGrad can be formulated as

$$v_t = \sum_{k=1}^{t-1} g(W_k, B_k(Z)) \odot g(W_k, B_k(Z)), W_t = W_{t-1} - \frac{\eta}{\sqrt{v_t}+\epsilon} \odot g(W_{t-1}, B_t(Z)), \tag{14}$$

where W_0 is the initial parameter and η, ϵ are two predefined hyper-parameters. Note that v_t is a random variable determined by $W^{[t-1]} \triangleq (W_0, \dots, W_{t-1})$. For the concise of notations we use $\Psi(W^{[t-1]}, Z) \triangleq \left(\eta/(\sqrt{v_t}+\epsilon)\right) \odot g(W_{t-1}, B_t(Z))$ to denote the "adaptive gradient", which is computed by normalizing the current gradient with accumulate squared gradient. Inspired by (Neu et al., 2021; Wang and Mao, 2022), we introduce the following auxiliary weight process $\{\widetilde{W}_t\}_{t\in[T]}$ for analysis:

$$\widetilde{W}_0 = W_0, \widetilde{W}_t = \widetilde{W}_{t-1} - \Psi(W^{[t-1]}, Z) + N_t, \tag{15}$$

where $N_t \triangleq \sigma_t N$. $\{\sigma_t\}_{t\in[T]}$ are predefined hyperparameters and N is a Gaussian random variable independent to $W^{[T]}$ and Z. For concise, we define $U_t \triangleq \sum_{k=1}^{t} N_k$. The upper bound for a transductive learner trained by AdaGrad is presented in the following theorem.

Theorem 10 *Suppose that (i) $\ell(\mathbf{w}, \mathbf{z}) \in [0, B]$ holds for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{z} \in \mathrm{D}$, where $B > 0$ is a constant and (ii) $\mathbb{E}_{U_T,Z}\left[R_u(w_T + U_T, Z) - R_u(w_T, Z)\right] \geq 0$ holds for any realization of $W_T = w_T$. Then we have*

$$
\begin{aligned}
&\mathbb{E}_{W_T,Z}\left[R_u(W_T, Z) - R_m(W_T, Z)\right] \\
\leq &\frac{1}{2}\sqrt{C_{m,u} d \left(\frac{1}{m} + \frac{1}{u}\right) \sum_{t=1}^{T} \log\left(\frac{1}{d\sigma_t^2}\mathbb{E}_{W^{[t-1]},Z}\left[\left\|\Psi(W^{[t-1]}, Z)\right\|_2^2\right] + 1\right)} \\
&+ \mathbb{E}_{Z,W_T,U_T}\left[R_m(W_T + U_T, Z) - R_m(W_T, Z)\right].
\end{aligned} \tag{16}
$$

Assumption (ii) in Theorem 10 is also used in Corollary 9 to establish the PAC-Bayesian bound. This assumption requires that adding random noise to the final parameter does not decrease the risk on unlabeled examples in expectation, which is also used in (Foret et al., 2021; Wang and Mao, 2022). The first term in Eq. (16) records the square of "adaptive gradient" norm along the training trajectory, and the second term measures the expected change of training risk after adding random noise. Compared with stability based (Cong et al., 2021) or complexity based (Tang and Liu, 2023) methods, our results do not contain any Lipschitz or smoothness constants and thus more easier to computed. Besides, the smoothness or Hölder smoothness assumption limit the application scope of previous results, e.g., they could not be applied to neural networks with ReLU as activation function. As a comparison, our result does not rely on these assumptions and has a wider applicability. Also, since the Lipschitz constant is the upper bound of the norm of gradient, our result can more finely depict the impact of optimization trajectories on generalization. Furthermore, the second term characterizes the flatness of the final parameter, which conveys the same insight as Theorem 8, namely a flat optima is beneficial to achieve smaller generalization gap. Considering the popularity of Adam (Kingma and Ba, 2015) in real-world applications, we also derive the corresponding results. The reflected insights are similar yet the formulations are more tedious, and we place the details in Appendix I.

5 Applications

5.1 Semi-supervised Learning

Due to the expensive cost of collecting high-quality labeled data, semi-supervised learning aims to train a learner with a few labeled examples and a large amount of unlabeled data. The analysis for the generalization of semi-supervised learner has been widely explored (Mey and Loog, 2023), and the theoretical results differ by the problem setting and assumptions. Here we focus on the transductive setting, which is also termed as setting 2 of transudctive learning (Vapnik, 1982). Formally, the labeled and unlabeled data is represented as $S_m \triangleq \{(X_i, Y_i)\}_{i\in[m]}$ and $S_u \triangleq \{(X_i, Y_i)\}_{i\in[u]}$, which are samples independently from certain distribution. The semi-supervised learner takes $S_m \cup S_u^X$ as input and outputs the hypothesis characterized by $W \in P_{W|S_m \cup S_u^X}$, where $S_u^X \triangleq \{X_i\}_{i\in[u]}$. Different from the setting we present in Subsection 3.2, here each example (X, Y) should be regarded as random variable rather than constant pair. Furthermore, the training and test risk are defined as $R(W, S_m) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(W, (X_i, Y_i))$ and $R(W, S_u) \triangleq \frac{1}{u}\sum_{i=m+1}^{m+u} \ell(W, (X_i, Y_i))$, respectively. Leveraging the theoretical results established in Section 4, the generalization of semi-supervised learner can be obtained.

Proposition 11 *Under the assumptions of Corollary 6, we have*

$$\left|\mathbb{E}_{S_m,S_u^X,W}\left[R(W,S_u)-R(W,S_m)\right]\right| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u},\widetilde{Z}}\sqrt{2I^{S_{m+u},\widetilde{Z}}(F_i;U_i)}, \tag{17}$$

$$\left|\mathbb{E}_{S_m,S_u^X,W}\left[R(W,S_u)-R(W,S_m)\right]\right| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u},\widetilde{Z}}\sqrt{2I^{S_{m+u},\widetilde{Z}}(L_i;U_i)}, \tag{18}$$

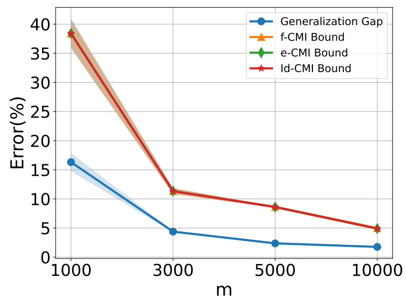
$$\left|\mathbb{E}_{S_m,S_u^X,W}\left[R(W,S_u)-R(W,S_m)\right]\right| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u},\widetilde{Z}}\sqrt{2I^{S_{m+u},\widetilde{Z}}(\Delta_i;U_i)}. \tag{19}$$

Compared with Eq. (8) in Corollary 6, the upper bound in Eq. (17) includes the disentangled mutual information $I^{s_{m+u},\widetilde{z}}(F_i;U_i)$ rather than $I^{\widetilde{z}}(F_i;U_i)$. The reason is that each element in S_{m+u} is random variable rather than constant, and the randomness of S_{m+u} should also be taken into consideration.

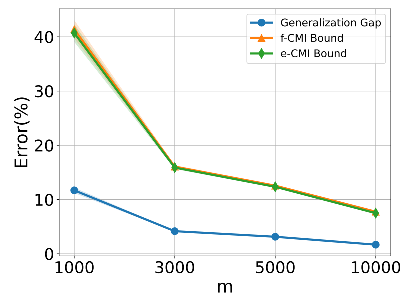
5.2 Transductive Graph Learning

Composed of several objects and their relationship, graph-structured data plays an important role in real-world applications, e.g., recommendation system (Wang et al., 2019b; He et al., 2020; Huang et al., 2021), drug discovery (Sun et al., 2020; Bongini et al., 2021), and traffic flow forecasting (Song et al., 2020; Li and Zhu, 2021; Lan et al., 2022). Recent years have witnessed the success of GNNs in various learning and inference tasks on graph-structured data. The graph learning tasks can be divided into transductive task and inductive task, and we only discuss the first one. Transductive graph learning tasks include graph-level task and node/edge-level task. The goal for node-level task is to predict the label of nodes. For edge-level task, the learner is required to predict whether there is a link between two nodes. Both of them fall into the category of transductive learning. Taking node classification as an example, all nodes are randomly divided into training and test nodes, and the labels of training nodes are revealed to the GNN model during training. Let D be the set containing all nodes and all edges (including positive and negative edges) respectively, our results can be applied on analyzing the generalization gap of GNNs on node classification task and link prediction task. Now we use node classification with GCN (Kipf and Welling, 2017) as an illustration. Denote by $\widetilde{\mathrm{A}} \in \mathbb{R}^{n\times n}$ and $\mathrm{X} \in \mathbb{R}^{n\times d}$ the normalized adjacent matrix with self-loops and feature matrix, respectively. The prediction of a two-layer GCN model is given by $\hat{Y} = \mathrm{Softmax}(\widetilde{\mathrm{A}}\mathrm{ReLU}(\widetilde{\mathrm{A}}\mathrm{X}W_1)W_2)$, where $W_1 \in \mathbb{R}^{d_0\times d_1}, W_2 \in \mathbb{R}^{d_1\times|\mathcal{Y}|}$ are parameters. Here we use $W \triangleq [\mathrm{vec}[W_1], \mathrm{vec}[W_2]]$ to denote the collection of all parameters. Without loss of generality we assume that $Z_j = \mathbf{z}_j$. Define $H^{(1)} \triangleq \widetilde{\mathrm{A}}\mathrm{X}W_1$, the gradient $g(W,Z)$ is formulated by $g(W,Z) = [g_1(W,Z), g_2(W,Z)]$ with

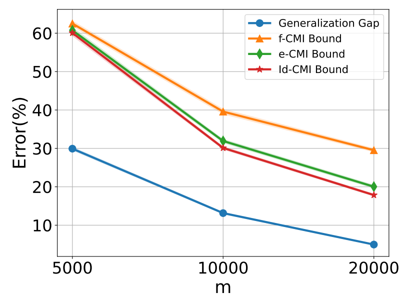
$$\begin{aligned}
g_1(W,Z) &\triangleq \frac{1}{m}\sum_{i=1}^{m}(\widehat{Y}_{i,:}-Y_{i,:})\otimes\left(\sum_{j=1}^{n}\widetilde{\mathrm{A}}_{ij}\mathrm{ReLU}(H^{(1)}_{j,:})\right),\\
g_2(W,Z) &\triangleq \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\widetilde{\mathrm{A}}_{ij}\left(\mathrm{ReLU}'\left(\sum_{k=1}^{n}\widetilde{\mathrm{A}}_{jk}\mathrm{X}_{k,:}W_1\right)\odot\left((\widehat{Y}_{i,:}-Y_{i,:})W_2^{\top}\right)\right)\otimes H^{(1)}_{j,:},
\end{aligned} \tag{20}$$



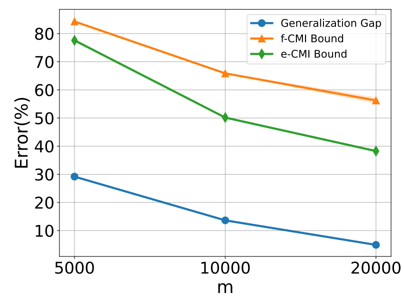
(a) MNIST with $k = 2$



(b) MNIST with $k = 3$



(c) CIFAR-10 with $k = 2$



(d) CIFAR-10 with $k = 3$

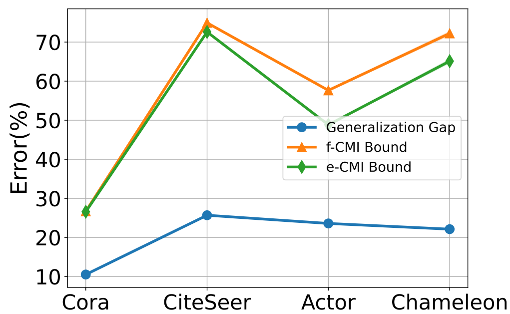
Figure 1: Estimation of the transductive generalization gap and the derived bounds on MNIST and CIFAR-10 with different values of m and k.

where $\mathrm{ReLU}'(\cdot)$ is the derivation of ReLU function and $Y \in \{0,1\}^{n\times|\mathcal{Y}|}$ is the label matrix. Notice that here we have assumed that the loss is computed on all labeled nodes, which is a common setting in practices (Kipf and Welling, 2017; Gasteiger et al., 2019; Chien et al., 2021). By plugging $g(W, Z)$ into Theorem 10 we can obtain the upper bound. Notably, the architecture of GNN model and graph-structured property are reflected in the gradient terms, by which their impacts on generalization is described. Following the analysis technique in (Cong et al., 2021; Tang and Liu, 2023), one can derive fine-grained upper bounds for other GNN models and gain insights on its generalization behavior.

