Information-Theoretic Generalization Bounds for Transductive Learning and its Applications

1 Introduction

In the standard supervised learning paradigm [28], we receive a set of instances sampling independently from unknown distribution that are composed of attribute and targets. Our task is to build a learner that maps examples to their targets based on limited instances, guided by specific optimization algorithm. A modern and popular practice of this paradigm is training a deep neural network for image classification via stochastic gradient descent [16]. Referring to the prediction performance of learner on unseen examples, generalization is one of the central problems in machine learning theory. The past decades have witnessed efforts devoted to characterizing and understanding the generalization of learners. Classical approaches use the complexity of hypothesis space [2, 1], the stability of algorithm [5, 24] or the KL divergence between the prior and posterior on hypothesis space [17, 23] to measure generalization. Recently, mutual information and its variants are shown to serve as a ideal generalization measure, since they reflect both the impact of dataset and optimization on generalization. Research of this viewpoint originates from works [21, 22, 29], and is further developed by subsequent studies [18, 14, 15, 13]. One of the key results is that the mutual information between weights returned by learning algorithms and training datasets is sufficient to characterize the generalization performance of the learner.

The aforementioned results for supervised learning are far not enough to cover all machine learning scenarios. In real-world scenario, the collected data could come from multiple domains, and some examples could lack of targets due to the expensive cost of annotations. This raises the need of exploring new theory to measure the generalization, where the key challenge is relaxing the identical and independent assumptions of instances. We move towards this direction by analyzing a classical but important regime termed as transductive learning [27]. In this learning paradigm, we are provided with a fixed set of instances containing both labeled example and unlabeled examples, and our task is to build a learner that make prediction for those unlabeled ones. Note that under this setting, the examples to be predicted are available to learner. As a comparison, the examples to be predicted are not available to learner in supervised learning. Popular practices of transductive learning are semi-supervised learning and graph learning, while the latter has been widely adopted in areas such as financial fraud detection, drag discovery and recommendation system.

Existing results for transductive generalization bounds fall into three categories: complexity-based bounds derived by VC dimension [7] or transductive Rademacher complexity [11, 26], stability-based bounds [10, 8] and PAC-Bayesian bounds [9, 4]. These results are sufficient to provide learning guarantee for classical learner or algorithms such as transductive support vector machine and unlabeled-labeled representation. However, they are far from the optimal approaches to understanding the generalization behavior of deep transducitve model such as Graph Neural Network. The reasons are three folds. First, [12] shows that VC dimension results in trivial generalization error bounds, and transductive Rademacher complexity is algorithm-independent and fail to characterize the impact of optimization algorithms on generalization. Second, stability-based bounds depend on Lipschitz and smoothness constant [6], which is difficulty to compute for deep models [19]. Third, PAC-Bayesian bounds have complicated forms that are difficulty to analyze and compute. Furthermore, it is unclear to what extent they reflect the impact of optimization algorithms. In a nutshell, effort to establish data-dependent and algorithm-dependent generalization bounds for transductive learning is limited.

Motivated by this, we establish information-theoretic generalization bounds for transductive learning.

2 Preliminaries

2.1 Notations

We stipulate that random variables and their realization are denoted by uppercase and lowercase letters, respectively. For given random variable X, we denote its distribution measure by P_X. Similarly, the conditional distribution measure of X given Y is given by $P_{X|Y}$. We use $\mathrm{D}_{\mathrm{KL}}(P||Q)$ to denote the KL divergence between two probability measure P and Q, where we have assumed that they are on the same measure space and the Radon-Nikodym derivative of P with respect to Q is well defined. With this notation, the mutual information between X and Y is defined as $I(X;Y) \triangleq \mathrm{D}_{\mathrm{KL}}(P_{X,Y}||P_X P_Y)$. Furthermore, the disintegrated mutual information is denoted by $I^z(X;Y) \triangleq \mathrm{D}_{\mathrm{KL}}(P_{X,Y|Z=z}||P_{X|Z=z}P_{Y|Z=z})$, whose expectation taking over $Z \sim P_Z$ is the conditional mutual information $I(X;Y|Z) = \mathbb{E}_Z[I^Z(X;Y)]$. Detail illustration of all notations are provided in the Appendix.

