Journal of Machine Learning Research 25 (2024) 1-69 Submitted 10/23; Revised 11/24; Published 12/24

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, 100872, China

Yong Liu* LIUYONGGSAI@RUC.EDU.CN
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, 100872, China

Editor: Pierre Alquier

Abstract

In this paper, we establish generalization bounds for transductive learning algorithms in the context of information theory and PAC-Bayes, covering both the random sampling and the random splitting setting. First, we show that the transductive generalization gap can be controlled by the mutual information between training label selection and the hypothesis. Next, we propose the concept of transductive supersample and use it to derive transductive information-theoretic bounds involving conditional mutual information and different information measures. We further establish transductive PAC-Bayesian bounds with weaker assumptions on the type of loss function and the number of training and test data points. Lastly, we use the theoretical results to derive upper bounds for adaptive optimization algorithms under the transductive learning setting. We also apply them to semi-supervised learning and transductive graph learning scenarios, meanwhile validating the derived bounds by experiments on synthetic and real-world datasets.

Keywords: transductive learning, generalization bounds, information theory, PAC-Bayes

1. Introduction

Supervised learning is one of the most common paradigms in the field of machine learning (Shalev-Shwartz and Ben-David, 2014; Mohri et al., 2018), where a limited number of data points coming from an unknown distribution are provided and each one is composed of some features and a unique associated label. Our task is to build a model with these data points and use it to predict the labels of new coming data points based solely on their features. To this end, we first need to specify the learning algorithm and the model class. Next, we feed the collected data to the learning algorithm, which selects a model from the model class and returns it to us. Before applying this model to real-world scenarios, we need to evaluate its capability from diverse perspectives. Generalization ability, the prediction performance of the model on unseen data, is one of the most important capabilities that we focus on. Over the past decades, researchers have been developing theories to analyze and explain the

*. Yong Liu is the corresponding author.

generalization ability of machine learning algorithms. Early findings in this field suggest that generalization can be connected to the complexity of hypothesis space (Koltchinskii and Panchenko, 2000; Koltchinskii, 2001; Bartlett and Mendelson, 2002; Bartlett et al., 2005) and the stability of learning algorithms (Rogers and Wagner, 1978; Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002; Shalev-Shwartz et al., 2010). Most recently, information theory is shown to be a promising theoretical framework to analyze the generalization ability of learning algorithms (Zhang, 2006; Russo and Zou, 2016, 2020; Xu and Raginsky, 2017; Negrea et al., 2019; Haghifam et al., 2020; Steinke and Zakynthinou, 2020; Haghifam et al., 2021; Sefidgaran et al., 2022; Wang and Mao, 2023a). Theoretical results derived under this framework convey a key insight. Specifically, a hypothesis that reveals less information about the training data tends to have better generalization ability. Moreover, Banerjee and Montúfar (2021); Grünwald et al. (2021) have revealed that studying generalization from the perspective of information theory is closely related to PAC-Bayes, an earlier research route that uses the divergence between two probability measures on the hypothesis space to depict the generalization ability of learning algorithms (Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999; Catoni, 2007). Notably, results derived from both information theory and PAC-Bayes are data-dependent and algorithm-dependent, thereby reflecting the impact of training data and learning algorithms on generalization.

So far, the theoretical analysis of the generalization ability of supervised learning algorithms is well-developed and productive. However, the supervised learning paradigm is not sufficient to cover all real-world application scenarios. First, real-world data may not be identically distributed. For example, a global server will exchange data with diverse clients in computer networks or distributed computing systems. Data received from different clients may not follow the same distribution. Second, real-world data may lack labels due to the expensive annotation cost or privacy protection requirements. Third, real-world data itself could be of low quality. Due to environmental interference or equipment failures, some data points could have noisy labels or incomplete features. These issues give rise to new learning paradigms and learning algorithms, prompting researchers to further develop new generalization theories for them. In this paper, we focus on the transductive learning paradigm (Vapnik, 1998, Chapter 8), where both labeled and unlabeled data points are provided. Our task is to build a model based on them and use it to make predictions for these unlabeled ones. Notably, in the transductive learning setting, the features of the unlabeled data points intended for prediction are accessible to the model. Further, two settings for transductive learning are proposed in Chapter 8 of Vapnik (1998) and they are respectively referred to as Setting 1 and Setting 2. In Setting 1, we sample partial data points from the collection of full data points without replacement and reveal their labels to the model. The learning goal is to minimize the risk of the model on the rest data points. Since we only concern the predictions of those unlabeled data points, Vapnik (1982, 1998) terms this setting *the problem of estimating the values of a function at given points*, and we term it *the random splitting setting* in this paper. In Setting 2, however, a sequence of data points are drawn from an unknown distribution and we could only observe the labels of the data points at the front of this sequence. Notice that the data points in this sequence could come from different distributions, or there exists dependency between theirs. The learning goal is to choose a model so as to minimize the expected risk on the remaining data points in this sequence whose labels are not visible. Particularly, if the data points in this sequence are

independent and identically distributed, the expected risk on the rest data points is exactly the expected risk in supervised learning, which depicts how well the hypothesis generalizes on unseen data points. Consequently, Vapnik (1982, 1998) terms this setting *the problem of estimating a function*, and we term it *the random sampling setting* in this paper. In contrast, within the supervised learning paradigm, the features of test data points are not accessible to the model during training. As a result, this scenario, where models must make predictions on previously unseen data points, is referred to by Vapnik (1982, 1998) as the inductive learning setting or inductive inference.

To understand these two kinds of transductive learning settings more intuitively, let us consider a scenario where we have a set of images and each of them is annotated by experts. We then randomly select some images without replacement from this set and reveal their labels together with all images to the model, and require it to make predictions for those images whose labels are not revealed. This process exemplifies the random splitting setting. Now suppose that we acquire new unlabeled images after a few days. By feeding both the previously collected labeled images and the new unlabeled images into the model, we ask it to predict the labels for the newly acquired images. This process exemplifies the random sampling setting. Compared with the random splitting setting, the random sampling setting is closer to real-world scenarios. However, we emphasize that the random splitting setting has its own unique value. Specifically, all the randomness in the random splitting setting is due to the partition of the full sample into training and test data. Consequently, we do not need to know the underlying distribution of the data or make assumptions about it, even in the presence of independence among the data points. For example, a fundamental task in graph learning is node classification, which involves a single graph composed of nodes and edges. Each node has associated features and a label, and our goal is to construct a model that can predict the labels of nodes in this graph. In this task, we assume that the graph structure is static. That is, one that does not add new nodes over time. Then, we can apply the random splitting setting by randomly selecting some nodes without replacement from the entire set of nodes and providing their labels along with all node features to the model. Indeed, this approach is widely used to train graph neural networks (GNNs) for node classification (Gilmer et al., 2017; Kipf and Welling, 2017; Veličković et al., 2018).

Existing results of the generalization upper bounds for transductive learning algorithms include complexity-based bounds derived from VC-dimension (Cortes and Mohri, 2006), transductive Rademacher complexity (El-Yaniv and Pechyony, 2009; Yang, 2023), permutational Rademacher complexity (Tolstikhin et al., 2015) and transductive local Rademacher complexity (Yang, 2023), stability-based bounds for transductive classification (El-Yaniv and Pechyony, 2006) and transductive regression (Cortes et al., 2008), and PAC-Bayesian bounds (Audibert and Bousquet, 2007; Derbeko et al., 2004; Bégin et al., 2014; Catoni, 2007, Chapter 3). However, these findings still face certain limitations. Under the random splitting setting, complexity-based bounds (Cortes and Mohri, 2006; El-Yaniv and Pechyony, 2009; Tolstikhin et al., 2015) and stability-based bounds (Cortes et al., 2008) are independent of the learning algorithm and training data selection, respectively. Consequently, these results fail to simultaneously reflect the influence of both the learning algorithm and the selection of training data on generalization performance. Moreover, applying these results to deep transductive models like GNNs presents additional challenges. Specifically, bounds derived from VC-dimension could become trivial (Esser et al., 2021), while stability based

bounds (Cong et al., 2021) involve Lipschitz or smoothness constants of the loss function that are hard to estimate or even be bounded (Neu et al., 2021). Additionally, for PAC-Bayesian bounds, results established by Derbeko et al. (2004) are of slow order, and results of Bégin et al. (2014) require strong assumptions about the loss function and the number of training and test data points. Furthermore, under the random sampling setting, existing results require that the number of test data points is a multiple of the number of training data points (Catoni, 2007, Chapter 3) or equals the number of training data points (Catoni, 2003; Audibert and Bousquet, 2007).

In this paper, we delve into the generalization ability analysis of transductive learning algorithms within the framework of information theory and PAC-Bayesian theory. First, under the random splitting setting, we establish average and single-draw bounds for the transductive generalization gap using tools from information theory. These results contain the mutual information between the hypothesis and the training label selection, which indicates that transductive learning algorithms generalize better when their output hypothesis is less dependent on the specific choice of training label. We then introduce the concept of transductive supersample and extend the framework of Steinke and Zakynthinou (2020) to the transductive learning setting, which allows us to transport results based on various information measures derived under the inductive learning setting to the transductive learning setting. Second, we establish PAC-Bayesian bounds for the transductive generalization gap under both the random splitting setting and the random sampling setting. These results allow the number of training and test data points to be arbitrary integers and only require the loss function to be bounded. Moreover, we demonstrate that a previous result, which states that a flatter loss landscape implies better generalization performance under the inductive learning setting, remains valid under the transductive learning setting. This result provides theoretical support to recent empirical observation of Chen et al. (2023). Third, we apply the above results to establish generalization upper bounds for adaptive optimization algorithms under the transductive learning setting. We also demonstrate the applications of our theoretical results on semi-supervised learning and transductive graph learning scenarios and validate our results with numerical experiments on both synthetic and real-world datasets. The key contributions of this work can be summarized as follows.

- We establish information-theoretic bounds for the transductive generalization gap and the largest eigenvalue of its Hessian under the random splitting setting.
- We propose the concept of transductive supersamples for the first time and derive new information-theoretic and PAC-Bayesian bounds under the random splitting setting.
- We establish PAC-Bayesian bounds with weaker assumptions on the number of data points and the type of loss function under the random sampling setting.

In the remainder of this paper, we begin with an overview of the literature related to our work in Section 2. We then introduce the mathematical notations used throughout this paper along with the random splitting setting and the sampling setting of transductive learning in Section 3. The main theoretical results are presented in Section 4, followed by their applications in Section 5. The experimental setting and results are detailed in Section 6. Finally, we conclude the paper in Section 7. Complete proofs of all theoretical results in the main text are provided in the appendix.

2. Related Work

2.1 Information-theoretic Generalization Theory

Russo and Zou (2016, 2020) and Xu and Raginsky (2017) link the expected generalization error with the mutual information between training examples and algorithm output. Subsequent studies can be categorized into four main groups: (i) deriving tighter upper bounds by introducing novel information measures (Harutyunyan et al., 2021; Hellström and Durisi, 2022; Wang and Mao, 2023a), problem settings (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; Rodríguez-Gálvez et al., 2021; Zhou et al., 2022; Clerico et al., 2022); (ii) establishing bounds characterized 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 derive upper bounds for optimization algorithms like stochastic gradient descent (SGD) (Neu et al., 2021; Wang and Mao, 2022) or stochastic gradient langevin dynamics (SGLD) (Pensia et al., 2018; Negrea et al., 2019; Wang et al., 2021); and (iv) extending above theoretical results to different 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., 2024), and domain adaption (Wang and Mao, 2023b). However, these studies fail to account for the transductive learning setting. Another related area is the information bottleneck theory (Tishby et al., 2000) and its applications in explaining 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 is independent of our works. For a thorough review of information-theoretic generalization theory, we refer readers to the recent monograph of Hellström et al. (2023).

2.2 PAC-Bayesian Generalization Theory

The classical results in PAC-Bayesian generalization theory include McAllester's bound (McAllester, 1999), Maurer-Langford-Seeger's bound (Langford and Seeger, 2001; Seeger, 2002; Maurer, 2004) and Catoni's bound (Catoni, 2007, Chapter 1). Building on these foundational works, a substantial body of research has emerged that applies or extends these results to analyze various models or algorithms, including computing non-vacuous bounds for deep neural networks (Dziugaite and Roy, 2017; Zhou et al., 2019; Pérez-Ortiz et al., 2021; Dziugaite et al., 2021; Lotfi et al., 2022) and establishing upper bounds for optimization algorithms (London, 2017; 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 network architectures (Neyshabur et al., 2018; Liao et al., 2021; Mbacke et al., 2023). For more details, see monographs of Guedj (2019); Alquier (2024) along with their references. However, the above studies do not yet encompass the random splitting setting of transductive learning.

2.3 Generalization Theory of Transductive Learning

The concept of transductive learning along with the earliest generalization bounds are introduced by Vapnik (1982). Under the random splitting setting, generalization bounds

are established by transductive algorithm stability (El-Yaniv and Pechyony, 2006) and transductive Rademacher complexity (El-Yaniv and Pechyony, 2009), respectively. Later, Tolstikhin et al. (2015) introduce permutational Rademacher complexity and demonstrate that it is more suitable for the transductive learning setting compared to transductive Rademacher complexity. Their difference is that the expectation is taken over the transductive Rademacher variables in transductive Rademacher Complexity, while in permutational Rademacher complexity, the expectation is taken over the selection of training labels. By incorporating the variance of functions, Tolstikhin et al. (2014) develop new concentration inequalities and derive tighter bounds. A novel complexity metric, termed transductive local Rademacher complexity, is introduced by Yang (2023) recently. The expectation is taken over both standard Rademacher variables and the selection of training labels in transductive local Rademacher complexity, which allows us to obtain tighter bounds in some scenarios. In contrast, we establish upper bounds based on information theory. Transductive PAC-Bayesian bounds under the random splitting setting are initially developed by Derbeko et al. (2004), which is subsequently improved by Bégin et al. (2014). We further improve their results and apply them to reveal the impact of loss landscape flatness on generalization. Under the random sampling setting, Catoni firstly establishes the generalization bound in the context of PAC-Bayesian theory (Catoni, 2007, Chapter 3), which is further improved by Audibert and Bousquet (2007) via the generic chaining technique. Different from these studies, we establish transductive PAC-Bayesian bounds through distinct techniques, and the results have weaker assumptions on the number of data points and loss function type. Besides, the above theoretical results have been applied to areas such as 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). We select semi-supervised learning and transductive graph learning as examples to demonstrate our theoretical results. Extensions to other domains are left 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 a given random variable X, we denote its distribution measure by P_X. The conditional distribution measure of X given Y is denoted by $P_{X|Y}$. We use $\mathrm{D}_{\mathrm{KL}}(P||Q)$ to denote the Kullback–Leibler (KL) divergence between two probability measures P and Q from the same probability space. Notice that $\mathrm{D}_{\mathrm{KL}}(P||Q)$ is well defined if P is absolutely continuous with respect to (w.r.t.) Q, which we denote by $P \ll Q$. The mutual information between X and Y is represented as $I(X;Y) = \mathrm{D}_{\mathrm{KL}}(P_{X,Y}||P_X P_Y)$. Furthermore, we use $I^z(X;Y) = \mathrm{D}_{\mathrm{KL}}(P_{X,Y|Z=z}||P_{X|Z=z}P_{Y|Z=z})$ to represent the disintegrated mutual information, whose expectation taken 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 sets and sequences, respective. Particularly, $[n]$ represents the set $\{1,\ldots,n\}$. $\mathbb{R}$ and $\mathbb{N}$ are the set of real numbers and natural numbers, respectively. $\mathbb{R}_{\geq 0}$ and $\mathbb{N}_+$ are the set of non-negative

real numbers and positive integers, respectively. The Hadamard Product and Kronecker Product are denoted by $\odot$ and $\otimes$, respectively. The spectral norm and L_2 norm are denoted by $\|\cdot\|$ and $\|\cdot\|_2$, respectively. We use $\mathrm{Perm}((\cdot))$ to denote the set containing the full permutations of a given sequence $(\cdot)$. For example, we have $\mathrm{Perm}((1,2,3)) = \{(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)\}$. Besides, $\mathrm{Unif}(\{\cdot\})$ means the uniform distribution over a fixed support set $\{\cdot\}$. For any fixed integers $m, u \in \mathbb{N}_+$, we define $C_{m,u} \triangleq \frac{2(m+u)\max(m,u)}{(m+u-1/2)(2\max(m,u)-1)}$ as a constant with regard to m and u. For $0 < p < 1$, we define $\Phi_a(p) \triangleq -a^{-1}\log\left(1-[1-e^{-a}]p\right)$ as a function of p, where $a \in \mathbb{R}$ is the parameter.

3.2 Random Splitting Setting for Transductive Learning

Let $\{s_i\}_{i=1}^n$ be a given set with finite cardinality. Each element $s \triangleq (x,y) \in \mathcal{X}\times\mathcal{Y}$ in this set is composed of feature $x \in \mathcal{X}$ and label $y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the feature domain and finite label domain, respectively. Denote by $m \in \{1,\ldots,n-1\}$ and $u \triangleq n-m$ the number of training data points with labels and test data points, respectively. The training data points are obtained by randomly sampled m elements without replacement from $\{s_i\}_{i=1}^n$. To this end, we continuously sample from $\{1,\ldots,n\}$ without replacement and sequentially place the obtained elements into a sequence $Z \triangleq (Z_1,\ldots,Z_n)$. In other words, Z_i is the element obtained from the i-th sampling process. Once the sampling process is finished and Z is obtained, the training labels are given by $\{y_{Z_i}\}_{i=1}^m$. That is, only partial selected labels $\{y_{Z_i}\}_{i=1}^m$ together with all features $\{x_i\}_{i=1}^m$ are revealed to the transductive model. Notice that the randomness comes from the selection of training labels and it is essentially contained in the sampling sequence Z. Now let us see a simple example to understand the above process. For simplicity, we let $n=3$ and $m=1$. Assuming that we have completed a sampling of $\{1,2,3\}$. The obtained elements from the first, second, and third times were 2, 3, and 1, respectively. Then the sampling sequence is given by $z=(2,3,1)$, which is a realization of Z. In this example, the training set provided to the transductive model is $\{(x_2,y_2)\}\bigcup\{x_1,x_3\}$, and it is required to predict the labels of s_1 and s_3.

Formally, a transductive model is defined as a mapping $f : \mathcal{X} \to \mathcal{Y}$. In this paper, we stipulate that each model is parameterized. That is, each model (hypothesis) is associated with a unique parameter $w \in \mathcal{W}$, where $\mathcal{W}$ is the parameter space. We denote by $\mathcal{P}(\mathcal{W})$ the set of distribution over $\mathcal{W}$. The model class (hypothesis space) is given by $\mathcal{F}_{\mathcal{W}} \triangleq \{f_w : \mathcal{X} \to \mathcal{Y} | w \in \mathcal{W}\}$. A transductive learning algorithm will take all features together with training labels determined by Z as input and return a hypothesis $f_W \in \mathcal{F}_{\mathcal{W}}$, whose randomness is depicted by a Markov kernel $P_{W|Z}$. Let $\ell : \mathcal{W}\times(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R}_{\geq 0}$ be the loss function, the transductive training error and test error of a hypothesis f_W are defined as $R_{\mathrm{train}}(W,Z) \triangleq \frac{1}{m}\sum_{i=1}^m \ell(W,s_{Z_i})$ and $R_{\mathrm{test}}(W,Z) = \frac{1}{u}\sum_{i=m+1}^{m+u}\ell(W,s_{Z_i})$, respectively. The transductive generalization gap is then defined as $\mathcal{E}(W,Z) \triangleq R_{\mathrm{test}}(W,Z) - R_{\mathrm{train}}(W,Z)$. Furthermore, we use $\mathbb{E}_{W,Z}[\mathcal{E}(W,Z)]$ to denote the expectation of $\mathcal{E}(W,Z)$ taken over $P_{W,Z} = P_Z \otimes P_{W|Z}$, which represents the transductive generalization gap of the hypothesis f_W over the randomness of training labels selection and the learning algorithm. In certain cases, the loss function $\ell(w,s)$ can also be represented as $r(f_w(x),y)$, where $f_w(x)$ is the prediction of the model 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 $\mathbb{1}\{\cdot\}$ is the indicator function.

3.3 Random Sampling Setting for Transductive Learning

Denote by $S = (S_1, \ldots, S_n) \sim P_{S_1,\ldots,S_n}$ a sequence of data points, where $S_i \triangleq (X_i, Y_i), i \in [n]$ is composed of feature $X_i \in \mathcal{X}$ and label $Y_i \in \mathcal{Y}$. Notice that S is a sequence of random variables rather than constants. After the data sequence S is obtained, partial labels $(Y_1, \ldots, Y_m)$ together with all features $(X_1, \ldots, X_n)$ are revealed to a transductive model, whose task is predicting the labels of $(X_{m+1}, \ldots, X_n)$. Recall that $m \in \{1, \ldots, n-1\}$ and $u \triangleq n - m$ are the number of training data points with labels and test data points, respectively. Similar to the notations used in the random splitting, the hypothesis space is denoted by $\mathcal{F}_{\mathcal{W}} \triangleq \{f_w : \mathcal{X} \to \mathcal{Y} | w \in \mathcal{W}\}$. For a hypothesis $f_W \in \mathcal{F}_{\mathcal{W}}$, its transductive training and test error are defined as $R_{\text{train}}(W, S) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(W, S_i)$ and $R_{\text{test}}(W, S) = \frac{1}{u}\sum_{i=m+1}^{m+u} \ell(W, S_i)$ respectively. Particularly, under the case that $u = km, k \in \mathbb{N}_+$, the transductive training and test errors are reformulated as $R_{\text{train}}(W, S) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(W, S_i)$ and $R_{\text{test}}(W, S) \triangleq \frac{1}{km}\sum_{i=m+1}^{(k+1)m} \ell(W, S_i)$, respectively. The randomness of learning algorithms is depicted by the Markov kernel $P_{W|S}$.

4. Main Results

4.1 Mutual Information Bounds for Transductive Learning

In this section, we establish generalization bounds for the transductive generalization gap $\mathcal{E}(W, Z)$ in the context of information theory. To this end, there exist two technical challenges. The first challenge is that the parameter W returned by a transductive learning algorithm is dependent on the sampling sequence Z. The reason is that the algorithm is run after the training labels are selected, which is determined by the sampling sequence Z. Under the framework of information theory (Xu and Raginsky, 2017; Hellström et al., 2023), the approach for this issue is applying the change of measure technique and shifting the distribution from $P_{W,Z}$ to $P_{W\otimes Z}$, by which we can decouple the dependence between W and Z. The second challenge is that the training and test data points are dependent since the labels of training data points are obtained by sampling without replacement. In other words, once the training labels are specified (or the first m entries of Z are observed), both $R_{\text{train}}(W, Z)$ and $R_{\text{test}}(W, Z)$ are uniquely determined. In this work, we adopt the martingale technique to address this issue, which is a widely used tool in statistics. The process is constructing a Doob's martingale and then showing that the corresponding martingale differences are bounded. Integrating the above steps, we obtain the following results.

Theorem 1 *Suppose that $\ell(w, s) \in [0, B]$ holds for all $w \in \mathcal{W}, s \in \{s_i\}_{i=1}^n$, where $B > 0$ is a constant. Also, suppose that $P_{W,Z} \ll P_W P_Z$. Then we have*

$$|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W, Z) - R_{\text{train}}(W, Z)\right]| \leq \sqrt{\frac{B^2 C_{m,u} I(W; Z)(m+u)}{2mu}}, \tag{1}$$

$$\mathbb{E}_{W,Z}\left[\left(R_{\text{test}}(W, Z) - R_{\text{train}}(W, Z)\right)^2\right] \leq \frac{B^2 C_{m,u}(I(W; Z) + \log 3)(m+u)}{mu}. \tag{2}$$

Notice that $C_{m,u} \approx 1$ holds when the values of m, u are sufficiently large. Theorem 1 shows that the expectation of transductive generalization gap is upper bounded by the mutual information between the sampling sequence Z and the hypothesis W returned by the learning algorithm. Since Z is obtained by sampling without replacement from $[n]$ and its first

m entries are indices of training data points, it essentially depicts the randomness of training label selection. Therefore, Theorem 1 demonstrates that for the hypothesis returned by a transductive learning algorithm, the less dependence it has on the selection of training labels, the smaller the generalization gap it will have. Intuitively, if the model only "memorizes" the obtained training labels (or strongly relies on training labels to predict those unlabeled data points), it fails to capture the underlying relation between data points and their labels, making it difficult to perform well on data points whose labels are not revealed. As a comparison, the result under the inductive learning setting (Xu and Raginsky, 2017, Theorem 1) says that the generalization error is upper bounded by the mutual information between the training set $S = \{S_i\}_{i=1}^n$ and the hypothesis W, where data points $S_i = (X_i, Y_i) \sim P_{X,Y}, i \in [n]$ are random variables. Since all features are available for the model, the randomness only comes from training label selection in the transductive learning setting, and the training set S is accordingly replaced by the sampling sequence Z. Moreover, the assumption of our result is slightly stronger than that used in the inductive learning setting. To be specific, Theorem 1 of Xu and Raginsky (2017) only requires the loss function is sub-Gaussian (that is, supposing that there exists a constant $\sigma > 0$ such that $\log \mathbb{E}_{P_{X,Y}}\left[e^{\lambda \ell(w,(X,Y))}\right] \leq \frac{\lambda^2\sigma^2}{2}$ for all $w \in \mathcal{W}$), while Theorem 1 requires the loss function to be bounded (that is, supposing that $\ell(w, s)$ is bounded for all $w \in \mathcal{W}$ and $s \in \{s_i\}_{i=1}^n$). The reason why we make a stronger reason is to ensure that martingale differences constructed in the proof have bounded differences. However, we point out that the bounded assumption of the loss function in Theorem 1 can be further relaxed, inspired by the work of Steinke and Zakynthinou (2020). Concretely, by replacing the loss bounded assumption in Theorem 1 with the following one: suppose that there exists a function $\widetilde{\Delta} : (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}_{\geq 0}$ such that $|\ell(w, s_1) - \ell(w, s_2)| \leq \widetilde{\Delta}(s_1, s_2)$ holds for all $s_1, s_2 \in \{s_i\}_{i=1}^n$ and $w \in \mathcal{W}$, we can obtain results analogous to Eqs. (1,2) where B is accordingly replaced by $\sup_{s_1,s_2\in\{s_i\}_{i=1}^n} \widetilde{\Delta}(s_1, s_2)$. If further assuming that $\sup_{s_1,s_2\in\{s_i\}_{i=1}^n} \widetilde{\Delta}(s_1, s_2) \leq B$ holds for a constant $B > 0$, we exactly recover the results presented in Theorem 1.

Eq. (1) in Theorem 1 is an upper bound for the average transductive generalization error, where the expectation is taken over the distribution $P_{W,Z}$. In real-world applications, particularly deep learning scenarios, only a few train-test splits Z are sampled to evaluate the performance of a transductive learning algorithm. And, restricted by computational costs, we will only repeat the algorithm several times. Therefore, a single-draw bound that holds with probability at least $1 - \delta$ under the distribution $P_{W,Z}$ can better depict the generalization ability of learning algorithms under this circumstance, which is derived by adopting the monitor technique proposed by Bassily et al. (2016).

Theorem 2 *Under the assumptions of Theorem 1, for any $0 < \delta < 1$, with probability at least $1 - \delta$ over the randomness of Z and W we have*

$$|R_{\text{test}}(W, Z) - R_{\text{train}}(W, Z)| \leq \sqrt{\frac{2B^2 C_{m,u}(m+u)}{mu}\left(\log\left(\frac{2}{\delta}\right) + \frac{I(W;Z)}{\delta}\right)}. \tag{3}$$

Moreover, we have

$$\mathbb{E}_{W,Z}[|R_{\text{test}}(W, Z) - R_{\text{train}}(W, Z)|] \leq \sqrt{\frac{B^2 C_{m,u}(m+u)(I(W;Z) + \log 2)}{2mu}}. \tag{4}$$

In addition to the single-draw bound, we obtain another result from the monitor technique, which is an expectation bound on the absolute value of the transductive generalization error and serves as a supplement of Theorem 1. Now we are in a place to compare Theorem 2 with a previous transductive generalization bound based on the complexity of hypothesis space (El-Yaniv and Pechyony, 2009, Theorem 1), which is restated as follows.