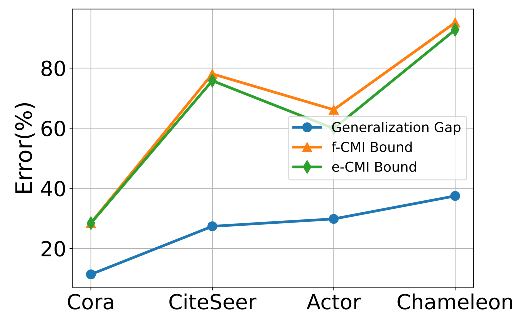
6 Experiments

6.1 Experimental Setup

Semi-supervised Learning. We choose image classification on MNIST and CIFAR-10 as the semi-supervised learning tasks. The semi-supervised learning loss for unlabeled images is defined as the mean square error between the prediction of the augmented images and the vanilla images by the model, which is also termed as consistency regularization in semi-supervised learning. Following (Harutyunyan et al., 2021; Guo et al., 2020), we adopt a four-layer CNN and Wide ResNet-28-10 (Zagoruyko and Komodakis, 2016) as the model for MNIST and CIFAR-10, respectively. For both these two experiments, we train the model on 1000 mini batches using Adam optimizer with learning rate 0.001, and the



(a) GAT



(b) GPR-GNN

Figure 2: Estimation of the transductive generalization gap and the derived bounds on real-world datasets with GAT and GPR-GNN.

number of images per mini batch is fixed to 128. The loss is set to zero-one loss. Following (Harutyunyan et al., 2021), we make the training process be deterministic by fixing the sequence of mini batch and the initialization of parameters via random seed. Please refer to Appendix K for the detail of network architecture and the estimation of the expected generalization gap and derived bounds.

Transductive Graph Learning. We choose semi-supervised node classification on synthetic and real world datasets as the learning tasks. Specifically, we select cSBMs (Deshpande et al., 2018) as the synthetic data, and Cora, CiteSeer (Sen et al., 2008; Yang et al., 2016), Actor and Chameleon to be the real-world dataset. For each of these datasets, we adopt GAT (Veličković et al., 2018) and GPR-GNN (Chien et al., 2021) as the learner, which are the representative of spatial and spectral GNNs. Please refer to Appendix K for more details. We train the model on all labeled nodes for 300 epochs with Adam optimizer with learning rate 0.01.

6.2 Experimental Results

Figure 1 shows the results of semi-supervised learning algorithms on MNIST and CIFAR-10 datasets, where m and k denote the number of labeled images and the ration of unlabeled images to labeled images respectively. Notice that we only extend the results of f-CMI and e-CMI to the case that $k \geq 3$. It can be observed that our established bounds are non-vacuous, and the difference between the estimated value and generalization gap decreases with the increase of m. Furthermore, this difference also increases when k becomes larger. The reason is that a larger value of k leads to the larger estimated error of the conditional mutual information, as we reveal in Subsection 4.2. Besides, the e-CMI bound is no larger than the f-CMI bound, and the Id-CMI bound is no larger than the e-CMI bound for $k = 2$. This result has been revealed in (Hellström and Durisi, 2022; Wang and Mao, 2023a) and it still holds in transductive learning setting. The results of transductive graph learning are presented in Figure 2 and Figure 3, respectively. The tendency is generally consistent with that of semi-supervised learning. Since heterophilic graphs (Actor and Chameleon) are more difficult to learn than homophilic graphs (Cora and CiteSeer) and the synthetic

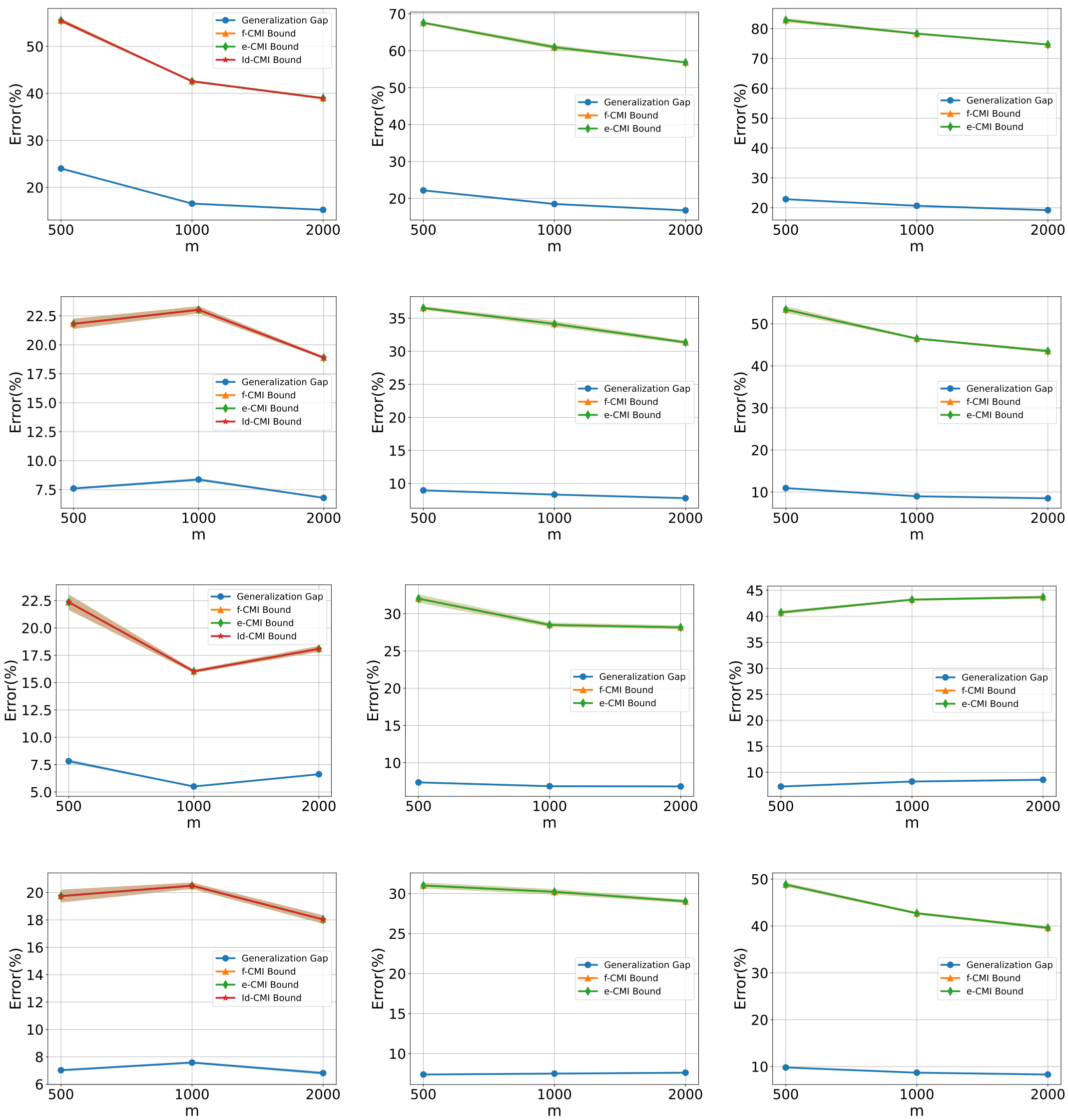


Figure 3: Estimation of the transductive generalization gap and the derived bounds on cSBMs. The first (third) and the second (fourth) row show the results of GAT (GPR-GNN). The first (second) and third (fourth) rows correspond to $\phi = -0.5$ ($\phi = 0.5$). The left, middle, and right figures in each row correspond to $k = 2, 3, 4$.

graphs, the difference between f-CMI bound and e-CMI bound for heterophilic graphs is more pronounced.

7 Conclusion

In this work, we study the generalization of transductive learning algorithms under the viewpoint of information theory, and establish upper bounds for general transductive algorithms and iterative algorithms in terms of different information measure. Furthermore, we demonstrate their applications in semi-supervised learning and transductive graph learning, and also empirically validate them by experiments. Promising future directions include applying our results to other scenarios and designing new information measures for transductive learning setting.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NO.62076234); the Beijing Natural Science Foundation (NO.4222029); the Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double First Class" Initiative, Renmin University of China; the Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098); the Public Computing Cloud, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (NO.2021030199); the Huawei-Renmin University joint program on Information Retrieval; the Unicom Innovation Ecological Cooperation Plan; the CCF Huawei Populus Grove Fund; and the National Key Research and Development Project (NO.2022YFB2703102).

Appendix A. Notations and Lemma

We introduce additional notations used throughout the paper. The combination number is denoted as $C_m^n = \frac{m!}{n!(m-n)!}$. Furthermore, we use $\mathbb{N}$ to denote the set of all non-negative integers, and $\mathbb{N}_+$ to denote the set of all positive integers. Also, $\mathbb{R}$ is the set of real number and $\mathbb{R}_{\geq 0}$ is the set of non-negative real numbers. The gamma function is denoted as $\Gamma(\cdot)$. We finish this section by introducing the following lemma, which is also termed as the Donsker-Varadhan dual characterization of KL divergence or Gibbs variational principle in the literature. This lemma is the foundation of most information and PAC-Bayesian theoretical results.

Lemma 12 (Theorem 4.6 in Polyanskiy and Wu (2022)) *Let P, and Q be two probability measure on $\mathcal{X}$ and $\mathcal{F} \triangleq \{f : \mathcal{X} \to \mathbb{R}\}$ the family of bounded measurable function. Then we have*

$$\mathrm{D}_{\mathrm{KL}}(P||Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P\left[f(X)\right] - \log \mathbb{E}_Q\left[\exp\left\{f(X)\right\}\right]. \tag{21}$$

Appendix B. Proof of Theorem 1

We firstly show that $\mathcal{E}(w, Z)$ satisfies sub-gaussian property for a fixed realization w of W. Inspired by (Cortes et al., 2008; El-Yaniv and Pechyony, 2007), we construct the following martingale difference sequences:

$$V_i \triangleq \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_i] - \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}], i \in [n]. \tag{22}$$

With this definition, one can verify that $\mathcal{E}(w, Z) - \mathbb{E}[\mathcal{E}(w, Z)] = \sum_{i=1}^n V_i$. Note that V_i is a function of $Z_1, \ldots, Z_i$. Define

$$\begin{aligned}
L_i &\triangleq \inf_z \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}, Z_i = z] - \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}],\\
U_i &\triangleq \sup_z \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}, Z_i = z] - \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}],
\end{aligned}$$

we have $L_i \leq V_i \leq U_i$. Then we show that $U_i - L_i$ is a bounded random variable when $Z_1, \ldots, Z_{i-1}$ are given:

$$\begin{aligned}
&U_i - L_i\\
=&\sup_z \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}, Z_i = z] - \inf_z \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}, Z_i = z]\\
=&\sup_{z,\tilde{z}} \Big\{\mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}, Z_i = z] - \mathbb{E}[\mathcal{E}(w, Z)|Z_1, \ldots, Z_{i-1}, Z_i = \tilde{z}]\Big\}\\
=&\frac{u!(m-i)!C_{n-i-1}^{m-i}}{(n-i)!} \cdot \frac{(m+u)B}{mu}\\
=&\frac{(m+u)B}{m(m+u-i)} \triangleq c_i.
\end{aligned} \tag{23}$$

Since $\mathbb{E}[V_i|Z_1,\dots,Z_{i-1}]=0$, by Hoeffding's lemma, $\mathbb{E}\left[e^{\lambda V_i}|Z_1,\dots,Z_{i-1}\right]\le e^{\frac{\lambda^2 c_i^2}{8}}$ holds for any $\lambda\in\mathbb{R}$. Notice that

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}V_i\right\}\right]=&\mathbb{E}\left[\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n-1}V_i\right\}\exp\left\{V_n\right\}\right]\bigg|Z_1,\dots,Z_{n-1}\right]\\ =&\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n-1}V_i\right\}\mathbb{E}\left[\exp\left\{V_n\right\}\right]\bigg|Z_1,\dots,Z_{n-1}\right]\\ \le&\exp\left\{\frac{\lambda^2c_n^2}{8}\right\}\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n-1}V_i\right\}\right].\end{aligned}\tag{24}$$

By recursively repeating the above process, we obtain

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}V_i\right\}\right]\\ \le&\exp\left\{\frac{\lambda^2}{8}\sum_{i=1}^{n}c_i^2\right\}=\exp\left\{\frac{\lambda^2(m+u)^2B^2}{8m^2}\sum_{i=1}^{n}\frac{1}{(m+u-i)^2}\right\}\\ \le&\exp\left\{\frac{\lambda^2B^2(m+u)^2}{8m(u-1/2)(m+u-1/2)}\right\}.\end{aligned}\tag{25}$$

Due to the symmetric of training and test partition, the final bound is obtained by taking the largest one among them:

$$\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}V_i\right\}\right]\le\exp\left\{\frac{\lambda^2B^2(m+u)^2}{8mu(m+u-1/2)}\cdot\frac{2\max(m,u)}{2\max(m,u)-1}\right\}\tag{26}$$