2.2 Transductive Learning

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

3 Transductive Generalization Bounds with Information Measure

3.1 Upper Bounds by Mutual Information

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

Theorem 1. *Suppose $\ell(w, \mathbf{z}) \leq B$ holds for any $w \in \mathcal{W}$ and $\mathbf{z} \in D$, then*

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

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

where $C_{m,u} \triangleq \frac{2B^2(m+u)\max(m,u)}{(m+u-1/2)(2\max(m,u)-1)}$.

Theorem 1 shows that the expectation of transductive generalization error is upper bounded by the mutual information between permutation Z and hypothesis W. This result implies that the less dependence the output hypothesis has on the selection of training data, the better generalization a transductive learning algorithms will have. One can image that if the algorithm only "memorize" the partition of training and testing examples (or heavily depends on the training samples it sees), we could not expect that it will have strong generalization ability. As a comparison, similar results [29] under supervised learning setting say that the generalization error is upper bounded by the mutual information between training set S and hypothesis W, which could be regarded as a special case of our result where u is infinite. In this event, the transductive training and test error corresponding to the supervised training and test error, respectively. Also, m is the number of examples in S, and the term $\frac{1}{u}$ vanish. Note that since the test error is computed over infinite examples, there is no dependence between it and the training error. Furthermore, the assumption of Thm. 1 is stronger than that in supervised learning setting, where they only requires the loss to be sub-Gaussian while we require the loss to be bounded. However, we believe that our result could be extend to the unbounded loss setting under proper assumptions.

The result presented in Theorem 1 is a expectation bound over all possible selection of training data. In real application particularly deep learning scenario, only a few partitions (determined by random seed) are adopted to verify the quality of a transductive learning algorithm, and the empirical results show that whose performance could generally be guaranteed. This urges us to establish the high probability bound in order to better describe the generalization behavior of deep transductive learners. Achieve this relies on the monitor technique proposed in [3]. Another derivant is the expectation bound on the absolute value of transductive generalization error, which serves as a supplement of Theorem 1. The aforementioned results are summarized in Theorem 2.

Theorem 2. *Suppose* $\ell(w, \mathbf{z}) \in [0, B]$ *holds for any* $w \in \mathcal{W}$ *and* $\mathbf{z} \in D$*, with probability at least* $1 - \delta$ *over all* Z*:*

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

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

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

Since $C_{m,u} \approx B^2$ when $m + u$ is large, the high probability bound presented in Eq. 3 is of order $(1/m + 1/u)^{\frac{1}{2}}$. Despite a degenerated constant factor from $\log(1/\delta)$ to $1/\delta$, our bound is sharper than that in existing studies [20, 12]. Although the mutual information term $I(Z; W)$ could not be easily computed, we will show in Sec. that it has a unique advantage when the learner is optimized by stochastic algorithms such as stochastic gradient descent and its variants. Besides, result similar to Eq. (4) can be derived from Eq. (2), despite the constant factor is slightly larger.

3.2 Upper Bounds by General Information Measures

So far, all the bounds we have established contains the mutual information term $I(Z; W)$, either in expectation or high probability. One unsatisfied property of mutual information is that it does not have a finite upper bound, which may lead to vacuous bounds under some circumstances. Fortunately, this issue could be addressed by adopting "supersamples" setting proposed in [25] under supervised learning setting. The key insight is that introducing another random variable to control the randomness of sampling training and test examples, which is independent of the instances. However, directly applying this technique on transductive learning setting is not feasible. The reason is that the training and test examples given returned by this setting is independent, which is yet dependent in transductive learning setting. To bridge this gap, we propose the following transductive supersamples under a specific condition that the number of training examples is equal to that of test examples.

Definition 1 (Transductive Supersamples). *Let* $D = \{\mathbf{z}_i\}_{i=1}^n$ *be a given set where* n *is a finite even number. Denote by* $m = \frac{1}{2}n$*, the transductive supersamples is a sequence* $\widetilde{Z}^m \triangleq (\widetilde{Z}_1, \ldots, \widetilde{Z}_m)$

generated by sampling without replacement from D, where $\widetilde{Z}_i \triangleq \left\{\widetilde{Z}_{i,0}, \widetilde{Z}_{i,1}\right\}$ for $i \in [m]$ is an unordered set with cardinality 2.