Theorem 3 (El-Yaniv and Pechyony, 2009, Theorem 1) *Let $B > 0$ be a constant. Suppose that $\ell(w, s) \in [0, B]$ holds for all $w \in \mathcal{W}, s \in \{s_i\}_{i=1}^{n}$. For any $0 < \delta < 1$, with probability $1 - \delta$ over the randomness of Z and W,*

$$R_{\text{test}}(W, Z) \leq R_{\text{train}}(W, Z) + \mathfrak{R}_{m+u}(\mathcal{W}) + \frac{c_0 B(m+u)\sqrt{\min(m,u)}}{mu} + \sqrt{\frac{C_{m,u}(m+u)\log(1/\delta)}{2mu}},$$

where $c_0 \triangleq \sqrt{\frac{32\log(4e)}{3}}$ and $\mathfrak{R}_{m+u}(\mathcal{W})$ is the transductive Rademacher complexity.

The single-draw bound in Eq. (3) of Theorem 2 and Theorem 3 is of order $\sqrt{(m+u)/mu}$ and $\sqrt{\min(m,u)}(m+u)/mu$, respectively. Since $\sqrt{\min(m,u)} > \sqrt{mu/(m+u)}$, the bound in Eq. (3) is sharper than that in Theorem 3 regarding the dependency of m and u, if

$$I(W; Z) \lesssim \frac{mu \cdot \min(m,u)}{m+u} \cdot \log(1/\delta) \cdot \delta. \tag{5}$$

Although the mutual information term $I(W; Z)$ could not be easily computed, we show in Section 4.5 that it has a unique advantage for analyzing iterative learning algorithms, such as SGD and its variants. Besides, a result similar to Eq. (4) can be derived by taking the square root on both sides of Eq. (2) and applying Jensen's inequality. But, the constant factor of the derived result is slightly larger than that of Eq. (4). We remark that Eq. (4) is not a deteriorated version of Eq. (1), as it is obtained by the monitor technique rather than directly derived from Eq. (4). Besides, one may notice a limitation of Eq. (3) that its dependency of δ deteriorates from $\log(1/\delta)$ to $1/\delta$ compared with Theorem 3. This issue can not be easily addressed without extra cost, that is, improving the dependency of δ may deteriorate the dependency of m and u. To see this, we adopt the technique proposed by Hellström and Durisi (2020) to derive another single draw bound analogous to Eq. (3), where the cost of performing a change of measure from $P_{W,Z}$ to $P_{W \otimes Z}$ is depicted by the information density $\log \frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z}$ instead of mutual information $I(W; Z)$. The derived result is given in Proposition 4. As can be seen, the confidence parameter δ in Eq. (6) is of order $\log(1/\delta)$, yet the dependency of m, u is worse than that of Eq. (3).

Proposition 4 *Under the assumptions of Theorem 1 and suppose that $m \geq 2, u \geq 3$ or $m \geq 3, u \geq 2$, with probability at least $1 - \delta$ over the randomness of Z and W, we have*

$$|R_{\text{test}}(W, Z) - R_{\text{train}}(W, Z)| \leq \sqrt{\frac{B^2 C_{m,u}(m+u)}{2(mu - m - u)} \left(\log\left(\frac{1}{\delta}\sqrt{\frac{mu}{m+u}} \right) + \log \frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z} \right)}. \tag{6}$$

4.2 Conditional Mutual Information Bounds for Transductive Learning

The established bounds in Section 4.1 contain either the mutual information or the information density. Both of them are unbounded and may result in the bounds being unable

to provide meaningful learning guarantees. Also, both W and Z are high dimensional random variables in real-world scenarios, which makes it challenging to compute the numerical value of $I(W;Z)$ with finite samples. To address this issue, under the inductive learning setting, Steinke and Zakynthinou (2020) propose the conditional mutual information (CMI) framework to disentangle the randomness of sampling data points and splitting them into training and test sets. Concretely, the randomness of sampling data points is depicted by the supersample $\widetilde{Z} \in (\mathcal{X} \times \mathcal{Y})^{m\times 2}$, which is an array containing $2m$ data points drawn independently from $P_{X,Y}$. The randomness of splitting training and test set is depicted by the selector sequence $U = (U_1, \ldots, U_m) \sim \mathrm{Unif}(\{0,1\})^m$, which is independent of $\widetilde{Z}$. In this way, we can use $I(W;U|\widetilde{Z})$ instead of $I(W;Z)$ to measure the information on training data carried by the hypothesis returned by learning algorithms. Now let us come back to the transductive learning setting. Notice that the training and test data points are *independent* in the inductive setting yet *dependent* in the transductive learning setting. Thus, the supersample setting of Steinke and Zakynthinou (2020) is not applicable to the transductive learning setting. To bridge this gap, we propose the concept of transductive supersample under the assumption that the number of test data points is a multiple of the number of training data points, namely $u = km, k \in \mathbb{N}_+$. As a warm-up example, we first discuss the case that the number of training examples is equal to that of test examples, namely $k = 1$.

Definition 5 (Transductive Supersample) *Let $m = u = \frac{n}{2}$. Performing m rounds of sampling without replacement from the set $\{1, \ldots, n\}$ and each time two elements are chosen. For each $i \in [m]$, arrange the elements from the i-th sampling into a sequence $\widetilde{Z}_i \triangleq (\widetilde{Z}_{i,0}, \widetilde{Z}_{i,1})$ in ascending order, that is, we have $\widetilde{Z}_{i,0} < \widetilde{Z}_{i,1}$. The transductive supersample is defined as a sequence $\widetilde{Z} \triangleq (\widetilde{Z}_1, \ldots, \widetilde{Z}_m)$.*

Notice that for each $i \in [m]$, the first and second elements of $\widetilde{Z}_i$ are $\widetilde{Z}_{i,0}$ and $\widetilde{Z}_{i,1}$, respectively. Now let us elaborate on Definition 5 through a simple example. For simplicity, we let $m = u = 2$. Assuming that we have completed a sampling of [4], and the elements obtained from the first and second rounds are $\{3,2\}$ and $\{1,4\}$. By arranging the elements $\{3,2\}$ into a sequence in ascending order, we have $\widetilde{z}_{1,0} = 2$ and $\widetilde{z}_{1,1} = 3$ and thus $\widetilde{z}_1 = (2,3)$. Similarly, we have $\widetilde{z}_2 = (1,4)$. Then the transductive supersample is denoted by $\widetilde{z} = (\widetilde{z}_1, \widetilde{z}_2) = ((2,3),(1,4))$. Clearly, each entry in $\widetilde{z}$ is a sequence of length 2, and we remark that $((2,3),(1,4)) \neq ((1,4),(2,3))$. Recall that the sequence Z defined in Section 3.2 is obtained by each time sampling one element from $[n]$ without replacement. By introducing the selector sequence U to permute elements in the transductive supersample $\widetilde{Z}$, we show that the sampling sequence Z can be recovered from $\widetilde{Z}$ and U.

Proposition 6 *Let $U \triangleq (U_1, \ldots, U_m) \sim \mathrm{Unif}(\{0,1\})^m$ be the sequence of random variables that is independent of a transductive supersamples $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_m)$. Define*

$$\mathscr{Z}(\widetilde{Z}, U) \triangleq (\widetilde{Z}_{1,U_1}, \ldots, \widetilde{Z}_{m,U_m}, \widetilde{Z}_{1,1-U_1}, \ldots, \widetilde{Z}_{m,1-U_m}) \tag{7}$$

as the sequence induced by $\widetilde{Z}$ and U. Define the transductive training and test error as

$$\begin{aligned} R_{\mathrm{train}}(W, \widetilde{Z}, U) &\triangleq \frac{1}{m}\sum_{i=1}^{m} \ell\Big(W, s_{\mathscr{Z}_i(\widetilde{Z},U)}\Big) = \frac{1}{m}\sum_{i=1}^{m} \ell\Big(W, s_{\widetilde{Z}_{i,U_i}}\Big), \\ R_{\mathrm{test}}(W, \widetilde{Z}, U) &\triangleq \frac{1}{m}\sum_{i=m+1}^{2m} \ell\Big(W, s_{\mathscr{Z}_i(\widetilde{Z},U)}\Big) = \frac{1}{m}\sum_{i=1}^{m} \ell\Big(W, s_{\widetilde{Z}_{i,1-U_i}}\Big), \end{aligned} \tag{8}$$

where $\mathscr{Z}_i(\widetilde{Z},U), i\in[2m]$ *is the* i*-th entry of the sequence* $\mathscr{Z}(\widetilde{Z},U)$*. We have*

$$\begin{aligned}&\mathbb{E}_{W,\widetilde{Z},U}[R_{\text{test}}(W,\widetilde{Z},U)-R_{\text{train}}(W,\widetilde{Z},U)]\\=&\mathbb{E}_{W,Z}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]=\mathbb{E}_{W,Z}[\mathcal{E}(W,Z)].\end{aligned}\tag{9}$$

Again, we use a simple example to explain our thought. As before, we let $m=u=2$. Assume that the realization of $\widetilde{Z}$ and U are $\widetilde{z}=((2,3),(1,4))$ and $u=(0,1)$ respectively. In this case, we have $\widetilde{z}_1=(2,3)$, $\widetilde{z}_2=(1,4)$, $u_1=0$ and $u_2=1$. Then the sequence Z they induce is given by $\mathscr{Z}(\widetilde{z},u)=(\widetilde{z}_{1,u_1},\widetilde{z}_{2,u_2},\widetilde{z}_{1,1-u_1},\widetilde{z}_{2,1-u_2})=(2,4,3,1)$. Proposition 6 shows that if $m=u$, we can replace Z by $\widetilde{Z}$ and U. Concretely, instead of directly sampling Z, we can firstly sample $\widetilde{Z}$ and U respectively and then permute the entries in $\widetilde{Z}$ according to Eq. (7). In this way, we can disentangle the randomness of Z into two parts, one contained in $\widetilde{Z}$ and the other contained in U. Notably, U is dependent to $\widetilde{Z}$. This enables us to tackle the dependence introduced by sampling without replacement and establish CMI bounds for transductive learning algorithms. Moreover, if $\ell(\cdot)$ is the zero-one loss, we can further derive PAC-Bayesian bounds under the CMI framework by using Catoni's technique (Catoni, 2007, Chapter 3.1). The aforementioned results are summarized in Theorem 7.

Theorem 7 *Suppose that* $\ell(w,s)\in[0,B]$ *holds for all* $w\in\mathcal{W}, s\in\{s_i\}_{i=1}^n$*. We have*

$$|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]|\leq\mathbb{E}_{\widetilde{Z}}\sqrt{\frac{2B^2}{m}I^{\widetilde{Z}}(W;U)},\tag{10}$$

$$\mathbb{E}_{W,Z}\left[\left(R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right)^2\right]\leq\frac{4B^2}{m}(I(W;U|\widetilde{Z})+\log 3).\tag{11}$$

Moreover, suppose that $\ell(\cdot)$ *is the zero-one loss. Let* $P\in\mathcal{P}(\mathcal{W})$ *be a prior distribution that is independent of* U*. For any* $0<\delta<1$*,* $\lambda>0$*, and* $Q\in\mathcal{P}(\mathcal{W})$ *such that* $P\ll Q$*, with probability at least* $1-\delta$ *over the randomness of* $\widetilde{Z}$ *and* U*,*

$$\sup_{Q\in\mathcal{P}(\mathcal{W})}\mathbb{E}_{W\sim Q}\left[\Phi_{\lambda/m}(R(W))-R_{\text{train}}(W,\widetilde{Z},U)\right]\leq\frac{\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{\lambda},\tag{12}$$

where $R(W)\triangleq\frac{1}{2}[R_{\text{train}}(W,\widetilde{Z},U)+R_{\text{test}}(W,\widetilde{Z},U)]=\frac{1}{2m}\sum_{i=1}^{2m}\ell(W,s_i)$*.*

By the property of mutual information that conditioning reduces uncertainty, we have $I(W;U|\widetilde{Z})\leq I(W;U)\leq m\log 2$ holds, suggesting that the conditional mutual information has a finite upper bound. Eq. (10) is consistent with the results of Steinke and Zakynthinou (2020) in formulation, and the only difference is that $\widetilde{Z}$ should be interpreted as the transductive supersamples. Eq. (12) is the extension of Theorem 3.1.2 given by Catoni (2007) to the random splitting setting. We remark that the prior P is independent of U, but may depend on $\widetilde{Z}$. Notice that the randomness of selecting training labels is determined by both U and $\widetilde{Z}$. Thus, the prior P can obtain some but not all information on the selection of training labels, which is similar to Catoni's result where P depends on the dataset S.

Although the mutual information term $I(W;U|\widetilde{Z})$ is bounded, computing its numerical value is still challenging since W may be high-dimensional in real-world application scenarios, particularly deep learning scenarios. To address this issue, various new information measures are proposed (Harutyunyan et al., 2021; Hellström and Durisi, 2022; Wang

and Mao, 2023a), which are theoretically smaller than $I(W;U|\widetilde{Z})$ and easier to estimate. Using the concept of transductive supersample, these results can be transported to the transductive learning setting.

Corollary 8 *Suppose that $r(\hat{y},y)\in[0,B]$ holds for all $\hat{y}\in\widehat{\mathcal{Y}}$ and $y\in\mathcal{Y}$, where $B>0$ is a constant. Denote by $f_w(x)\in\mathbb{R}^K$ the prediction of the model and $F_i\triangleq(f_W(x_{\widetilde{Z}_{i,0}}),f_W(x_{\widetilde{Z}_{i,1}}))$ the sequence of predictions. Denote by $L_i\triangleq(\ell(W,s_{\widetilde{Z}_{i,0}}),\ell(W,s_{\widetilde{Z}_{i,1}}))$ the sequence of loss values and $\Delta_i\triangleq\ell(W,s_{\widetilde{Z}_{i,1}})-\ell(W,s_{\widetilde{Z}_{i,0}})$ the difference of loss value. We have*

$$|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]|\leq\frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(F_i;U_i)},\tag{13}$$

$$|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]|\leq\frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(L_i;U_i)},\tag{14}$$

$$|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]|\leq\frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta_i;U_i)}.\tag{15}$$

According to the type of conditional mutual information they contain, the bounds in Eqs. (13), (14), and (15) are termed as transductive f-CMI bound (Harutyunyan et al., 2021), transductive e-CMI bound (Hellström and Durisi, 2022), and transductive Id-CMI bound (Wang and Mao, 2023a), respectively. The only difference between these results and the previous one is that here $\widetilde{Z}$ represents the *transductive supersamples*. In applications, the prediction of the model is a low-dimension vector and thus reduces the difficulty of computing the conditional mutual information $I(W;U|\widetilde{Z})$. Note that L_i in Eq. (14) and Δ_i in Eq. (15) are two-dimensional and one-dimensional random variables, yielding more computationally convenient and sharper bounds. However, these bounds can not explicitly depict some factors that affect generalization such as the norm of weights or the sharpness of the loss landscape, and they are accordingly called black-box algorithms generalization bounds (Harutyunyan et al., 2021). As a comparison, the bounds in Theorem 7 and Section 4.1 are more informative to understanding generalization (see Section 4.5 for more details), yet their numerical values are difficult to compute. To summarize, different types of results provide different perspectives for us to understand the generalization abilities of transductive learning algorithms. Now we are in a place to discuss more general cases that $u=km, k\in\mathbb{N}_+$ by extending the definition of transductive supersample.

Definition 9 (k-Transductive Supersample) *For $k\in\mathbb{N}_+$, let $m=\frac{n}{k+1}$ and $u=km$. Performing m rounds of sampling without replacement from the set $\{1,\dots,n\}$ and each time $(k+1)$ elements are chosen. For each $i\in[m]$, arrange the elements from the i-th sampling into a sequence $\widetilde{Z}_i\triangleq(\widetilde{Z}_{i,0},\dots,\widetilde{Z}_{i,k})$ in ascending order, that is, $\widetilde{Z}_{i,0}<\cdots<\widetilde{Z}_{i,k}$. The transductive supersample is defined as a sequence $\widetilde{Z}\triangleq(\widetilde{Z}_1,\dots,\widetilde{Z}_m)$.*

Clearly, Definition 5 is a special case of Definition 9 with $k=1$. Similarly, we extend the definition of the selector sequence as $U\triangleq(U_1,\dots,U_m)\sim\text{Unif}(\text{Perm}((0,\dots,k)))^m$, where $U_i\triangleq(U_{i,0},\dots,U_{i,k}), i\in[m]$ is a sequence of length $(k+1)$. With these notations, we define

$$\mathscr{Z}(\widetilde{Z},U)=(\widetilde{Z}_{1,U_{1,0}},\dots,\widetilde{Z}_{m,U_{m,0}},\widetilde{Z}_{1,U_{1,1}},\dots,\widetilde{Z}_{m,U_{m,1}},\dots,\widetilde{Z}_{1,U_{m,k}},\dots,\widetilde{Z}_{m,U_{m,k}})\tag{16}$$

as the sequence induced by $\widetilde{Z}$ and U. Then the first m elements in $\mathscr{Z}(\widetilde{Z}, U)$ are the indices of training data points, and others are the indices of test data points. Now let us see a short example of this definition. For simplicity, we let $m = 2$ and $k = 3$. Assuming that we have completed a sampling of [6]. The obtained elements from the first and second times are $\{3, 5, 2\}$ and $\{1, 6, 4\}$, respectively. For the elements $\{3, 5, 2\}$, we put it into a sequence and get $\widetilde{z}_1 = (2, 3, 5)$. Similarly we have $\widetilde{z}_2 = (1, 4, 6)$. Also, assuming that the realization of U is $u = ((1, 0, 2), (2, 1, 0))$, that is, $u_1 = (1, 0, 2)$ and $u_2 = (2, 1, 0)$. Then the sequence $\mathscr{Z}(\widetilde{Z}, U)$ induced by $\widetilde{Z}$ and U is $\mathscr{Z}(\widetilde{z}, u) = (\widetilde{z}_{1,u_{1,0}}, \widetilde{z}_{2,u_{2,0}}, \widetilde{z}_{1,u_{1,1}}, \widetilde{z}_{2,u_{2,1}}, \widetilde{z}_{1,u_{1,2}}, \widetilde{z}_{1,u_{2,2}}) = (3, 6, 2, 4, 5, 1)$. In this way, Z can be replaced by $\widetilde{Z}$ and U, which enables us to extend the results in Theorem 7 and Corollary 8 to the case that $u = km, k \in \mathbb{N}_+$. Since the formulations are similar, we place the details in Appendix F. Since U has more possible values to take with the increase of k, we need to accordingly increase the samples of U to reduce the estimated error of conditional mutual information.

We close this part by discussing the connection between the transductive supersample and the supersample under the leave-one-out CMI framework (Haghifam et al., 2022), where the selector variable is defined as $U = \mathrm{Unif}([n+1])$. The supersample under this framework is denoted by $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_{n+1}) \in (\mathcal{X} \times \mathcal{Y})^{n+1}$, where $\widetilde{Z}_i \sim P_{X,Y}, i \in [n+1]$ are independent and identically distributed (i.i.d.) data points. The role of U is selecting an entry from $\widetilde{Z}$, which will be used as a test data point. And the remaining entries in $\widetilde{Z}$ are used as training data. Formally, the training set and test set are given by $\{\widetilde{Z}_U\}$ and $\{\widetilde{Z}_j | j \neq U\}$ respectively. Recall that the random variable U_i is defined by $U_i \sim \mathrm{Unif}(\mathrm{Perm}((0, \ldots, k)))$ for a fixed index $i \in [m]$ under the transductive learning setting that $u = km$. If we redefine $U_i, i \in [m]$ as $U_i \sim \mathrm{Unif}(\mathrm{Perm}((1, \ldots, n+1)))$, one can verify that $U_{i,j} \sim \mathrm{Unif}([n+1])$ for $j \in [n+1]$, thereby recovering the definition of U used in leave-one-out CMI framework. And, recall that $\widetilde{Z}_i, i \in [m]$ is a sequence of length $(k+1)$, which plays a role similar to the arbitrarily-sized subset used by Haghifam et al. (2022). However, this connection only lies in the case that $u = km, k \in \mathbb{N}_+$. For common cases where m and u are arbitrary integers, we do not yet know how to extend the CMI framework to the transductive learning setting.

4.3 Transductive PAC-Bayesian Bounds under the Random Splitting Setting

In this section, we establish generalization bounds for the random splitting setting of transductive learning algorithms in the context of PAC-Bayesian theory, which is closely connected with information theory since the foundation of both is the change of measure technique. The overall procedure is similar to that used in Section 4.1, where two main ingredients are applying the change of measure technique and deriving an upper bound for the moment-generating function of the transductive generalization gap via the martingale approach. Similar to the PAC-Bayesian bounds derived in the inductive setting, the transductive PAC-Bayesian bounds also include the KL divergence between a posterior Q and a prior P. The theoretical results are summarized in the following theorem.

Theorem 10 *Suppose that $\ell(w, s) \in [0, B]$ holds for all $w \in \mathcal{W}$ and $s \in \{s_i\}_{i=1}^n$, where $B > 0$ is a constant. Let $P \in \mathcal{P}(\mathcal{W})$ be a prior distribution independent of Z. For any $0 < \delta < 1$, $\lambda > 0$, and $Q \in \mathcal{P}(\mathcal{W})$ such that $Q \ll P$, we have*

$$\mathbb{E}_Z \mathbb{E}_{W \sim Q}[R_{\mathrm{test}}(W, Z) - R_{\mathrm{train}}(W, Z)] \leq \frac{\lambda B^2 C_{m,u}(m+u)}{8mu} + \frac{\mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}(Q||P)\right]}{\lambda}. \tag{17}$$

With probability at least $1-\delta$ *over the randomness of* Z*, we have*

$$\sup_{Q\in\mathcal{P}(\mathcal{W})} \mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)] \leq \frac{\lambda B^2 C_{m,u}(m+u)}{8mu} + \frac{\mathrm{D}_{\mathrm{KL}}(Q||P) + \log(1/\delta)}{\lambda}. \tag{18}$$

With probability at least $1-\delta$ *over the randomness of* Z *and* $W\sim Q$ *we have*

$$R_{\text{test}}(W,Z) \leq R_{\text{train}}(W,Z) + \frac{\lambda B^2 C_{m,u}(m+u)}{8mu} + \frac{1}{\lambda}\left(\log\frac{\mathrm{d}Q}{\mathrm{d}P} + \log(1/\delta)\right). \tag{19}$$

Eq. (17) is a bound for the average transductive generalzation gap $\mathbb{E}_Z\mathbb{E}_{W\sim P}[\mathcal{E}(W,Z)]$. Eq. (18) is an average bound that holds with high probability under the distribution P_Z. Eq. (19) is a singe-draw bound that holds with high probability under the distribution $P_{W,Z}$. Following Alquier (2024), denote by $\mathbb{E}_Z[Q]$ the probability distributions

$$\frac{\mathrm{d}\mathbb{E}_Z[Q]}{\mathrm{d}P}(w) \triangleq \mathbb{E}_Z\left[\frac{\mathrm{d}Q}{\mathrm{d}P}(w)\right], w\in\mathcal{W}. \tag{20}$$

By the equality $\mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}(Q||P)\right] = \mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}(Q||\mathbb{E}_Z[Q])\right] + \mathrm{D}_{\mathrm{KL}}(\mathbb{E}_Z[Q]||P)$ and the fact that $I(W;Z) = \mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}(Q||\mathbb{E}_Z[Q])\right]$, choosing $P = \mathbb{E}_Z[Q]$ as the prior and minimizing the right hand side (r.h.s.) of Eq. (17) w.r.t. λ enable us to recover the information-theoretic bound in Eq. (1). Now let us turn to Eq. (18), whose r.h.s. holds for any probability measure $Q\in\mathcal{P}(\mathcal{W})$. For a fixed prior P, Lemma 22 shows that the optimal posterior Q minimizing the term $\lambda\mathbb{E}_{W\sim Q}[R_{\text{train}}(W,Z)] + \mathrm{D}_{\mathrm{KL}}(Q||P)$ is the Gibbs distribution $P_{e^{-\lambda R_{\text{train}}}}$ given by

$$\frac{\mathrm{d}P_{e^{-\lambda R_{\text{train}}}}}{\mathrm{d}P}(w) = \frac{e^{-\lambda R_{\text{train}}(w,Z)}}{\int_{\mathcal{W}} e^{-\lambda R_{\text{train}}(w',Z)}\,\mathrm{d}P(w')}. \tag{21}$$

This is consistent with the result under the random sampling setting of transductive learning (Catoni, 2007, Theorem 3.1.2). Notice that Eqs. (18) and (19) are Catoni-typed bounds since they contain an extra parameter λ. By minimizing the right hand side of them, we obtain a formulation of the optimal value of λ, which contains the complexity term $\mathrm{D}_{\mathrm{KL}}(Q||P)$. Since λ needs to be specified before observing Z and the posterior Q generally depends on Z, this approach is not feasible in practice. An alternative approach that enables λ to be optimized, proposed by Alquier (2024); Catoni (2007), is to optimize λ over a predefined set Λ with finite cardinality, as shown in the following corollary.

Corollary 11 *Denote by* Λ *a predefined set that satisfies* $|\Lambda| < \infty$*. Under the assumption of Theorem 10, for any* $0<\delta<1$ *and* $Q\in\mathcal{P}(\mathcal{W})$ *such that* $Q \ll P$*, with probability at least* $1-\delta$ *over the randomness of* Z*, we have*

$$\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)] \leq \inf_{\lambda\in\Lambda}\left\{\frac{\lambda C_{m,u}(m+u)}{8mu} + \frac{\mathrm{D}_{\mathrm{KL}}(Q||P) + \log(|\Lambda|/\delta)}{\lambda}\right\}.$$

Particularly, for $\Lambda \triangleq \left\{e^i, i\in\mathbb{N}\right\} \cap [1, mu/(m+u)]$*, we have*

$$\begin{aligned}&\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)]\\ \leq &\inf_{\lambda\in\left[1,\frac{mu}{m+u}\right]}\left\{\frac{\lambda C_{m,u}(m+u)}{8mu} + \frac{e}{\lambda}\left(\mathrm{D}_{\mathrm{KL}}(Q||P) + \log\left(\frac{1}{\delta}\log\left(\frac{mu}{m+u}\right)\right)\right)\right\}.\end{aligned}$$

Corollary 11 shows that demonstrates λ to be optimized resulting in an additional term $\log\log(mu/(m+u))$ appearing in the bound. Now we are in a place to compare our result to the latest explicit transductive PAC-Bayesian bound derived under the random splitting setting, which is given by Bégin et al. (2014) and we restate it as follows.