Combining Eq. (26) and the facts that $\mathcal{E}(w,Z)-\mathbb{E}[\mathcal{E}(w,Z)]=\sum_{i=1}^{n}V_i$ and $\mathbb{E}[\mathcal{E}(w,Z)]=0$, we obtain

$$\begin{aligned}&\mathbb{E}_Z\left[\exp\left\{\lambda(R_u(w,Z)-R_m(w,Z))\right\}\right]\\ \le&\exp\left\{\frac{\lambda^2B^2(m+u)^2}{8mu(m+u-1/2)}\cdot\frac{2\max(m,u)}{2\max(m,u)-1}\right\}\\ =&\exp\left\{\frac{\lambda^2(m+u)C_{m,u}}{8mu}\right\}\end{aligned}\tag{27}$$

where $C_{m,u}\triangleq\frac{2B^2(m+u)\max(m,u)}{(m+u-1/2)(2\max(m,u)-1)}$. Denote by Z' the independent copy of Z, which is independent from W and has the same distribution as Z. Then we have

$$\begin{aligned}&\log\mathbb{E}_{W,Z'}\left[\exp\left\{\lambda(R_u(W,Z')-R_m(W,Z'))\right\}\right]\\ =&\log\left(\int_w\mathbb{E}_{Z'}\left[\exp\left\{\lambda(R_u(w,Z')-R_m(w,Z'))\right\}\right]\mathrm{d}P_W(w)\right)\\ \le&\log\left(\int_w\exp\left\{\frac{\lambda^2(m+u)C_{m,u}}{8mu}\right\}\mathrm{d}P_W(w)\right)\\ =&\frac{\lambda^2C_{m,u}}{8}\left(\frac{1}{m}+\frac{1}{u}\right).\end{aligned}\tag{28}$$

By Lemma 12, for any $\lambda \in \mathbb{R}$:

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(P_{Z,W}||P_{Z',W}) \\
\geq&\mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z)-R_m(W,Z))\right]-\log\mathbb{E}_{Z',W}\left[\exp\left\{\lambda(R_u(W,Z')-R_m(W,Z'))\right\}\right] \\
\geq&\mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z)-R_m(W,Z))\right]-\frac{\lambda^2 C_{m,u}}{8}\left(\frac{1}{m}+\frac{1}{u}\right),
\end{aligned} \tag{29}
$$

which implies that

$$
\left|\mathbb{E}_{Z,W}\left[R_u(W,Z)-R_m(W,Z)\right]\right| \leq \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)I(W;Z)}. \tag{30}
$$

This finishes the proof for the first part. For the second part, note that Eq. (27) can be rewritten as $\mathbb{E}_Z[\exp\{\lambda\mathcal{E}(w,Z)\}] \leq \exp\{\lambda^2\sigma_{m,u}\}$, where $\sigma_{m,u} \triangleq C_{m,u}(1/m+1/u)/8$. Similarly we have $\mathbb{E}_Z[\exp\{-\lambda\mathcal{E}(w,Z)\}] \leq \exp\{\lambda^2\sigma_{m,u}\}$. Then we have

$$
\mathbb{P}\left\{|\mathcal{E}(w,Z)|\geq t\right\} \leq \mathbb{P}\left\{\mathcal{E}(w,Z)\geq t\right\}+\mathbb{P}\left\{\mathcal{E}(w,Z)\leq -t\right\} \leq 2\exp\left\{-\frac{t^2}{4\sigma_{m,u}}\right\}, \tag{31}
$$

where the first and the second inequality are due to the Boole's inequality and the Chernoff technique, respectively. For any $k \in \mathbb{N}_+$, we have

$$
\begin{aligned}
\mathbb{E}\left[|\mathcal{E}(w,Z)|^k\right] =& \int_0^\infty \mathbb{P}\{|\mathcal{E}(w,Z)|^k \geq u\}\,\mathrm{d}u \\
=& k\int_0^\infty \mathbb{P}\{|\mathcal{E}(w,Z)|\geq t\}\,t^{k-1}\,\mathrm{d}t \\
\leq& 2k\int_0^\infty \exp\left\{-\frac{t^2}{4\sigma_{m,u}}\right\}t^{k-1}\,\mathrm{d}t = (4\sigma_{m,u})^{\frac{k}{2}}k\Gamma(k/2),
\end{aligned} \tag{32}
$$

which implies that

$$
\mathbb{E}\left[\exp\{\lambda\mathcal{E}^2(w,Z)\}\right] = 1+\sum_{k=1}^\infty \frac{\lambda^k}{k!}\mathbb{E}\left[|\mathcal{E}(w,Z)|^{2k}\right] \leq 1+2\sum_{k=1}^\infty(4\lambda\sigma_{m,u})^k. \tag{33}
$$

By Lemma 12, for any $\lambda \in \mathbb{R}$:

$$
\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(P_{Z,W}||P_{Z,W'}) \\
\geq&\mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z)-R_m(W,Z))^2\right]-\log\mathbb{E}_{Z,W'}\left[\exp\left\{\lambda(R_u(W',Z)-R_m(W',Z))^2\right\}\right] \\
\geq&\mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z)-R_m(W,Z))^2\right]-\log\left(1+2\sum_{k=1}^\infty(4\lambda\sigma_{m,u})^k\right),
\end{aligned} \tag{34}
$$

Let $\lambda = 1/8\sigma_{m,u}$ and plugging into $\sigma_{m,u} \triangleq C_{m,u}(1/m+1/u)/8$, we obtain

$$
\mathbb{E}_{Z,W}\left[(R_u(W,Z)-R_m(W,Z))^2\right] \leq C_{m,u}\left(\frac{1}{m}+\frac{1}{u}\right)(I(Z;W)+\log 3). \tag{35}
$$

Appendix C. Proof of Theorem 2

Denote by $Z^{(1)}, \ldots, Z^{(k)}$ the k independent copy of Z. By running a transductive algorithm $\mathcal{A}$ on each $Z^{(j)}$ respectively, we obtain the corresponding output $W^{(j)} = \mathcal{A}(Z^{(j)})$ for $j \in [k]$. By this way, $(Z^{(j)}, W^{(j)})$ can be regarded as independent copy of (Z, W) for $j \in [k]$. Now assume that there is a monitor that returns

$$(J^*, R^*) \triangleq \underset{j \in [k], r \in \{\pm 1\}}{\operatorname{argmax}} \; r\mathcal{E}(W^{(j)}, Z^{(j)}), \;\; W^* = W_{J^*}.$$

One can verify that

$$R^* \mathcal{E}(W^{(J^*)}, Z^{(J^*)}) = \max_{j \in [k]} |\mathcal{E}(Z^{(j)}, W^{(j)})|.$$

Now taking expectation on both side, we have

$$\mathbb{E}_{Z^{(1)}, \ldots, Z^{(k)}, J^*, R^*, W^*} \left[R^* \mathcal{E}(W^{(J^*)}, Z^{(J^*)}) \right] = \mathbb{E}_{Z^{(1)}, \ldots, Z^{(k)}, W_1, \ldots, W_k} \left[\max_{j \in [k]} |\mathcal{E}(Z^{(j)}, W^{(j)})| \right].$$

Using the same procedure used in the proof of Theorem 1, we have

$$\log \mathbb{E}_{J^*, R^*, W^*} \mathbb{E}_{Z^{(1)}, \ldots, Z^{(k)}} \left[\exp \left\{ \lambda R^* \mathcal{E}(W^{(J^*)}, Z^{(J^*)}) \right\} \right] \leq \frac{\lambda^2 C_{m,u}}{8} \left(\frac{1}{m} + \frac{1}{u} \right).$$

By Donsker-Varadhan's variational formula, the following inequality holds for any $\lambda \in \mathbb{R}$:

$$\begin{aligned}
& D(P_{Z^{(1)}, \ldots, Z^{(k)}, J^*, R^*, W^*} || P_{Z^{(1)}, \ldots, Z^{(k)}} \otimes P_{J^*, R^*, W^*}) \\
\geq & \mathbb{E}_{Z^{(1)}, \ldots, Z^{(k)}, J^*, R^*, W^*} \left[\lambda R^* \mathcal{E}(W^{(J^*)}, Z^{(J^*)}) \right] \\
& - \log \mathbb{E}_{J^*, R^*, W^*} \mathbb{E}_{Z^{(1)}, \ldots, Z^{(k)}} \left[\exp \left\{ \lambda R^* \mathcal{E}(W^{(J^*)}, Z^{(J^*)}) \right\} \right] \\
\geq & \lambda \mathbb{E}_{Z^{(1)}, \ldots, Z^{(k)}, J^*, R^*, W^*} \left[R^* \mathcal{E}(W^{(J^*)}, Z^{(J^*)}) \right] - \frac{\lambda^2 C_{m,u}}{8} \left(\frac{1}{m} + \frac{1}{u} \right).
\end{aligned}$$

which implies

$$\mathbb{E}_{Z^{(1)}, \ldots, Z^{(k)}, J^*, R^*, W^*} \left[R^* \mathcal{E}(W^{(J^*)}, Z^{(J^*)}) \right] \leq \sqrt{\frac{C_{m,u}}{2} \left(\frac{1}{m} + \frac{1}{u} \right) I(Z^{(1)}, \ldots, Z^{(k)}; J^*, R^*, W^*)}. \tag{36}$$

Next we provide a upper bound for the mutual information. Note that

$$\begin{aligned}
& I(Z^{(1)}, \ldots, Z^{(k)}; J^*, R^*, W^*) \\
\leq & I(Z^{(1)}, \ldots, Z^{(k)}; J^*, R^*, W^*, W_1, \ldots, W_k) \\
= & I(Z^{(1)}, \ldots, Z^{(k)}; W_1, \ldots, W_k) + I(Z^{(1)}, \ldots, Z^{(k)}; J^*, R^*, W^* | W_1, \ldots, W_k) \\
= & \sum_{j=1}^{k} I(Z^{(j)}; W_j) + I(Z^{(1)}, \ldots, Z^{(k)}; J^*, R^*, W^* | W_1, \ldots, W_k) \\
\leq & k I(Z; W) + \log(2k).
\end{aligned} \tag{37}$$

where we have used the fact that $(Z^{(k)}, W_j), j \in [k]$ are independent. Plugging Eq. (37) into Eq. (36) yields

$$\mathbb{E}_{Z^{(1)},\dots,Z^{(k)},W^{(1)},\dots,W^{(k)}}\left[\max_{j\in[k]}|\mathcal{E}(Z^{(j)},W^{(j)})|\right] \le \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)(\log(2k)+kI(Z,W))}.$$

Since $(Z^{(j)}, W^{(j)})$ are independent copy of (Z, W), for any $\alpha > 0$ we have

$$\mathbb{P}_{Z^{(1)},W^{(1)},\dots,Z^{(k)},W^{(k)}}\left\{\max_{j\in[k]}|\mathcal{E}(Z^{(j)},W^{(j)})| < \alpha\right\} = \left(\mathbb{P}_{Z,W}\left\{|\mathcal{E}(Z,W)|<\alpha\right\}\right)^k.$$

By Markov's inequality:

$$\begin{aligned}
&\mathbb{P}_{Z^{(1)},W^{(1)},\dots,Z^{(k)},W^{(k)}}\left\{\max_{j\in[k]}|\mathcal{E}(Z^{(j)},W^{(j)})| \ge \alpha\right\}\\
\le&\frac{1}{\alpha}\mathbb{E}_{Z^{(1)},\dots,Z^{(k)},W^{(1)},\dots,W^{(k)}}\left[\max_{j\in[k]}|\mathcal{E}(Z^{(j)},W^{(j)})|\right]\\
\le&\frac{1}{\alpha}\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)(\log(2k)+kI(Z,W))}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}_{Z,W}\left\{|\mathcal{E}(Z,W)|\ge\alpha\right\}\\
=&1-\mathbb{P}_{Z,W}\left\{|\mathcal{E}(Z,W)|<\alpha\right\}\\
=&1-\left(\mathbb{P}_{Z^{(1)},W^{(1)},\dots,Z^{(k)},W^{(k)}}\left\{\max_{j\in[k]}|\mathcal{E}(Z^{(j)},W^{(j)})|<\alpha\right\}\right)^{\frac{1}{k}}\\
=&1-\left(1-\mathbb{P}_{Z^{(1)},W^{(1)},\dots,Z^{(k)},W^{(k)}}\left\{\max_{j\in[k]}|\mathcal{E}(Z^{(j)},W^{(j)})|<\alpha\right\}\right)^{\frac{1}{k}}\\
\le&1-\left(1-\frac{1}{\alpha}\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)(\log(2k)+kI(Z,W))}\right)^{\frac{1}{k}}.
\end{aligned}$$

Let $\alpha = 2\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)(\log(2k)+kI(Z,W))}$ and $k=\lfloor\frac{1}{\delta}\rfloor$, we have obtained the result. Let $k=1$, we obtain

$$\mathbb{E}_{Z,W}\left[|\mathcal{E}(Z,W)|\right] \le \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)(\log 2+I(Z,W))}.$$