Def. 1 shows that the transductive supersamples is obtained by continuously sampling an unorderd instance pairs from a fixed set until there is no remained instances. Please refer to Appendix for a illustrated example of this definition. As a comparison, transductive samples Z^n is obtained by each time sampling an instance from D. The deeper relationship between Z^m and $\widetilde{Z}^m$ is given by the following lemma.

Lemma 1. *Denote by $\widetilde{\mathcal{Z}}$ the set containing all $\widetilde{Z}^m$. Let $S = (S_1, \ldots, S_m) \sim \mathrm{Unif}(\{0,1\})^m$ be the sequence of random variables that is independent of $\widetilde{Z}^m$. Sampling without replacement from D is equivalent to firstly sampling $\widetilde{Z}$ from $\widetilde{\mathcal{Z}}$ and applying S to permute $\widetilde{Z}$.*

Lemma 1 implies that there is another way to obtain the random permutation vector Z^n with the help of transductive supersamples. Let $\widetilde{Z}^m$ and S be the random variables described in Lemma 1, Z can be expressed by $Z^n = (\widetilde{Z}_{1,S_1}, \ldots, \widetilde{Z}_{m,S_m}, \widetilde{Z}_{1,1-S_1}, \ldots, \widetilde{Z}_{m,1-S_m})$. Let $\mathcal{E}(W, \widetilde{Z}, S)$ be the transductive generalization risk under supersampling setting defined by

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

we have $\mathbb{E}[\mathcal{E}(W, \widetilde{Z}, S)] = \mathbb{E}[\mathcal{E}(W, Z)]$. This enables us to characterize the generalization bounds using conditional mutual information, as presented in Theorem 3.

Theorem 3. *Suppose $\ell(w, \mathbf{z}) \in [0, B]$ holds for any $w \in \mathcal{W}$ and $\mathbf{z} \in D$, then*

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

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

By [25], we have $I(S; W|\widetilde{Z}) \leq H(S; Z) \leq m \log 2$ holds, suggesting that the conditional mutual information has a finite upper bound and thus provides a non-vacuous generalization bound. Eq. (6) is consistent with the results in supervised learning setting [25] in form, and the only difference is that $\widetilde{Z}$ should be interpreted as the transductive supersamples. Also, we can recover the result in supervised learning (Theorem 1.2 in [25]) whereas the full sample set D has a infinite cardinality. In this event, D is exactly the space containing all instances, implying that entries in the sequence $\widetilde{Z}^m$ are independent to each other.

Although the mutual information term $I(S; W|\widetilde{Z})$ in Theorem 3 is bounded, computing its numerical value is still difficult, due to W is always a high-dimensional random variable. Thanks to the transductice supersampling setting, various information-theoretical measures adopted in supervised learning setting can be extended to transductive learning setting, as shown in the following Corollary.

Corollary 1. *Suppose the criterion l satisfies $l(\hat{y}, y) \in [0, B]$ holds for any $\hat{y} \in \widehat{\mathcal{Y}}, y \in \mathcal{Y}$. Let $f_w(\mathbf{x}) \in \mathbb{R}^d$ be the prediction given by learner. Denote by $F \in \mathbb{R}^{m \times 2d}$ the prediction matrix where the i-th row is given by $F_{i,:} \triangleq (f_W(\widetilde{\mathcal{X}}_{i,0}), f_W(\widetilde{\mathcal{X}}_{i,1}))$. We have*

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

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

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

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

In applications, the prediction of learner is a low-dimension vector and thus reduce the difficulty of computing the conditional mutual information $I(F;S|\widetilde{Z})$. Note that L in Eq. (9) and δ in Eq. (10) are two-dimensional and one-dimensional random variable, yielding more convenient computation and sharper bounds. Here we point out that each result has its advantage. The vanilla bound in Eq. (6) is more informative to understanding generalization, and the expense is is difficulty of calculating numerical value. In contrast, the other bounds in Corollary 1 has computation convenience, yet they are inferior in reflecting factors that affect generalization. Despite the exist of this trade-off, these results are sufficient for us to understand the generalization behavior of transductive learner or establish non-vacuous bounds for them.