Theorem 12 (Bégin et al., 2014, Corollary 7(b)) *Suppose that the number of training and test data points satisfies $m, u \geq 20$. Let $P \in \mathcal{P}(\mathcal{W})$ be a prior distribution independent of Z. For any $0 < \delta < 1$ and $Q \in \mathcal{P}(\mathcal{W})$ such that $Q \ll P$, with probability at least $1-\delta$ over the randomness of Z, we have*

$$\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] \leq \sqrt{\frac{m+u}{2mu}\left(\mathrm{D}_{\mathrm{KL}}(Q||P) + \log\left(\frac{3\log(m)}{\delta}\sqrt{\frac{mu}{m+u}}\right)\right)}.$$

Different from our result, Theorem 12 presents a Langford-Seeger-Maurer-typed PAC-Bayesian bound and does not include an additional parameter λ. Accordingly, the overall formulation of the bound is more concise than ours, and the order of the bound w.r.t. the number of training data points m and test data points u is clear. However, the physical meaning of the bound in Theorem 12 is implicit since the optimal posterior can not be reflected from the bound. In contrast, the physical meaning of our result (Eq. 18) is explicit as it reveals that the optimal posterior is the Gibbs distribution. To directly compare the order of our result with the bound given in Theorem 12, inspired by the work of Alquier (2024), we reformulate the latter by introducing the parameter λ by using the inequality $\sqrt{(\lambda a/2)(2b/\lambda)} \leq \lambda a/4 + b/\lambda$ that holds for any $a, b, \lambda > 0$:

$$\begin{aligned}
&\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)] \\
\leq&\mathbb{E}_{W\sim Q}[R_{\text{train}}(W,Z)] + \inf_{\lambda>0}\left\{\frac{\lambda(m+u)}{8emu} + \frac{e}{\lambda}\left(\mathrm{D}_{\mathrm{KL}}(Q||P) + \log\left(\frac{3\log(m)}{\delta}\sqrt{\frac{mu}{m+u}}\right)\right)\right\}.
\end{aligned}$$

Using the inequality $\log x \leq \sqrt{x}$ that holds for all $x > 0$, we have $\log(mu/(m+u)) \leq \sqrt{mu/(m+u)}$. Therefore, Theorem 10 with a geometric grid (Corollary 11) gives better result than Theorem 12 by saving a factor $\log(3\log(m))$ when m and u are large. However, Theorem 12 may provide a tighter bound if m and u is small since its constant of the first term is smaller than that in Theorem 10. However, we emphasize that Theorem 10 is essentially different than Theorem 12, and its assumption is much weaker. Specifically, Theorem 12 requires that $\ell(\cdot)$ is zero-one loss, and the values of m and u must satisfy $m, u \geq 20$. In contrast, Theorem 10 applies to any bounded loss, and there are no constraints on the value of m and u. The reason is that we derive the upper bound for the moment-generating function by the martingale technique, while Bégin et al. (2014) firstly obtain an implicit bound by introducing a variant of the binary relative entropy (Germain et al., 2009) and then convert it to an explicit one by the Pinsker's inequality. As for the implicit one (Bégin et al., 2014, Corollary 7 (a)), we do not know how to compare it with ours, and we are accordingly not sure which one is sharper.

One of the most important insights delivered by PAC-Bayesian bounds is that the generalization ability of a model is related to the flatness of its loss landscape, and a flat minimum is beneficial for generalization. With the help of Theorem 10, this result can be extended to the transductive learning setting when $\ell(\cdot)$ is the zero-one loss.

Corollary 13 *Suppose that (i) $m \geq 2, u \geq 3$ or $m \geq 3, u \geq 2$ and (ii) $R_{\text{test}}(W, Z) \leq \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)}[R_{\text{test}}(W + \boldsymbol{\epsilon}, Z)]$ holds for all sampling sequence Z, where $W \in \mathbb{R}^d$ is the parameter returned by a given transductive learning algorithm and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. For any $\lambda > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$ over the randomness of Z we have*

$$R_{\text{test}}(W, Z) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\text{train}}(W + \boldsymbol{\epsilon}, Z) + \frac{(\lambda C_{m,u} + 8)(m+u)}{8mu} + \frac{1}{\lambda}\left(\frac{1}{2}\left[1 + d \log\left(1 + \frac{\left(1 + \widetilde{C}_{m,u}\right)^2 \|W\|_2^2}{\rho^2}\right)\right] + \log\left(\frac{1}{6\delta}\right) + 2\log\left(\frac{6\pi mu}{m+u}\right)\right),$$

where $\widetilde{C}_{m,u} \triangleq \sqrt{2\log(mu/(m+u))/d}$ and $\rho \triangleq \sigma(1 + \widetilde{C}_{m,u})\sqrt{d}$.

The assumption $R_{\text{test}}(W, Z) \leq \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)}[R_{\text{test}}(W + \boldsymbol{\epsilon}, Z)]$ used in Corollary 13 means that adding random Gaussian noise to the parameter of the model after training does not decrease its performance on unlabeled data points. We remark that the parameter λ can not be directly optimized since its optimal value depends on $\|W\|_2$. The term $\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\text{train}}(W + \boldsymbol{\epsilon}, Z)$ characterizes the change of loss landscape within a ball with W as the center and ρ as the radius. Formally, we call W as sharp minima if the loss values around it differ significantly from itself, namely $R_{\text{train}}(W + \boldsymbol{\epsilon}, Z)$ is significantly larger than $R_{\text{train}}(W, Z)$. Therefore, Corollary 13 suggests that a flat optima could have better transductive generalization performance. A classical approach (Foret et al., 2021) to ensure the flatness of loss landscape is solving a minimax optimization problem $\min_w \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\text{train}}(w + \boldsymbol{\epsilon}, Z)$. By converting the minimax optimization into a bi-level optimization problem and solving it via the hypergradient algorithm, Chen et al. (2023) show that in terms of the recommendation task, GNNs with flatter minima have a better generalization ability than those with sharper minima. This observation serves as strong evidence of Corollary 13 and also demonstrates that we can study the generalization ability of deep transductive learning models (such as GNNs) by analyzing the flatness of its loss landscape. Particularly, recent work of Tang and Liu (2023) reveals that the initial residual and identity mapping techniques 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 the loss landscape and ultimately affect the generalization of the model is worth exploring. Besides, our results can also enhance the applicability of existing generalization analysis for GNNs. For example, combining Theorem 10 with the technique used by Neyshabur et al. (2018), results of Liao et al. (2021) could be extended to the transductive learning setting. Recently, Lee et al. (2024) have combined the techniques of Neyshabur et al. (2018) and Theorem 12 to establish a generalization guarantee for a deterministic triplet classifier. They also derive the first generalization bounds for knowledge graph representation learning. The posterior distribution used by Neyshabur et al. (2018); Lee et al. (2024) is constructed by adding random Gaussian noise to the parameter of the model, which is similar to ours. Besides, the assumption of result given by Lee et al. (2024) is weaker than ours since it does not require that adding Gaussian perturbation should not decrease the test error. However, one can easily obtain analogous results of Neyshabur et al. (2018); Lee et al. (2024) by applying their technique to our new transductive PAC-Bayesian bound in Theorem 10.

We close this part by giving some extra comments on Corollary 13. Firstly, we have that $\rho = \mathcal{O}(\sqrt{d})$, and its numerical value is generally not available. The reason is that ρ depends

on σ and we are unable to obtain the exact numerical value of σ in the assumption. And, minimizing the term $\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\text{train}}(W+\boldsymbol{\epsilon}, Z)$ will result in the loss landscape around W being excessively flat, since $\rho = \mathcal{O}(\sqrt{d})$. Therefore, ρ is regarded as a hyperparameter in practice (Foret et al., 2021). Secondly, Corollary 13 still suffers from some issues. On the one hand, if d is sufficiently large, the term $d\log(1+1/(\sigma^2 d)) \approx O(1))$ may be independent of d, which implies that there could be no impact of the dimension in the first term of the bound when the dimension is sufficiently large. On the other hand, σ that appears in the assumption is data-free, which results in the variance of the posterior $Q \triangleq \mathcal{N}(O, \sigma^2 \mathbf{I}_d)$ being data-independent. We point out here that these issues essentially come from the proof of Foret et al. (2021) and they naturally appear since our proof for Corollary 13 generally follows theirs. However, we would like to clarify that the reason for presenting this corollary is to briefly reveal the application potential of our theoretical results. Indeed, the aforementioned issues could be tackled by using new proof techniques. For example, to address the shortcoming that the variance of the posterior is data-independent, we can adopt the Fisher information matrix as the variance of perturbed noise (Kim et al., 2022), which will make the derived result more in line with real-world scenarios.

4.4 Transductive PAC-Bayesian Bounds under the Random Sampling Setting

So far, all results are derived under the random splitting setting (Section 3.2). In this section, we turn to the random sampling setting (Section 3.3) where a sequence of data points is provided. Establishing risk bounds under this setting is challenging because the data points could come from different distributions and there could exist dependence among them. Theorem 10.1 of Vapnik (1998) shows that if the data points are i.i.d., the risk bounds derived under the random splitting setting can be transported to the random sampling setting. Later, Catoni (2003, 2007) proposes another approach that directly derives risk bounds without transporting results from the random splitting setting, which only requires that the data distribution and the prior be exchangeable (Catoni, 2003, Chapter 2.1) or partially exchangeable (Catoni, 2007, Chapter 3.1). Here, the prior is defined as a function $P : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{P}(\mathcal{W})$. To proceed, we first introduce the following definitions.

Definition 14 (Exchangeable and Partially Exchangeable Distribution) *Let Π be the set of all bijections $\pi : [n] \to [n]$. We say $P_{S_1,\ldots,S_n}$ is an exchangeable distribution if $P_{S_1,\ldots,S_n} = P_{S_{\pi(1)},\ldots,S_{\pi(n)}}$ holds for all $\pi \in \Pi$. If $u = km, k \in \mathbb{N}_+$, let Π be the set of all bijections $\pi : \{0,\ldots,k\} \to \{0,\ldots,k\}$. We say $P_{S_1,\ldots,S_{(k+1)m}}$ is a partially exchangeable distribution if for all $\pi \in \Pi, i \in [m]$, the following holds:*

$$P_{S_1,\ldots,S_i,\ldots,S_m,\ldots,S_{1+mk},\ldots,S_{i+mk},\ldots,S_{(k+1)m}} = P_{S_1,\ldots,S_{i+m\pi(0)},\ldots,S_m,\ldots,S_{1+mk},\ldots,S_{i+m\pi(k)},\ldots,S_{(k+1)m}}.$$

Definition 15 (Exchangeable and Partially Exchangeable Prior) *Let Π be the set of all bijections $\pi : [n] \to [n]$. We say P is an exchangeable prior if $P((s_1,\ldots,s_n)) = P((s_{\pi(1)},\ldots,s_{\pi(n)}))$ holds for all $(s_1,\ldots,s_n) \in (\mathcal{X}\times\mathcal{Y})^n$ and $\pi \in \Pi$. If $u = km, k \in \mathbb{N}_+$, let Π be the set of all bijections $\pi : \{0,\ldots,k\} \to \{0,\ldots,k\}$. We say P is a partially exchangeable prior if for all $(s_1,\ldots,s_n) \in (\mathcal{X}\times\mathcal{Y})^n, \pi \in \Pi, i \in [m]$, the following holds:*

$$\begin{aligned} &P\left((s_1,\ldots,s_i,\ldots,s_m,\ldots,s_{1+mk},\ldots,s_{i+mk},\ldots,s_{(k+1)m})\right) \\ =&P\left((s_1,\ldots,s_{i+m\pi(0)},\ldots,s_m,\ldots,s_{1+mk},\ldots,s_{i+m\pi(k)},\ldots,s_{(k+1)m})\right). \end{aligned}$$

Clearly, exchangeable distribution (or partially exchangeable distribution) refers to those distributions $P_{S_1,\ldots,S_n}$ that remain unchanged after applying a permutation on the full indices $(1,\ldots,n)$ (or partial indices $(i, m+i, \ldots, km+i), i \in [m]$) of the random variable sequence $(S_1,\ldots,S_n)$. Similarly, applying permutation on the indices $(1,\ldots,n)$ (or partial indices $(i, m+i, \ldots, km+i), i \in [m]$) of input variables $(s_1,\ldots,s_n)$ does not change the value of an exchangeable prior (or partially exchangeable prior). Now we elucidate why these concepts are needed in assumptions. The concept of exchangeable distribution (or partially exchangeable distribution) is to handle the dependency between training error and test error. It allows us to implicitly take expectation over the random permutation on the data sequence and thus either connect the transductive training error with total error on full data or transport the result from the random splitting setting to the random sampling setting. To see this, denote by $Z = (Z_1,\ldots,Z_n)$ the sequence obtained by sampling without replacement from $[n]$. Assuming that $S \sim P_{(S_1,\ldots,S_n)}$ and $P_{(S_1,\ldots,S_n)}$ is an exchangeable distribution, for any $w \in \mathcal{W}$ we have

$$\mathbb{E}_S\left[e^{\frac{1}{u}\sum_{i=m+1}^{m+u}\ell(w,S_i)-\frac{1}{m}\sum_{i=1}^{m}\ell(w,S_i)}\right] = \mathbb{E}_S\mathbb{E}_Z\left[e^{\frac{1}{u}\sum_{i=m+1}^{m+u}\ell(w,S_{Z_i})-\frac{1}{m}\sum_{i=1}^{m}\ell(w,S_{Z_i})}\right]. \tag{22}$$

Notice that the term $\frac{1}{u}\sum_{i=m+1}^{m+u}\ell(w,S_{Z_i})-\frac{1}{m}\sum_{i=1}^{m}\ell(w,S_{Z_i})$ in the exponential function of Eq. (22) is exactly the transductive generalization gap that we analyze in Section 4.1 and Section 4.3. Thereby, we can apply the martingale technique to derive an upper bound. If the number of training data points m and test data points u satisfy a certain relation, for example, $u = km, k \in \mathbb{N}_+$, we can use $U = (U_1,\ldots,U_m) \sim \mathrm{Unif}(\mathrm{Perm}((0,\ldots,k)))^m$ instead of Z to permute the data sequence and thus establish a connection between training error and total error. Concretely, assume that $S = (S_1,\ldots,S_{(k+1)m}) \sim P_{S_1,\ldots,S_{(k+1)m}}$ and $P_{S_1,\ldots,S_{(k+1)m}}$ is a partially exchangeable distribution, for any $w \in \mathcal{W}$ we have

$$\begin{aligned}
&\mathbb{E}_S\left[e^{\frac{1}{m}\sum_{i=1}^{m}\ell(w,S_i)}\right] = \mathbb{E}_S\mathbb{E}_U\left[e^{\frac{1}{m}\sum_{i=1}^{m}\ell(w,S_{i+mU_{i,0}})}\right]\\
=&\mathbb{E}_S\mathbb{E}_U\left[\prod_{i=1}^{m}e^{\frac{1}{m}\ell(w,S_{i+mU_{i,0}})}\right] = \mathbb{E}_S\left[\prod_{i=1}^{m}\mathbb{E}_U\left[e^{\frac{1}{m}\ell(w,S_{i+mU_{i,0}})}\right]\right]\\
=&\mathbb{E}_S\left[\exp\left\{\sum_{i=1}^{m}\log\left(\mathbb{E}_U\left[e^{\frac{1}{m}\ell(w,S_{i+mU_{i,0}})}\right]\right)\right\}\right]\\
\le&\mathbb{E}_S\left[\exp\left\{m\log\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_U\left[e^{\frac{1}{m}\ell(w,S_{i+mU_{i,0}})}\right]\right)\right\}\right] \le \mathbb{E}_S\left[\exp\{\Phi_{1/m}(\bar{R}(w,S))\right],
\end{aligned} \tag{23}$$

where $\bar{R}(w,S) \triangleq \frac{1}{(k+1)m}\sum_{i=1}^{(k+1)m}\ell(w,S_i)$ is the total error. In this way, we can build a connection between the training error $\frac{1}{m}\sum_{i=1}^{m}\ell(w,S_i)$ and the total error $\bar{R}(w,S)$. The above technique first appeared in the proof of Theorem 3.1.2 given by Catoni (2007) but was presented with different notations. Specifically, in Definition 3.1.1 of Catoni (2007), he introduces m circular permutations to permute the sequence $(S_1, S_2, \ldots, S_{(k+1)m})$, where the i-th one only applied to partial indices $(S_i, S_{i+m}, \ldots, S_{i+km})$ for $i \in [m]$ and leaving the rest indices unchanged, which is mathematically equivalent to use $U_i \sim \mathrm{Unif}(\mathrm{Perm}((0,\ldots,k)))$ to permute the subsequence $(S_i, S_{i+m}, \ldots, S_{i+km})$ for $i \in [m]$. Further, if $m = u$, the

selector sequence can be simplified as $U = (U_1, \ldots, U_m) \sim \mathrm{Unif}(\{0,1\})^m$. Assume that $S \sim P_{(S_1,\ldots,S_{2m})}$ and $P_{(S_1,\ldots,S_{2m})}$ is a partially exchangeable distribution, for any $w \in \mathcal{W}$,

$$\begin{aligned}
&\mathbb{E}_S\left[e^{\frac{1}{m}\sum_{i=m+1}^{2m}\ell(w,S_i)-\frac{1}{m}\sum_{i=1}^{m}\ell(w,S_i)}\right] = \mathbb{E}_S\left[e^{\frac{1}{m}\sum_{i=1}^{m}(\ell(w,S_{m+i})-\ell(w,S_i))}\right] \\
=&\mathbb{E}_S\mathbb{E}_U\left[e^{\frac{1}{m}\sum_{i=1}^{m}(-1)^{U_i}(\ell(w,S_{i+m})-\ell(w,S_i))}\right] \\
=&\mathbb{E}_S\mathbb{E}_U\left[\prod_{i=1}^{m}e^{\frac{(-1)^{U_i}}{m}(\ell(w,S_{i+m})-\ell(w,S_i))}\right] = \mathbb{E}_S\left[\prod_{i=1}^{m}\mathbb{E}_{U_i}\left[e^{\frac{(-1)^{U_i}}{m}(\ell(w,S_{i+m})-\ell(w,S_i))}\right]\right] \qquad (24) \\
=&\mathbb{E}_S\left[\prod_{i=1}^{m}\cosh\left\{\frac{1}{m}\big(\ell(w,S_{i+m})-\ell(w,S_i)\big)\right\}\right] \le \mathbb{E}_S\left[e^{\frac{1}{2m^2}\sum_{i=1}^{m}(\ell(w,S_{m+i})-\ell(w,S_i))^2}\right].
\end{aligned}$$

The last term in Eq. (24) can be bounded via various techniques. For the case that $\ell(\cdot)$ is the zero-one loss, Catoni (2003) shows that it equals to $\mathbb{E}_S\big[e^{\frac{1}{2m^2}\sum_{i=1}^{2m}\ell(w,S_i)}\big]$ in the proof of Lemma 2.1 given by Catoni (2003). Further, Audibert and Bousquet (2007) show that for any type of losses, it can be upper bounded by $\mathbb{E}_S\big[e^{\frac{1}{m^2}\sum_{i=1}^{2m}\ell^2(w,S_i)}\big]$ via the AM–GM inequality. To summarize, exchangeable distribution or partially exchangeable distribution enables us to handle the dependence between training error and test error, and a stronger assumption on the relation between m and u allows us to use more sophisticated techniques. Moreover, the concept of exchangeable prior (or partially exchangeable prior) is to ensure that the information about selecting training labels is not carried by the prior P. Recall that applying permutation on the indices of input variables does not change the value of the prior P. Although the prior P is allowed to depend on the full dataset S, it will always return the same parameter regardless of the selection of training labels and thus could not carry any information on the training label selection. We remark that this assumption plays the same role as that requiring P to be independent of Z in the random splitting setting.

Now we can extend Theorem 10 to the random sampling setting, inspired by the connection between the random splitting setting and the random sampling setting (Vapnik, 1998, Theorem 10.1). For the convenience of comparison, we restate results of Catoni (2007) and Audibert and Bousquet (2007) in Theorem 17 and Theorem 18, respectively. The first one requires that $u = km, k \in \mathbb{N}_+$ while the second one requires that $m = u$.

Corollary 16 *Suppose that $P_{(S_1,\ldots,S_n)}$ is an exchangeable distribution, and $\ell(w,s) \in [0,B]$ holds for all $w \in \mathcal{W}$ and $s \in (\mathcal{X}\times\mathcal{Y})^n$, where $B > 0$ is a constant. For any $0 < \delta < 1$, $\lambda > 0$, and $Q \in \mathcal{P}(\mathcal{W})$ such that $Q \ll P$, with probability $1-\delta$ over the randomness of S,*

$$\sup_{Q\in\mathcal{P}(\mathcal{W})} \mathbb{E}_{W\sim Q}\left[R_{\mathrm{test}}(W,S) - R_{\mathrm{train}}(W,S)\right] \le \frac{\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{\lambda} + \frac{\lambda B^2 C_{m,u}(m+u)}{8mu}.$$

Theorem 17 (Catoni, 2007, Theorem 3.1.2) *Suppose that $P_{(S_1,\ldots,S_{(k+1)m})}$ is a partially exchangeable distribution, and $\ell(\cdot)$ is the zero-one loss. For any $0 < \delta < 1$, $\lambda > 0$, and $Q \in \mathcal{P}(\mathcal{W})$ such that $Q \ll P$, with probability at least $1-\delta$ over the randomness of S,*

$$\sup_{Q\in\mathcal{P}(\mathcal{W})} \mathbb{E}_{W\sim Q}\left[\Phi_{\lambda/m}(R(W,S)) - R_{\mathrm{train}}(W,S)\right] \le \frac{\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{\lambda}, \qquad (25)$$

where $R(W,S) \triangleq \frac{mR_{\mathrm{train}}(W,S)+kmR_{\mathrm{test}}(W,S)}{(k+1)m}$ is the total error computed on full data.

Theorem 18 (Audibert and Bousquet, 2007, Lemma 10) *Suppose that $P_{S_1,\ldots,S_{2m}}$ is a partially exchangeable distribution. For any $0 < \delta < 1$, $\lambda > 0$, and $Q \in \mathcal{P}(\mathcal{W})$ such that $Q \ll P$, with probability at least $1-\delta$ over the randomness of S,*

$$\sup_{Q\in\mathcal{P}(\mathcal{W})} \mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,S) - R_{\text{train}}(W,S)\right] \leq \mathbb{E}_{S,W\sim Q}\left[\frac{\lambda}{m^2}\sum_{i=1}^{2m}\ell^2(W,S_i)\right] + \frac{\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{\lambda}.$$

Overall, Corollary 16 is a non-trivial extension of Theorem 17 to more common cases. Theorem 17 requires that the data distribution and prior be partially exchangeable, which are weaker than Corollary 16. However, Theorem 17 only applies to the case that $u = km, k \in \mathbb{N}_+$ and $\ell(\cdot)$ is the zero-one loss, yet Corollary 16 applies to arbitrary values of m, u and only requires $\ell(\cdot)$ be bounded. The main difference between our proof and Catoni's proof is the way to tackle the dependence between training error and test error. Specifically, we use the assumption of exchangeable distribution to build a connection between the random sampling setting and random splitting setting and then tackle this dependence by the martingale approach, as we show in Eq. (22). Different from ours, Theorem 3.1.2 of Catoni (2007) requires an extra assumption that $u = km, k \in \mathbb{N}_+$, and uses the assumption of partially exchangeable distribution to connect the training error with total error, by which the dependence is eliminated. His key insight is that under the case $u = km, k \in \mathbb{N}_+$, permuting the sequence that contains $(k+1)m$ data points can be achieved by firstly dividing them into k subsequences and then permuting the data points within each subsequence separately. Finally, merging these subsequences together gives a permutation on the entire sequence, as we show in Eq. (23). Further, more techniques can be adopted under the assumption that $m = u$, which enables us to connect the transductive generalization gap with the empirical measure associated with the full data points, as shown in Theorem 18. However, for more general cases where u and m are arbitrary integers, adopting the selector sequence U is not sufficient to depict the randomness, resulting in the technique of Catoni (2003, 2007); Audibert and Bousquet (2007) not applicable. The reason is that the number of errors observed in S follows a hypergeometric distribution rather than a binomial distribution (Bégin et al., 2014). In contrast, the martingale approach we use still works under these cases, and it can also be applied to scenarios where the objective is a function of matrices (See Section 4.5 for more details). To summarize, we improve previous results by relaxing the assumption on both the type of loss function and the values of m and u. Additionally, the technique we adopt has a broad applicability.

It is worth pointing out that introducing a selector sequence $(U_1, \ldots, U_m)$ to depict the randomness of selecting training labels is motivated from the CMI framework (Steinke and Zakynthinou, 2020), which itself could be tracked back to Catoni's circular permutation technique (Catoni, 2007, Chapter 3.1), as revealed recently in Chapter 6.7 of Hellström et al. (2023). However, we emphasize that the transductive supersample $\widetilde{Z}$ we propose is essentially a new concept. Our key motivation is that, under the random split setting, we could disentangle the uncertainty of training label selection into two parts. One is contained in $\widetilde{Z}$, and the other is contained in U, and $\widetilde{Z}$ is independent of U. Thus, we only need to perform sampling without replacement to obtain $\widetilde{Z}$ and then apply U to permute it. This allows us to tackle the dependence induced by sampling without replacement. Also, for fixed realizations of $\widetilde{Z}$ that is given by $\widetilde{z}_i = (i, m+i, \ldots, km+i), i \in [m]$, we recover the definition of transductive risk under the random sampling setting. Therefore,

the transductive supersample concept serves as a connection between the random sampling setting and random splitting setting and allows us to transport results from the former to the latter. Moreover, we remark that many techniques used in the inductive learning setting are still applicable in the transductive learning setting, for example, using chaining technique (Audibert and Bousquet, 2007) or Bernstein assumption (Grünwald et al., 2021) to derive bounds with a faster rate. Correspondingly, our derived bounds can be further improved by applying these techniques. Recently, Rodríguez-Gálvez et al. (2024) extend Catoni's results under mild assumptions on the loss function. However, these results are derived under the inductive learning setting. It remains unclear whether they apply to the transductive learning setting.

4.5 Upper Bounds for Adaptive Optimization Algorithms

As noted previously, one of the advantages of our theoretical results over previous results is that they fully account for the impact of the optimization algorithm on generalization. In this section, we demonstrate this advantage by analyzing AdaGrad (Duchi et al., 2011) in the context of transductive learning. Unlike SGD, AdaGrad features an adaptively adjusted learning rate throughout the training process. Let $\{W_t\}_{t=1}^T$ represent the weight sequences on the training trajectory of AdaGrad. Following the work of Wang and Mao (2022), we adopt the setting that mini-batches data points are fixed. Denote $(B_1, \ldots, B_T)$ as the sequence of mini-batches, with $B_t, t \in [T]$ representing the set of data points used in the t-th iteration. For simplicity, we assume that the model only minimizes the loss on labeled data points, that is, $B_t = B_t(Z) \subseteq \{s_{Z_i}\}_{i=1}^m$ for $t \in [T]$. And, we assume that each mini-batch contains exactly b data points. The average gradient on $B_t, t \in [T]$ is defined as

$$g(w, B_t(Z)) \triangleq \frac{1}{b} \sum_{s \in B_t(Z)} \nabla_w \ell(w, s),$$

where $\ell(\cdot)$ represents the objective function. Notice that the loss is computed exclusively on the labeled data point. It might appear that the features of unlabeled data points are not used. However, we point out that in certain scenarios, the features of unlabeled data are implicitly used by the model during loss computation, as seen in transductive graph learning (see Section 5.2 for more details). Additionally, proxy loss on unlabeled data points, such as using pseudo labels generated by the model itself to compute the loss, can be included in the objective function. This would not affect the formulation of the final theoretical result. Recall that the update rule of AdaGrad can be formulated as

$$V_t = \sum_{k=0}^{t-1} g(W_k, B_k(Z)) \odot g(W_k, B_k(Z)), W_t = W_{t-1} - \frac{\eta}{\sqrt{V_t} + \epsilon} \odot g(W_{t-1}, B_t(Z)), t \in [T],$$

where W_0 is the initial parameter and η, ϵ are two predefined hyper-parameters. Notice that V_t is a random variable determined by the sequence $W^{[t-1]} \triangleq (W_0, \ldots, W_{t-1})$ and Z. For notational simplicity, we denote the "adaptive gradient" as $\Psi(W^{[t-1]}, Z) \triangleq \big(\eta/(\sqrt{V_t} + \epsilon)\big) \odot g(W_{t-1}, B_t(Z))$, which is computed by normalizing the current gradient with accumulate squared gradient. Inspired by the work of Neu et al. (2021) and Wang and Mao (2022), we introduce the following auxiliary weight process $\{\widetilde{W}_t\}_{t=1}^T$ defined as

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

where $\{\sigma_t\}_{t\in[T]}$ are predefined hyperparameters and N is a standard Gaussian random variable independent of $W^{[T]}$ and Z. Denote by $N_{1:t} \triangleq \sum_{k=1}^{t} N_k$, then the expectation of the transductive generalization gap $\mathbb{E}_{Z,W_T}\left[R_{\text{test}}(W_T, Z) - R_{\text{train}}(W_T, Z)\right]$ can be decomposed into three terms:

$$\begin{aligned}
&\mathbb{E}_{Z,W_T}\left[R_{\text{test}}(W_T, Z) - R_{\text{train}}(W_T, Z)\right] \\
=&\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{test}}(W_T + N_{1:T}, Z) - R_{\text{train}}(W_T + N_{1:T}, Z)\right] \\
&+ \mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{train}}(W_T + N_{1:T}, Z) - R_{\text{train}}(W_T, Z)\right] \\
&- \mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{test}}(W_T + N_{1:T}, Z) - R_{\text{test}}(W_T, Z)\right].
\end{aligned} \tag{26}$$

We first discuss how to derive upper bounds for the second and third terms in Eq. (26), which require additional assumptions. The first approach relies on the following assumption: $\mathbb{E}_{Z,N_{1:T}}\left[R_{\text{test}}(w_T + N_{1:T}, Z) - R_{\text{test}}(w_T, Z)\right] \geq 0$ holds for any realization w_T of W_T. This assumption means that adding random noise to the final learned parameter W_T does not, on average, reduce the transductive error on unlabeled data points. A similar assumption is also used by Foret et al. (2021); Wang and Mao (2022) and in Corollary 13 to establish information-theoretic or PAC-Bayesian bounds. Under this assumption, the third term in Eq. (26) can be omitted since

$$\begin{aligned}
&\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{test}}(W_T + N_{1:T}, Z) - R_{\text{test}}(W_T, Z)\right] \\
=&\int_{w_T} \mathbb{E}_{Z,N_{1:T}}\left[R_{\text{test}}(w_T + N_{1:T}, Z) - R_{\text{test}}(w_T, Z)|W_T = w_T\right] \mathrm{d}P_{W_T}(w_T) \geq 0.
\end{aligned} \tag{27}$$

The second term $\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{train}}(W_T + N_{1:T}, Z) - R_{\text{train}}(W_T, Z)\right]$ is directly remained, which reflects the flatness of the training loss landscape around W_T. The second approach relies on the assumption that the spectral norm of the Hessian is bounded, which enables us to provide an upper bound for the trace of the Hessian. By applying Taylor's expansion on R_{test} and R_{train} respectively and neglecting high order terms, we can derive upper bounds for the second and third terms in Eq. (26). We present this result in the following theorem.