Appendix D. Proof of Proposition 4

Denote by $\mathcal{U}$ the set containing all values of U. For any $\widetilde{z}\in\widetilde{\mathcal{Z}}$ and $u\in\mathcal{U}$, we use

$$\bar{z}(\widetilde{z},u) \triangleq (\widetilde{z}_{1,u_1},\dots,\widetilde{z}_{m,u_m},\widetilde{z}_{1,1-u_1},\dots,\widetilde{z}_{m,1-u_m}) \tag{38}$$

to denote the random permutation vector induced by $\widetilde{z}$ and u. Let $\bar{Z}:\{\bar{z}(\widetilde{z},u)|\widetilde{z}\in\widetilde{\mathcal{Z}},u\in\mathcal{U}\}$, it is sufficient to proof that $\bar{Z}=\mathfrak{Z}$, where $\mathfrak{Z}$ is the set includes all possible values of

Z defined in Subsection 3.2. Note that each element $\widetilde{z}$ in $\widetilde{\mathcal{Z}}$ differs by each other, since $\widetilde{z}$ is a partitions of $2m$ elements into m subsets, where each subset contains 2 elements. Thus, the cardinality of $\widetilde{\mathcal{Z}}$ is $\frac{(2m)!}{2^m}$. Furthermore, for fixed $\widetilde{z}$ and $u_1, u_2 \in \mathcal{U}$, it is clear that $u_1 \neq u_2$ implies $\bar{z}(\widetilde{z}, u_1) \neq \bar{z}(\widetilde{z}, u_2)$, due to the fact that $\bar{z}(\widetilde{z}, u_1) = \bar{z}(\widetilde{z}, u_2)$ if and only if $\bar{z}(\widetilde{z}, u_1)_j = \bar{z}(\widetilde{z}, u_2)_j$ for $j \in [2m]$. Here $\bar{z}(\widetilde{z}, u)_j$ represents the j-th entry in the sequence. Now we claim that for any two element $\bar{z}_1(\widetilde{z}_1, u_1), \bar{z}_2(\widetilde{z}_2, u_2) \in \bar{Z}$, if $\bar{z}_1(\widetilde{z}_1, u_1) = \bar{z}_2(\widetilde{z}_2, u_2)$, then $\widetilde{z}_1 = \widetilde{z}_2$ holds. To see this, we show that $\widetilde{z}$ can be uniquely determined when seeing $\bar{z}(\widetilde{z}, u)$. Recall that $\widetilde{z}$ is a sequence containing m element, and each element $\widetilde{z}_j$ is a set. By Eq. (38), we conclude that $\widetilde{z}_{1,u_1}$ comes from $\widetilde{z}_1$, $\widetilde{z}_{2,u_2}$ comes from $\widetilde{z}_2$, and so on. Similarly, $\widetilde{z}_{1,1-u_1}$ comes from $\widetilde{z}_1$, $\widetilde{z}_{2,1-u_2}$ comes from $\widetilde{z}_2$, and so on. By this way, we have recovered $\widetilde{z}$ from $\bar{z}(\widetilde{z}, u)$ and it is unique. Together with the fact $\bar{z}(\widetilde{z}, u_1) = \bar{z}(\widetilde{z}, u_2) \implies u_1 = u_2$ that we have just shown, we conclude that $\bar{z}_1(\widetilde{z}_1, u_1) = \bar{z}_2(\widetilde{z}_2, u_2) \implies \widetilde{z}_1 = \widetilde{z}_2, u_1 = u_2$, which suggest that $(\widetilde{z}, u) \to \bar{z}(\widetilde{z}_1, u_1)$ is a one-to-one mapping. Since $|\mathcal{S}| = 2^m$, we have $|\bar{Z}| = |\widetilde{\mathcal{Z}}||\mathcal{S}| = (2m)! = |\mathfrak{Z}|$. Combining this with the fact that $\bar{Z} \subseteq \mathfrak{Z}$, we conclude that $\bar{Z} = \mathfrak{Z}$.

Appendix E. Proof of Theorem 5

We firstly present a warm-up example for illustrating the proposed transductive supersamples. Suppose $m = 2$, all entries in $\widetilde{Z}$ are as follows:

- $(\{\mathbf{z}_1, \mathbf{z}_2\}, \{\mathbf{z}_3, \mathbf{z}_4\})$
- $(\{\mathbf{z}_1, \mathbf{z}_3\}, \{\mathbf{z}_2, \mathbf{z}_4\})$
- $(\{\mathbf{z}_1, \mathbf{z}_4\}, \{\mathbf{z}_2, \mathbf{z}_3\})$
- $(\{\mathbf{z}_2, \mathbf{z}_3\}, \{\mathbf{z}_1, \mathbf{z}_4\})$
- $(\{\mathbf{z}_2, \mathbf{z}_4\}, \{\mathbf{z}_1, \mathbf{z}_3\})$
- $(\{\mathbf{z}_3, \mathbf{z}_4\}, \{\mathbf{z}_1, \mathbf{z}_2\})$

Now we give the formal proof. Denote by w and $\widetilde{z}$ the fixed realizations of W and $\widetilde{Z}$. For any $\lambda \in \mathbb{R}$, by Hoeffding's Lemma:

$$\begin{aligned}
&\mathbb{E}_U\left[\exp\left\{\lambda \mathcal{E}(w, \widetilde{z}, U)\right\}\right] \\
=&\mathbb{E}_U\left[\exp\left\{\frac{\lambda}{m}\sum_{i=1}^{m} \ell(w, \widetilde{z}_{i,U_i}) - \ell(w, \widetilde{z}_{i,1-U_i})\right\}\right] \leq \exp\left\{\frac{\lambda^2 B^2}{2m}\right\}.
\end{aligned} \tag{39}$$

Let U' be the independent copy of U, we have

$$\begin{aligned}
&\log \mathbb{E}_{U',W|\widetilde{Z}=\widetilde{z}}\left[\exp\left\{\lambda \mathcal{E}(W, \widetilde{z}, U')\right\}\right] \\
=&\log\left(\int_w \mathbb{E}_{U'}\left[\exp\left\{\lambda \mathcal{E}(w, \widetilde{z}, U')\right\}\right] \mathrm{d}P_{W|\widetilde{Z}=\widetilde{z}}(w)\right) \leq \frac{\lambda^2 B^2}{2m}.
\end{aligned} \tag{40}$$

where we have used the fact that $P_{U',W|\widetilde{Z}=\widetilde{z}} = P_{W|\widetilde{Z}=\widetilde{z}}P_{U'}$, due to U' is independent from both $\widetilde{Z}$ and W. By Lemma 12, for any $\lambda \in \mathbb{R}$:

$$\begin{aligned}
&I^{\widetilde{z}}(U;W) = \mathrm{D}_{\mathrm{KL}}(P_{S,W|\widetilde{Z}=\widetilde{z}}||P_{S',W|\widetilde{Z}=\widetilde{z}}) \\
\geq&\mathbb{E}_{S,W|\widetilde{Z}=\widetilde{z}}\left[\lambda\mathcal{E}(W,\widetilde{z},S)\right] - \log\mathbb{E}_{S',W|\widetilde{Z}=\widetilde{z}}\left[\exp\left\{\lambda\mathcal{E}(W,z,S')\right\}\right] \\
\geq&\lambda\mathbb{E}_{S,W|\widetilde{Z}=\widetilde{z}}\left[\mathcal{E}(W,\widetilde{z},S)\right] - \frac{\lambda^2B^2}{2m},
\end{aligned} \tag{41}$$

which implies that

$$\left|\mathbb{E}_{U,W|\widetilde{Z}=\widetilde{z}}\left[\mathcal{E}(W,\widetilde{z},U)\right]\right| \leq \sqrt{\frac{2B^2}{m}I^{\widetilde{z}}(U;W)}.$$

Taking expectation over $\widetilde{Z}$ on both side, we have obtain

$$\begin{aligned}
&\left|\mathbb{E}_{Z,W}\left[R_u(W,Z) - R_m(W,Z)\right]\right| = \left|\mathbb{E}_{\widetilde{Z},U,W}\left[\mathcal{E}(W,\widetilde{Z},U)\right]\right| \\
\leq&\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{U,W|\widetilde{Z}}\left[\mathcal{E}(W,\widetilde{Z},U)\right]\right| \leq \mathbb{E}_{\widetilde{Z}}\sqrt{\frac{2B^2}{m}I^{\widetilde{Z}}(U;W)}.
\end{aligned} \tag{42}$$

For the second part, note that Eq. (39) can be rewritten as $\mathbb{E}_U\left[\exp\left\{\lambda\mathcal{E}(w,\widetilde{z},U)\right\}\right] \leq \exp\{\lambda^2B^2/2m\}$. Similarly we have $\mathbb{E}_U\left[\exp\left\{-\lambda\mathcal{E}(w,\widetilde{z},U)\right\}\right] \leq \exp\{\lambda^2B^2/2m\}$. By the same technique in Section B, we have

$$\mathbb{E}_{U,W|\widetilde{Z}=\widetilde{z}}\left[\mathcal{E}^2(W,\widetilde{z},U)\right] \leq \frac{4B^2}{m}(I^{\widetilde{z}}(U;W) + \log 3). \tag{43}$$

Taking expectation on both side, we have obtain

$$\mathbb{E}_{U,W}\left[(R_u(W,Z) - R_m(W,Z))^2\right] = \mathbb{E}_{\widetilde{Z},U,W}\left[\mathcal{E}^2(W,\widetilde{Z},U)\right] \leq \frac{4B^2}{m}(I(U;W|\widetilde{Z})+\log 3). \tag{44}$$

We close this proof by presenting the results for more ordinary cases where $u = km$ with $k \in \mathbb{N}_+$. By Definition 7, the transductive generalization under this setting is defined by

$$\mathcal{E}(W,\widetilde{Z},U) \triangleq \frac{1}{m}\sum_{i=1}^{m}\ell(W,\widetilde{Z}_{i,S_i}) - \frac{1}{km}\sum_{i=1}^{m}\sum_{j=0}^{k-1}\ell(W,(\widetilde{Z}_i \setminus \widetilde{Z}_{i,U_i})_j). \tag{45}$$

Notice that here V does not appear in the formulation. The reason is that V controls the sequence of test examples, and the loss is independent to this randomness. One can verify that $\mathbb{E}_{W,Z}\left[R_u(W,Z) - R_m(W,Z)\right] = \mathbb{E}_{\widetilde{Z},U,W}\left[\mathcal{E}(W,\widetilde{Z},U)\right]$ holds. Following the same procedure we have

$$\begin{aligned}
&\left|\mathbb{E}_{W,Z}\left[R_u(W,Z) - R_m(W,Z)\right]\right| \leq \mathbb{E}_{\widetilde{Z}}\sqrt{\frac{2(k+1)B^2}{n}I^{\widetilde{Z}}(U;W)} \\
&\mathbb{E}_{W,Z}\left[(R_u(W,Z) - R_m(W,Z))^2\right] \leq \frac{4(k+1)B^2}{n}(I(U;W|\widetilde{Z}) + \log 3).
\end{aligned} \tag{46}$$

Appendix F. Proof of Corollary 6

Denote by $g(F_i, U_i, \widetilde{Y}_i) \triangleq r(F_{i,U_i}, \widetilde{Y}_{i,U_i}) - r(F_{i,1-U_i}, \widetilde{Y}_{i,1-U_i})$ the function of $(F_i, U_i, \widetilde{Y}_i)$. Let f_i and $\widetilde{z}_i = (\widetilde{x}_i, \widetilde{y}_i)$ be the fixed realizations of F_i and $\widetilde{Z}_i$. For any $\lambda \in \mathbb{R}$ and $i \in [m]$, by Hoeffding's Lemma:

$$\mathbb{E}_{U_i}\left[\exp\left\{\lambda g(f_i, U_i, \widetilde{y}_i)\right\}\right] \leq \exp\left\{\frac{\lambda^2 B^2}{2}\right\}. \tag{47}$$

Let U_i' be the independent copy of U_i, by Lemma 12:

$$\begin{aligned} I(F_i; U_i|\widetilde{Z} = \widetilde{z}) &\geq \lambda \mathbb{E}_{F_i, U_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i, U_i, \widetilde{y}_i)\right] - \log \mathbb{E}_{F_i, U_i'|\widetilde{Z}=\widetilde{z}}\left[\exp\left\{\lambda g(F_i, U_i', \widetilde{y}_i)\right\}\right] \\ &\geq \lambda \mathbb{E}_{F_i, U_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i, U_i, \widetilde{y}_i)\right] - \frac{\lambda^2 B^2}{2}. \end{aligned} \tag{48}$$

Then we have

$$\left|\mathbb{E}_{U,W|\widetilde{Z}=\widetilde{z}}\left[\ell(W, \widetilde{z}_{i,U_i}) - \ell(W, \widetilde{z}_{i,1-U_i})\right]\right| = \left|\mathbb{E}_{F_i, U_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i, U_i, \widetilde{y}_i)\right]\right| \leq B\sqrt{2I^{\widetilde{z}}(F_i; U_i)}, \tag{49}$$

which implies that

$$\begin{aligned} &\left|\mathbb{E}_{W,Z}\left[R_u(W,Z) - R_m(W,Z)\right]\right| = \left|\mathbb{E}_{W,\widetilde{Z},U}\left[\mathcal{E}(W, \widetilde{Z}, U)\right]\right| \\ \leq &\mathbb{E}_{\widetilde{Z}}\left|\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{U,W|\widetilde{Z}}\left[\ell(W, \widetilde{Z}_{i,U_i}) - \ell(W, \widetilde{Z}_{i,1-U_i})\right]\right| \\ \leq &\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{U,W|\widetilde{Z}}\left[\ell(W, \widetilde{Z}_{i,U_i}) - \ell(W, \widetilde{Z}_{i,1-U_i})\right]\right| \\ = &\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{F_i,U_i|\widetilde{Z}}\left[g(F_i, U_i, \widetilde{Y}_i)\right]\right| \leq \frac{B}{m}\sum_{i=1}^m \mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(F_i; U_i)}. \end{aligned} \tag{50}$$