We close this part by briefly discuss how to extend the aforementioned results to more ordinary cases. Due to the definition of transductive supersamples, the results in this part only apply to $m = u$. Here we point out that they can be extended to cases that $m = ku$ or $u = km$ where $k \in \mathbb{N}_+$. We only discuss the case that $u = km$ according to the symmetry. To this end, the transductive supersamples are extended to the following k-transductive supersamples.

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

Note that Definition 1 is a special case of Definition 2 where $k = 1$. Similarly we need to extend the definition of the indicator variable S. Let $S = (S_1, \ldots, S_m) \sim \mathrm{Unif}(\{0, \ldots, k\})^m$ be the sequence of random variables that is independent of $\widetilde{Z}^m$. Suppose $S_i = s_i$, we use $\bar{S}_i \triangleq (\bar{S}_{i,1}, \ldots, \bar{S}_{i,k})$ to represent the sequence that has the same order of $(0, \ldots, k)$ and does not contain s_i. Without loss of generality, assuming that $s_i = 0$, then $\bar{S}_i = (1, \ldots, k)$ and thus $\bar{S}_{i,j} = j$. With this definition, the permutation vector Z can be expressed by

$$Z^n = (\widetilde{Z}_{1,S_1}, \ldots, \widetilde{Z}_{m,S_m}, \widetilde{Z}_{1,\bar{S}_{1,1}}, \ldots, \widetilde{Z}_{1,\bar{S}_{1,k}}, \ldots, \widetilde{Z}_{1,\bar{S}_{m,1}}, \ldots, \widetilde{Z}_{1,\bar{S}_{m,k}}).$$

By this way, all results in Theorem 3 and Corollary 1 can be extended to the case that $u = km$. Please refer to Appendix for concrete formulations.

3.3 Connection with Transductive PAC-Bayesian Bounds

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

Theorem 4. *Suppose $\ell(w, \mathbf{z}) \le B$ holds for any $w \in \mathcal{W}$ and $\mathbf{z} \in D$. Let P be a prior distribution P on $\mathcal{W}$. With probability at least $1 - \delta$ over the randomness of Z, for any distribution Q on $\mathcal{W}$,*

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

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

Compared with previous transductive PAC-Bayesian bound (Corollary 7(b) in in [4]), the result in Theorem 4 has the following advantages: (1)

References

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

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

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

[4] Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-bayesian theory for transductive learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33, pages 105–113, 2014.

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

[6] 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.

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

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

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

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

[20] 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.

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

[22] 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.

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

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

[25] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, volume 125, pages 3437–3452, 2020.

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

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

[28] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[29] 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.

Appendix

Theorem 5. *Suppose $\ell(w,\mathbf{z}) \leq B$ holds for any $w \in \mathcal{W}$ and $\mathbf{z} \in D$, then*

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

where $C_{m,u} \overset{def}{=} \frac{2B^2(m+u)\max(m,u)}{(m+u-1/2)(2\max(m,u)-1)}$.

4 Proof of Theorem 1

We denote by $Z=(Z_1,\dots,Z_n)$ the sequence of random variables that indicate the instances draw without replacement from $D=\{\mathbf{z}_1,\dots,\mathbf{z}_n\}$. Taking $n=3$ for example, a possible Z could be $Z=(\mathbf{z}_2,\mathbf{z}_3,\mathbf{z}_1)$, *i.e.*, the first, second and third sampled instances are $\mathbf{z}_2$, $\mathbf{z}_3$, and $\mathbf{z}_1$ respectively. Then the transductive train error and test error are defined as

$$R_m(W,Z) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m}\ell(W,Z_i),$$

and

$$R_u(W,Z) \overset{\text{def}}{=} \frac{1}{u}\sum_{i=m+1}^{m+u}\ell(W,Z_i),$$

respectively. Then the transductive generalization gap $\mathcal{E}(W,Z)$ is defined as $\mathcal{E}(W,Z) \overset{\text{def}}{=} R_u(W,Z) - R_m(W,Z)$. We firstly show that $\mathcal{E}(w,Z)$ satisfies sub-gaussian property for fixed $w \in \mathcal{W}$. Inspired by [8, 11], we construct the following martingale difference sequences:

$$V_i \overset{\text{def}}{=} \mathbb{E}[\mathcal{E}(w,Z)|Z_1,\dots,Z_i] - \mathbb{E}[\mathcal{E}(w,Z)|Z_1,\dots,Z_{i-1}], i \in [n]. \tag{12}$$