Theorem 19 *Suppose that (i) the loss $\ell(w, s)$ is second order differential w.r.t. w and (ii) $\sup_s \left\|\frac{\partial^2 \ell(w,s)}{\partial w^2}\right\| \leq B_H$ that holds for all $w \in \mathcal{W}$ and $s \in \{s_i\}_{i=1}^n$. Then we have*

$$\begin{aligned}
&\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{train}}(W_T + N_{1:T}, Z) - R_{\text{train}}(W_T, Z)\right] \\
\leq&\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{test}}(W_T + N_{1:T}, Z) - R_{\text{test}}(W_T, Z)\right] \\
&+ \left(\sum_{t=1}^{T} \sigma_t^2\right) \sqrt{\frac{32 d^2 B_H^2 C_{m,u}(I(W_T; Z) + \log(d))(m+u)}{mu}} + \mathcal{O}\left(\left(\sum_{t=1}^{T} \sigma_t^2\right)^2\right).
\end{aligned}$$

In the proof of Theorem 19, we establish a concentration inequality for the largest eigenvalue of the Hessian associated with the transductive generalization gap by the matrix martingale technique. Once this concentration inequality is obtained, various methods such as covering numbers (Ju et al., 2022) or transductive Rademacher complexity (El-Yaniv and Pechyony, 2009) can be applied to bound the second and the third terms in Eq. (26). To align with the motivation of this paper, we derive an upper bound within the framework of information theory. Notice that B_H represents the maximum sharpness that may occur

across the entire loss landscape. Therefore, although the assumption of the second approach differs from that of the first, it conveys the same geometric meaning: a flatter optima implies a smaller transductive generalization gap. This interpretation is consistent with both the first approach and Theorem 10. Finally, inspired by the work of Neu et al. (2021); Wang and Mao (2022), we derive the following upper bound for the first term in Eq. (26).

Theorem 20 *Under the assumption of Theorem 1, we have*

$$\begin{aligned}&\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{test}}(W_T+N_{1:T},Z)-R_{\text{train}}(W_T+N_{1:T},Z)\right]\\ \leq&\frac{1}{2}\sqrt{\frac{dC_{m,u}(m+u)}{mu}\sum_{t=1}^{T}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}_{W_0,\ldots,W_{t-1},Z}\left[\left\|\Psi(W_0,\ldots,W_{t-1},Z)\right\|_2^2\right]+1\right)}.\end{aligned}\tag{28}$$

The right hand of Eq. (28) is described by the norm square of the "adaptive gradient" on the training trajectory. The expectation is bounded if the gradients at each point on this training trajectory are bounded, a condition that can generally be satisfied in practice. The behavior of this expectation depends on the specific learning algorithms and training data we use. Although Theorem 20 does not provide a convergence guarantee towards a minimum, it allows us to estimate the generalization behavior of the model from the perspective of optimization trajectory. Analyzing the optimization property of adaptive optimization algorithms under the transductive learning setting goes beyond the scope of this paper, and we left it for future work. Compared with results derived from algorithm stability (Cong et al., 2021) or complexity of hypothesis space (Tang and Liu, 2023), Theorem 20 does not include any Lipschitz or smoothness constants. Since the Lipschitz constant serves as the upper bound of the norm of gradients, our result can more finely depict the impact of optimization algorithms on generalization ability. Another advantage of our result is that it does not require the activation function to be smooth (Cong et al., 2021) or Hölder smooth (Tang and Liu, 2023). Therefore, it can be applied to neural networks using ReLU as the activation function. Moreover, we also derive analogous results for Adam (Kingma and Ba, 2015), given its widespread use in practice. Since their formulations and reflected insights are similar to those of Theorem 20, we place them in Appendix M.

5. Applications

5.1 Semi-supervised Learning

Due to the expensive cost of collecting high-quality labeled data, semi-supervised learning (Shahshahani and Landgrebe, 1994; Blum and Mitchell, 1998; Joachims, 1999; Zhu et al., 2003) has attracted increasing attention, whose goal is to train a model using a small amount of labeled data and a large volume of unlabeled data. Establishing generalization guarantees for semi-supervised learning algorithms has been extensively studied (Mey and Loog, 2023), with theoretical results varying based on the problem setting and underlying assumptions. In this section, we adopt the random sampling setting of transductive learning as outlined in Section 4.4, assuming that $u = km, k \in \mathbb{N}_+$. Recall that under this setting, the full data points are defined as $S = (S_1, \ldots, S_{(k+1)m})$, where $S_i \triangleq (X_i, Y_i), i \in [(k+1)m]$. To transport the results derived under the random splitting setting to the random sampling setting, we follow the work of Vapnik (1982) and assume that $S_i \sim P_S, i \in [(k+1)m]$,

meaning that each data point is independently drawn from the same distribution. This assumption about the data distribution is slightly stronger than the partially exchangeable assumption but allows us to derive results more conveniently. And, the numerical values of them are easy to estimate. Let us start from the case where $u = m$. Denote by $Z = (Z_1, \ldots, Z_{2m})$ the sequence obtained by sampling without replacement from $[2m]$, and let $U \triangleq (U_1, \ldots, U_m) \sim \mathrm{Unif}(\{0,1\})^m$ represent the selector sequence. We have

$$\begin{aligned}
\mathbb{E}_{W,S}\left[\mathcal{E}(W,S)\right] =& \mathbb{E}_{W,S}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell(W,S_i) - \frac{1}{m}\sum_{i=1}^{m}\ell(W,S_i)\right] \\
=& \mathbb{E}_{W,S,Z}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell(W,S_{Z_i}) - \frac{1}{m}\sum_{i=1}^{m}\ell(W,S_{Z_i})\right] \\
=& \mathbb{E}_{W,S,\widetilde{Z},U}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell\big(W,S_{\mathscr{Z}_i(\widetilde{Z},U)}\big) - \frac{1}{m}\sum_{i=1}^{m}\ell\big(W,S_{\mathscr{Z}_i(\widetilde{Z},U)}\big)\right],
\end{aligned} \tag{29}$$

where the second and third equality are obtained by the assumption $S_i \sim P_S, i \in [2m]$ and Proposition 6, respectively. Notice that Eq. (29) can be viewed as the expectation of the transductive generalization gap $\mathbb{E}_{W,\widetilde{Z},U}[\mathcal{E}(W,\mathscr{Z}(\widetilde{Z},U))]$ over the data points $(S_1,\ldots,S_{2m})$. In other words, the transductive generalization gap studied in Section 4.2 can be expressed as $\mathbb{E}_{W,\widetilde{Z},U|S=s}[\mathcal{E}(W,\mathscr{Z}(\widetilde{Z},U))]$. Based on the results derived in Section 4.2, we obtain the following information-theoretic bounds for semi-supervised learning algorithms.

Proposition 21 *Suppose that $r(\hat{y},y) \in [0,B]$ holds for all $\hat{y} \in \widehat{\mathcal{Y}}$ and $y \in \mathcal{Y}$, where $B > 0$ is a constant. Also, suppose that $S_1,\ldots,S_{2m} \sim P_S$. Denote by $f_w(X) \in \mathbb{R}^K$ the prediction of the model and $F_i \triangleq (f_W(X_{\widetilde{Z}_{i,0}}), f_W(X_{\widetilde{Z}_{i,1}}))$ the sequence of predictions. Denote by $L_{i,:} \triangleq (\ell(W,S_{\widetilde{Z}_{i,0}}), \ell(W,S_{\widetilde{Z}_{i,1}}))$ the sequence of loss values and $\Delta_i \triangleq \ell(W,S_{\widetilde{Z}_{i,0}}) - \ell(W,S_{\widetilde{Z}_{i,1}})$ the difference of loss value. We have*

$$\left|\mathbb{E}_{W,S}\left[R_{\mathrm{test}}(W,S) - R_{\mathrm{train}}(W,S)\right]\right| \le \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S,\widetilde{Z}}\sqrt{2I^{S,\widetilde{Z}}(F_i;U_i)}, \tag{30}$$

$$\left|\mathbb{E}_{W,S}\left[R_{\mathrm{test}}(W,S) - R_{\mathrm{train}}(W,S)\right]\right| \le \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S,\widetilde{Z}}\sqrt{2I^{S,\widetilde{Z}}(L_i;U_i)}, \tag{31}$$

$$\left|\mathbb{E}_{W,S}\left[R_{\mathrm{test}}(W,S) - R_{\mathrm{train}}(W,S)\right]\right| \le \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{S,\widetilde{Z}}\sqrt{2I^{S,\widetilde{Z}}(\Delta_i;U_i)}. \tag{32}$$

Proposition 21 extends Corollary 8 to the random sampling setting. Notice that the mutual information terms in Proposition 21 are conditioned on two random variables S and $\widetilde{Z}$, whereas in Corollary 8, they are conditioned only on $\widetilde{Z}$. The reason for this difference is that Corollary 8 is derived under the random splitting setting, where the training data is selected on a fixed set of data points. Thus, the only source of randomness comes from the sampling without replacement process represented by $\widetilde{Z}$. In contrast, for the random sampling setting, we first sample a sequence of data points from P_S and then divide them into training and test sets according to $\widetilde{Z}$, which is obtained by sampling

without replacement from $[2m]$. Therefore, the randomness of training data is depicted by both $(S_1, \ldots, S_{2m})$ and Z. Additionally, we point out that the bounds converge as the number of test data points u increases, given our assumption that $u = m$. Following the same procedure, we can obtain results similar to Eqs. (30) and (31) for the common case where $u = km, k \in \mathbb{N}_+$. These results and detailed derivations are provided in Appendix N.

5.2 Transductive Graph Learning

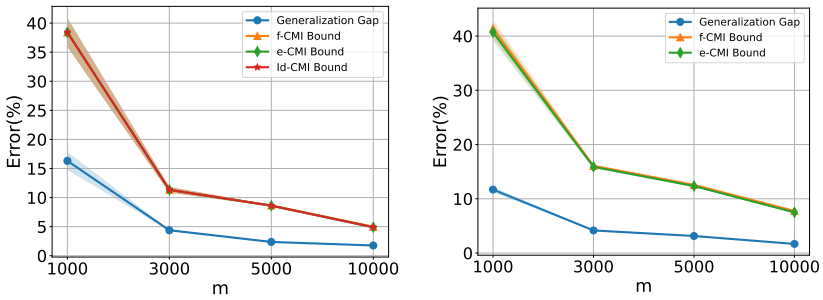
Graph-structured data, composed of multiple objects and their relationships, plays a crucial role in various real-world applications, such as recommendation systems (Wang et al., 2019b; He et al., 2020; Huang et al., 2021), drug discovery (Sun et al., 2020; Bongini et al., 2021), and traffic flow forecasting (Song et al., 2020; Li and Zhu, 2021; Lan et al., 2022). Graph learning tasks can be categorized into transductive tasks and inductive tasks. In this paper, we focus on transductive graph tasks, which can be further divided into node-level tasks and edge-level tasks. For transductive node-level tasks, some nodes are randomly selected from a fixed graph and their labels are revealed to the model. The learning goal is to predict the labels of the remaining nodes based on the revealed labels, features of all nodes, and the graph structure. For transductive edge-level tasks, sub-graphs (that is, subsets of full nodes along with their edges) are sampled from a fixed graph and provided to the model. The learning objective is to predict the connections between nodes that are not included in the sampled sub-graphs. Both types of transductive learning tasks fit within the random splitting setting of transductive learning, allowing us to apply the results derived in this work to analyze the generalization abilities of GNNs on these tasks.

We now demonstrate the application of the derived results on the transductive node classification task. Assume that the model is a two-layer GCN (Kipf and Welling, 2017). Let $\widetilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ be the normalized adjacent matrix with self-loops and $\mathbf{X} \in \mathbb{R}^{n\times d_0}$ be the node feature matrix. Let $\mathbf{W}_1 \in \mathbb{R}^{d_0\times d_1}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_1\times|\mathcal{Y}|}$ be the learnable parameters, whose collection is denoted by $\mathbf{W} \triangleq [\mathrm{vec}[\mathbf{W}_1], \mathrm{vec}[\mathbf{W}_2]]$. Here, vec$[\cdot]$ is the vectorization operator that reshapes a matrix into a column vector. The prediction of the model is $\widehat{\mathbf{Y}} = \mathrm{Softmax}(\widetilde{\mathbf{A}}\mathrm{ReLU}(\mathbf{H})\mathbf{W}_2) \in \mathbb{R}^{n\times|\mathcal{Y}|}$ with $\mathbf{H} \triangleq \widetilde{\mathbf{A}}\mathbf{X}\mathbf{W}_1$. Without loss of generality, we assume that the sequence sampled without replacement from $[n]$ is given by $Z = (1, \ldots, n)$. Let $\mathbf{Y} \in \{0,1\}^{n\times|\mathcal{Y}|}$ be the label matrix where each row represents the one-hot vector of the corresponding node's label. Suppose that $\ell(\cdot)$ is the cross-entropy loss, then the transductive training error is $R_{\mathrm{train}}(\mathbf{W}, Z) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{|\mathcal{Y}|}\mathbf{Y}_{ij}\log\widehat{\mathbf{Y}}_{ij}$. The gradient of $R_{\mathrm{train}}(\mathbf{W}, Z)$ is given by $g(\mathbf{W}, Z) = \left[\frac{\partial R_{\mathrm{train}}(\mathbf{W},Z)}{\partial\mathrm{vec}[\mathbf{W}_1]}, \frac{\partial R_{\mathrm{train}}(\mathbf{W},Z)}{\partial\mathrm{vec}[\mathbf{W}_2]}\right]^\top$ with

$$\frac{\partial R_{\mathrm{train}}(\mathbf{W}, Z)}{\partial\mathrm{vec}\left[\mathbf{W}_1\right]} = \frac{1}{m}\sum_{i=1}^{m}(\widehat{\mathbf{Y}}_{i,:} - \mathbf{Y}_{i,:}) \otimes \left(\sum_{j=1}^{n}\widetilde{\mathbf{A}}_{ij}\mathrm{ReLU}(\mathbf{H}_{j,:})\right),$$

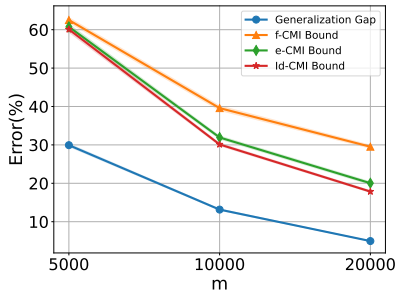
$$\frac{\partial R_{\mathrm{train}}(\mathbf{W}, Z)}{\partial\mathrm{vec}\left[\mathbf{W}_2\right]} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\widetilde{\mathbf{A}}_{ij}\left(\mathrm{ReLU}'\left(\sum_{k=1}^{n}\mathbf{H}_{k,:}\right) \odot \left((\widehat{\mathbf{Y}}_{i,:} - \mathbf{Y}_{i,:})\mathbf{W}_2^\top\right)\right) \otimes \mathbf{H}_{j,:},$$

where $\mathrm{ReLU}'(\cdot)$ is the derivative of the ReLU function. Here we assume that the transductive training error is computed over all labeled nodes, which is a common setting in practice (Kipf and Welling, 2017; Gasteiger et al., 2019; Chien et al., 2021). By substituting $g(\mathbf{W}, Z)$ into

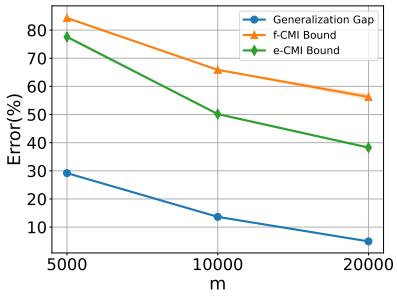


(a) MNIST with $k = 1$

(b) MNIST with $k = 2$



(c) CIFAR-10 with $k = 1$



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

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

Theorem 20, we can derive an upper bound for GCNs trained with adaptive optimization algorithms. Notably, the influence of GNN architectures and graph-structured data on model generalization is captured by the norm of the gradients. Following the analysis of Cong et al. (2021); Tang and Liu (2023), one can derive fine-grained upper bounds for other GNN models and gain insights into their generalization behavior.

6. Experiments

6.1 Experimental Setup

For semi-supervised learning, we select image classification on the MNIST and CIFAR-10 datasets as learning tasks. Following the work of Guo et al. (2020), the loss for unlabeled images is defined as the mean square error between the prediction of augmented images and their vanilla (non-augmented) counterparts by the model, which is known as a consistency regularization in the field of semi-supervised learning. The loss for labeled images is the cross-entropy loss. Following the work of Harutyunyan et al. (2021); Guo et al. (2020), we utilize a four-layer CNN and a Wide ResNet-28-10 (Zagoruyko and Komodakis, 2016) as the model for MNIST and CIFAR-10, respectively. For both datasets, we train the model for 1000 iterations using the Adam optimizer. The learning rate is set to 0.001, and each mini-batch contains 128 images. The objective function of training is the summation of losses on labeled and unlabeled losses. For evaluation, we use the zero-one loss as the

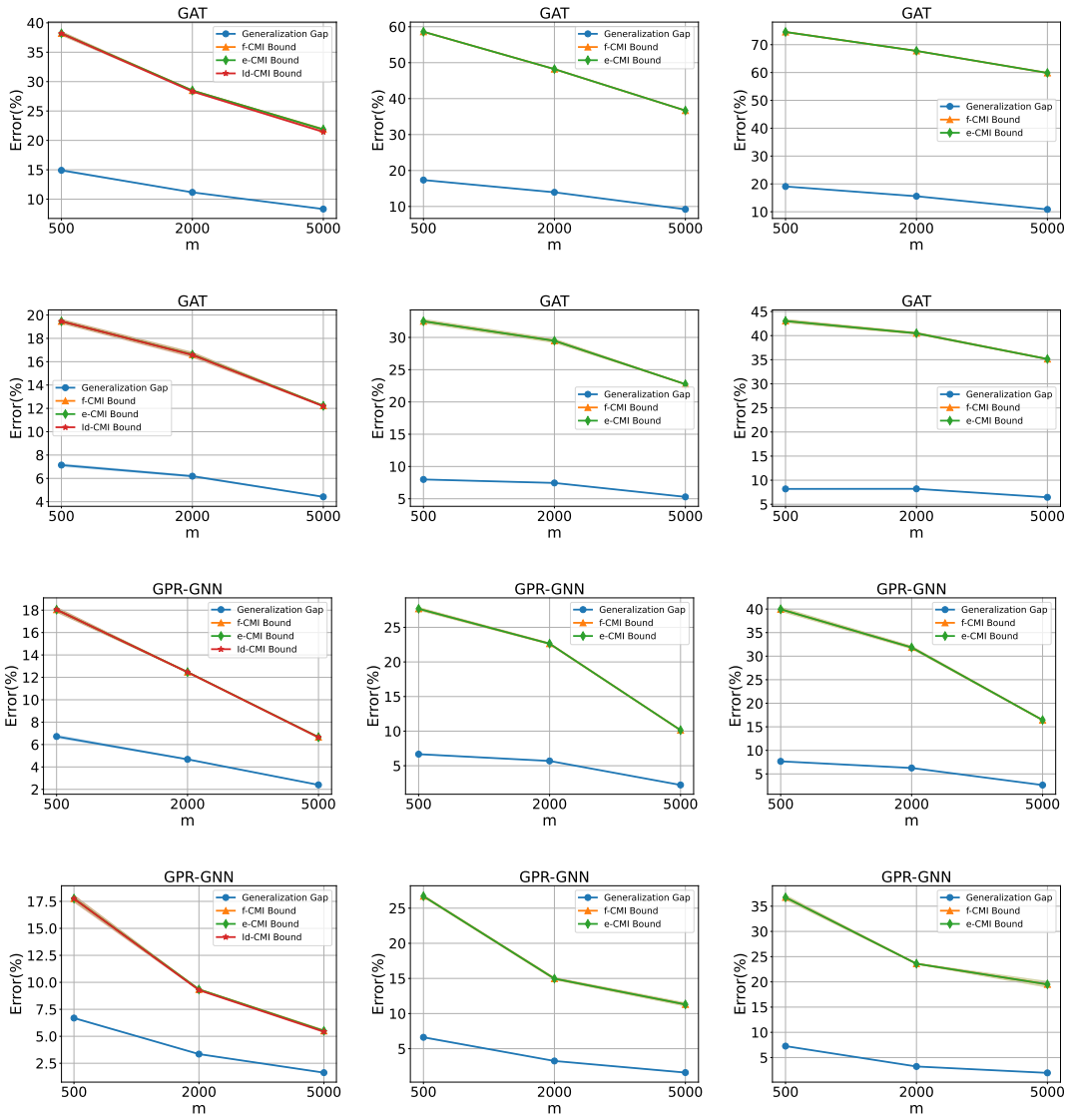
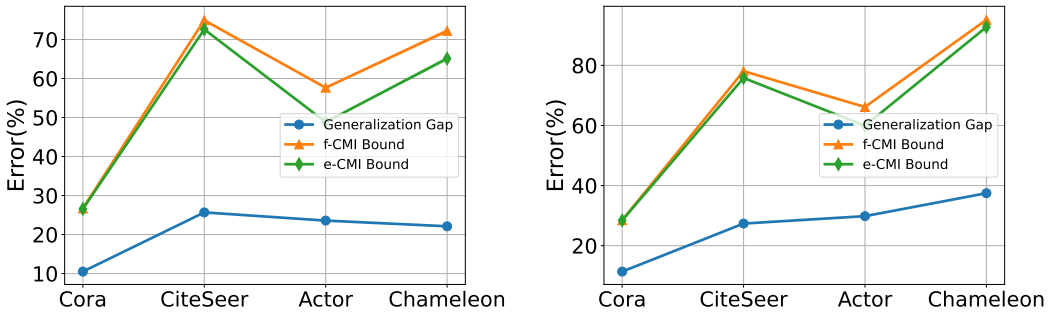


Figure 2: Estimations of the transductive generalization gap and the derived bounds on cSBMs with GAT and GPR-GNN. The first (second) and third (fourth) rows correspond to $\phi = -0.5$ ($\phi = 0.5$). The left, middle, and right figures in each row correspond to $k = 1$, $k = 2$ and $k = 3$.

criterion. Following the work of Harutyunyan et al. (2021), we make the training process deterministic by fixing the sequence of mini-batch and the initialization of model parameters through random seeds. For transductive graph learning, we select semi-supervised node classification on both synthetic and real-world datasets as learning tasks. Specifically, we use cSBMs (Deshpande et al., 2018) as synthetic data and Cora, CiteSeer (Sen et al., 2008; Yang et al., 2016), Actor, and Chameleon as real-world data. For each dataset, we adopt GAT (Veličković et al., 2018) and GPR-GNN (Chien et al., 2021) as the models, since



(a) GAT (b) GPR-GNN

Figure 3: Estimations of the transductive generalization gap and the derived bounds on real-world datasets with GAT and GPR-GNN.

they are regarded as representative models of spatial and spectral GNNs, respectively. In experiments on both synthetic and real-world datasets, we train the model for 300 iterations using the Adam optimizer with a learning rate of 0.01 without weight decay. Following the setting of Kipf and Welling (2017); Gasteiger et al. (2019); Chien et al. (2021), the GNN model passes over all labeled nodes during each iteration. Additional experimental details, including methods for estimating the expected generalization gap and the derived bounds, are provided in Appendix O.

6.2 Experimental Results

Figure 1 illustrates the results of semi-supervised learning algorithms on the MNIST and CIFAR-10 datasets, where m represents the number of labeled images, and k denotes the ratio of unlabeled to labeled images. The value of f-CMI, e-CMI, and Id-CMI bounds are calculated using Eqs. (30), (31) and (32), respectively. For cases where $k \geq 2$, we report only the f-CMI and e-CMI bound values. It can be observed that the numerical values of our established bounds are non-vacuous, with the discrepancy between the estimated value and the generalization gap decreasing as m increases. Conversely, this discrepancy tends to increase when k grows larger. This phenomenon occurs because a higher k value results in greater estimation error for the conditional mutual information, as discussed in Section 4.2. Additionally, it is observed that the e-CMI bound does not exceed the f-CMI bound, and the Id-CMI bound does not surpass the e-CMI bound for $k = 2$. These observations align with the findings of Hellström and Durisi (2022); Wang and Mao (2023a), and they remain valid in the transductive learning setting. The results of transductive graph learning are presented in Figures 2 and 3. Here, the f-CMI, e-CMI and Id-CMI bounds are computed according to Eqs. (13), (14), and (15). Generally, the trends observed are consistent with those seen in semi-supervised learning. Notably, in Figure 2, the estimated values of the f-CMI, e-CMI, and Id-CMI bounds are nearly identical, which could be due to the small magnitude of the generalization gap. In contrast, as shown in Figures 1(c) and 1(d), when the generalized gap is large, the differences among these information-theoretic bounds become more apparent.

7. Conclusion

In this work, we investigate the generalization ability of transductive learning algorithms. We develop a variety of generalization upper bounds under both the random sampling setting and the random splitting setting within the framework of information theory and PAC-Bayes. To address the dependency between training and test data points due to sampling without replacement, our key techniques involve employing martingale methods and introducing the concept of transductive supersamples. We anticipate that the findings and methodologies presented in this work will shed light on understanding and analyzing the generalization performance of transductive learning algorithms. Promising avenues for future research include extending our theoretical results to additional transductive learning scenarios and devising novel information measures tailored to the transductive learning setting. It is worth noting that while our study emphasizes establishing generalization upper bounds for transductive learning algorithms, further efforts to explore lower bounds remain a worthwhile area of investigation.