Denote by $g(L_i, U_i) = L_{i,U_i} - L_{i,1-U_i}$ and $g(\Delta_i, U_i) \triangleq (-1)^{S_i}\Delta_i$, by the same technique we have

$$\begin{aligned} \left|\mathbb{E}_{Z,W}\left[R_u(W,Z) - R_m(W,Z)\right]\right| &\leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(L_i; S_i)} \\ \left|\mathbb{E}_{Z,W}\left[R_u(W,Z) - R_m(W,Z)\right]\right| &\leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta_i; S_i)}. \end{aligned} \tag{51}$$

Appendix G. Proof of Theorem 8

By Markov's inequality, for any distribution P that is independent to Z and $\delta \in (0,1)$:

$$\mathbb{P}\left\{\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z) - R_m(W,Z))}\right] \geq \frac{1}{\delta}\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z) - R_m(W,Z))}\right]\right\} \leq \delta. \tag{52}$$

By Lemma 12 and Eq. (27), for any distribution Q, with probability at least $1-\delta$ over the randomness of Z:

$$\begin{aligned}
&\lambda\mathbb{E}_{W\sim Q}\left[R_u(W,Z)-R_m(W,Z)\right]\\
\leq&\mathrm{D_{KL}}(Q||P)+\log\left(\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z)-R_m(W,Z))}\right]\right)\\
\leq&\mathrm{D_{KL}}(Q||P)+\log\left(\frac{1}{\delta}\right)+\log\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z)-R_m(W,Z))}\right]\\
=&\mathrm{D_{KL}}(Q||P)+\log\left(\frac{1}{\delta}\right)+\log\left(\int_w\mathbb{E}_{Z'}\left[\exp\left\{\lambda(R_u(w,Z')-R_m(w,Z'))\right\}\right]\,\mathrm{d}P(w)\right)\\
\leq&\mathrm{D_{KL}}(Q||P)+\log\left(\frac{1}{\delta}\right)+\frac{\lambda^2C_{m,u}}{8}\left(\frac{1}{m}+\frac{1}{u}\right),
\end{aligned}\tag{53}$$

which implies that: for any distribution Q, with probability at least $1-\delta$ over the randomness of Z,

$$|\mathbb{E}_{W\sim Q}\left[R_u(W,Z)-R_m(W,Z)\right]|\leq\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\mathrm{D_{KL}}(Q||P)+\log\left(\frac{1}{\delta}\right)\right)}.\tag{54}$$

Note that by setting $Q=P_{W|Z}$ and $P=P_W$, we recover a degenerated version of Theorem 1 holds with probability $1-\delta$:

$$\begin{aligned}
&|\mathbb{E}_{W,Z}\left[R_u(W,Z)-R_m(W,Z)\right]|\leq\mathbb{E}_Z\left[|\mathbb{E}_{W\sim Q}\left[R_u(W,Z)-R_m(W,Z)\right]|\right]\\
\leq&\mathbb{E}_Z\left[\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\mathrm{D_{KL}}(P_{W|S}||P_W)+\log\left(\frac{1}{\delta}\right)\right)}\right]\\
\leq&\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\mathbb{E}_Z\left[\mathrm{D_{KL}}(P_{W|S}||P_W)\right]+\log\left(\frac{1}{\delta}\right)\right)}\\
=&\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(I(S;W)+\log\left(\frac{1}{\delta}\right)\right)}.
\end{aligned}\tag{55}$$

It is worth mentioning that we can also recover another degenerated version of Theorem 1, following the technique used in (Bégin et al., 2014). Denote by $R_{m+u}(W,Z)\triangleq\frac{1}{m+u}\sum_{i=1}^{m+u}\ell(W,Z_i)=\frac{m}{m+u}R_m(W,Z)+\frac{u}{m+u}R_u(W,Z)$ the error on Z. Denote by $\mathcal{D}(p,q)\triangleq p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}$ the KL divergence between two Bernoulli distributions with success probability p and q. Then the $\mathcal{D}$-function introduced in (Bégin et al., 2014) is expressed by $\mathcal{D}^*_\beta(p,q)\triangleq\mathcal{D}(p,q)+\frac{1-\beta}{\beta}\mathcal{D}(\frac{q-\beta p}{1-\beta},q)$. By Theorem 5 and Theorem 6 in (Bégin et al., 2014), for fixed realization w of W:

$$\mathbb{E}_Z\left[\exp\left\{m\mathcal{D}^*_\beta(R_m(w,Z),R_{m+u}(w,Z))\right\}\right]\leq3\log(m)\sqrt{\frac{mu}{m+u}},\tag{56}$$

which implies that

$$\begin{aligned}
&\log\mathbb{E}_{W\otimes Z}\left[\exp\left\{m\mathcal{D}^*_\beta(R_m(W,Z),R_{m+u}(W,Z))\right\}\right]\\
=&\log\left(\int_w\mathbb{E}_Z\left[\exp\left\{m\mathcal{D}^*_\beta(R_m(w,Z),R_{m+u}(w,Z))\right\}\right]\,\mathrm{d}P_W(w)\right)\\
\leq&\log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right).
\end{aligned}\tag{57}$$

By Lemma 12 we have

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(P_{Z,W}||P_{Z,W'})\\
\geq&\mathbb{E}_{Z,W}\left[m\mathcal{D}_{\beta}^{*}(R_m(W,Z),R_{m+u}(W,Z))\right]-\log\mathbb{E}_{W\otimes Z}\left[e^{m\mathcal{D}_{\beta}^{*}(R_m(W,Z),R_{m+u}(W,Z))}\right]\\
\geq&m\mathbb{E}_{Z,W}\left[\mathcal{D}_{\beta}^{*}(R_m(W,Z),R_{m+u}(W,Z))\right]-\log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right).
\end{aligned}\tag{58}$$

By Pinsker's inequality and plugging in $\beta=\frac{m}{m+u}$, the expectation term can be lower bounded by

$$\begin{aligned}
&\mathbb{E}_{Z,W}\left[\mathcal{D}_{\beta}^{*}(R_m(W,Z),R_{m+u}(W,Z))\right]\\
=&\mathbb{E}_{Z,W}\left[\mathcal{D}(R_m(W,Z),R_{m+u}(W,Z))\right]\\
&+\frac{u}{m}\mathbb{E}_{Z,W}\left[\mathcal{D}\left(\frac{m+u}{u}R_{m+u}(W,Z)-\frac{m}{u}R_m(W,Z),R_{m+u}(W,Z)\right)\right]\\
\geq&2\mathbb{E}_{Z,W}\left[(R_m(W,Z)-R_{m+u}(W,Z))^2\right]+2\frac{m}{u}\mathbb{E}_{Z,W}\left[(R_m(W,Z)-R_{m+u}(W,Z))^2\right]\\
=&2\frac{m+u}{u}\mathbb{E}_{Z,W}\left[\left(R_m(W,Z)-\frac{m}{m+u}R_m(W,Z)+\frac{u}{m+u}R_u(W,Z)\right)^2\right]\\
=&\frac{2u}{m+u}\mathbb{E}_{Z,W}\left[(R_m(W,Z)-R_u(W,Z))^2\right]\\
\geq&\frac{2u}{m+u}\left(\mathbb{E}_{Z,W}[R_m(W,Z)-R_u(W,Z)]\right)^2.
\end{aligned}\tag{59}$$

Combining Eq. (58) and Eq. (59) we obtain a degenerated bound compared with that provided in Theorem 1 with extra factors $\log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right)$:

$$|\mathbb{E}_{Z,W}[R_m(W,Z)-R_u(W,Z)]|\leq\sqrt{\frac{1}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left[I(Z;W)+\log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right)\right]}.$$

Appendix H. Proof of Corollary 9

The proof generally follows the proof of Theorem 2 in (Foret et al., 2021). For given posterior distribution Q, we need to properly select the optimal prior distribution P^* such that the KL divergence term $\mathrm{D}_{\mathrm{KL}}(Q||P)$ can be minimized. However, this solution is not applicable. The reason is that P will depend on Q and Z, yet we require that P should be chosen before observing Z. Therefore, the most widely adopted method is to construct a predefined set of prior distribution $\mathcal{P}=\{P_j\}_{j\in\mathbb{N}}$, and then establishing a high probability guarantee for each $P\in\mathcal{P}$. After that, we can establish a high probability guarantee for the optimal prior P^* by using union bound inequality.

Formally, denote by $\mathbf{w}\in\mathbb{R}^d$ the parameter return by the learning algorithm and σ a predefined hyper-parameter, we define the posterior distribution as $Q\triangleq\mathcal{N}(\mathbf{w}+\boldsymbol{\epsilon},\sigma^2\mathbf{I}_d)$. Let $c=\sigma_Q^2(1+e^{4n/d})$, the predefined set $\mathcal{P}$ is constructed as

$$\mathcal{P}\triangleq\left\{\mathcal{N}(\boldsymbol{\epsilon},\sigma_j^2\mathbf{I})\middle|\sigma_j=ce^{(1-j)/d}\right\}.$$

Here c is a constant depends on m, n, d, σ, whose value will be discussed later. For any $P \triangleq \mathcal{N}(\boldsymbol{\epsilon}, \sigma_P^2 \mathbf{I}) \in \mathcal{P}$, by calculating the KL divergence term, we have

$$\mathrm{D}_{\mathrm{KL}}(Q||P) = \frac{1}{2}\left[\frac{d\sigma^2 + \|\mathbf{w}\|_2^2}{\sigma_P^2} - d + d\log\left(\frac{\sigma_P^2}{\sigma^2}\right)\right],$$

which implies that

$$\operatorname*{argmin}_{\sigma_P} \mathrm{D}_{\mathrm{KL}}(Q||P) = \sqrt{\sigma^2 + \|\mathbf{w}\|_2^2/d}.$$

Therefore, we can define the optimal prior distribution as $P^* = \mathcal{N}(\boldsymbol{\epsilon}, \sigma_{j^*}^2 \mathbf{I})$ where

$$j^* = \left\lfloor 1 - d\log\left(\frac{\sigma^2 + \|\mathbf{w}\|_2^2/d}{c}\right)\right\rfloor,$$

which implies that

$$-d\log\left(\frac{\sigma^2 + \|\mathbf{w}\|_2^2/d}{c}\right) \le j^* \le 1 - d\log\left(\frac{\sigma^2 + \|\mathbf{w}\|_2^2/d}{c}\right), \tag{60}$$

and

$$\sigma^2 + \frac{\|\mathbf{w}\|_2^2}{d} \le \sigma_{j^*}^2 \le e^{1/d}\left(\sigma^2 + \frac{\|\mathbf{w}\|_2^2}{d}\right). \tag{61}$$

Here we have used a fact that $\sigma^2 + \|\mathbf{w}\|_2^2/d < c$, which will be shown later. Therefore, we have

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(Q||P^*) &= \frac{1}{2}\left[\frac{d\sigma_Q^2 + \|\mathbf{w}\|_2^2}{\sigma_{j^*}^2} - d + d\log\left(\frac{\sigma_{j^*}^2}{\sigma_Q^2}\right)\right] \\
&\le \frac{1}{2}\left[\frac{d(\sigma_Q^2 + \|\mathbf{w}\|_2^2/d)}{\sigma_Q^2 + \|\mathbf{w}\|_2^2/d} - d + d\log\left(\frac{e^{1/d}\left(\sigma_Q^2 + \|\mathbf{w}\|_2^2/d\right)}{\sigma^2}\right)\right] \\
&= \frac{1}{2}\left[d\log\left(\frac{e^{\frac{1}{d}}\left(\sigma^2 + \|\mathbf{w}\|_2^2/d\right)}{\sigma^2}\right)\right] = \frac{1}{2}\left[1 + d\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{d\sigma^2}\right)\right].
\end{aligned} \tag{62}$$

Denote by A_j the event that

$$A_j \triangleq \left\{\left|\mathbb{E}_{W\sim Q}\left[R_u(W,Z) - R_m(W,Z)\right]\right| \ge \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m} + \frac{1}{u}\right)\left(\mathrm{D}_{\mathrm{KL}}(Q||P_j) + \log\left(\frac{1}{\delta_j}\right)\right)}\right\}.$$

Let $\delta_j \triangleq \frac{6\delta}{\pi^2 j^2}$, by Theorem 8, for any distribution Q we have

$$\mathbb{P}\left\{A_{j^*}\right\} \le \mathbb{P}\left\{\cup_{j=1}^{\infty} A_j\right\} \le \sum_{j=1}^{\infty} \mathbb{P}\left\{A_j\right\} = \sum_{j=1}^{\infty} \delta_j = \sum_{j=1}^{\infty} \frac{6\delta}{\pi^2 j^2} = \delta. \tag{63}$$