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

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

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

$$\begin{aligned}
&U_i - L_i\\
=&\sup_z \mathbb{E}[\mathcal{E}(w,Z)|Z_1,\dots,Z_{i-1},Z_i=z] - \inf_z \mathbb{E}[\mathcal{E}(w,Z)|Z_1,\dots,Z_{i-1},Z_i=z]\\
=&\sup_{z,\tilde{z}}\Big\{\mathbb{E}[\mathcal{E}(w,Z)|Z_1,\dots,Z_{i-1},Z_i=z] - \mathbb{E}[\mathcal{E}(w,Z)|Z_1,\dots,Z_{i-1},Z_i=\tilde{z}]\Big\}\\
=&\frac{u!(m-i)!C_{n-i-1}^{m-i}}{(n-i)!}\cdot\frac{(m+u)B}{mu}\\
=&\frac{(m+u)B}{m(m+u-i)} \overset{\text{def}}{=} c_i.
\end{aligned}\tag{13}$$

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

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}V_i\right\}\right] =&\mathbb{E}\left[\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n-1}V_i\right\}\exp\{V_n\}\right]\middle|Z_1,\dots,Z_{n-1}\right]\\
=&\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n-1}V_i\right\}\mathbb{E}\left[\exp\{V_n\}\right]\middle|Z_1,\dots,Z_{n-1}\right]\\
\leq&\exp\left\{\frac{\lambda^2 c_n^2}{8}\right\}\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n-1}V_i\right\}\right].
\end{aligned}\tag{14}$$

By recursively repeating the above process, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}V_i\right\}\right]\\
\leq&\exp\left\{\frac{\lambda^2}{8}\sum_{i=1}^{n}c_i^2\right\}=\exp\left\{\frac{\lambda^2(m+u)^2B^2}{8m^2}\sum_{i=1}^{n}\frac{1}{(m+u-i)^2}\right\}\\
\leq&\exp\left\{\frac{\lambda^2B^2(m+u)^2}{8m(u-1/2)(m+u-1/2)}\right\}.
\end{aligned}\tag{15}$$

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

$$\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{n}V_i\right\}\right]\leq\exp\left\{\frac{\lambda^2B^2(m+u)^2}{8mu(m+u-1/2)}\cdot\frac{2\max(m,u)}{2\max(m,u)-1}\right\}\tag{16}$$

Combining Eq. (16) and the facts that $\mathcal{E}(w,Z)-\mathbb{E}[\mathcal{E}(w,Z)]=\sum_{i=1}^{n}V_i$ and $\mathbb{E}[\mathcal{E}(w,Z)]=0$, we obtain

$$\begin{aligned}
&\mathbb{E}_Z\left[\exp\left\{\lambda(R_u(w,Z)-R_m(w,Z))\right\}\right]\\
\leq&\exp\left\{\frac{\lambda^2B^2(m+u)^2}{8mu(m+u-1/2)}\cdot\frac{2\max(m,u)}{2\max(m,u)-1}\right\}\\
=&\exp\left\{\frac{\lambda^2(m+u)C_{m,u}}{8mu}\right\}
\end{aligned}\tag{17}$$

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

$$\begin{aligned}
&\log\mathbb{E}_{W,Z'}\left[\exp\left\{\lambda(R_u(W,Z')-R_m(W,Z'))\right\}\right]\\
=&\log\left(\int_w\mathbb{E}_{Z'}\left[\exp\left\{\lambda(R_u(w,Z')-R_m(w,Z'))\right\}\right]\,\mathrm{d}P_W(w)\right)\\
\leq&\log\left(\int_w\exp\left\{\frac{\lambda^2(m+u)C_{m,u}}{8mu}\right\}\,\mathrm{d}P_W(w)\right)\\
=&\frac{\lambda^2C_{m,u}}{8}\left(\frac{1}{m}+\frac{1}{u}\right),
\end{aligned}\tag{18}$$

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

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(P_{Z,W}||P_{Z,W'})\\
\geq&\mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z)-R_m(W,Z))\right]-\log\mathbb{E}_{Z,W'}\left[\exp\left\{\lambda(R_u(W',Z)-R_m(W',Z))\right\}\right]\\
\geq&\mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z)-R_m(W,Z))\right]-\frac{\lambda^2C_{m,u}}{8}\left(\frac{1}{m}+\frac{1}{u}\right),
\end{aligned}\tag{19}$$

which implies that

$$\mathbb{E}_{Z,W}\left[R_u(W,Z)-R_m(W,Z)\right]\leq\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)I(Z;W)}.\tag{20}$$

This finishes the proof for the first part.