Acknowledgements

We gratefully appreciate the editor and anonymous referees for their valuable and constructive comments. This research was supported by the National Natural Science Foundation of China (No.62476277), the National Key Research and Development Program of China (No.2024YFE0203200), the CCF-ALIMAMA TECH Kangaroo Fund (No.CCF-ALIMAMA OF 2024008), and the Huawei-Renmin University joint program on Information Retrieval. We also acknowledge the support provided by the fund for building worldclass universities (disciplines) of Renmin University of China and by the funds from Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, from Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, from Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "DoubleFirst Class" Initiative, Renmin University of China, from Public Policy and Decision-making Research Lab of Renmin University of China, and from Public Computing Cloud, Renmin University of China.

Appendix A. Lemma

Lemma 22 (Polyanskiy and Wu, 2025, Theorem 4.6) *Let P and Q be two probability measures on the same measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra on $\mathcal{X}$. Denote by $\mathcal{F} \triangleq \{f : \mathcal{X} \to \mathbb{R}\}$ the family of bounded measurable function on $\mathcal{X}$. Then we have*

$$\mathrm{D}_{\mathrm{KL}}(P||Q) = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_P \left[f(X) \right] - \log \mathbb{E}_Q \left[\exp \left\{ f(X) \right\} \right] \right\}. \tag{33}$$

Denote by $\mathcal{P}(\mathcal{X})$ the set that contains all probability distributions on $(\mathcal{X}, \mathcal{A})$. For any $f \in \mathcal{F}$ we have

$$\log \left(\mathbb{E}_Q[\exp(f(X))] \right) = \sup_{P \in \mathcal{P}(\mathcal{X})} \left\{ \mathbb{E}_P[f(X)] - \mathrm{D}_{\mathrm{KL}}(P||Q) \right\}. \tag{34}$$

Appendix B. Proof of Theorem 1

Proof. We first establish an upper bound for the moment-generating function $\mathbb{E}_Z \left[e^{\lambda \mathcal{E}(w,Z)} \right]$ by the martingale technique for any $w \in \mathcal{W}$. Without loss of generality, we assume that $u \geq m$. Inspired by the work of El-Yaniv and Pechyony (2006); Cortes et al. (2008); El-Yaniv and Pechyony (2009), we construct the following Doob's martingale difference sequences

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

One can verify that $\mathcal{E}(w,Z) - \mathbb{E}[\mathcal{E}(w,Z)] = \sum_{i=1}^{n} \xi_i$ and $\mathbb{E}[\xi_i|Z_1, \ldots, Z_{i-1}] = 0$ holds. Notice that ξ_i is a function of $Z_1, \ldots, Z_i$. For $i \in [n]$, define

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

we have $\xi_i^{\inf} \leq \xi_i \leq \xi_i^{\sup}$. One can find that for $i \in [n]$,

$$\begin{aligned}
&\xi_i^{\sup} - \xi_i^{\inf} \\
=&\sup_z \mathbb{E}[\mathcal{E}(w,Z)|Z_1, \ldots, Z_{i-1}, Z_i = z] - \inf_z \mathbb{E}[\mathcal{E}(w,Z)|Z_1, \ldots, Z_{i-1}, Z_i = z] \\
=&\sup_{z,\tilde{z}} \Big\{ \mathbb{E}[\mathcal{E}(w,Z)|Z_1, \ldots, Z_{i-1}, Z_i = z] - \mathbb{E}[\mathcal{E}(w,Z)|Z_1, \ldots, Z_{i-1}, Z_i = \tilde{z}] \Big\}.
\end{aligned} \tag{35}$$

Notice that once $Z_1, \ldots, Z_m$ are given, the values of $Z_{m+1}, \ldots, Z_n$ do not affect the value of $\mathcal{E}(w,Z)$. Thus, $\xi_i^{\sup} - \xi_i^{\inf} = 0$ holds for $i = m+1, \ldots, n$. Now we discuss the case that $i \in [m]$. To proceed with the proof, we first illustrate the meaning of Eq. (35) by considering a fixed realization z_j of Z_j with $j \in [i-1]$. Denote by z_i and $\tilde{z}_i$ the realizations of Z_i and $\widetilde{Z}_i$, respectively. Let $Z_{i+1:n} = (z_1, \ldots, z_{i-1}, z_i, Z_{i+1}, \ldots, Z_n)$ be the sequence where $Z_{i+1}, \ldots, Z_n$ are obtained by sampling without replacement from $[n] \backslash \{z_1, \ldots, z_{i-1}, z_i\}$, and $\widetilde{Z}_{i+1:n} = (z_1, \ldots, z_{i-1}, \tilde{z}_i, \widetilde{Z}_{i+1}, \ldots, \widetilde{Z}_n)$ be the sequence where $\widetilde{Z}_{i+1}, \ldots, \widetilde{Z}_n$ are obtained by sampling without replacement from $[n] \backslash \{z_1, \ldots, z_{i-1}, \tilde{z}_i\}$. The meaning of Eq. (35) is the maximum value of the expectation $\mathbb{E}[\mathcal{E}(w, Z_{i+1:n}) - \mathcal{E}(w, \widetilde{Z}_{i+1:n})]$ over any possible values $z_1, \ldots, z_{i-1}, z_i, \tilde{z}_i$. To this end, it is sufficient to consider two cases: (1) z_i appears at positions $i+1$ to m of $\widetilde{Z}_{i+1:n}$, and (2) z_i appears at positions $m+1$ to n of $\widetilde{Z}_{i+1:n}$.

For case (1), the expectation is equal to zero since for each realization of $Z_{i+1:n}$ we can always find a corresponding and unique realization of $\widetilde{Z}_{i+1:n}$ such that they are equal to each other. For case (2), the maximum of the expectation is $\frac{m+u}{mu}[\ell(w, s_{\tilde{z}_i}) - \ell(w, s_{z_i})]$. To see this, notice that for each realization of $Z_{i+1:n}$, we can always find a corresponding and unique realization of $\widetilde{Z}_{i+1:n}$ such that $Z_{i+1:n}$ and $\widetilde{Z}_{i+1:n}$ only differs at the i-th position. The last step is to compute the probability of z_i appearing at positions $m+1$ to n of $\widetilde{Z}_{i+1:n}$. To this end, we need to sample $m-i$ elements among the rest $n-i-1$ elements (that is, the set $[n]\backslash\{z_1, \ldots, z_i, \tilde{z}_i\}$), and then apply permutation on them. Thus, the probability is given by $\frac{u!(m-i)!C_{n-i-1}^{m-i}}{(n-i)!}$. Putting all above ingredients together and notice that $\frac{m+u}{mu}[\ell(w, s_{\tilde{z}_i}) - \ell(w, s_{z_i})] \leq \frac{(m+u)B}{mu}$, we obtain

$$\xi_i^{\sup} - \xi_i^{\inf} = \begin{cases} \frac{u!(m-i)!C_{n-i-1}^{m-i}}{(n-i)!} \cdot \frac{(m+u)B}{mu} = \frac{(m+u)B}{m(m+u-i)}, & i = 1, \ldots, m, \\ 0, & i = m+1, \ldots, n. \end{cases} \tag{36}$$

Then we have

$$\begin{aligned} \mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n} \xi_i\right\}\right] =& \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n-1} \xi_i\right\} \exp\{\lambda \xi_n\} \,\middle|\, Z_1, \ldots, Z_{n-1}\right]\right] \\ =& \mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n-1} \xi_i\right\} \mathbb{E}\left[\exp\{\lambda \xi_n\} \,|Z_1, \ldots, Z_{n-1}\right]\right] \\ \leq& \exp\left\{\frac{\lambda^2 (\xi_n^{\sup} - \xi_n^{\inf})^2}{8}\right\} \mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n-1} \xi_i\right\}\right], \end{aligned} \tag{37}$$

where the second line is due to the tower property of conditional expectation and the last line is obtained by Hoeffding's inequality (Hoeffding, 1963). By iteration we obtain

$$\begin{aligned} & \mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n} \xi_i\right\}\right] \\ \leq& \exp\left\{\frac{\lambda^2}{8} \sum_{i=1}^{n} (\xi_i^{\sup} - \xi_i^{\inf})^2\right\} = \exp\left\{\frac{\lambda^2 B^2 (m+u)^2}{8m^2} \sum_{i=1}^{m} \frac{1}{(m+u-i)^2}\right\} \\ =& \exp\left\{\frac{\lambda^2 B^2 (m+u)^2}{8m^2} \sum_{i=u}^{m+u-1} \frac{1}{i^2}\right\} < \exp\left\{\frac{\lambda^2 B^2 (m+u)^2}{8m^2} \sum_{i=u}^{m+u-1} \frac{1}{i^2 - 1/4}\right\} \\ =& \exp\left\{\frac{\lambda^2 B^2 (m+u)^2}{8m^2} \sum_{i=u}^{m+u-1} \left(\frac{1}{i-1/2} - \frac{1}{i+1/2}\right)\right\} \\ =& \exp\left\{\frac{\lambda^2 B^2 (m+u)^2}{8m(u-1/2)(m+u-1/2)}\right\}. \end{aligned} \tag{38}$$

For the case that $u \leq m$, by the symmetry of m and u, we have

$$\mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^{n} \xi_i\right\}\right] \leq \exp\left\{\frac{\lambda^2 B^2 (m+u)^2}{8u(m-1/2)(m+u-1/2)}\right\}. \tag{39}$$

Accordingly, the final bound is obtained by the smallest one of these two bounds,

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}\xi_i\right\}\right] &\leq \exp\left\{\frac{\lambda^2B^2(m+u)^2}{8mu(m+u-1/2)}\cdot\frac{2\max(m,u)}{2\max(m,u)-1}\right\}\\ &= \exp\left\{\frac{\lambda^2B^2(m+u)C_{m,u}}{8mu}\right\}.\end{aligned} \tag{40}$$

Combining Eq. (40) and the facts that $\mathcal{E}(w,Z)-\mathbb{E}[\mathcal{E}(w,Z)]=\sum_{i=1}^{n}\xi_i$ and $\mathbb{E}[\mathcal{E}(w,Z)]=0$, for any $w\in\mathcal{W}$ we have

$$\mathbb{E}_Z\left[\exp\left\{\lambda(R_{\text{test}}(w,Z)-R_{\text{train}}(w,Z))\right\}\right]\leq\exp\left\{\frac{\lambda^2B^2(m+u)C_{m,u}}{8mu}\right\}. \tag{41}$$

Denote by Z' the independent copy of Z, which is independent of W and has the same distribution as Z. Then we have

$$\begin{aligned}&\log\mathbb{E}_{W,Z'}\left[\exp\left\{\lambda(R_{\text{test}}(W,Z')-R_{\text{train}}(W,Z'))\right\}\right]\\ =&\log\left(\int_w\mathbb{E}_{Z'}\left[\exp\left\{\lambda(R_{\text{test}}(w,Z')-R_{\text{train}}(w,Z'))\right\}\right]\mathrm{d}P_W(w)\right)\\ \leq&\log\left(\int_w\exp\left\{\frac{\lambda^2(m+u)C_{m,u}}{8mu}\right\}\mathrm{d}P_W(w)\right)\\ =&\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}.\end{aligned} \tag{42}$$

By Lemma 22, for any $\lambda\in\mathbb{R}$ we have

$$\begin{aligned}&\mathrm{D}_{\mathrm{KL}}(P_{Z,W}||P_{Z',W})\\ \geq&\mathbb{E}_{Z,W}\left[\lambda(R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z))\right]-\log\mathbb{E}_{Z',W}\left[\exp\left\{\lambda(R_{\text{test}}(W,Z')-R_{\text{train}}(W,Z'))\right\}\right]\\ \geq&\mathbb{E}_{Z,W}\left[\lambda(R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z))\right]-\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu},\end{aligned} \tag{43}$$

which implies that

$$\left|\mathbb{E}_{Z,W}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]\right|\leq\sqrt{\frac{B^2C_{m,u}I(W;Z)(m+u)}{2mu}}. \tag{44}$$

This finishes the proof for Eq. (1). For Eq. (2), notice that Eq. (41) can be rewritten as $\mathbb{E}_Z[\exp\{\lambda\mathcal{E}(w,Z)\}]\leq\exp\{\lambda^2\sigma_{m,u}\}$, where $\sigma_{m,u}\triangleq\frac{B^2C_{m,u}(m+u)}{8mu}$. Replacing $\mathcal{E}(w,Z)$ with $-\mathcal{E}(w,Z)$, we have $\mathbb{E}_Z[\exp\{-\lambda\mathcal{E}(w,Z)\}]\leq\exp\{\lambda^2\sigma_{m,u}\}$. Therefore, one can find that

$$\mathbb{P}\left\{|\mathcal{E}(w,Z)|\geq t\right\}\leq\mathbb{P}\left\{\mathcal{E}(w,Z)\geq t\right\}+\mathbb{P}\left\{\mathcal{E}(w,Z)\leq -t\right\}\leq 2\exp\left\{-\frac{t^2}{4\sigma_{m,u}}\right\}, \tag{45}$$

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

$$\begin{aligned}\mathbb{E}\left[|\mathcal{E}(w,Z)|^k\right]&=\int_0^\infty\mathbb{P}\{|\mathcal{E}(w,Z)|^k\geq u\}\,\mathrm{d}u=k\int_0^\infty\mathbb{P}\{|\mathcal{E}(w,Z)|\geq t\}\,t^{k-1}\,\mathrm{d}t\\ &\leq 2k\int_0^\infty\exp\left\{-\frac{t^2}{4\sigma_{m,u}}\right\}t^{k-1}\,\mathrm{d}t=(4\sigma_{m,u})^{\frac{k}{2}}k\Gamma(k/2),\end{aligned} \tag{46}$$

which implies that

$$\mathbb{E}\left[\exp\{\lambda\mathcal{E}^2(w,Z)\}\right] = 1+\sum_{k=1}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}\left[|\mathcal{E}(w,Z)|^{2k}\right] \le 1+2\sum_{k=1}^{\infty}(4\lambda\sigma_{m,u})^k. \tag{47}$$

By Lemma 22, for any $\lambda\in\mathbb{R}$ we have

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

Let $\lambda\to 1/(8\sigma_{m,u})$ and plug into $\sigma_{m,u}=\frac{B^2C_{m,u}(m+u)}{8mu}$, we obtain

$$\mathbb{E}_{Z,W}\left[(R_{\mathrm{test}}(W,Z)-R_{\mathrm{train}}(W,Z))^2\right]\le\frac{B^2C_{m,u}(m+u)(I(W;Z)+\log 3)}{mu}. \tag{48}$$

This finishes the proof.

Appendix C. Proof of Theorem 2

Proof. For any $k\in\mathbb{N}_+$, denote by $Z^{(1)},\dots,Z^{(k)}$ the k independent copies of Z. For $j\in[k]$ we run a transductive algorithm on $Z^{(j)}$ and obtain a output $W^{(j)}\sim P_{W|Z^{(j)}}$. By this way, $(Z^{(j)},W^{(j)})$ can be regarded as an independent copy of (Z,W) for $j\in[k]$. Now suppose that there is a monitor that returns

$$(J^*,R^*)\triangleq\operatorname*{argmax}_{j\in[k],r\in\{\pm1\}} r\mathcal{E}(W^{(j)},Z^{(j)}),\ W^*\triangleq W^{(J^*)}.$$

One can verify that

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

Taking expectation on both side, we have

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

Following the same procedure as that in Appendix B, we have

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

By Lemma 22, the following inequality holds for any $\lambda\in\mathbb{R}$

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

which implies that

$$\mathbb{E}_{Z^{(1)},\ldots,Z^{(k)},J^*,R^*,W^*}\left[R^*\mathcal{E}(W^{(J^*)},Z^{(J^*)})\right]\leq\sqrt{\frac{B^2C_{m,u}I(Z^{(1)},\ldots,Z^{(k)};J^*,R^*,W^*)(m+u)}{2mu}}. \tag{49}$$

Next we provide an upper bound for the mutual information term. Notice that

$$\begin{aligned}
&I(Z^{(1)},\ldots,Z^{(k)};J^*,R^*,W^*)\\
\leq&I(Z^{(1)},\ldots,Z^{(k)};J^*,R^*,W^*,W^{(1)},\ldots,W^{(k)})\\
=&I(Z^{(1)},\ldots,Z^{(k)};W^{(1)},\ldots,W^{(k)})+I(Z^{(1)},\ldots,Z^{(k)};J^*,R^*,W^*|W^{(1)},\ldots,W^{(k)})\\
=&\sum_{j=1}^{k}I(Z^{(j)};W^{(j)})+I(Z^{(1)},\ldots,Z^{(k)};J^*,R^*,W^*|W^{(1)},\ldots,W^{(k)})\\
\leq&kI(Z;W)+\log(2k).
\end{aligned} \tag{50}$$

where we have used the fact that $(Z^{(j)},W^{(j)}),j\in[k]$ are independent copies of (Z,W), and mutual information is determined only by the distribution of two probability measures. Plugging Eq. (50) into Eq. (49) yields

$$\mathbb{E}_{Z^{(1)},\ldots,Z^{(k)},W^{(1)},\ldots,W^{(k)}}\left[\max_{j\in[k]}|\mathcal{E}(Z^{(j)},W^{(j)})|\right]\leq\sqrt{\frac{B^2C_{m,u}(m+u)}{2mu}(\log(2k)+kI(Z,W))}. \tag{51}$$

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

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

By Markov's inequality,

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

Therefore,

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

Let $\alpha = 2\sqrt{\frac{B^2 C_{m,u}(m+u)(\log(2k)+kI(Z,W))}{2mu}}$ and $k = \lfloor \frac{1}{\delta} \rfloor$, we get Eq. (3) using the inequality $(1/2)^\delta \geq 1-\delta$ that holds for any $\delta \in (0,1)$. Let $k=1$, we get Eq. (4) from Eq. (51).

Appendix D. Proof of Proposition 4

Proof. We start from Eq. (41). Using Theorem 2.6 (IV) of Wainwright (2019) and let $\lambda = 1 - \frac{m+u}{mu} \in \left[\frac{1}{6}, 1\right)$, for any $w \in \mathcal{W}$ we have

$$\mathbb{E}_Z \left[\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(w,Z)}{B^2 C_{m,u}(m+u)} \right\} \right] \leq \sqrt{\frac{mu}{m+u}}. \tag{52}$$

Denote by Z' the independent copy of Z, we have

$$\begin{aligned} &\mathbb{E}_{Z\otimes W} \left[\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} \right\} \right] \\ =&\mathbb{E}_{Z'}\mathbb{E}_W \left[\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z')}{B^2 C_{m,u}(m+u)} \right\} \right] \\ =&\int_w \mathbb{E}_{Z'} \left[\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(w,Z')}{B^2 C_{m,u}(m+u)} \right\} \right] \mathrm{d}P_W(w) \leq \sqrt{\frac{mu}{m+u}}. \end{aligned} \tag{53}$$

Then we have

$$\begin{aligned} &\mathbb{E}_{W\otimes Z} \left[\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} \right\} \right] \\ \geq&\mathbb{E}_{W\otimes Z} \left[\mathbb{1}\left\{ \frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z} > 0 \right\} \exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} \right\} \right] \\ =&\mathbb{E}_{W,Z} \left[\left(\frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z} \right)^{-1} \exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} \right\} \right] \\ =&\mathbb{E}_{W,Z} \left[\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} - \log\frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z} \right\} \right]. \end{aligned} \tag{54}$$

Combining Eq. (54) with Eq. (53) and applying Markov's inequlity gives

$$\begin{aligned} &\mathbb{P}_{W,Z} \left(\frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} - \log\left(\sqrt{\frac{mu}{m+u}}\right) - \log\frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z} \geq \log(1/\delta) \right) \\ =&\mathbb{P}_{W,Z} \left(\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} - \log\left(\sqrt{\frac{mu}{m+u}}\right) - \log\frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z} \right\} \geq 1/\delta \right) \\ \leq&\delta \mathbb{E}_{W,Z} \left[\exp\left\{ \frac{2(mu-m-u)\mathcal{E}^2(W,Z)}{B^2 C_{m,u}(m+u)} - \log\left(\sqrt{\frac{mu}{m+u}}\right) - \log\frac{\mathrm{d}P_{W,Z}}{\mathrm{d}P_W P_Z} \right\} \right] \\ \leq&\delta. \end{aligned} \tag{55}$$

This finishes the proof.

Appendix E. Proof of Proposition 6

Proof. Denote by $\mathcal{U}$ and $\widetilde{\mathcal{Z}}$ the set containing all values of U and the transductive supersample $\widetilde{Z}$, respectively. Let $\overline{\mathscr{Z}} \triangleq \{\mathscr{Z}(\widetilde{z},u)|\widetilde{z}\in\widetilde{\mathcal{Z}},u\in\mathcal{U}\}$ be the set containing all sequences induce by $\widetilde{Z}$ and U. Now we show that $(\widetilde{z},u)\mapsto\mathscr{Z}(\widetilde{z},u)$ is a bijection. Clearly, it can be verified that $(\widetilde{z},u)\mapsto\mathscr{Z}(\widetilde{z},u)$ is a surjection. So it is sufficient to show that this mapping is injective. For any two element $\mathscr{Z}(\widetilde{z}_1,u_1),\mathscr{Z}(\widetilde{z}_2,u_2)\in\overline{\mathscr{Z}}$, we claim that $\mathscr{Z}(\widetilde{z}_1,u_1)=\mathscr{Z}(\widetilde{z}_2,u_2)\Rightarrow\widetilde{z}_1=\widetilde{z}_2$ holds. If $\widetilde{z}_1\neq\widetilde{z}_2$, by Eq. (7) one can find that there must exist $i\in[m]$ such that either $\mathscr{Z}_i(\widetilde{z}_1,u_1)\neq\mathscr{Z}_i(\widetilde{z}_2,u_2)$ or $\mathscr{Z}_{m+i}(\widetilde{z}_1,u_1)\neq\mathscr{Z}_{m+i}(\widetilde{z}_2,u_2)$. Recall that u_i only determines the relative position of $\mathscr{Z}_i(\widetilde{z},u)$ and $\mathscr{Z}_{m+i}(\widetilde{z},u)$ for $i\in[m]$. Since two sequences are equal if and only if the element at each position is equal, we have $\mathscr{Z}(\widetilde{z},u_1)=\mathscr{Z}(\widetilde{z},u_2)\Rightarrow u_1=u_2$, which implies that $(\widetilde{z},u)\mapsto\mathscr{Z}(\widetilde{z},u)$ is a injection. Since $\widetilde{Z}$ is obtained by sampling without replacement from $[n]$, elements in $\widetilde{\mathcal{Z}}$ are different from each other, and $|\widetilde{\mathcal{Z}}|=(2m)!/2^m$ holds. Clearly, we have $|\mathcal{U}|=2^m$. For each $\widetilde{z}\in\widetilde{\mathcal{Z}}$, applying U to permute it generates a different sequence. Thus, there are $(2m)!$ elements in $\overline{\mathscr{Z}}$ and they are different from each other, which implies that $\overline{\mathscr{Z}}=\mathrm{Perm}((1,\ldots,2m))$ holds. Let $\mathcal{Z}$ be the set containing all values of the sequence Z, then we have

$$\begin{aligned}
&\mathbb{E}_{W,Z}\left[\mathcal{E}(W,Z)\right]=\mathbb{E}_{W,Z}\left[R_{\mathrm{test}}(W,Z)-R_{\mathrm{train}}(W,Z)\right]\\
=&\frac{1}{(2m)!}\sum_{z\in\mathcal{Z}}\mathbb{E}_{W|Z=z}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell(W,s_{z_i})-\frac{1}{m}\sum_{i=1}^{m}\ell(W,s_{z_i})\right]\\
=&\frac{1}{(2m)!}\sum_{u\in\mathcal{U}}\sum_{\widetilde{z}\in\widetilde{\mathcal{Z}}}\mathbb{E}_{W|\widetilde{Z}=\widetilde{z},U=u}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell(W,s_{\mathscr{Z}_i(\widetilde{z},u)})-\frac{1}{m}\sum_{i=1}^{m}\ell(W,s_{\mathscr{Z}_i(\widetilde{z},u)})\right]\\
=&\mathbb{E}_{W,\widetilde{Z},U}[R_{\mathrm{test}}(W,\widetilde{Z},U)-R_{\mathrm{train}}(W,\widetilde{Z},U)].
\end{aligned}\tag{56}$$

For the case that $u=km,k\in\mathbb{N}_+$, one can verify that $(\widetilde{z},u)\mapsto\mathscr{Z}(\widetilde{z},u)$ is also a bijection and thus $\overline{\mathscr{Z}}=\mathrm{Perm}((1,\ldots,(k+1)m))$, following the above process. Then we have

$$\begin{aligned}
&\mathbb{E}_{W,Z}\left[\mathcal{E}(W,Z)\right]=\mathbb{E}_{W,Z}\left[R_{\mathrm{test}}(W,Z)-R_{\mathrm{train}}(W,Z)\right]\\
=&\frac{1}{(km+m)!}\sum_{z\in\mathcal{Z}}\mathbb{E}_{W|Z=z}\left[\frac{1}{km}\sum_{i=m+1}^{(k+1)m}\ell(W,s_{z_i})-\frac{1}{m}\sum_{i=1}^{m}\ell(W,s_{z_i})\right]\\
=&\frac{1}{(km+m)!}\sum_{u\in\mathcal{U}}\sum_{\widetilde{z}\in\widetilde{\mathcal{Z}}}\mathbb{E}_{W|\widetilde{Z}=\widetilde{z},U=u}\left[\frac{1}{km}\sum_{i=m+1}^{(k+1)m}\ell(W,s_{\mathscr{Z}_i(\widetilde{z},u)})-\frac{1}{m}\sum_{i=1}^{m}\ell(W,s_{\mathscr{Z}_i(\widetilde{z},u)})\right]\\
=&\mathbb{E}_{W,\widetilde{Z},U}[R_{\mathrm{test}}(W,\widetilde{Z},U)-R_{\mathrm{train}}(W,\widetilde{Z},U)],
\end{aligned}\tag{57}$$

where

$$\begin{aligned}
R_{\mathrm{test}}(W,\widetilde{Z},U)&\triangleq\frac{1}{km}\sum_{i=m+1}^{(k+1)m}\ell\left(W,s_{\mathscr{Z}_i(\widetilde{Z},U)}\right)=\frac{1}{km}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(W,s_{\widetilde{Z}_{i,U_{i,j}}}\right),\\
R_{\mathrm{train}}(W,\widetilde{Z},U)&\triangleq\frac{1}{m}\sum_{i=1}^{m}\ell\left(W,s_{\mathscr{Z}_i(\widetilde{Z},U)}\right)=\frac{1}{m}\sum_{i=1}^{m}\left(W,s_{\widetilde{Z}_{i,U_{i,0}}}\right).
\end{aligned}\tag{58}$$

This completes the proof.