Therefore, with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_{W\sim Q}\left[R_u(W,Z)-R_m(W,Z)\right]\\
=&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}\left[R_u(\mathbf{w}+\boldsymbol{\epsilon},Z)-R_m(\mathbf{w}+\boldsymbol{\epsilon},Z)\right]\\
\leq&\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\mathrm{D_{KL}}(Q||P_{j^*})+\log\left(\frac{1}{\delta_j}\right)\right)}\\
\leq&\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\frac{1}{2}\left[1+d\log\left(1+\frac{\|\mathbf{w}\|_2^2}{d\sigma^2}\right)\right]+\log\left(\frac{1}{6\delta}\right)+2\log\left(\pi j^*\right)\right)}.
\end{aligned}\tag{64}$$

Since $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$, Lemma 1 in (Laurent and Massart, 2000) suggests that

$$\mathbb{P}\{\|\boldsymbol{\epsilon}\|_2^2\geq d\sigma^2+2\sigma^2\sqrt{dt}+2t\sigma^2\}\leq e^{-t}.\tag{65}$$

Let $\widetilde{C}_{m,u}\triangleq\sqrt{\log(4mu/(m+u))/d}$, with probability at least $1-\sqrt{(m+u)/4mu}$ we have

$$\begin{aligned}
\|\boldsymbol{\epsilon}\|_2^2\leq&d\sigma^2+2\sigma^2\sqrt{d\log(\sqrt{4mu/(m+u)})}+\sigma^2\log(4mu/(m+u))\\
\leq&d\sigma^2+2\sigma^2\sqrt{d\log(4mu/(m+u))}+\sigma^2\log(4mu/(m+u))\\
=&\sigma^2 d\left(1+\sqrt{\frac{\log(4mu/(m+u))}{d}}\right)^2=\sigma^2 d\left(1+\widetilde{C}_{m,u}\right)^2\triangleq\rho^2,
\end{aligned}\tag{66}$$

Denote by $A=\{\|\boldsymbol{\epsilon}\|_2^2\leq\rho^2\}$ the event that the Euclidean norm of $\boldsymbol{\epsilon}$ is not larger than ρ, with probability at least $1-\delta$:

$$\begin{aligned}
&R_u(\mathbf{w},Z)\\
\leq&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}\left[R_u(\mathbf{w}+\boldsymbol{\epsilon},Z)\right]\\
=&\mathbb{P}\{A\}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}\left[R_u(\mathbf{w}+\boldsymbol{\epsilon},Z)|A\right]+\mathbb{P}\{\bar{A}\}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}\left[R_u(\mathbf{w}+\boldsymbol{\epsilon},Z)|\bar{A}\right]\\
=&\left(1-\sqrt{\frac{m+u}{4mu}}\right)\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}\left[R_u(\mathbf{w}+\boldsymbol{\epsilon},Z)|A\right]+\sqrt{\frac{m+u}{4mu}}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}\left[R_u(\mathbf{w}+\boldsymbol{\epsilon},Z)|\bar{A}\right]\\
\leq&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}\left[R_u(\mathbf{w}+\boldsymbol{\epsilon},Z)|A\right]+\sqrt{\frac{m+u}{4mu}}\\
\leq&\mathbb{E}_{W\sim Q}R_m(W,Z)+\sqrt{\frac{m+u}{4mu}}\\
&+\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\frac{1}{2}\left[1+d\log\left(1+\frac{\|\mathbf{w}\|_2^2}{d\sigma^2}\right)\right]+\log\left(\frac{1}{6\delta}\right)+2\log\left(\pi j^*\right)\right)}\\
\leq&\max_{\|\boldsymbol{\epsilon}\|_2\leq\rho}R_m(\mathbf{w}+\boldsymbol{\epsilon},Z)\\
&+\sqrt{C_{m,u}\left(\frac{1}{m}+\frac{1}{u}\right)\left(1+\frac{d}{2}\log\left(1+\frac{\|\mathbf{w}\|_2^2}{\rho^2}\left(1+\widetilde{C}_{m,u}\right)^2\right)+\log\left(\frac{1}{6\delta}\right)+2\log\left(\pi j^*\right)\right)}.
\end{aligned}\tag{67}$$

Here we use the assumption to obtain the first inequality, and the second line is due to law of total expectation. The second inequality is due to the fact that $R_u(\mathbf{w},Z)\leq 1$ for any

$\mathbf{w}$ and Z since the loss is 0-1 loss. We use the fact $\sqrt{a}+\sqrt{a+b}\leq\sqrt{2(2a+b)}$ in the last inequality. The remaining step is to specify the value of c. Note that if

$$\|\mathbf{w}\|_2^2 \geq \frac{\rho^2}{(1+\widetilde{C}_{m,u})^2}\left(\exp\left\{\frac{2mu}{(m+u)d}\right\}-1\right), \tag{68}$$

the slack term in Eq. (67) will exceed 1 and the inequality holds trivially. Therefore, we only need to consider the case that

$$\|\mathbf{w}\|_2^2 < \frac{\rho^2}{(1+\widetilde{C}_{m,u})^2}\left(\exp\left\{\frac{2mu}{(m+u)d}\right\}-1\right). \tag{69}$$

which implies that

$$\sigma^2+\frac{\|\mathbf{w}\|_2^2}{d} < \frac{\rho^2}{(1+\widetilde{C}_{m,u})^2 d}\exp\left\{\frac{2mu}{(m+u)d}\right\} = \sigma^2\exp\left\{\frac{2mu}{(m+u)d}\right\} \triangleq c. \tag{70}$$

Here we have used Eq. (66) since we only need to consider the case that $\|\boldsymbol{\epsilon}\|_2^2\leq\rho^2$. One can verify that j^* is an valid integer under this definition. Note that

$$\begin{aligned}\log(j^*) &\leq \log\left(1+d\log\left(\frac{c}{\sigma^2+\|\mathbf{w}\|_2^2/d}\right)\right)\\ &\leq \log\left(1+d\log\left(\frac{c}{\sigma^2}\right)\right) = \log\left(1+\frac{2mu}{(m+u)}\right)\\ &\leq \log\left(\frac{4mu}{(m+u)}\right).\end{aligned} \tag{71}$$

Plugging Eq. (71) into Eq. (67), with probability at least $1-\delta$ over the randomness of Z:

$$\begin{aligned}&R_u(\mathbf{w},Z)\\ &\leq \max_{\|\boldsymbol{\epsilon}\|_2\leq\rho} R_m(\mathbf{w}+\boldsymbol{\epsilon},Z)\\ &+\sqrt{\frac{C_{m,u}(m+u)\left(2+d\log\left(1+\frac{\|\mathbf{w}\|_2^2}{\rho^2}\left(1+\widetilde{C}_{m,u}\right)^2\right)+2\log\left(\frac{1}{6\delta}\right)+4\log\left(\frac{4\pi mu}{m+u}\right)\right)}{2mu}}.\end{aligned} \tag{72}$$

Appendix I. Proof of Theorem 10

This proof is inspired by (Neu et al., 2021; Wang and Mao, 2022). By the assumption that $\mathbb{E}\left[R_u(w_T+U_T,Z)-R_u(w_T,Z)\right]\geq 0$ holds for any realization of $W_T=w_T$, we have

$$\begin{aligned}&\mathbb{E}_{Z,W_T,U_T}\left[R_u(\widetilde{W}_T,Z)-R_u(W_T,Z)\right]\\ &=\int_{w_T}\left[\mathbb{E}_{Z,U_T}\left[R_u(w_T+U_T,Z)-R_u(w_T,Z)|W_T=w_T\right]\right]\mathrm{d}P_{W_T}(w_T)\geq 0.\end{aligned} \tag{73}$$

Therefore, the transductive generalization error can be bounded by

$$\begin{aligned}
&\mathbb{E}_{Z,W_T}\left[R_u(W_T,Z)-R_m(W_T,Z)\right] \\
=&\mathbb{E}_{Z,W_T,U_T}\left[R_m(\widetilde{W}_T,Z)-R_m(W_T,Z)\right]-\mathbb{E}_{Z,W_T,U_T}\left[R_u(\widetilde{W}_T,Z)-R_u(W_T,Z)\right] \\
&+\mathbb{E}_{Z,W_T,U_T}\left[R_u(\widetilde{W}_T,Z)-R_m(\widetilde{W}_T,Z)\right] \\
\le&\mathbb{E}_{Z,W_T,U_T}\left[R_m(\widetilde{W}_T,Z)-R_m(W_T,Z)\right]+\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)I(Z;\widetilde{W}_T)}.
\end{aligned} \tag{74}$$

Now the last step is to provide a upper bound for $I(Z;\widetilde{W}_T)$. Following (Wang and Mao, 2022), the mutual information term is decomposed by

$$\begin{aligned}
&I(Z;\widetilde{W}_T) \\
=&I\left(Z;\widetilde{W}_{T-1}-\frac{\eta}{\sqrt{v_T}+\epsilon}\odot g(W_{T-1},B_T(Z))+N_T\right) \\
\le&I\left(Z;\widetilde{W}_{T-1},-\frac{\eta}{\sqrt{v_T}+\epsilon}\odot g(W_{T-1},B_T(Z))+N_T\right) \\
=&I(Z;\widetilde{W}_{T-1})+I\left(-\frac{\eta}{\sqrt{v_T}+\epsilon}\odot g(W_{T-1},B_T(Z))+N_T;Z\middle|\widetilde{W}_{T-1}\right).
\end{aligned}$$

Recursively repeating the above process, we obtain

$$\begin{aligned}
I(Z;\widetilde{W}_T)\le&\sum_{t=1}^{T}I\left(-\frac{\eta}{\sqrt{v_t}+\epsilon}\odot g(W_{t-1},B_t(Z))+N_t;Z\middle|\widetilde{W}_{t-1}\right) \\
=&\sum_{t=1}^{T}I\left(-\frac{\eta}{\sqrt{v_t}+\epsilon}\odot g(\widetilde{W}_{t-1}-U_{t-1},B_t(Z))+N_t;Z\middle|\widetilde{W}_{t-1}\right).
\end{aligned} \tag{75}$$

Then we need to provide a upper bound for the conditional mutual information. Let V,X,U be random variables that are independent of $N\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)$. Define Ψ as a function of random variables U,V,X,Y. Denote by $h(\cdot)$ the differential entropy, then

$$\begin{aligned}
&I(\Psi(V,y-U,X)+\sigma N;X|Y=y) \\
=&h(\Psi(V,y-U,X)+\sigma N|Y=y)-h(\Psi(V,y-U,X)+\sigma N|X,Y=y).
\end{aligned} \tag{76}$$

For the first term, using the fact that Gaussian minimizes entropy, we have

$$\begin{aligned}
&h(\Psi(V,y-U,X)+\sigma N|Y=y) \\
\le&\frac{d}{2}\log\left(2\pi e\frac{\mathbb{E}\left[\|\Psi(V,X,y-U)+\sigma N\|_2^2|Y=y\right]}{d}\right) \\
=&\frac{d}{2}\log\left(2\pi e\frac{\mathbb{E}\left[\|\Psi(U,V,X,y-U)\|_2^2|Y=y\right]+\sigma^2\mathbb{E}\left[\|N\|_2^2\right]}{d}\right) \\
=&\frac{d}{2}\log\left(2\pi e\frac{\mathbb{E}\left[\|\Psi(V,X,y-U)\|^2|Y=y\right]+d\sigma^2}{d}\right).
\end{aligned} \tag{77}$$

For the second term, we have:

$$\begin{aligned}
&h(\Psi(V, y-U, X)+\sigma N|X, Y=y) \geq h(\Psi(V, y-U, X)+\sigma N|U, V, X, Y=y) \\
=&h(\sigma N)=\frac{d}{2}\log 2\pi e\sigma^2.
\end{aligned} \tag{78}$$

Let $V=W^{[t-2]} \triangleq (W_0,\dots,W_{t-2})$, $X=Z$, $Y=\widetilde{W}_{t-1}$, $U=U_{t-1}$ and

$$\begin{aligned}
&\Psi(V, y-U, X)=\Psi(W^{[t-2]}, \widetilde{w}_{t-1}-U_{t-1}, Z) \\
=&-\frac{\eta}{\sqrt{v_t(W^{[t-2]}, \widetilde{w}_{t-1}-U_{t-1})+\epsilon}} \odot g(\widetilde{w}_{t-1}-U_{t-1}, B_t(Z)),
\end{aligned}$$

plugging Eqs. (77,78) into Eq. (76), we have

$$\begin{aligned}
&I\left(\Psi(W^{[t-2]}, \widetilde{w}_{t-1}-U_{t-1}, Z)+N_t; Z\middle|\widetilde{W}_{T-1}=\widetilde{w}_{t-1}\right) \\
=&I\left(\Psi(W^{[t-2]}, \widetilde{w}_{t-1}-U_{t-1}, Z)+N_t; Z\middle|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}\right) \\
\leq&\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\Psi(W^{[t-2]}, \widetilde{w}_{t-1}-U_{t-1}, Z)\right\|_2^2\middle|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}\right]+1\right),
\end{aligned}$$

which implies that

$$\begin{aligned}
&I\left(\Psi(W^{[t-2]}, \widetilde{w}_{t-1}-U_{t-1}, Z)+N_t; Z\middle|\widetilde{W}_{t-1}\right) \\
\leq&\int_{\widetilde{w}_{t-1}}\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\Psi(W^{[t-2]}, \widetilde{w}_{t-1}-U_{t-1}, Z)\right\|_2^2\middle|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}\right]+1\right)\mathrm{d}P_{\widetilde{W}_{T-1}}(\widetilde{w}_{T-1}) \\
\leq&\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\Psi(W^{[t-2]}, \widetilde{W}_{t-1}-U_{t-1}, Z)\right\|_2^2\right]+1\right).
\end{aligned} \tag{79}$$