Now we turn to the second part. Following the same process, we have $\mathbb{E}_Z[\exp\{-\lambda\mathcal{E}(w,Z)\}]\leq\exp\{\lambda^2\sigma_{m,u}\}$, where $\sigma_{m,u}\triangleq C_{m,u}(1/m+1/u)/8$. Therefore,

$$\mathbb{P}\left\{|\mathcal{E}(w,Z)|\geq t\right\}\leq\mathbb{P}\left\{\mathcal{E}(w,Z)\geq t\right\}+\mathbb{P}\left\{\mathcal{E}(w,Z)\leq -t\right\}\leq 2\exp\left\{-\frac{t^2}{4\sigma_{m,u}}\right\}.\tag{21}$$

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

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{E}(w,Z)|^k\right] =& \int_0^\infty \mathbb{P}\{|\mathcal{E}(w,Z)|^k \geq u\}\,\mathrm{d}u \\
=& k\int_0^\infty \mathbb{P}\{|\mathcal{E}(w,Z)| \geq t\}\, t^{k-1}\,\mathrm{d}t \\
\leq& 2k\int_0^\infty \exp\left\{-\frac{t^2}{4\sigma_{m,u}}\right\} t^{k-1}\,\mathrm{d}t = (4\sigma_{m,u})^{\frac{k}{2}} k\Gamma(k/2),
\end{aligned} \tag{22}$$

which implies that

$$\mathbb{E}\left[\exp\{\lambda\mathcal{E}^2(w,Z)\}\right] = 1 + \sum_{k=1}^\infty \frac{\lambda^k}{k!}\mathbb{E}\left[|\mathcal{E}(w,Z)|^{2k}\right] \leq 1 + 2\sum_{k=1}^\infty (4\lambda\sigma_{m,u})^k. \tag{23}$$

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

$$\begin{aligned}
&\mathrm{D_{KL}}(P_{Z,W}||P_{Z,W'}) \\
\geq& \mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z) - R_m(W,Z))^2\right] - \log\mathbb{E}_{Z,W'}\left[\exp\left\{\lambda(R_u(W',Z) - R_m(W',Z))^2\right\}\right] \\
\geq& \mathbb{E}_{Z,W}\left[\lambda(R_u(W,Z) - R_m(W,Z))^2\right] - \log\left(1 + 2\sum_{k=1}^\infty (4\lambda\sigma_{m,u})^k\right),
\end{aligned} \tag{24}$$

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

$$\mathbb{E}_{Z,W}\left[(R_u(W,Z) - R_m(W,Z))^2\right] \leq C_{m,u}\left(\frac{1}{m} + \frac{1}{u}\right)(I(Z;W) + \log 3). \tag{25}$$

Our last step is to show that we could obtain sharper PAC-Bayesian bounds. By Markov inequality, for any distribution P that independent to Z:

$$\mathbb{P}\left\{\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z)-R_m(W,Z))}\right] \geq \frac{1}{\delta}\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z)-R_m(W,Z))}\right]\right\} \leq \delta. \tag{26}$$

By Donsker-Varadhan's variational formula, for any distribution Q, with probability at least $1-\delta$ over the randomness of Z:

$$\begin{aligned}
&\lambda\mathbb{E}_{W\sim Q}\left[R_u(W,Z) - R_m(W,Z)\right] \\
\leq& \mathrm{D_{KL}}(Q||P) + \log\left(\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z)-R_m(W,Z))}\right]\right) \\
\leq& \mathrm{D_{KL}}(Q||P) + \log\left(\frac{1}{\delta}\right) + \log\mathbb{E}_Z\mathbb{E}_{W\sim P}\left[e^{\lambda(R_u(W,Z)-R_m(W,Z))}\right] \\
=& \mathrm{D_{KL}}(Q||P) + \log\left(\frac{1}{\delta}\right) + \log\left(\int_w \mathbb{E}_{Z'}\left[\exp\left\{\lambda(R_u(w,Z') - R_m(w,Z'))\right\}\right]\mathrm{d}P(w)\right) \\
\leq& \mathrm{D_{KL}}(Q||P) + \log\left(\frac{1}{\delta}\right) + \frac{\lambda^2 C_{m,u}}{8}\left(\frac{1}{m} + \frac{1}{u}\right),
\end{aligned} \tag{27}$$