Appendix F. Proof of Theorem 7

Proof. By Proposition 6, we have

$$\begin{aligned}
&\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]\\
=&\mathbb{E}_{W,\widetilde{Z},U}\left[R_{\text{test}}(W,\widetilde{Z},U)-R_{\text{train}}(W,\widetilde{Z},U)\right]\triangleq\mathbb{E}_{W,\widetilde{Z},U}\left[\mathcal{E}(W,\widetilde{Z},U)\right]\\
=&\mathbb{E}_{W,\widetilde{Z},U}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell\left(W,s_{\widetilde{Z}_{i,1-U_i}}\right)-\frac{1}{m}\sum_{i=1}^{m}\ell\left(W,s_{\widetilde{Z}_{i,U_i}}\right)\right].
\end{aligned}\tag{59}$$

where we use $\mathcal{E}(W,\widetilde{Z},U)$ as the abbreviation of $R_{\text{test}}(W,\widetilde{Z},U)-R_{\text{train}}(W,\widetilde{Z},U)$. Denote by w and $\widetilde{z}$ the fixed realizations of W and $\widetilde{Z}$. For any $\lambda\in\mathbb{R}$, by Hoeffding's inequality (Hoeffding, 1963) we have

$$\begin{aligned}
&\mathbb{E}_U\left[\exp\left\{\lambda\mathcal{E}(w,\widetilde{z},U)\right\}\right]\\
=&\mathbb{E}_U\left[\exp\left\{\frac{\lambda}{m}\sum_{i=1}^{m}\ell(w,s_{\widetilde{z}_{i,1-U_i}})-\ell(w,s_{\widetilde{z}_{i,U_i}})\right\}\right]\leq\exp\left\{\frac{\lambda^2B^2}{2m}\right\}.
\end{aligned}\tag{60}$$

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

$$\begin{aligned}
&\log\mathbb{E}_{U',W|\widetilde{Z}=\widetilde{z}}\left[\exp\left\{\lambda\mathcal{E}(W,\widetilde{z},U')\right\}\right]\\
=&\log\left(\int_w\mathbb{E}_{U'}\left[\exp\left\{\lambda\mathcal{E}(w,\widetilde{z},U')\right\}\right]\,\mathrm{d}P_{W|\widetilde{Z}=\widetilde{z}}(w)\right)\leq\frac{\lambda^2B^2}{2m},
\end{aligned}\tag{61}$$

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 to both $\widetilde{Z}$ and W. By Lemma 22, for any $\lambda\in\mathbb{R}$,

$$\begin{aligned}
I^{\widetilde{z}}(U;W)=&\mathrm{D}_{\text{KL}}(P_{U,W|\widetilde{Z}=\widetilde{z}}||P_{U',W|\widetilde{Z}=\widetilde{z}})\\
\geq&\mathbb{E}_{U,W|\widetilde{Z}=\widetilde{z}}\left[\lambda\mathcal{E}(W,\widetilde{z},U)\right]-\log\mathbb{E}_{U',W|\widetilde{Z}=\widetilde{z}}\left[\exp\left\{\lambda\mathcal{E}(W,z,U')\right\}\right]\\
\geq&\lambda\mathbb{E}_{U,W|\widetilde{Z}=\widetilde{z}}\left[\mathcal{E}(W,\widetilde{z},U)\right]-\frac{\lambda^2B^2}{2m},
\end{aligned}\tag{62}$$

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

$$\begin{aligned}
&\left|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]\right|=\left|\mathbb{E}_{W,\widetilde{Z},U}\left[\mathcal{E}(W,\widetilde{Z},U)\right]\right|\\
\leq&\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{U,W|\widetilde{Z}}\left[\mathcal{E}(W,\widetilde{Z},U)\right]\right|\leq\mathbb{E}_{\widetilde{Z}}\sqrt{\frac{2B^2}{m}I^{\widetilde{Z}}(U;W)}.
\end{aligned}\tag{63}$$

For the second part, note that Eq. (60) 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 $\mathbb{E}_U\left[\exp\left\{-\lambda\mathcal{E}(w,\widetilde{z},U)\right\}\right]\leq\exp\{\lambda^2B^2/2m\}$. Following the same procedure as that in Appendix B, we have

$$\mathbb{E}_{U,W|\widetilde{Z}=\widetilde{z}}\left[\mathcal{E}^2(W,\widetilde{z},U)\right]\leq\frac{4B^2}{m}(I^{\widetilde{z}}(U;W)+\log 3).\tag{64}$$

Taking expectation on both side, we have obtain

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

Now let us consider a special case that $\ell(\cdot)$ is the zero-one loss. Denote by w and $\widetilde{z}$ the fixed realizations of W and $\widetilde{Z}$. By Proposition 6, for any $\lambda>0$ we have

$$\begin{aligned}
\log\left(\mathbb{E}_U\left[e^{-\lambda R_{\text{train}}(w,\widetilde{z},U)}\right]\right)&=\log\left(\mathbb{E}_U\left[e^{-\frac{\lambda}{m}\sum_{i=1}^m\ell(w,s_{\widetilde{z}_{i,U_i}})}\right]\right)\\
&=\log\left(\mathbb{E}_U\left[\prod_{i=1}^m e^{-\frac{\lambda}{m}\ell(w,s_{\widetilde{z}_{i,U_i}})}\right]\right)\\
&=\log\left(\prod_{i=1}^m\mathbb{E}_{U_i}\left[e^{-\frac{\lambda}{m}\ell(w,s_{\widetilde{z}_{i,U_i}})}\right]\right)\\
&=\sum_{i=1}^m\log\left(\mathbb{E}_{U_i}\left[e^{-\frac{\lambda}{m}\ell(w,s_{\widetilde{z}_{i,U_i}})}\right]\right)\\
&\leq m\log\left(\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{U_i}\left[e^{-\frac{\lambda}{m}\ell(w,s_{\widetilde{z}_{i,U_i}})}\right]\right).
\end{aligned}\tag{65}$$

The fact that $\ell(\cdot)$ is the zero-one loss implies

$$e^{-\frac{\lambda}{m}\ell(w,s_{\widetilde{z}_{i,U_i}})}=1-\ell(w,s_{\widetilde{z}_{i,U_i}})+e^{-\frac{\lambda}{m}}\ell(w,s_{\widetilde{z}_{i,U_i}}).\tag{66}$$

Recall that $\Phi_a(p)=-a^{-1}\log(1-[1-e^{-a}]p)$ where $a\in\mathbb{R}$ and $0<p<1$, define

$$R(W)\triangleq\frac{1}{2}\left(R_{\text{train}}(W,Z)+R_{\text{test}}(W,Z)\right)=\frac{1}{2m}\sum_{i=1}^{2m}\ell(W,s_{Z_i}),\tag{67}$$

we have

$$\begin{aligned}
&m\log\left(\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{U_i}\left[e^{-\frac{\lambda}{m}\ell(w,s_{\widetilde{z}_{i,U_i}})}\right]\right)\\
=&m\log\left(\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{U_i}\left[1-\ell(w,s_{\widetilde{z}_{i,U_i}})+e^{-\frac{\lambda}{m}}\ell(w,s_{\widetilde{z}_{i,U_i}})\right]\right)\\
=&m\log\left(1-\frac{1}{m}\sum_{i=1}^m\mathbb{E}_{U_i}\left[\ell(w,s_{\widetilde{z}_{i,U_i}})\right]+\frac{e^{-\frac{\lambda}{m}}}{m}\sum_{i=1}^m\mathbb{E}_{U_i}\left[\ell(w,s_{\widetilde{z}_{i,U_i}})\right]\right)\\
=&m\log\left(1-\frac{1}{2m}\sum_{i=1}^{2m}\ell(w,s_i)+\frac{e^{-\frac{\lambda}{m}}}{2m}\sum_{i=1}^{2m}\ell(w,s_i)\right)=-\lambda\Phi_{\lambda/m}(R(w)).
\end{aligned}\tag{68}$$

Combining the above results and rearranging the term we get

$$\mathbb{E}_U\left[e^{\lambda(\Phi_{\lambda/m}(R(w))-R_{\text{train}}(w,\widetilde{z},U))}\right]\leq 1.\tag{69}$$

By Lemma 22, for any posterior Q and prior P we have

$$\begin{aligned}
&\mathbb{E}_U\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\mathbb{E}_{W\sim Q_{W|\widetilde{z},U}}\left[\lambda\left(\Phi_{\lambda/m}(R(W))-R_{\text{train}}(W,\widetilde{z},U)\right)\right]-\mathrm{D}_{\mathrm{KL}}(Q||P)}\right]\\
=&\mathbb{E}_U\mathbb{E}_{W\sim P}\left[e^{\lambda\left(\Phi_{\lambda/m}(R(W))-R_{\text{train}}(W,\widetilde{z},U)\right)}\right]\\
=&\int_w\mathbb{E}_U\left[e^{\lambda\left(\Phi_{\lambda/m}(R(w))-R_{\text{train}}(w,\widetilde{z},U)\right)}\right]\mathrm{d}P(w)\leq 1.
\end{aligned}\tag{70}$$

Taking the expectation over $\widetilde{Z}$ on both side and applying the Markov's inequality, for any $0<\delta<1$ and $\lambda>0$, with probability over the randomness of $\widetilde{Z}$ and U:

$$\sup_{Q\in\mathcal{P}(\mathcal{W})}\mathbb{E}_{W\sim Q}\left[\Phi_{\lambda/m}(R(W))-R_{\text{train}}(W,\widetilde{Z},U)\right]\leq\frac{\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{\lambda}.\tag{71}$$

We close this proof by presenting the results for the cases that $u=km$ with $k\in\mathbb{N}_+$. Recall that the transductive training and test error under these cases are defined as

$$\begin{aligned}
R_{\text{test}}(W,\widetilde{Z},U)&\triangleq\frac{1}{km}\sum_{i=m+1}^{(k+1)m}\ell\left(W,s_{\mathscr{Z}_i(\widetilde{Z},U)}\right)=\frac{1}{km}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(W,s_{\widetilde{Z}_{i,U_{i,j}}}\right),\\
R_{\text{train}}(W,\widetilde{Z},U)&\triangleq\frac{1}{m}\sum_{i=1}^{m}\ell\left(W,s_{\mathscr{Z}_i(\widetilde{Z},U)}\right)=\frac{1}{m}\sum_{i=1}^{m}\left(W,s_{\widetilde{Z}_{i,U_{i,0}}}\right).
\end{aligned}\tag{72}$$

By Proposition 6 we have

$$\begin{aligned}
&\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]=\mathbb{E}_{W,\widetilde{Z},U}\left[R_{\text{test}}(W,\widetilde{Z},U)-R_{\text{train}}(W,\widetilde{Z},U)\right],\\
&\mathbb{E}_{W,Z}\left[(R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z))^2\right]=\mathbb{E}_{W,\widetilde{Z},U}\left[(R_{\text{test}}(W,\widetilde{Z},U)-R_{\text{train}}(W,\widetilde{Z},U))^2\right].
\end{aligned}$$

Following the proof in this part and plugging into $m=\frac{n}{k+1}$ we obtain

$$\begin{aligned}
&\left|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]\right|\leq\mathbb{E}_{\widetilde{Z}}\sqrt{\frac{2(k+1)B^2}{n}I^{\widetilde{Z}}(U;W)},\\
&\mathbb{E}_{W,Z}\left[(R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z))^2\right]\leq\frac{4(k+1)B^2}{n}(I(U;W|\widetilde{Z})+\log 3).
\end{aligned}\tag{73}$$

This completes the proof.

Appendix G. Proof of Corollary 8

Proof. Denote by

$$g(F_i,U_i,\widetilde{Z})\triangleq r(F_{i,1-U_i},y_{\widetilde{Z}_{i,1-U_i}})-r(F_{i,U_i},y_{\widetilde{Z}_{i,U_i}})\tag{74}$$

the function of $(F_i,U_i,\widetilde{Z})$. Recall that $F_{i,U_i}=f_W(x_{\widetilde{Z}_{i,U_i}})$ and $F_{i,1-U_i}=f_W(x_{\widetilde{Z}_{i,1-U_i}})$ hold. Let f_i and $\widetilde{z}$ be the fixed realizations of F_i and $\widetilde{Z}$. For any $\lambda\in\mathbb{R}$ and $i\in[m]$, by Hoeffding's inequality (Hoeffding, 1963) we have

$$\mathbb{E}_{U_i}\left[\exp\left\{\lambda g(f_i,U_i,\widetilde{z})\right\}\right]\leq\exp\left\{\frac{\lambda^2B^2}{2}\right\}.\tag{75}$$

Let U_i' be the independent copy of U_i, by Lemma 22,

$$\begin{aligned} I(F_i;U_i|\widetilde{Z}=\widetilde{z}) &\geq \lambda \mathbb{E}_{F_i,U_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i,U_i,\widetilde{z})\right] - \log \mathbb{E}_{F_i,U_i'|\widetilde{Z}=\widetilde{z}}\left[\exp\left\{\lambda g(F_i,U_i',\widetilde{z})\right\}\right] \\ &\geq \lambda \mathbb{E}_{F_i,U_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i,U_i,\widetilde{z})\right] - \frac{\lambda^2 B^2}{2}, \end{aligned} \tag{76}$$

which implies that

$$\left|\mathbb{E}_{F_i,U_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i,U_i,\widetilde{z})\right]\right| \leq B\sqrt{2I^{\widetilde{z}}(F_i;U_i)}. \tag{77}$$

Taking expectation over $\widetilde{Z}$ on both side yields

$$\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{F_i,U_i|\widetilde{Z}}\left[g(F_i,U_i,\widetilde{Z})\right]\right| \leq B\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(F_i;U_i)}. \tag{78}$$

Then we have

$$\begin{aligned} &\left|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]\right| = \left|\mathbb{E}_{W,\widetilde{Z},U}\left[\mathcal{E}(W,\widetilde{Z},U)\right]\right| \\ =&\left|\mathbb{E}_{W,\widetilde{Z},U}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell\left(W,s_{\mathscr{Z}_i(\widetilde{Z},U)}\right)-\frac{1}{m}\sum_{i=1}^{m}\ell\left(W,s_{\mathscr{Z}_i(\widetilde{Z},U)}\right)\right]\right| \\ =&\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\mathbb{E}_{W,U_i|\widetilde{Z}}\left[\ell\left(W,s_{\widetilde{Z}_{i,1-U_i}}\right)-\ell\left(W,s_{\widetilde{Z}_{i,U_i}}\right)\right]\right| \\ \leq&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{W,U_i|\widetilde{Z}}\left[\ell\left(W,s_{\widetilde{Z}_{i,1-U_i}}\right)-\ell\left(W,s_{\widetilde{Z}_{i,U_i}}\right)\right]\right| \\ =&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{W,U_i|\widetilde{Z}}\left[r\big(f_W(x_{\widetilde{Z}_{i,1-U_i}}),y_{\widetilde{Z}_{i,1-U_i}}\big)-r\big(f_W(x_{\widetilde{Z}_{i,U_i}}),y_{\widetilde{Z}_{i,U_i}}\big)\right]\right| \\ =&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{F_i,U_i|\widetilde{Z}}\left[r(F_{i,1-U_i},y_{\widetilde{Z}_{i,1-U_i}})-r(F_{i,U_i},y_{\widetilde{Z}_{i,U_i}})\right]\right| \\ =&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{F_i,U_i|\widetilde{Z}}\left[g(F_i,U_i,\widetilde{Z})\right]\right| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(F_i;U_i)}. \end{aligned} \tag{79}$$

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

$$\left|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]\right| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(L_i;U_i)}, \tag{80}$$

$$\left|\mathbb{E}_{W,Z}\left[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)\right]\right| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta_i;U_i)}. \tag{81}$$

For the cases that $u=km, k\in\mathbb{N}_+$, define $F_i\triangleq(f_W(X_{\widetilde{Z}_{i,0}}),\ldots,f_W(X_{\widetilde{Z}_{i,k}})), i\in[m]$ and $L_{i,:}\triangleq(\ell(W,S_{\widetilde{Z}_{i,0}}),\ldots,\ell(W,S_{\widetilde{Z}_{i,k}})), i\in[m]$ be the prediction of model and the sequence of loss values. Denote by $g(F_i,U_i,\widetilde{Z})\triangleq\frac{1}{k}\sum_{j=1}^{k}r\big(F_{i,U_{i,j}},y_{\widetilde{Z}_{i,U_{i,j}}}\big)-r\big(F_{i,U_{i,0}},y_{\widetilde{Z}_{i,U_{i,0}}}\big)$ and

$g(F_i, U_i, \widetilde{Z}) \triangleq \frac{1}{k}\sum_{j=1}^{k} L_{i,U_{i,j}}, -L_{i,U_{i,0}}$ the function of $(F_i, U_i, \widetilde{Z})$ and $(L_i, U_i, \widetilde{Z})$, respectively. Following the above process one can verify that Eqs. (79,80) still hold under the cases that $u = km, k \in \mathbb{N}_+$. This finishes the proof.

Appendix H. Proof of Theorem 10

Proof. We start from Eq. (41). By Lemma 22, for any $\lambda > 0$ we have

$$\begin{aligned}
&\mathbb{E}_Z\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]-\mathrm{D}_{\mathrm{KL}}(Q||P)}\right]\\
=&\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[e^{\lambda(R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z))}\right]\\
=&\int_w \mathbb{E}_Z\left[e^{\lambda(R_{\text{test}}(w,Z)-R_{\text{train}}(w,Z))}\right]\mathrm{d}P(w)\\
\leq& e^{\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}}.
\end{aligned} \tag{82}$$

By Jensen's inequality we have

$$\begin{aligned}
&\mathbb{E}_Z\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]-\mathrm{D}_{\mathrm{KL}}(Q||P)}\right]\\
\geq& e^{\mathbb{E}_Z\left[\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]-\mathrm{D}_{\mathrm{KL}}(Q||P)\right]}.
\end{aligned} \tag{83}$$

Combining Eq. (82) with Eq. (83), for any posterior $Q \in \mathcal{P}(\mathcal{W})$ and $\lambda > 0$ we have

$$\mathbb{E}_Z\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] \leq \frac{\lambda B^2 C_{m,u}(m+u)}{8mu} + \frac{\mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}(Q||P)\right]}{\lambda}. \tag{84}$$

This gives Eq. (17). Using Markov's inequality, for any $\lambda > 0$ and $0 < \delta < 1$:

$$\begin{aligned}
&\mathbb{P}_Z\left(\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \mathrm{D}_{\mathrm{KL}}(Q||P) - \frac{\lambda^2B^2C_{m,u}(m+u)}{8mu} \geq \log(1/\delta)\right)\\
=&\mathbb{P}_Z\left(\exp\left\{\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \mathrm{D}_{\mathrm{KL}}(Q||P) - \frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}\right\} \geq \frac{1}{\delta}\right)\\
\leq&\delta\mathbb{E}_Z\left[\exp\left\{\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \mathrm{D}_{\mathrm{KL}}(Q||P) - \frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}\right\}\right] \leq \delta,
\end{aligned}$$

where the last inequality is due to Eq. (82). By rearranging the term we obtain Eq. (18). To obtain Eq. (19), firstly we have

$$\begin{aligned}
&\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[\exp\left\{\lambda\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right]\right\}\right]\\
\geq&\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[\mathbb{1}\left\{\frac{\mathrm{d}Q}{\mathrm{d}P} > 0\right\}\exp\left\{\lambda\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right]\right\}\right]\\
=&\mathbb{E}_Z\mathbb{E}_{W\sim Q}\left[\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^{-1}\exp\left\{\lambda\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right]\right\}\right]\\
=&\mathbb{E}_Z\mathbb{E}_{W\sim Q}\left[\exp\left\{\lambda\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \log\frac{\mathrm{d}Q}{\mathrm{d}P}\right\}\right].
\end{aligned} \tag{85}$$

Combining Eq. (85) with Eq. (82) and using Markov's inequality, for any $0 < \delta < 1$:

$$\begin{aligned}
&\mathbb{P}_{Z,W\sim Q}\left(\lambda\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \log\frac{\mathrm{d}Q}{\mathrm{d}P} - \frac{\lambda^2 B^2 C_{m,u}(m+u)}{8mu} \geq \log(1/\delta)\right)\\
=&\mathbb{P}_{Z,W\sim Q}\left(\exp\left\{\lambda\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \log\frac{\mathrm{d}Q}{\mathrm{d}P} - \frac{\lambda^2 B^2 C_{m,u}(m+u)}{8mu}\right\} \geq \frac{1}{\delta}\right)\\
\leq&\delta\mathbb{E}_{Z,W\sim Q}\left[\exp\left\{\lambda\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \log\frac{\mathrm{d}Q}{\mathrm{d}P} - \frac{\lambda^2 B^2 C_{m,u}(m+u)}{8mu}\right\}\right] \leq \delta.
\end{aligned} \tag{86}$$

By rearranging the term we obtain Eq. (19). Interestingly, we can also get a degenerated version of Theorem 1, following the technique of Bégin et al. (2014). Denote by $R(W) \triangleq \frac{1}{m+u}\sum_{i=1}^{m+u}\ell(W, s_{Z_i}) = \frac{m}{m+u}R_{\text{train}}(W,Z) + \frac{u}{m+u}R_{\text{test}}(W,Z)$ the average error on $\{s_i\}_{i=1}^n$. 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 by 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 of Bégin et al. (2014), for fixed realization w of W we have

$$\mathbb{E}_Z\left[\exp\left\{m\mathcal{D}^*_\beta(R_{\text{train}}(w,Z), R(w))\right\}\right] \leq 3\log(m)\sqrt{\frac{mu}{m+u}}, \tag{87}$$

which implies that

$$\log\mathbb{E}_{W\otimes Z}\left[\exp\left\{m\mathcal{D}^*_\beta(R_{\text{train}}(W,Z), R(W))\right\}\right] \leq \log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right). \tag{88}$$

By Lemma 22 we have

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(P_{Z,W}||P_{Z\otimes W})\\
\geq&\mathbb{E}_{Z,W}\left[m\mathcal{D}^*_\beta(R_{\text{train}}(W,Z), R(W))\right] - \log\mathbb{E}_{Z\otimes W}\left[e^{m\mathcal{D}^*_\beta(R_{\text{train}}(W,Z),R(W))}\right]\\
\geq&m\mathbb{E}_{Z,W}\left[\mathcal{D}^*_\beta(R_{\text{train}}(W,Z), R(W))\right] - \log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right).
\end{aligned} \tag{89}$$

By Pinsker's inequality (Boucheron et al., 2013, Theorem 4.19) and plugging in $\beta = \frac{m}{m+u}$, the expectation term can be lower bounded by

$$\begin{aligned}
&\mathbb{E}_{Z,W}\left[\mathcal{D}^*_\beta(R_{\text{train}}(W,Z), R(W))\right]\\
=&\frac{u}{m}\mathbb{E}_{Z,W}\left[\mathcal{D}\left(\frac{m+u}{u}R(W) - \frac{m}{u}R_{\text{train}}(W,Z), R(W)\right)\right] + \mathbb{E}_{Z,W}\left[\mathcal{D}(R_{\text{train}}(W,Z), R(W))\right]\\
\geq&\frac{2m}{u}\mathbb{E}_{Z,W}\left[(R_{\text{train}}(W,Z) - R(W))^2\right] + 2\mathbb{E}_{Z,W}\left[(R_{\text{train}}(W,Z) - R(W))^2\right]\\
=&\frac{2(m+u)}{u}\mathbb{E}_{Z,W}\left[\left(R_{\text{train}}(W,Z) - \frac{m}{m+u}R_{\text{train}}(W,Z) - \frac{u}{m+u}R_{\text{test}}(W,Z)\right)^2\right]\\
=&\frac{2u}{m+u}\mathbb{E}_{Z,W}\left[(R_{\text{train}}(W,Z) - R_{\text{test}}(W,Z))^2\right]\\
\geq&\frac{2u}{m+u}\left(\mathbb{E}_{Z,W}[R_{\text{train}}(W,Z) - R_{\text{test}}(W,Z)]\right)^2.
\end{aligned} \tag{90}$$

Combining Eq. (89) with Eq. (90) we obtain

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

Compared with the one presented in Theorem 1, this bound is degenerated since it contains a extra factor $\log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right)$. This finishes the proof.

Appendix I. Proof of Corollary 11

Proof. We start from Eq. (82): for a fixed $\lambda \in \Lambda$ we have

$$\mathbb{E}_Z\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]-\mathrm{D}_{\mathrm{KL}}(Q||P)-\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}}\right] \leq 1. \tag{91}$$

Then

$$\begin{aligned}
&\mathbb{E}_Z\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W}),\lambda\in\Lambda}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]-\mathrm{D}_{\mathrm{KL}}(Q||P)-\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}}\right] \\
=&\mathbb{E}_Z\left[\sup_{\lambda\in\Lambda} e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]-\mathrm{D}_{\mathrm{KL}}(Q||P)-\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}}\right] \\
\leq&\sum_{\lambda\in\Lambda}\mathbb{E}_Z\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W}),\lambda\in\Lambda}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,Z)-R_{\text{train}}(W,Z)]-\mathrm{D}_{\mathrm{KL}}(Q||P)-\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}}\right] = |\Lambda|.
\end{aligned} \tag{92}$$

By Markov's inequality, for any $\lambda > 0$ and $0 < \delta < 1$:

$$\begin{aligned}
&\mathbb{P}_Z\left(\sup_{\substack{Q\in\mathcal{P}(\mathcal{W})\\ \lambda\in\Lambda}}\lambda\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \mathrm{D}_{\mathrm{KL}}(Q||P) - \frac{\lambda^2B^2C_{m,u}(m+u)}{8mu} \geq \log(|\Lambda|/\delta)\right) \\
=&\mathbb{P}_Z\left(\exp\left\{\sup_{\substack{Q\in\mathcal{P}(\mathcal{W})\\ \lambda\in\Lambda}}\lambda\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \mathrm{D}_{\mathrm{KL}}(Q||P) - \frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}\right\} \geq \frac{|\Lambda|}{\delta}\right) \\
\leq&\frac{\delta}{|\Lambda|}\mathbb{E}_Z\left[\exp\left\{\sup_{\substack{Q\in\mathcal{P}(\mathcal{W})\\ \lambda\in\Lambda}}\lambda\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] - \mathrm{D}_{\mathrm{KL}}(Q||P) - \frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}\right\}\right] \\
\leq&\delta.
\end{aligned}$$

This finishes the proof of the first inequality. For the second inequality, let $\Lambda \triangleq \left\{e^i, i \in \mathbb{N}\right\} \cap [1, mu/(m+u)]$, we have

$$\begin{aligned}
&\inf_{\lambda\in\Lambda}\left\{\frac{\lambda C_{m,u}(m+u)}{8mu} + \frac{\mathrm{D}_{\mathrm{KL}}(Q||P) + \log(|\Lambda|/\delta)}{\lambda}\right\} \\
\leq&\inf_{\lambda\in\left[1,\frac{mu}{m+u}\right]}\left\{\frac{e^{\lfloor\log\lambda\rfloor}C_{m,u}(m+u)}{8mu} + \frac{1}{e^{\lfloor\log\lambda\rfloor}}\left(\mathrm{D}_{\mathrm{KL}}(Q||P) + \log\left(\frac{1}{\delta}\log\left(\frac{mu}{m+u}\right)\right)\right)\right\} \\
\leq&\inf_{\lambda\in\left[1,\frac{mu}{m+u}\right]}\left\{\frac{\lambda C_{m,u}(m+u)}{8mu} + \frac{e}{\lambda}\left(\mathrm{D}_{\mathrm{KL}}(Q||P) + \log\left(\frac{1}{\delta}\log\left(\frac{mu}{m+u}\right)\right)\right)\right\},
\end{aligned} \tag{93}$$

where the second line is due to the fact that $|\Lambda| \leq \log\left(\frac{mu}{m+u}\right)$, and the third line is obtained by the inequality $\log(\lambda) - 1 \leq \lfloor\log\lambda\rfloor \leq \log\lambda$. This finishes the proof.

Appendix J. Proof of Corollary 13

Proof. This proof is motivated by the proof of Theorem 2 of Foret et al. (2021). The core idea is that, for a given posterior distribution Q, we aim to properly select the optimal prior distribution P^* such that the KL divergence term $\mathrm{D}_{\mathrm{KL}}(Q||P)$ is minimized. However, this solution is not applicable because the prior P obtained by this approach will depend on Z. However, P is required to be chosen before observing Z. The approach to address this issue is to construct a predefined set of prior distribution $\{P_j\}_{j\in\mathbb{N}_+}$, and then establish a high probability guarantee for each measure in this set. After that, we can establish a high probability guarantee for the optimal prior P^* by using union bound inequality.