Let $w^{[k]} \triangleq (w_1,\dots,w_k)$ and

$$\zeta(W^{[t-2]}, \widetilde{W}_{t-1}-U_{t-1}, Z) \triangleq \left\|\Psi(W^{[t-2]}, \widetilde{W}_{t-1}-U_{t-1}, Z)\right\|_2^2,$$

we have

$$\begin{aligned}
&\mathbb{E}_{W^{[t-2]}, \widetilde{W}_{t-1}, U_{t-1}, Z}\left[\zeta(W^{[t-2]}, \widetilde{W}_{t-1}-U_{t-1}, Z)\right] \\
=&\int_z\int_{w^{[t-2]}}\int_{\widetilde{w}_{t-1}}\int_u \zeta(w^{[t-2]}, \widetilde{w}_{t-1}-u, z)\,\mathrm{d}P_{W_{t-1}|Z, W_{t-2}}(\widetilde{w}_{t-1}-u)\,\mathrm{d}P_{U_{t-1}}(u)\,\mathrm{d}P_{W^{[t-2]}|Z}(w^{[t-2]})\,\mathrm{d}P_Z(z) \\
=&\int_z\int_{w^{[t-2]}}\int_{\widetilde{w}_{t-1}}\int_{w_{t-1}} \zeta(w^{[t-2]}, w_{t-1}, z)\,\mathrm{d}P_{W_{t-1}|Z, W_{t-2}}(w_{t-1})\,\mathrm{d}P_{U_{t-1}}(\widetilde{w}_{t-1}-w_{t-1})\,\mathrm{d}P_{W^{[t-2]}|Z}(w^{[t-2]})\,\mathrm{d}P_Z(z) \\
=&\int_z\int_{w^{[t-2]}}\int_{w_{t-1}} \zeta(w^{[t-2]}, w_{t-1}, z)\,\mathrm{d}P_{W_{t-1}|Z, W_{t-2}}(w_{t-1})\,\mathrm{d}P_{W^{[t-2]}|Z}(w^{[t-2]})\,\mathrm{d}P_Z(z) \\
=&\int_z\int_{w^{[t-1]}} \zeta(w^{[t-2]}, w_{t-1}, z)\,\mathrm{d}P_{W^{[t-1]}, Z}(w^{[t-1]}, z) \\
=&\mathbb{E}_{W^{[t-1]}, Z}[\zeta(W^{[t-1]}, Z)].
\end{aligned} \tag{80}$$

Here the first inequality is due to the convolution formulation, and we use $w_{t-1} \triangleq \widetilde{w}_{t-1} - u$ to obtain the second inequality. The third inequality is due to $P_{U_{t-1}}$ is Normal distribution. Plugging Eqs. (79, 80) into Eq. (75), we have

$$\begin{aligned} &I(Z; \widetilde{W}_T) \\ \leq &\sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E} \left[\left\| \Psi(W^{[t-2]}, W_{t-1}, Z) \right\|_2^2 \right] + 1 \right) \\ = &\sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E} \left[\left\| \frac{\eta}{\sqrt{v_t(W^{[t-1]})} + \epsilon} \odot g(W_{t-1}, B_t(Z)) \right\|_2^2 \right] + 1 \right). \end{aligned} \tag{81}$$

Combining Eq. (81) with Eq. (74), we have

$$\begin{aligned} &\mathbb{E}_{Z, W_T} \left[R_u(W_T, Z) - R_m(W_T, Z) \right] \\ \leq &\sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E} \left[\left\| \frac{\eta}{\sqrt{v_t(W^{[t-1]})} + \epsilon} \odot g(W_{t-1}, B_t(Z)) \right\|_2^2 \right] + 1 \right) \\ &+ \mathbb{E}_{Z, W_T, U_T} \left[R_m(W_T + U_T, Z) - R_m(W_T, Z) \right] \end{aligned} \tag{82}$$

Now we discuss how to extend this result to Adam optimization algorithm. For $t \in [T]$, the update rule of Adam is

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g(W_{t-1}, B_t(Z)), \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g(W_{t-1}, Z) \odot g(W_{t-1}, B_t(Z)), \\ \hat{v}_t &= \frac{v_t}{1 - \beta_2^t}, \hat{m}_t = \frac{m_t}{1 - \beta_1^t}, W_t = W_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \odot \hat{m}_t. \end{aligned}$$

Define

$$\Psi(W^{[t-1]}, Z) \triangleq - \sum_{\tau=0}^{t-1} \frac{\eta(1 - \beta_1)\beta_1^{t-\tau-1}}{\sqrt{\hat{v}_t} + \epsilon} \odot g(W_\tau, B_{\tau+1}(Z)),$$

we have $W_t = W_{t-1} + \Psi(W^{[t-1]}, Z)$. Similarly, we construct the weight process as

$$\widetilde{W}_0 = W_0, \widetilde{W}_t = \widetilde{W}_{t-1} + \Psi(W^{[t-1]}, Z) + N_t.$$

By the same technique, one can find that

$$\begin{aligned} &I(Z; \widetilde{W}_T) \\ \leq &\sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E} \left[\left\| \Psi(W^{[t-2]}, W_{t-1}, Z) \right\|_2^2 \right] + 1 \right) \\ = &\sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E} \left[\left\| \sum_{\tau=0}^{t-1} \frac{\eta(1 - \beta_1)\beta_1^{t-\tau-1}}{\sqrt{\hat{v}_t} + \epsilon} \odot g(W_\tau, B_{\tau+1}(Z)) \right\|_2^2 \right] + 1 \right). \end{aligned}$$

Then the upper bound is given by

$$\begin{aligned} &\mathbb{E}_{Z, W_T} \left[R_u(W_T, Z) - R_m(W_T, Z) \right] \\ \leq &\sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E} \left[\left\| \sum_{\tau=0}^{t-1} \frac{\eta(1 - \beta_1) beta_1^{t-\tau-1}}{\sqrt{\hat{v}_t} + \epsilon} \odot g(W_\tau, B_{\tau+1}(Z)) \right\|_2^2 \right] + 1 \right) \\ &+ \mathbb{E}_{Z, W_T, U_T} \left[R_m(W_T + U_T, Z) - R_m(W_T, Z) \right]. \end{aligned}$$

Appendix J. Proof of Proposition 11

Following the proof in Section F, we have

$$\left|\mathbb{E}_{F_i,U_i|\widetilde{Z}=\widetilde{z},S_{m+u}=s_{m+u}}\left[g(F_i,U_i,\widetilde{Y}_i)\right]\right| \leq B\sqrt{2I^{\widetilde{z},s_{m+u}}(F_i;U_i)}. \tag{83}$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}_{S_m,S_u^X,W}\left[R(W,S_u)-R(W,S_m)\right]\right| \\
=&\left|\mathbb{E}_{S_m,S_u^X,W}\left[\frac{1}{u}\sum_{i=m+1}^{m+u}\ell(W,(X_i,Y_i))-\frac{1}{m}\sum_{i=1}^{m}\ell(W,(X_i,Y_i))\right]\right| \\
=&\left|\mathbb{E}_{S_{m+u}}\mathbb{E}_{Z,W|S_{m+u}}\left[R_u(W,Z)-R_m(W,Z)\right]\right| \\
=&\left|\mathbb{E}_{S_{m+u}}\mathbb{E}_{\widetilde{Z},U,W|S_{m+u}}\left[R_u(W,Z)-R_m(W,Z)\right]\right| \\
\leq&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u}}\mathbb{E}_{\widetilde{Z}|S_{m+u}}\left|\mathbb{E}_{W,U|\widetilde{Z},S_{m+u}}\left[\ell(W,\widetilde{Z}_{i,U_i})-\ell(W,\widetilde{Z}_{i,1-U_i})\right]\right| \\
=&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u}}\mathbb{E}_{\widetilde{Z}|S_{m+u}}\left|\mathbb{E}_{F_i,U_i|\widetilde{Z},S_{m+u}}\left[g(F_i,U_i,\widetilde{Y}_i)\right]\right| \\
\leq&\frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u},\widetilde{Z}}\sqrt{2I^{S_{m+u},\widetilde{Z}}(F_i;U_i)}.
\end{aligned} \tag{84}$$

Similarly we have

$$\begin{aligned}
\left|\mathbb{E}_{S_m,S_u^X,W}\left[R(W,S_u)-R(W,S_m)\right]\right| &\leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u},\widetilde{Z}}\sqrt{2I^{S_{m+u},\widetilde{Z}}(L_i;U_i)}, \\
\left|\mathbb{E}_{S_m,S_u^X,W}\left[R(W,S_u)-R(W,S_m)\right]\right| &\leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{m+u},\widetilde{Z}}\sqrt{2I^{S_{m+u},\widetilde{Z}}(\Delta_i;U_i)}.
\end{aligned} \tag{85}$$

Appendix K. Experiment Details

Estimating the Expected Transductive Generalization Gaps and the Derived Bounds. Notice that compute the accurate value of the expected transductive generalization gap (and also the derived upper bounds) is not applicable since we need to run the algorithm on $(m+u)!$ partitions in total. Therefore we use Monte Carlo simulation to estimate these expectations based on finite samples.

For semi-supervised learning, the sampling process are as follows: (i) randomly draw t_1 full samples set s_{m+u} by each time sampling $m+u$ images from the raw images set, (ii) randomly draw t_2 transductive supersamples $\widetilde{z}$ based on Definition 7, (iii) randomly draw t_3 train/test split variables s and obtain the training and test samples set according to Subsection 4.2. Notice that here we do not consider the randomness of U for the case $k \geq 2$. The reason is that U control the permutation of test samples, and the learning algorithm we consider is independent of this permutation. Now we discuss the estimation

Layer Type	Parameter
Conv	16 filters, 3×3 kernels, stride 1, padding 1, BatchNormalization, ReLU
Conv	32 filters, 3×3 kernels, stride 1, padding 1, BatchNormalization, ReLU
Conv	32 filters, 3×3 kernels, stride 1, padding 1
ConvShortcut	32 filters, 1×1 kernels, stride 1, BatchNormalization, ReLU
Conv$\times$6	32 filters, 3×3 kernels, stride 1, padding 1, BatchNormalization, ReLU
Conv	64 filters, 3×3 kernels, stride 1, padding 1, BatchNormalization, ReLU
Conv	64 filters, 3×3 kernels, stride 1, padding 1
ConvShortcut	64 filters, 1×1 kernels, stride 1, BatchNormalization, ReLU
Conv$\times$6	64 filters, 3×3 kernels, stride 1, padding 1, BatchNormalization, ReLU
Conv	128 filters, 3×3 kernels, stride 1, padding 1, BatchNormalization, ReLU
Conv	128 filters, 3×3 kernels, stride 1, padding 1
ConvShortcut	128 filters, 1×1 kernels, stride 1, BatchNormalization, ReLU
Conv$\times$6	128 filters, 3×3 kernels, stride 1, padding 1, BatchNormalization, ReLU
FC	10 units, linear activation

Table 1: The architecture of the convolutional neural network used in CIFAR-10 classification tasks. $\times 6$ means repeating the layer for 6 times.

of transductive generalization gap and the upper bounds established in Corollary 6. Taking Eq. (17) as an example, for each $(s_{m+u}, \widetilde{z})$ we use the mean value over t_3 samples of S to estimate the conditional term expectation term $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{F_i,U_i|\widetilde{z},s_{m+u}}g(F_i,S_i,\widetilde{y}_i)$. After that we use t_1t_2 samples of S_{m+u} and $\widetilde{Z}$ to estimate the expected generalization gap, whose mean and standard deviation are shown in Figure. Similarly, we use a plug-in estimator (Paninski, 2003) to estimate the disentangled mutual information $I^{s_{m+u},\widetilde{z}}(F_i;S_i)$ over the t_3 samples of S. Then the upper bounds in Proposition. (11) are estimated by the t_1t_2 samples of S_{m+u} and $\widetilde{Z}$, whose mean and standard deviation are shown in Figure.

For transductive graph learning, the estimation process is generally follows that of semi-supervised learning, expect that we do not need to consider the sample of S_{m+u}. Concretely, the sampling process is only composed of (ii) and (iii). Accordingly, we use t_2 samples of $\widetilde{Z}$ to estimate the expected bounds and the conditional mutual information.

Network Architecture and Hyperparameter Setting. For semi-supervised learning, the network architecture on MNIST and CIFAR-10 are presented in Table 1 of (Harutyunyan et al., 2021) and Table 1 respectively. On both these two datasets, we set $t_1 = t_2 = 2$ and $t_3 = 50$. For transductive graph learning, the architecture of GAT and GPR-GNN follows the settings in (Chien et al., 2021), and we set $t_2 = 5$ and $t_3 = 50$ for evaluation. We adopt the released code [1] to generate the cSBMs datasets with $n \in \{500, 1000, 2000\}$ and $\phi \in \{-0.5, 0.5\}$. The number of training nodes are defined by $m := \frac{n}{k}$ for real-world graph datasets, where n is the total number of nodes. To ensure that n can be evenly divided by k, we set $k = 2$ for Cora and Actor, and $k = 3$ for CiteSeer and Chameleon.

1. https://github.com/jianhao2016/GPRGNN

References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

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

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

Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. An information-theoretical approach to semi-supervised learning under covariate-shift. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 7433–7449, 2022.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, pages 254–263, 2018.

Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, page 7245–7254, 2018.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1046–1059, 2016.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-bayesian theory for transductive learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 105–113, 2014.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

Pietro Bongini, Monica Bianchini, and Franco Scarselli. Molecular generative graph neural networks for drug discovery. *Neurocomputing*, 450:242–252, 2021.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020.

Yuheng Bu, Gholamali Aminian, Laura Toni, Gregory W. Wornell, and Miguel Rodrigues. Characterizing and understanding the generalization error of transfer learning with gibbs algorithm. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 8673–8699, 2022.

Olivier Catoni. Pac-bayesian supervised classification: The thermodynamics of statistical learning. *Institute of Mathematical Statistics Lecture Notes Monograph Series*, 56:1–163, 2007.

Huiyuan Chen, Chin-Chia Michael Yeh, Yujie Fan, Yan Zheng, Junpeng Wang, Vivian Lai, Mahashweta Das, and Hao Yang. Sharpness-aware graph collaborative filtering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2369–2373, 2023.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1725–1735, 2020.

Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. In *Advances in Neural Information Processing Systems*, 2021.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021.

Eugenio Clerico, Amitis Shidani, George Deligiannidis, and Arnaud Doucet. Chained generalisation bounds. In *Proceedings of 35th Conference on Learning Theory*, pages 4212–4257, 2022.

Weilin Cong, Morteza Ramezani, and Mehrdad Mahdavi. On provable benefits of depth in training graph convolutional networks. In *Advances in Neural Information Processing Systems*, 2021.

Corinna Cortes and Mehryar Mohri. On transductive regression. In *Advances in Neural Information Processing Systems*, pages 305–312, 2006.

Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability of transductive regression algorithms. In *Proceedings of the 25th International Conference on Machine Learning*, page 176–183, 2008.

Jaydeep De, Xiaowei Zhang, Feng Lin, and Li Cheng. Transduction on directed graphs via absorbing random walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1770–1784, 2018.

Leyan Deng, Defu Lian, Chenwang Wu, and Enhong Chen. Graph convolution network based recommender systems: Learning guarantee and item mixture powered strategy. In *Advances in Neural Information Processing Systems*, 2022.

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22:117–142, 2004.

Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. In *Advances in Neural Information Processing Systems*, pages 8590–8602, 2018.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence*, 2017.

Ran El-Yaniv and Dmitry Pechyony. Stable transductive learning. In *Conference on Learning Theory*, pages 35–49, 2006.

Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. In *Conference on Learning Theory*, pages 157–171, 2007.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004, 2021.

Pascal Mattia Esser, Leena C. Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. In *Advances in Neural Information Processing Systems*, pages 27043–27056, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Borja Rodríguez Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. In *2020 IEEE Information Theory Workshop*, page 1–5, 2021.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272, 2017.

Pere Giménez-Febrer, Alba Pagès-Zamora, and Georgios B. Giannakis. Generalization error bounds for kernel matrix completion and extrapolation. *IEEE Signal Processing Letters*, 27:326–330, 2020.

Chen Gong, Xiaojun Chang, Meng Fang, and Jian Yang. Teaching semi-supervised classifier via generalized distillation. In Jérôme Lang, editor, *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2156–2162, 2018.

M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, pages 729–734, 2005.

Lan-Zhe Guo, Zhenyu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 3897–3906, 2020.

Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. In *Advances in Neural Information Processing Systems*, pages 16457–16467, 2020a.

Hassan Hafez-Kolahi, Shohreh Kasaei, and Mahdiyeh Soleymani-Baghshah. Sample complexity of classification with compressed input. *Neurocomputing*, 415:286–294, 2020b.

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. In *Advances in Neural Information Processing Systems*, 2020.

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Daniel M. Roy. Towards a unified information-theoretic framework for generalization. In *Advances in Neural Information Processing Systems*, pages 26370–26381, 2021.

Mahdi Haghifam, Shay Moran, Daniel M. Roy, and Gintare Karolina Dziugaite. Understanding generalization via leave-one-out conditional mutual information. *arXiv preprint arXiv:2206.14800*, 2022.

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, pages 24670–24682, 2021.

Haiyun He, Hanshu Yan, and Vincent Y. F. Tan. Information-theoretic characterization of the generalization error for iterative semi-supervised learning. *Journal of Machine Learning Research*, 23(287):1–52, 2022.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 639–648, 2020.

Fredrik Hellström and Giuseppe Durisi. A new family of generalization bounds using samplewise evaluated CMI. In *Advances in Neural Information Processing Systems*, 2022.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and pac-bayes. *arXiv preprint arXiv:2309.04381*, 2023.

Tinglin Huang, Yuxiao Dong, Ming Ding, Zhen Yang, Wenzheng Feng, Xinyu Wang, and Jie Tang. Mixgcf: An improved training method for graph neural network-based recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, page 665–674, 2021.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, 1999.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1), 2021a.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic bounds on transfer generalization gap based on jensen-shannon divergence. In *2021 29th European Signal Processing Conference*, pages 1461–1465, 2021b.

Sharu Theresa Jose, Osvaldo Simeone, and Giuseppe Durisi. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization. *IEEE Transactions on Information Theory*, 68(1):474–501, 2022.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *Proceedings of the 40th International Conference on Machine Learning*, pages 16049–16096, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, page 443–457, 2000.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, page 1097–1105, 2012.

Shiyong Lan, Yitong Ma, Weikang Huang, Wenwu Wang, Hongyu Yang, and Pyang Li. DSTAGNN: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, pages 11906–11917, 2022.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 4189–4196, 2021.

Renjie Liao, Raquel Urtasun, and Richard Zemel. A PAC-bayesian approach to generalization bounds for graph neural networks. In *International Conference on Learning Representations*, 2021.

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, page 2935–2944, 2017.

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

Sanae Lotfi, Marc Anton Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-bayes compression bounds so tight that they can explain generalization. In *Advances in Neural Information Processing Systems*, 2022.

Xuanyuan Luo, Bei Luo, and Jian Li. Generalization bounds for gradient methods via discrete and continuous prior. In *Advances in Neural Information Processing Systems*, 2022.

Mohammad Saeed Masiha, Amin Gohari, Mohammad Hossein Yassaee, and Mohammad Reza Aref. Learning under distribution mismatch and model misspecification. In *2021 IEEE International Symposium on Information Theory*, page 2912–2917, 2021.

Andreas Maurer. A note on the pac bayesian theorem. *arXiv preprint arXiv:cs/0411099*, 2004.

Yury Maximov, Massih-Reza Amini, and Zaid Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm (extended abstract). In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5637–5641, 2018.

Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. Pac-bayesian generalization bounds for adversarial generative models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24271–24290, 2023.

David A. McAllester. Some pac-bayesian theorems. *Maching Learning*, 37(3):355–363, 1999.

Alexander Mey and Marco Loog. Improved generalization in semi-supervised learning: A survey of theoretical results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4747–4767, 2023.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Proceedings of the 31st Conference On Learning Theory*, pages 605–638, 2018.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pages 11013–11023, 2019.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, volume 134, pages 3526–3545, 2021.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Kenta Oono and Taiji Suzuki. Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6): 1191–1253, 2003.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory*, pages 546–550, 2018.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227): 1–40, 2021.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022.

Mohamad Rida Rammal, Alessandro Achille, Aditya Golatkar, Suhas Diggavi, and Stefano Soatto. On leave-one-out conditional mutual information for generalization. In *Advances in Neural Information Processing Systems*, pages 10179–10190, 2022.

Arezou Rezazadeh, Sharu Theresa Jose, Giuseppe Durisi, and Osvaldo Simeone. Conditional mutual information-based generalization bound for meta learning. In *2021 IEEE International Symposium on Information Theory*, pages 1176–1181, 2021.

Omar Rivasplata, Emilio Parrado-Hernández, John Shawe-Taylor, Shiliang Sun, and Csaba Szepesvári. Pac-bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems*, page 9234–9244, 2018.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 1232–1240, 2016.

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, 2020.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20 (1):61–80, 2009.

Matthias Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

Milad Sefidgaran, Amin Gohari, Gaël Richard, and Umut Simsekli. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In *Proceedings of 35th Conference on Learning Theory*, pages 4416–4463, 2022.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

B.M. Shahshahani and D.A. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, 1994.

Shalev-Shwartz Shai and Ben-David Shai. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 46–54, 2016.

Ohad Shamir and Shai Shalev-Shwartz. Matrix completion with the trace norm: Learning, bounding, and transducing. *Journal of Machine Learning Research*, 15(98):3401–3423, 2014.

John Shawe-Taylor and Robert C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, page 2–9, 1997.

Rakesh Shivanna and Chiranjib Bhattacharyya. Learning on graphs using orthonormal representation is statistically consistent. In *Advances in Neural Information Processing Systems*, pages 3635–3643, 2014.

Rakesh Shivanna, Bibaswan K. Chatterjee, Raman Sankaran, Chiranjib Bhattacharyya, and Francis R. Bach. Spectral norm regularization of orthonormal representations for graph transduction. In *Advances in Neural Information Processing Systems*, pages 2215–2223, 2015.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 914–921, 2020.

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

Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento, Jiayu Zhou, and Fei Wang. Graph convolutional networks for computational drug development and discovery. *Briefings in bioinformatics*, 21(3):919—935, 2020.

Huayi Tang and Yong Liu. Towards understanding generalization of graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, pages 33674–33719, 2023.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop*, pages 1–5, 2015.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint arXiv:physics/0004057*, 2000.

Ilya Tolstikhin, Gilles Blanchard, and Marius Kloft. Localized complexities for transductive learning. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 857–884, 2014.

Ilya O. Tolstikhin, Nikita Zhivotovskiy, and Gilles Blanchard. Permutational rademacher complexity - A new complexity measure for transductive learning. In *Proceedings of the 26th International Conference on Algorithmic Learning Theory*, pages 209–223, 2015.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data: Empirical Inference Science*. Springer, New York, 1982.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

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*, pages 577–581, 2019a.

Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of SGLD using properties of gaussian channels. In *Advances in Neural Information Processing Systems*, 2021.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 165–174, 2019b.

Zifeng Wang, Shao-Lun Huang, Ercan Engin Kuruoglu, Jimeng Sun, Xi Chen, and Yefeng Zheng. PAC-bayes information bottleneck. In *International Conference on Learning Representations*, 2022.

Ziqiao Wang and Yongyi Mao. On the generalization of models trained with SGD: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2022.

Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. In *Proceedings of the 40th International Conference on Machine Learning*, pages 36111–36137, 2023a.

Ziqiao Wang and Yongyi Mao. Information-theoretic analysis of unsupervised domain adaptation. In *International Conference on Learning Representations*, 2023b.

Xuetong Wu, Jonathan H. Manton, Uwe Aickelin, and Jingge Zhu. Information-theoretic analysis for transfer learning. In *2020 IEEE International Symposium on Information Theory*, pages 2819–2824, 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.

Chao Xu, Hong Tao, Jing Zhang, Dewen Hu, and Chenping Hou. Label distribution changing learning with sample space expanding. *Journal of Machine Learning Research*, 24 (36):1–48, 2023.

Da Xu, Chuanwei Ruan, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. Rethinking neural vs. matrix-factorization collaborative filtering: the theoretical perspectives. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11514–11524, 2021.

Jun Yang, Shengyang Sun, and Daniel M. Roy. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. In *Advances in Neural Information Processing Systems*, 2019.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 40–48, 2016.

Yige Yuan, Bingbing Xu, Huawei Shen, Qi Cao, Keting Cen, Wen Zheng, and Xueqi Cheng. Towards generalizable graph contrastive learning: An information theory perspective. *arXiv preprint arXiv:2211.10929*, 2022.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016*, 2016.

Ruida Zhou, Chao Tian, and Tie Liu. Stochastic chaining and strengthened information-theoretic generalization bounds. In *2022 IEEE International Symposium on Information Theory*, pages 690–695, 2022.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach. In *International Conference on Learning Representations*, 2019.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, 2003.