Formally, denote by $W \in \mathbb{R}^d$ the parameter returned by the learning algorithm, we define the posterior distribution as $Q \triangleq \mathcal{N}(W, \sigma^2\mathbf{I}_d)$. Let c be a constant that depends on m, u, d, σ, whose value will be specified later. The probability measure in the predefined set $\{P_j\}_{j\in\mathbb{N}_+}$ is defined as $P_j = \mathcal{N}(O, \sigma_j^2\mathbf{I}_d)$, where $\sigma_j^2 = ce^{(1-j)/d}$. For any $j \in \mathbb{N}_+$, by calculating the KL divergence term, we have

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

which implies that

$$\operatorname*{argmin}_{\sigma_j} \mathrm{D}_{\mathrm{KL}}(Q||P) = \sqrt{\frac{d\sigma^2 + \|W\|_2^2}{d}}.$$

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

$$j^* = \left\lfloor 1 - d\log\left(\frac{d\sigma^2 + \|W\|_2^2}{dc}\right)\right\rfloor.$$

One can verify that the following inequalities hold

$$-d\log\left(\frac{d\sigma^2 + \|W\|_2^2}{dc}\right) \le j^* \le 1 - d\log\left(\frac{d\sigma^2 + \|W\|_2^2}{dc}\right). \tag{94}$$

Combining Eq. (94) with the equality $\sigma_{j^*}^2 = ce^{(1-j^*)/d}$ yields

$$\sigma^2 + \frac{\|W\|_2^2}{d} \le \sigma_{j^*}^2 \le e^{1/d}\left(\sigma^2 + \frac{\|W\|_2^2}{d}\right). \tag{95}$$

Then we have

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(Q||P^*) =& \frac{1}{2}\left[\frac{d\sigma^2 + \|W\|_2^2}{\sigma_{j^*}^2} - d + d\log\left(\frac{\sigma_{j^*}^2}{\sigma^2}\right)\right] \\
\le& \frac{1}{2}\left[\frac{d(d\sigma^2 + \|W\|_2^2)}{d\sigma^2 + \|W\|_2^2} - d + d\log\left(\frac{e^{1/d}\left(d\sigma^2 + \|W\|_2^2\right)}{d\sigma^2}\right)\right] \\
=& \frac{d}{2}\log\left(\frac{e^{1/d}\left(d\sigma^2 + \|W\|_2^2\right)}{d\sigma^2}\right) = \frac{1}{2}\left[1 + d\log\left(1 + \frac{\|W\|_2^2}{d\sigma^2}\right)\right].
\end{aligned} \tag{96}$$

For any $\lambda > 0$ and $j \in \mathbb{N}_+$, denote by A_j the event that

$$A_j \triangleq \left\{ \left| \mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right]\right| \geq \frac{\lambda C_{m,u}(m+u)}{8mu} + \frac{\mathrm{D}_{\mathrm{KL}}(Q||P_j) + \log(1/\delta_j)}{\lambda} \right\}.$$

Let $\delta_j = \frac{6\delta}{\pi^2 j^2}$, by Theorem 10, for any posterior distribution Q we have

$$\mathbb{P}\left\{A_{j^*}\right\} \leq \mathbb{P}\left\{ \bigcup_{j=1}^{\infty} A_j \right\} \leq \sum_{j=1}^{\infty} \mathbb{P}\left\{A_j\right\} = \sum_{j=1}^{\infty} \delta_j = \sum_{j=1}^{\infty} \frac{6\delta}{\pi^2 j^2} = \delta. \tag{97}$$

Therefore, with probability at least $1-\delta$, for any $\lambda > 0$:

$$\begin{aligned}
&\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,Z) - R_{\text{train}}(W,Z)\right] \\
=&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)}\left[R_{\text{test}}(W+\boldsymbol{\epsilon},Z) - R_{\text{train}}(W+\boldsymbol{\epsilon},Z)\right] \\
\leq&\frac{\lambda C_{m,u}(m+u)}{8mu} + \frac{\mathrm{D}_{\mathrm{KL}}(Q||P_{j^*}) + \log(1/\delta_j)}{\lambda} \\
\leq&\frac{\lambda C_{m,u}(m+u)}{8mu} + \frac{1}{\lambda}\left(\frac{1}{2}\left[1 + d\log\left(1 + \frac{\|W\|_2^2}{d\sigma^2}\right)\right] + \log\left(\frac{1}{6\delta}\right) + 2\log\left(\pi j^*\right)\right).
\end{aligned} \tag{98}$$

Since $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$, Lemma 1 of Laurent and Massart (2000) shows that $\mathbb{P}(\|\boldsymbol{\epsilon}\|_2^2 \geq d\sigma^2 + 2\sigma^2\sqrt{dt} + 2t\sigma^2) \leq e^{-t}$. Let $\widetilde{C}_{m,u} \triangleq \sqrt{2\log(mu/(m+u))/d}$, with probability at least $1-(m+u)/mu$ we have

$$\begin{aligned}
\|\boldsymbol{\epsilon}\|_2^2 \leq&d\sigma^2 + 2\sigma^2\sqrt{d\log(mu/(m+u))} + 2\sigma^2\log(mu/(m+u)) \\
\leq&d\sigma^2 + 2\sigma^2\sqrt{2d\log(mu/(m+u))} + 2\sigma^2\log(mu/(m+u)) \\
=&\sigma^2 d\left(1 + \sqrt{\frac{2\log(mu/(m+u))}{d}}\right)^2 = \sigma^2 d\left(1+\widetilde{C}_{m,u}\right)^2 = \rho^2.
\end{aligned} \tag{99}$$

Combining the above ingredients together, with probability at least $1-\delta$, for any $\lambda > 0$:

$$\begin{aligned}
&R_{\text{test}}(W,Z) \\
\leq&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)}\left[R_{\text{test}}(W+\boldsymbol{\epsilon},Z)\right] \\
=&\mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2^2 \leq \rho^2\right)\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)}\left[R_{\text{test}}(W+\boldsymbol{\epsilon},Z)|\|\boldsymbol{\epsilon}\|_2^2 \leq \rho^2\right] \\
&+\mathbb{P}\left(\|\boldsymbol{\epsilon}\|_2^2 > \rho^2\right)\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)}\left[R_{\text{test}}(W+\boldsymbol{\epsilon},Z)|\|\boldsymbol{\epsilon}\|_2^2 > \rho^2\right] \\
\leq&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)}\left[R_{\text{test}}(W+\boldsymbol{\epsilon},Z)|\|\boldsymbol{\epsilon}\|_2^2 \leq \rho^2\right] + \frac{m+u}{mu} \\
=&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_d)}\left[R_{\text{test}}(W+\boldsymbol{\epsilon},Z) - R_{\text{train}}(W+\boldsymbol{\epsilon},Z) + R_{\text{train}}(W+\boldsymbol{\epsilon},Z)|\|\boldsymbol{\epsilon}\|_2^2 \leq \rho^2\right] + \frac{m+u}{mu} \\
\leq&\max_{\|\boldsymbol{\epsilon}\|_2\leq\rho} R_{\text{train}}(W+\boldsymbol{\epsilon},Z) + \frac{m+u}{mu} + \frac{\lambda C_{m,u}(m+u)}{8mu} \\
&+\frac{1}{\lambda}\left(\frac{1}{2}\left[1 + d\log\left(1 + \frac{\|W\|_2^2}{d\sigma^2}\right)\right] + \log\left(\frac{1}{6\delta}\right) + 2\log\left(\pi j^*\right)\right) \\
\leq&\max_{\|\boldsymbol{\epsilon}\|_2\leq\rho} R_{\text{train}}(W+\boldsymbol{\epsilon},Z) + \frac{(\lambda C_{m,u}+8)(m+u)}{8mu} \\
&+\frac{1}{\lambda}\left(\frac{1}{2}\left[1 + d\log\left(1 + \frac{\|W\|_2^2}{\rho^2}\left(1+\widetilde{C}_{m,u}\right)^2\right)\right] + \log\left(\frac{1}{6\delta}\right) + 2\log\left(\pi j^*\right)\right).
\end{aligned} \tag{100}$$

Here we use the assumption to obtain the second line. The third and the forth line are due to law of total expectation and the fact that $R_{\text{test}}(w,z) \leq 1$ for any w and z. The remaining step is to specify the value of c. Note that if

$$\|W\|_2^2 \geq \sigma^2 d\left(e^{\frac{4mu}{(m+u)d}}-1\right) = \frac{\rho^2}{(1+\widetilde{C}_{m,u})^2}\left(e^{\frac{4mu}{(m+u)d}}-1\right), \tag{101}$$

the slack term in Eq. (100) will exceed 1 and the inequality holds trivially. Therefore, we only need to consider the case that

$$\|W\|_2^2 < \frac{\rho^2}{(1+\widetilde{C}_{m,u})^2}\left(e^{\frac{4mu}{(m+u)d}}-1\right) = d\sigma^2\left(e^{\frac{4mu}{(m+u)d}}-1\right), \tag{102}$$

which implies that

$$\sigma^2 + \frac{\|W\|_2^2}{d} < \sigma^2 e^{\frac{4mu}{(m+u)d}} \triangleq c. \tag{103}$$

Here we have used Eq. (99) 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

$$\begin{aligned}\log(j^*) &\leq \log\left(1+d\log\left(\frac{dc}{d\sigma^2+\|W\|_2^2}\right)\right)\\ &\leq \log\left(1+d\log\left(\frac{c}{\sigma^2}\right)\right)\\ &= \log\left(1+\frac{4mu}{(m+u)}\right) \leq \log\left(\frac{6mu}{m+u}\right).\end{aligned} \tag{104}$$

Plugging Eq. (104) into Eq. (100), with probability at least $1-\delta$ over the randomness of Z, for any $\lambda > 0$:

$$\begin{aligned}R_{\text{test}}(W,Z) \leq \max_{\|\boldsymbol{\epsilon}\|_2\leq\rho} R_{\text{train}}(W+\boldsymbol{\epsilon},Z) + \frac{(\lambda C_{m,u}+8)(m+u)}{8mu}\\ + \frac{1}{\lambda}\left(\frac{1}{2}\left[1+d\log\left(1+\frac{\|W\|_2^2}{\rho^2}\left(1+\widetilde{C}_{m,u}\right)^2\right)\right] + \log\left(\frac{1}{6\delta}\right) + 2\log\left(\frac{6\pi mu}{m+u}\right)\right).\end{aligned}$$

This finishes the proof.

Appendix K. Proof of Corollary 16 and Theorem 17

Proof. We start from Eq. (82) and rewrite it as an inequality under the random sampling setting. Notice that here Z represents the randomness of selecting training labels after the dataset S is observed. For any $\lambda > 0$ and any exchangeable prior P:

$$\begin{aligned}&\mathbb{E}_S\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\lambda\mathbb{E}_{W\sim Q}[R_{\text{test}}(W,S)-R_{\text{train}}(W,S)]-\mathrm{D}_{\mathrm{KL}}(Q\|P)}\right]\\ =&\mathbb{E}_S\mathbb{E}_{W\sim P}\left[e^{\lambda(R_{\text{test}}(W,S)-R_{\text{train}}(W,S))}\right]\\ =&\int_w \mathbb{E}_S\left[e^{\frac{\lambda}{u}\sum_{i=m+1}^{m+u}\ell(w,S_i)-\frac{\lambda}{m}\sum_{i=1}^{m}\ell(w,S_i)}\right]\mathrm{d}P(w)\\ =&\int_w \mathbb{E}_S\mathbb{E}_Z\left[e^{\frac{\lambda}{u}\sum_{i=m+1}^{m+u}\ell(w,S_{Z_i})-\frac{\lambda}{m}\sum_{i=1}^{m}\ell(w,S_{Z_i})}\right]\mathrm{d}P(w)\\ \leq& e^{\frac{\lambda^2B^2C_{m,u}(m+u)}{8mu}}.\end{aligned} \tag{105}$$

By Markov's inequality, for any $0 < \delta < 1$ and $\lambda > 0$, with probability $1 - \delta$ over the randomness of Z and S:

$$\sup_{Q \in \mathcal{P}(\mathcal{W})} \mathbb{E}_{W \sim Q} \left[R_{\text{test}}(W, S) - R_{\text{train}}(W, S) \right] \leq \frac{\mathrm{D}_{\mathrm{KL}}(Q||P) + \log(1/\delta)}{\lambda} + \frac{\lambda B^2 C_{m,u}(m+u)}{8mu}.$$

By Jensen's inequality, Eq. (105) implies that for any $\lambda > 0$:

$$\sup_{Q \in \mathcal{P}(\mathcal{W})} \mathbb{E}_{W \sim Q} \mathbb{E}_S \left[\lambda \left[R_{\text{test}}(W, S) - R_{\text{train}}(W, S) \right] - \mathrm{D}_{\mathrm{KL}}(Q||P) \right] \leq \frac{\lambda^2 B^2 C_{m,u}(m+u)}{8mu}.$$

Denote by S' the independent copy of S. Let $Q = P_{W|S}$, $P = \mathbb{E}_{S'} \left[P_{W|S'} \right]$, we have

$$\mathbb{E}_{W,S} \left[R_{\text{test}}(W, S) - R_{\text{train}}(W, S) \right] \leq \frac{\mathbb{E}_S \left[\mathrm{D}_{\mathrm{KL}} \left(P_{W|S} || \mathbb{E}_{S'} \left[P_{W|S'} \right] \right) \right]}{\lambda} + \frac{\lambda B^2 C_{m,u}(m+u)}{8mu}.$$

Using the fact that $\mathbb{E}_S \left[\mathrm{D}_{\mathrm{KL}} \left(P_{W|S} || \mathbb{E}_{S'} \left[P_{W|S'} \right] \right) \right] = I(W; S)$, optimizing λ on the r.h.s. yields the following information-theoretic bound,

$$\mathbb{E}_{W,S} \left[R_{\text{test}}(W, S) - R_{\text{train}}(W, S) \right] \leq \sqrt{\frac{B^2 C_{m,u} I(W; S)(m+u)}{2mu}}. \tag{106}$$

For the self-consistency of this paper, we provide the proof of Catoni's result (Catoni, 2007, Theorem 3.1.2). From now on we assume that $\ell(\cdot)$ is the zero-one loss and $u = km$. For any $w \in \mathcal{W}$ and $\lambda > 0$:

$$\begin{aligned}
\log \left(\mathbb{E}_U \left[e^{-\frac{\lambda}{m} \sum_{i=1}^m \ell\left(w, s_{i+mU_{i,0}}\right)} \right] \right) &= \log \left(\mathbb{E}_U \left[\prod_{i=1}^m e^{-\frac{\lambda}{m} \ell\left(w, s_{i+mU_{i,0}}\right)} \right] \right) \\
&= \log \left(\prod_{i=1}^m \mathbb{E}_{U_i} \left[e^{-\frac{\lambda}{m} \ell\left(w, s_{i+mU_{i,0}}\right)} \right] \right) \\
&= \sum_{i=1}^m \log \left(\mathbb{E}_{U_i} \left[e^{-\frac{\lambda}{m} \ell\left(w, s_{i+mU_{i,0}}\right)} \right] \right) \\
&\leq m \log \left(\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{U_i} \left[e^{-\frac{\lambda}{m} \ell\left(w, s_{i+mU_{i,0}}\right)} \right] \right).
\end{aligned} \tag{107}$$

Since $\ell(\cdot)$ is the zero-one loss, it can only take values in $\{0, 1\}$, which implies that

$$e^{-\frac{\lambda}{m} \ell\left(w, s_{i+mU_{i,0}}\right)} = 1 - \ell\left(w, s_{i+mU_{i,0}}\right) + e^{-\frac{\lambda}{m}} \ell\left(w, s_{i+mU_{i,0}}\right), i \in [m]. \tag{108}$$

Recall that $\Phi_a(p) = -a^{-1} \log\left(1 - [1 - e^{-a}]p\right)$ with $a \in \mathbb{R}$ and $0 < p < 1$, we have

$$\begin{aligned}
&\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{U_i} \left[e^{-\frac{\lambda}{m} \ell\left(w, s_{i+mU_{i,0}}\right)} \right] \\
=&\frac{1}{m} \sum_{i=1}^m \mathbb{E}_{U_i} \left[1 - \ell\left(w, s_{i+mU_{i,0}}\right) + e^{-\frac{\lambda}{m}} \ell\left(w, s_{i+mU_{i,0}}\right) \right] \\
=&1 - \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{U_i} \left[\ell\left(w, s_{i+mU_{i,0}}\right) \right] + \frac{e^{-\frac{\lambda}{m}}}{m} \sum_{i=1}^m \mathbb{E}_{U_i} \left[\ell\left(w, s_{i+mU_{i,0}}\right) \right] \\
=&1 - \bar{R}(w, s) + e^{-\frac{\lambda}{m}} \bar{R}(w, s),
\end{aligned} \tag{109}$$

where $\bar{R}(w,s) \triangleq \frac{1}{(k+1)m}\sum_{i=1}^{(k+1)m}\ell(w,s_i)$. Combining the above results and taking expectation on both side over S gives

$$\mathbb{E}_S\left[e^{\lambda\Phi_{\lambda/m}(\bar{R}(w,S))}\mathbb{E}_U\left[e^{-\frac{\lambda}{m}\sum_{i=1}^{m}\ell\left(w,S_{i+mU_{i,0}}\right)}\right]\right] \leq 1. \tag{110}$$

For any $\lambda > 0$ and any partially exchangeable prior P, by Lemma 22 we have:

$$\begin{aligned}
&\mathbb{E}_S\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\mathbb{E}_{W\sim Q}\left[\lambda\left(\Phi_{\lambda/m}(\bar{R}(W,S))-R_{\text{train}}(W,S)\right)\right]-\mathrm{D}_{\mathrm{KL}}(Q||P)}\right]\\
=&\mathbb{E}_S\mathbb{E}_{W\sim P}\left[e^{\lambda\left(\Phi_{\lambda/m}(\bar{R}(W,S))-R_{\text{train}}(W,S)\right)}\right]\\
=&\int_w \mathbb{E}_S\left[e^{\lambda\Phi_{\lambda/m}(\bar{R}(w,S))}e^{-\frac{\lambda}{m}\sum_{i=1}^{m}\ell(w,S_i)}\right]\mathrm{d}P(w)\\
=&\int_w \mathbb{E}_S\mathbb{E}_U\left[e^{\lambda\Phi_{\lambda/m}(\bar{R}(w,S))}e^{-\frac{\lambda}{m}\sum_{i=1}^{m}\ell(w,S_i)}\right]\mathrm{d}P(w)\\
=&\int_w \mathbb{E}_S\left[e^{\lambda\Phi_{\lambda/m}(\bar{R}(w,S))}\mathbb{E}_U\left[e^{-\frac{\lambda}{m}\sum_{i=1}^{m}\ell(w,S_{i+mU_{i,0}})}\right]\right]\mathrm{d}P(w) \leq 1.
\end{aligned} \tag{111}$$

By Markov's inequality, for any $0 < \delta < 1$ and $\lambda > 0$, with probability $1-\delta$ over the randomness of S:

$$\sup_{Q\in\mathcal{P}(\mathcal{W})}\mathbb{E}_{W\sim Q}\left[\Phi_{\lambda/m}(\bar{R}(W,S)) - R_{\text{train}}(W,S)\right] \leq \frac{\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{\lambda}. \tag{112}$$

For any fixed posterior Q, the fact that $\Phi_{\lambda/m}(\cdot)$ is a convex function gives

$$\Phi_{\lambda/m}\left(\mathbb{E}_{W\sim Q}\left[\bar{R}(W,S)\right]\right) \leq \mathbb{E}_{W\sim Q}\left[R_{\text{train}}(W,S)\right] + \frac{\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{\lambda}, \tag{113}$$

which implies that with probability at least $0 < \delta < 1$ over the randomness of S:

$$\begin{aligned}
&\mathbb{E}_{W\sim Q}\left[R_{\text{test}}(W,S)\right]\\
\leq&\frac{(k+1)\left(1-\exp\left(-\frac{\lambda\mathbb{E}_{W\sim Q}[R_{\text{train}}(W,S)]+\mathrm{D}_{\mathrm{KL}}(Q||P)+\log(1/\delta)}{m}\right)\right)}{k(1-e^{-\lambda/m})} - \frac{1}{k}\mathbb{E}_{W\sim Q}\left[R_{\text{train}}(W,S)\right].
\end{aligned} \tag{114}$$

This finishes the proof.

Appendix L. Proof of Theorem 19

Proof. Assume that the parameter $w \in \mathbb{R}^d$. Denote by

$$\boldsymbol{H}(w,Z) \triangleq \frac{1}{u}\sum_{i=m+1}^{m+u}\frac{\partial^2\ell(w,s_{Z_i})}{\partial w^2} - \frac{1}{m}\sum_{i=1}^{m}\frac{\partial^2\ell(w,s_{Z_i})}{\partial w^2} \triangleq \boldsymbol{H}_{\text{test}}(w,Z) - \boldsymbol{H}_{\text{train}}(w,Z)$$

a function that maps n random variables $Z_1,\ldots,Z_n$ to a self-joint matrix, namely, the Hessian of the transductive generalization gap $\frac{\partial^2\mathcal{E}(w,Z)}{\partial w^2}$. We remark that $\boldsymbol{H}(w,Z) \in \mathbb{R}^{d\times d}$ and $\mathbb{E}[\boldsymbol{H}(w,Z)] = \boldsymbol{O}$. We firstly establish an upper bound for the moment-generating

function $\mathbb{E}_Z\left[e^{\theta \mathrm{Tr}(\boldsymbol{H}(w,Z))}\right]$ by the matrix martingale technique, where w is a fixed realization of W. To this end, we construct the following Doob's martingale difference sequences

$$\xi_i \triangleq \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_i] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1}], i \in [n]. \tag{115}$$

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

$$\begin{aligned}
\xi_i^{\inf} &\triangleq \inf_z \left\|\mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=z] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1}]\right\|,\\
\xi_i^{\sup} &\triangleq \sup_z \left\|\mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=z] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1}]\right\|,
\end{aligned}$$

we have $\xi_i^{\inf} \le \|\xi_i\| \le \xi_i^{\sup}$. One can find that

$$\begin{aligned}
\xi_i^{\sup} - \xi_i^{\inf} =& \sup_z \left\|\mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=z] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1}]\right\|\\
&- \inf_z \left\|\mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=z] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1}]\right\|\\
=& \sup_{z,\tilde{z}} \Big\{ \left\|\mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=z] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1}]\right\|\\
&- \left\|\mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=\tilde{z}] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1}]\right\| \Big\}\\
\le& \sup_{z,\tilde{z}} \left\{\left\|\mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=z] - \mathbb{E}[\boldsymbol{H}(w,Z)|Z_1,\ldots,Z_{i-1},Z_i=\tilde{z}]\right\|\right\}.
\end{aligned}$$

After $Z_1,\ldots,Z_m$ are given, the values of $Z_{m+1},\ldots,Z_n$ do not affect the value of $\boldsymbol{H}(w,Z)$. Thus, $\xi_i^{\sup} - \xi_i^{\inf} = 0$ holds for $i = m+1,\ldots,n$. Now we discuss the case that $i \in [m]$. Similar to the proof in Appendix B, we consider a realization z_j of Z_j with $j \in [i-1]$. Denote by z_i and $\widetilde{z}_i$ the realizations of Z_i and $\widetilde{Z}_i$, respectively. Let $Z_{i+1:n} = (z_1,\ldots,z_{i-1},z_i,Z_{i+1},\ldots,Z_n)$ be the sequence where $Z_{i+1},\ldots,Z_n$ are obtained by sampling without replacement from $[n]\backslash\{z_1,\ldots,z_{i-1},z_i\}$, and $\widetilde{Z}_{i+1:n} = (z_1,\ldots,z_{i-1},\tilde{z}_i,\widetilde{Z}_{i+1},\ldots,\widetilde{Z}_n)$ be the sequence where $\widetilde{Z}_{i+1},\ldots,\widetilde{Z}_n$ are obtained by sampling without replacement from $[n]\backslash\{z_1,\ldots,z_{i-1},\tilde{z}_i\}$. Now we need to compute the maximum value of the expectation $\mathbb{E}[\boldsymbol{H}(w,Z_{i+1:n}) - \boldsymbol{H}(w,\widetilde{Z}_{i+1:n})]$ over any possible values $z_1,\ldots,z_{i-1},z_i,\widetilde{z}_i$. To this end, it is sufficient to consider two cases: (1) z_i appears at positions $i+1$ to m of $\widetilde{Z}_{i+1:n}$, and (2) z_i appears at positions $m+1$ to n of $\widetilde{Z}_{i+1:n}$. For case (1), the expectation is equal to zero, since for each realization of $Z_{i+1:n}$, we can always find a corresponding and unique realization of $\widetilde{Z}_{i+1:n}$ such that they are equal to each other. For case (2), for each realization of $Z_{i+1:n}$, we can always find a corresponding and unique realization of $\widetilde{Z}_{i+1:n}$ such that Z and $\widetilde{Z}$ only differs at the i-th position. Thus, the maximum value of the expectation is given by

$$\frac{m+u}{mu}\left\|\frac{\partial^2 \ell(w,s_{\widetilde{z}_i})}{\partial w^2} - \frac{\partial^2 \ell(w,s_{z_i})}{\partial w^2}\right\| \le \frac{2(m+u)}{mu}\sup_z \left\|\frac{\partial^2 \ell(w,z)}{\partial w^2}\right\| = \frac{2(m+u)B_H}{mu}. \tag{116}$$

The last step is to compute the probability of z_i appearing at positions $m+1$ to n of $\widetilde{Z}_{i+1:n}$. To this end, we need to sample $m-i$ elements among the rest $n-i-1$ elements (that is, the set $[n]\backslash\{z_1,\ldots,z_i,\widetilde{z}_i\}$), and then apply permutation on them. Thus, the probability is

given by $\frac{u!(m-i)!C_{n-i-1}^{m-i}}{(n-i)!}$. Put all the above ingredients together, we obtain

$$\xi_i^{\sup} - \xi_i^{\inf} = \begin{cases} \frac{u!(m-i)!C_{n-i-1}^{m-i}}{(n-i)!} \cdot \frac{2(m+u)B_H}{mu} = \frac{2(m+u)B_H}{m(m+u-i)}, & i = 1, \ldots, m, \\ 0, & i = m+1, \ldots, n. \end{cases} \tag{117}$$

Denote by

$$\mathbf{A}_i \triangleq \begin{cases} \frac{2(m+u)B_H}{m(m+u-i)}\mathbf{I}_d, & i = 1, \ldots, m, \\ \boldsymbol{O}, & i = m+1, \ldots, n. \end{cases} \tag{118}$$

Since $\lambda_{\max}\left(\xi_i^2\right) = \|\xi_i\|^2 \leq (\xi_i^{\sup} - \xi_i^{\inf})^2$, we conclude that for $i \in [n]$, $\mathbf{A}_i^2 - \xi_i^2$ is a semi-positive definite matrix. Denote by ε the standard Rademacher variable, that is, $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$, which is independent to Z. For any $\theta \in \mathbb{R}$ we have

$$\begin{aligned} &\mathbb{E}\left[\operatorname{Tr}\left(\exp\left\{\theta\sum_{i=1}^n \xi_i\right\}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\operatorname{Tr}\left(\exp\left\{\theta\sum_{i=1}^{n-1}\xi_i + \theta\xi_n\right\}\right)\middle| Z_1, \ldots, Z_{n-1}\right]\right] \\ \leq &\mathbb{E}\left[\mathbb{E}\left[\operatorname{Tr}\left(\exp\left\{\theta\sum_{i=1}^{n-1}\xi_i + 2\varepsilon\theta\xi_n\right\}\right)\middle| Z_1, \ldots, Z_{n-1}\right]\right] \\ \leq &\mathbb{E}\left[\operatorname{Tr}\left(\exp\left\{\theta\sum_{i=1}^{n-1}\xi_i + \log\mathbb{E}\left[\exp\left\{2\varepsilon\theta\xi_n\right\}|Z_1, \ldots, Z_{n-1}\right]\right\}\right)\right] \\ \leq &\mathbb{E}\left[\operatorname{Tr}\left(\exp\left\{\theta\sum_{i=1}^{n-1}\xi_i + 2\theta^2\mathbf{A}_n^2\right\}\right)\right], \end{aligned} \tag{119}$$

where the first line is due to the tower property of conditional expectation. The second, the third, and the forth line is obtained by Lemma 7.6, Corollary 3.3 and Lemma 7.7 of Tropp (2012), respectively. By iteration we obtain

$$\begin{aligned} &\mathbb{E}\left[\operatorname{Tr}\left(\exp\left\{\theta\sum_{i=1}^n \xi_i\right\}\right)\right] \leq \operatorname{Tr}\left(\exp\left\{2\theta^2\sum_{i=1}^n \mathbf{A}_i^2\right\}\right) \\ = &\operatorname{Tr}\left(\exp\left\{\frac{8\theta^2B_H^2(m+u)^2}{m^2}\left(\sum_{i=1}^m \frac{1}{(m+u-i)^2}\right)\mathbf{I}_d\right\}\right) \\ = &\operatorname{Tr}\left(\sum_{k=0}^\infty \frac{1}{k!}\cdot\left[\frac{8\theta^2B_H^2(m+u)^2}{m^2}\left(\sum_{i=1}^m \frac{1}{(m+u-i)^2}\right)\mathbf{I}_d\right]^k\right) \\ = &\sum_{k=0}^\infty \frac{d}{k!}\cdot\left[\frac{8\theta^2B_H^2(m+u)^2}{m^2}\left(\sum_{i=1}^m \frac{1}{(m+u-i)^2}\right)\right]^k \\ = &d\exp\left\{\frac{8\theta^2B_H^2(m+u)^2}{m^2}\left(\sum_{i=1}^m \frac{1}{(m+u-i)^2}\right)\right\} \leq d\exp\left\{\frac{8d^2\theta^2B_H^2(m+u)^2}{m(u-1/2)(m+u-1/2)}\right\}. \end{aligned}$$

Due to the symmetry of m and u, we have

$$\mathbb{E}\left[\operatorname{Tr}\left(\exp\left\{\theta\sum_{i=1}^n \xi_i\right\}\right)\right] \leq d\exp\left\{\frac{8\theta^2B_H^2(m+u)^2}{u(m-1/2)(m+u-1/2)}\right\}. \tag{120}$$

Then the final bound is obtained by taking the smaller one of these two bounds,

$$\begin{aligned}\mathbb{E}\left[\mathrm{Tr}\left(\exp\left\{\theta\sum_{i=1}^{n}\xi_i\right\}\right)\right] &\leq d\exp\left\{\frac{8\theta^2B_H^2(m+u)^2}{mu(m+u-1/2)}\cdot\frac{2\max(m,u)}{2\max(m,u)-1}\right\}\\ &= d\exp\left\{\frac{8\theta^2B_H^2C_{m,u}(m+u)}{mu}\right\}.\end{aligned} \tag{121}$$

By noticing that $\mathrm{Tr}(\boldsymbol{H}(w,z)) \leq d\lambda_{\max}(\boldsymbol{H}(w,z))$ that holds for any w and z, for any $\theta > 0$ we have

$$\begin{aligned}&\mathbb{E}_Z\left[e^{\theta\mathrm{Tr}(\boldsymbol{H}(w,Z))}\right] \leq \mathbb{E}_Z\left[e^{d\theta\lambda_{\max}(\boldsymbol{H}(w,Z))}\right]\\ =&\mathbb{E}_Z\left[\lambda_{\max}(e^{d\theta\boldsymbol{H}(w,Z)})\right] \leq \mathbb{E}_Z\left[\mathrm{Tr}\left(e^{d\theta\boldsymbol{H}(w,Z)}\right)\right]\\ =&\mathbb{E}\left[\mathrm{Tr}\left(\exp\left\{d\theta\sum_{i=1}^{n}\xi_i\right\}\right)\right] \leq d\exp\left\{\frac{8\theta^2B_H^2C_{m,u}(m+u)}{mu}\right\}.\end{aligned} \tag{122}$$

By Lemma 22, for any $\theta > 0$:

$$\begin{aligned}&\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[e^{\theta\mathrm{Tr}(\boldsymbol{H}(W,Z))}\right] = \mathbb{E}_Z\left[e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\theta\mathbb{E}_{W\sim Q}[\mathrm{Tr}(\boldsymbol{H}(W,Z))]-\mathrm{D_{KL}}(Q||P)}\right]\\ \geq& e^{\mathbb{E}_Z\left[\sup_{Q\in\mathcal{P}(\mathcal{W})}\theta\mathbb{E}_{W\sim Q}[\mathrm{Tr}(\boldsymbol{H}(W,Z))]-\mathrm{D_{KL}}(Q||P)\right]}\\ \geq& e^{\sup_{Q\in\mathcal{P}(\mathcal{W})}\mathbb{E}_Z\left[\mathbb{E}_{W\sim Q}[\theta\mathrm{Tr}(\boldsymbol{H}(W,Z))]-\mathrm{D_{KL}}(Q||P)\right]}.\end{aligned} \tag{123}$$

Combining Eq. (122) with Eq. (123), for any $\theta > 0$ we have

$$\sup_{Q\in\mathcal{P}(\mathcal{W})}\mathbb{E}_Z\left[\mathbb{E}_{W\sim Q}\left[\theta\mathrm{Tr}(\boldsymbol{H}(W,Z))\right]-\mathrm{D_{KL}}(Q||P)\right] \leq \frac{8d^2\theta^2B_H^2C_{m,u}(m+u)}{mu}+\log(d).$$

Now we put $Q = P_{W_T|Z}$ and $P = \mathbb{E}[P_{W_T|Z}] = P_{W_T}$ and get

$$\theta\mathbb{E}_{W_T,Z}\left[\mathrm{Tr}(\boldsymbol{H}(W,Z))\right] \leq I(W_T;Z)+\frac{8d^2\theta^2B_H^2C_{m,u}(m+u)}{mu}+\log(d), \tag{124}$$

which implies that

$$\left|\mathbb{E}_{W_T,Z}\left[\mathrm{Tr}(\boldsymbol{H}(W_T,Z))\right]\right| \leq \sqrt{\frac{32d^2B_H^2C_{m,u}(I(W_T;Z)+\log(d))(m+u)}{mu}}. \tag{125}$$

By Taylor's expansion on $R(w,z)$ w.r.t. w and using $N_{1:T} \sim \mathcal{N}(0, \sum_{t=1}^{T}\sigma_t^2) \in \mathbb{R}^d$, we have

$$\begin{aligned}&R_{\mathrm{train}}(w_T+N_{1:T},z)-R_{\mathrm{train}}(w_T,z)\\ =&\sum_{i=1}^{d}(N_{1:T})_i\left(\frac{\partial R_{\mathrm{train}}(W,Z)}{\partial w}\Big|_{w=w_T}\right)_i+\sum_{i,j=1}^{d}(N_{1:T})_i(N_{1:T})_j\left(\frac{\partial^2R_{\mathrm{train}}(W,Z)}{\partial w^2}\Big|_{w=w_T}\right)_{ij}\\ &+\sum_{i,j,k=1}^{d}(N_{1:T})_i(N_{1:T})_j(N_{1:T})_k\left(\frac{\partial^3R_{\mathrm{train}}(W,Z)}{\partial w^3}\Big|_{w=w_T}\right)_{ijk}+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right).\end{aligned} \tag{126}$$

Taking expectation on both side over $N_{1:T}$, we have

$$\begin{aligned}
&\mathbb{E}_{N_{1:T}}\left[R_{\text{train}}(w_T+N_{1:T},z)-R_{\text{train}}(w_T,z)\right]\\
=&\sum_{i=1}^{d}\mathbb{E}\left[\left(N_{1:T}^2\right)_i\right]\left(\frac{\partial^2 R_{\text{train}}(W,Z)}{\partial w^2}\bigg|_{w=w_T}\right)_{ii}\\
=&\text{Tr}\left(\mathbb{E}[N_{1:T}]\mathbb{E}[N_{1:T}^\top]\boldsymbol{H}_{\text{train}}(w_T,z)\right)+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right),
\end{aligned}\tag{127}$$

which implies that

$$\begin{aligned}
&\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{train}}(W_T+N_{1:T},Z)-R_{\text{train}}(W_T,Z)\right]\\
=&\text{Tr}\left(\mathbb{E}[N_{1:T}]\mathbb{E}[N_{1:T}^\top]\left(\mathbb{E}_{Z,W_T}\left[\boldsymbol{H}_{\text{train}}(W_T,Z)\right]\right)\right)+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right).
\end{aligned}\tag{128}$$

Similarly we have

$$\begin{aligned}
&\mathbb{E}_{Z,W_T,N_{1:T}}\left[R_{\text{test}}(W_T+N_{1:T},Z)-R_{\text{test}}(W_T,Z)\right]\\
=&\text{Tr}\left(\mathbb{E}[N_{1:T}]\mathbb{E}[N_{1:T}^\top]\left(\mathbb{E}_{Z,W_T}\left[\boldsymbol{H}_{\text{test}}(W_T,Z)\right]\right)\right)+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right).
\end{aligned}\tag{129}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{Z,W_T,N_{1:T}}\left[\left[R_{\text{train}}(W_T+N_{1:T},Z)-R_{\text{train}}(W_T,Z)\right]-\left[R_{\text{test}}(W_T+N_{1:T},Z)-R_{\text{test}}(W_T,Z)\right]\right]\\
=&\text{Tr}\left(\mathbb{E}[N_{1:T}]\mathbb{E}[N_{1:T}^\top]\left(\mathbb{E}_{Z,W_T}[\boldsymbol{H}_{\text{train}}(W_T,Z)-\boldsymbol{H}_{\text{test}}(W_T,Z)]\right)\right)+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right)\\
=&\left(\sum_{t=1}^{T}\sigma_t^2\right)\text{Tr}\left(\mathbb{E}_{Z,W_T}[\boldsymbol{H}_{\text{train}}(W_T,Z)-\boldsymbol{H}_{\text{test}}(W_T,Z)]\right)+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right)\\
\le&\left(\sum_{t=1}^{T}\sigma_t^2\right)\left|\mathbb{E}_{W_T,Z}\left[\text{Tr}(\boldsymbol{H}(W,Z))\right]\right|+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right)\\
\le&\left(\sum_{t=1}^{T}\sigma_t^2\right)\sqrt{\frac{32d^2B_H^2C_{m,u}(I(W_T;Z)+\log(d))(m+u)}{mu}}+\mathcal{O}\left(\left(\sum_{t=1}^{T}\sigma_t^2\right)^2\right).
\end{aligned}$$

This finishes the proof.

Appendix M. Proof of Theorem 20

Proof. This proof is inspired by the work of Neu et al. (2021); Wang and Mao (2022). We use $\widetilde{W}_T$ as the abbreviation of $W_T+N_{1:T}$. Following the work of Wang and Mao (2022), the mutual information term is decomposed by

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

By iteration we obtain

$$\begin{aligned} I(Z;\widetilde{W}_T) \leq & \sum_{t=1}^{T} I\left(-\frac{\eta}{\sqrt{V_t}+\epsilon}\odot g(W_{t-1},B_t(Z))+N_t;Z\middle|\widetilde{W}_{t-1}\right) \\ = & \sum_{t=2}^{T} I\left(-\frac{\eta}{\sqrt{V_t}+\epsilon}\odot g(\widetilde{W}_{t-1}-N_{1:t-1},B_t(Z))+N_t;Z\middle|\widetilde{W}_{t-1}\right) \\ & + I\left(-\frac{\eta}{\sqrt{V_1}+\epsilon}\odot g(\widetilde{W}_0,B_t(Z))+N_1;Z\middle|\widetilde{W}_0\right). \end{aligned} \tag{130}$$

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

$$\begin{aligned} & I(\Psi(V, y-U, X)+\sigma N; X|Y=y) \\ =& h(\Psi(V, y-U, X)+\sigma N|Y=y) - h(\Psi(V, y-U, X)+\sigma N|X, Y=y). \end{aligned} \tag{131}$$

For the first term in Eq. (131), by Theorem 2.7 of Polyanskiy and Wu (2025) we have

$$\begin{aligned} & h(\Psi(V, y-U, X)+\sigma N|Y=y) \\ \leq & \frac{d}{2}\log\left(\frac{2\pi e\mathbb{E}\left[\|\Psi(V, y-U, X)+\sigma N\|_2^2|Y=y\right]}{d}\right) \\ = & \frac{d}{2}\log\left(\frac{2\pi e\left(\mathbb{E}\left[\|\Psi(V, y-U, X)\|_2^2|Y=y\right]+\sigma^2\mathbb{E}\left[\|N\|_2^2\right]\right)}{d}\right) \\ = & \frac{d}{2}\log\left(\frac{2\pi e\left(\mathbb{E}\left[\|\Psi(V, y-U, X)\|_2^2|Y=y\right]+d\sigma^2\right)}{d}\right). \end{aligned} \tag{132}$$

For the second term in Eq. (131), we have

$$\begin{aligned} & h(\Psi(V, y-U, X)+\sigma N|X, Y=y) \geq h(\Psi(V, y-U, X)+\sigma N|U, V, X, Y=y) \\ = & h(\sigma N) = \frac{d}{2}\log\left(2\pi e\sigma^2\right). \end{aligned} \tag{133}$$

For $t=2,\ldots,T$, let $\sigma=\sigma_t$, $V=W^{[t-2]}\triangleq(W_0,\ldots,W_{t-2})$, $X=Z$, $Y=\widetilde{W}_{t-1}$, $U=N_{1:t-1}$ and

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

Plugging Eqs. (132,133) into Eq. (131), for $t=2,\ldots,T$ we have

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

which implies that

$$\begin{aligned}
&I\left(\Psi(W^{[t-2]},\widetilde{w}_{t-1}-N_{1:t-1},Z)+N_t;Z\middle|\widetilde{W}_{t-1}\right)\\
\leq&\int_{\widetilde{w}_{t-1}}\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\Psi(W^{[t-2]},\widetilde{w}_{t-1}-N_{1:t-1},Z)\right\|_2^2\middle|\widetilde{W}_{t-1}=\widetilde{w}_{t-1}\right]+1\right)\mathrm{d}P_{\widetilde{W}_{t-1}}(\widetilde{w}_{t-1})\\
\leq&\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\Psi(W^{[t-2]},\widetilde{W}_{t-1}-N_{1:t-1},Z)\right\|_2^2\right]+1\right)\\
=&\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\Psi(W^{[t-1]},Z)\right\|_2^2\right]+1\right).
\end{aligned}\tag{134}$$

By the same way, we have

$$I\left(-\frac{\eta}{\sqrt{V_1}+\epsilon}\odot g(\widetilde{W}_0,B_t(Z))+N_1;Z\middle|\widetilde{W}_0\right)\leq\frac{d}{2}\log\left(\frac{1}{d\sigma_1^2}\mathbb{E}\left[\left\|\Psi(W_0,Z)\right\|_2^2\middle|\widetilde{W}_0=\widetilde{w}_0\right]+1\right).\tag{135}$$

Plugging Eqs. (134,135) into Eq. (130), we have

$$\begin{aligned}
I(Z;\widetilde{W}_T)\leq&\sum_{t=1}^T\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\Psi(W^{[t-2]},W_{t-1},Z)\right\|_2^2\right]+1\right)\\
=&\sum_{t=1}^T\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\left\|\frac{\eta}{\sqrt{V_T(W^{[t-1]})}+\epsilon}\odot g(W_{t-1},B_t(Z))\right\|_2^2\right]+1\right).
\end{aligned}\tag{136}$$

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

$$\begin{aligned}
&M_t=\beta_1M_{t-1}+(1-\beta_1)g(W_{t-1},B_t(Z)),\\
&V_t=\beta_2V_{t-1}+(1-\beta_2)g(W_{t-1},Z)\odot g(W_{t-1},B_t(Z)),\\
&\hat{V}_t=\frac{V_t}{1-\beta_2^t},\hat{M}_t=\frac{M_t}{1-\beta_1^t},W_t=W_{t-1}-\frac{\eta}{\sqrt{\hat{V}_t}+\epsilon}\odot\hat{M}_t,
\end{aligned}$$

where β_1, β_2, η and ϵ are hyperparameters. 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

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

Following the above technique, one can find that

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

This completes the proof.

Appendix N. Proof of Proposition 21

Proof. Denote by $g(F_i, U_i, \widetilde{Z}, S) \triangleq r(F_{i,U_i}, Y_{\widetilde{Z}_{i,U_i}}) - r(F_{i,1-U_i}, Y_{\widetilde{Z}_{i,1-U_i}})$ the function of $(F_i, U_i, \widetilde{Z}, S)$. Recall that $F_{i,U_i} = f_W(X_{\widetilde{Z}_{i,U_i}})$ and $F_{i,1-U_i} = f_W(X_{\widetilde{Z}_{i,1-U_i}})$. Let f_i, $\widetilde{z}$ and s be the fixed realizations of F_i, $\widetilde{Z}$ and S, respectively. For any $\lambda \in \mathbb{R}$, by Hoeffding's inequality (Hoeffding, 1963),

$$\mathbb{E}_{U_i}\left[\exp\left\{\lambda g(f_i, U_i, \widetilde{z}, s)\right\}\right] \leq \exp\left\{\frac{\lambda^2 B^2}{2}\right\}, i \in [m]. \tag{137}$$

Let U_i' be the independent copy of U_i, by Lemma 22,

$$\begin{aligned}
&I(F_i; U_i | \widetilde{Z} = \widetilde{z}, S = s) \\
\geq &\lambda \mathbb{E}_{F_i, U_i | \widetilde{Z}=\widetilde{z}, S=s}\left[g(F_i, U_i, \widetilde{z}, s)\right] - \log \mathbb{E}_{F_i, U_i' | \widetilde{Z}=\widetilde{z}, S=s}\left[\exp\left\{\lambda g(F_i, U_i', \widetilde{z}, s)\right\}\right] \\
\geq &\lambda \mathbb{E}_{F_i, U_i | \widetilde{Z}=\widetilde{z}, S=s}\left[g(F_i, U_i, \widetilde{z})\right] - \frac{\lambda^2 B^2}{2},
\end{aligned} \tag{138}$$

which implies that

$$\left|\mathbb{E}_{F_i, U_i | \widetilde{Z}=\widetilde{z}, S=s}\left[g(F_i, U_i, \widetilde{z}, s)\right]\right| \leq B\sqrt{2 I^{\widetilde{z}, s}(F_i; U_i)}. \tag{139}$$

Taking expectation over the joint distribution of $\widetilde{Z}$ and S on both side yields

$$\mathbb{E}_{\widetilde{Z}, S}\left|\mathbb{E}_{F_i, U_i | \widetilde{Z}, S}\left[g(F_i, U_i, \widetilde{Z})\right]\right| \leq B \mathbb{E}_{\widetilde{Z}, S}\sqrt{2 I^{\widetilde{Z}, S}(F_i; U_i)}. \tag{140}$$

Then we have

$$\begin{aligned}
&\left|\mathbb{E}_{W,S}\left[R_{\text{test}}(W, S) - R_{\text{train}}(W, S)\right]\right| \\
=&\left|\mathbb{E}_{W,S,\widetilde{Z},U}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\ell\left(W, S_{\mathscr{Z}_i(\widetilde{Z},U)}\right) - \frac{1}{m}\sum_{i=1}^{m}\ell\left(W, S_{\mathscr{Z}_i(\widetilde{Z},U)}\right)\right]\right| \\
=&\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\mathbb{E}_{W,U_i|\widetilde{Z},S}\left[\ell\left(W, S_{\widetilde{Z}_{i,1-U_i}}\right) - \ell\left(W, S_{\widetilde{Z}_{i,U_i}}\right)\right]\right| \\
\leq&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\left|\mathbb{E}_{W,U_i|\widetilde{Z},S}\left[\ell\left(W, S_{\widetilde{Z}_{i,1-U_i}}\right) - \ell\left(W, S_{\widetilde{Z}_{i,U_i}}\right)\right]\right| \\
=&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\left|\mathbb{E}_{W,U_i|\widetilde{Z},S}\left[r\big(f_W(X_{\widetilde{Z}_{i,1-U_i}}), Y_{\widetilde{Z}_{i,1-U_i}}\big) - r\big(f_W(X_{\widetilde{Z}_{i,U_i}}), Y_{\widetilde{Z}_{i,U_i}}\big)\right]\right| \\
=&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\left|\mathbb{E}_{F_i,U_i|\widetilde{Z},S}\left[r(F_{i,1-U_i}, Y_{\widetilde{Z}_{i,1-U_i}}) - r(F_{i,U_i}, Y_{\widetilde{Z}_{i,U_i}})\right]\right| \\
=&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\left|\mathbb{E}_{F_i,U_i|\widetilde{Z},S}\left[g(F_i, U_i, \widetilde{Z}, S)\right]\right| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\sqrt{2 I^{\widetilde{Z},S}(F_i; U_i)}.
\end{aligned} \tag{141}$$

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

$$|\mathbb{E}_{W,S}\left[R_{\text{test}}(W,S) - R_{\text{train}}(W,S)\right]| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\sqrt{2I^{\widetilde{Z},S}(L_i;U_i)}, \tag{142}$$

$$|\mathbb{E}_{W,S}\left[R_{\text{test}}(W,S) - R_{\text{train}}(W,S)\right]| \leq \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S}\sqrt{2I^{\widetilde{Z},S}(\Delta_i;U_i)}. \tag{143}$$

For the cases that $u = km, k \in \mathbb{N}_+$, define $F_i \triangleq (f_W(X_{\widetilde{Z}_{i,0}}), \ldots, f_W(X_{\widetilde{Z}_{i,k}})), i \in [m]$ and $L_{i,:} \triangleq (\ell(W, S_{\widetilde{Z}_{i,0}}), \ldots, \ell(W, S_{\widetilde{Z}_{i,k}})), i \in [m]$ be the prediction of model and the sequence of loss values. Denote by $g(F_i, U_i, \widetilde{Z}, S) \triangleq \frac{1}{k}\sum_{j=1}^{k} r\big(F_{i,U_{i,j}}, Y_{\widetilde{Z}_{i,U_{i,j}}}\big) - r\big(F_{i,U_{i,0}}, Y_{\widetilde{Z}_{i,U_{i,0}}}\big)$ and $g(F_i, U_i, \widetilde{Z}, S) \triangleq \frac{1}{k}\sum_{j=1}^{k} L_{i,U_{i,j}}, -L_{i,U_{i,0}}$ the function of $(F_i, U_i, \widetilde{Z}, S)$ and $(L_i, U_i, \widetilde{Z}, S)$, respectively. Following the above process one can verify that Eqs. (141,142) still hold under the cases that $u = km, k \in \mathbb{N}_+$. This finishes the proof.

Appendix O. Experimental Details

We first discuss how to estimate the expected transductive generalization gaps and the derived bounds. Notice that computing their accurate values is not applicable since we need to run the algorithm on $(m+u)!$ splits in total. Therefore we use Monte Carlo simulation to estimate these expectations based on finite samples. For semi-supervised learning, the sampling process is as follows: (i) randomly draw t_1 full data points S by each time sampling $m+u$ images from the raw images set, (ii) randomly draw t_2 transductive supersamples $\widetilde{Z}$ based on Definition 9, (iii) randomly draw t_3 selector sequence U and obtain the training and test samples set according to Section 4.2. Now we discuss the estimation of the transductive generalization gap and the upper bounds established in Proposition 21. Take Eq. (30) as an example, for each $(s, \widetilde{z})$ we use the mean value over t_3 samples of U to estimate the conditional term expectation term $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{F_i,U_i|s,\widetilde{z}}\left[g(F_i, U_i, s, \widetilde{z})\right]$. After that, we use t_1 samples of S and t_2 samples of $\widetilde{Z}$ to estimate the expected generalization gap, whose mean and standard deviation are shown in Figure 1. Similarly, we use a plug-in estimator (Paninski, 2003) to estimate the disentangled mutual information $I^{s,\widetilde{z}}(F_i;U_i)$ over the t_3 samples of U. Then the upper bounds in Proposition 21 are estimated by the t_1 samples of S and t_2 samples of $\widetilde{Z}$, whose mean and standard deviation are shown in Figure 1. For transductive graph learning, the estimation process generally follows that of semi-supervised learning. The only difference is that we do not need to consider the sampling of S. 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. The results are shown in Figure 2 and Figure 3.

Now we detail the settings of network architecture and hyperparameters. For the semi-supervised learning task, the network architectures 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 the transductive graph learning task, the architecture of GAT and GPR-GNN follows the ones used by Chien et al. (2021). We set $t_2 = 5$ and $t_3 = 50$ on all datasets. We adopt the code released by Chien et al. (2021) to

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 for CIFAR-10. $\times 6$ means repeating the layer for 6 times.

generate the cSBMs datasets with $n \in \{500, 1000, 2000\}$ and $\phi \in \{-0.5, 0.5\}$. Recall that the number of training data points is defined by $m \triangleq \frac{n}{k+1}$. To ensure that $m \in \mathbb{N}_+$, we set $k = 1$ for Cora and Actor, and $k = 2$ for CiteSeer and Chameleon.

References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Gholamali Aminian, Laura Toni, and Miguel R. D. Rodrigues. Jensen-Shannon information based characterization of the generalization error of learning algorithms. In *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 *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 *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 *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.

Jean-Yves Audibert and Olivier Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8(32):863–889, 2007.

Pradeep Kr. Banerjee and Guido Montúfar. Information complexity and generalization bounds. In *IEEE International Symposium on Information Theory*, pages 676–681, 2021.

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 *Annual ACM 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 *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 *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.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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 *International Conference on Artificial Intelligence and Statistics*, pages 8673–8699, 2022.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint Laboratoire de Probabilités et Modèles Aléatoires 840*, 2003.

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 *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 *International Conference on Machine Learning*, 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 *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 *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.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 604–612, 2021.

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

Ran El-Yaniv and Dmitry Pechyony. Transductive Rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35(1):193–234, 2009.

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.

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

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*, page 353–360, 2009.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *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 *International Joint Conference on Artificial Intelligence*, pages 2156–2162, 2018.

Peter Grünwald, Thomas Steinke, and Lydia Zakynthinou. PAC-Bayes, MAC-Bayes and conditional mutual information: Fast rate bounds that handle general VC classes. In *Conference on Learning Theory*, pages 2217–2247, 2021.

Benjamin Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the Second Congress of the French Mathematical Society*, 2019.

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 *International Conference on Machine Learning*, 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. In *IEEE International Symposium on Information Theory*, pages 2487–2492, 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 *International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 639–648, 2020.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 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.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

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 *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, page 665–674, 2021.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *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 *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.

Haotian Ju, Dongyue Li, and Hongyang R. Zhang. Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. In *International Conference on Machine Learning*, pages 10431–10461, 2022.

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

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pages 11148–11161, 2022.

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. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

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

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Uncertainty in Artificial Intelligence*, pages 275–282, 2002.

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 *International Conference on Machine Learning*, pages 11906–11917, 2022.

John Langford and Matthias Seeger. *Bounds for Averaging Classifiers*. Technical Report CMU-CS-01-102, Carnegie Mellon University, 2001.

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

Jaejun Lee, Minsung Hwang, and Joyce Jiyoung Whang. PAC-Bayesian generalization bounds for knowledge graph representation learning. In *International Conference on Machine Learning*, pages 26589–26620, 2024.

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 *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 *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 *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. *Journal of Artificial Intelligence Research*, 61(1):761–786, 2018.

Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. PAC-Bayesian generalization bounds for adversarial generative models. In *International Conference on Machine Learning*, pages 24271–24290, 2023.

David A. McAllester. Some PAC-Bayesian theorems. In *Annual Conference on Computational Learning Theory*, pages 230–234, 1998.

David A. McAllester. PAC-Bayesian model averaging. In *Annual Conference on Computational Learning Theory*, pages 164–170, 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 *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*, 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 Computation*, 15 (6):1191–1253, 2003.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *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, 2025.

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 *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.

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

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110):1–43, 2024.

William H. Rogers and Terry J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3):506–514, 1978.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *International Conference on Artificial Intelligence and Statistics*, 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.

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 *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.

Behzad M. Shahshahani and David 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.

Shai Shalev-Shwartz and Shai Ben-David. *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 *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 *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 *International Conference on Machine Learning*, pages 33674–33719, 2023.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *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 *Conference on Learning Theory*, pages 857–884, 2014.

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

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

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

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Hao Wang, Mario Diaz, José Cândido S. Santos Filho, and Flavio P. Calmon. An information-theoretic view of generalization via wasserstein distance. In *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 *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 *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 *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 *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.

Yingzhen Yang. Sharp generalization of transductive learning: A transductive local rademacher complexity approach. *arXiv preprint arXiv:2309.16858*, 2023.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *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. *Neural Networks*, 172:106125, 2024.

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

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Ruida Zhou, Chao Tian, and Tie Liu. Stochastic chaining and strengthened information-theoretic generalization bounds. In *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 *International Conference on Machine Learning*, pages 912–919, 2003.