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

$$|\mathbb{E}_{W\sim Q}\left[R_u(W,Z) - R_m(W,Z)\right]| \leq \sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m} + \frac{1}{u}\right)\left(\mathrm{D_{KL}}(Q||P) + \log\left(\frac{1}{\delta}\right)\right)}. \tag{28}$$

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

$$\begin{aligned}
&|\mathbb{E}_{W,Z}\left[R_u(W,Z) - R_m(W,Z)\right]| \leq \mathbb{E}_Z\left[|\mathbb{E}_{W\sim Q}\left[R_u(W,Z) - R_m(W,Z)\right]|\right] \\
\leq&\mathbb{E}_Z\left[\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\mathrm{D}_{\mathrm{KL}}(P_{W|S}||P_W) + \log\left(\frac{1}{\delta}\right)\right)}\right] \\
\leq&\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(\mathbb{E}_Z\left[\mathrm{D}_{\mathrm{KL}}(P_{W|S}||P_W)\right] + \log\left(\frac{1}{\delta}\right)\right)} \\
=&\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left(I(S;W) + \log\left(\frac{1}{\delta}\right)\right)}.
\end{aligned} \tag{29}$$

Also, we can derive a expectation bound for PAC-Bayesian:

Remark 1. *Another proof idea is from the concentration inequality provided in , and following the ideas in that this implies the sub-gaussian property. This idea has also been used in previous work . However, the constant factors in the bound obtained by this technique is larger than that in Theorem 1. We provide the detail proof in Section .*

Theorem 6. *Suppose $\ell(w, \mathbf{z}) \leq B$ holds for any $w \in \mathcal{W}$ and $\mathbf{z} \in D$, then with probability at least $1-\delta$, we have*

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

where $C_{m,u} \overset{def}{=} \frac{2B^2(m+u)\max(m,u)}{(m+u-1/2)(2\max(m,u)-1)}$.

5 Proof of Theorem 2

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

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

One can verify that

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

Now taking expectation on both side, we have

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

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

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

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

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

which implies

$$\begin{aligned}&\mathbb{E}_{Z^{(1)},\dots,Z^{(k)},J^*,R^*,W^*}\left[R^*\mathcal{E}(W^{(J^*)},Z^{(J^*)})\right]\\ \leq&\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)I(Z^{(1)},\dots,Z^{(k)};J^*,R^*,W^*)}.\end{aligned} \tag{30}$$

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

$$\begin{aligned}&I(Z^{(1)},\dots,Z^{(k)};J^*,R^*,W^*)\\ \leq&I(S(\pi_1),\dots,S(\pi_k);J^*,R^*,W^*,W_1,\dots,W_k)\\ =&I(Z^{(1)},\dots,Z^{(k)};W_1,\dots,W_k)+I(S(\pi_1),\dots,S(\pi_k);J^*,R^*,W^*|W_1,\dots,W_k)\\ =&\sum_{j=1}^{k}I(S(\pi_j);W_j)+I(Z^{(1)},\dots,Z^{(k)};J^*,R^*,W^*|W_1,\dots,W_k)\\ \leq&kI(Z;W)+\log(2k).\end{aligned} \tag{31}$$

where we have use that $(S(\pi_j),W_j),j\in[k]$ are independent. Plugging Eq. (31) into Eq. (30) yields

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

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

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

By Markov's inequality:

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

Therefore,

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

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

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

6 Proof of Theorem 3

We show that it is possible to derive Theorem 1 from PAC-Bayesian perspective. Denote by $R_{m+u}(W,Z) \triangleq \frac{1}{m+u}\sum_{i=1}^{m+u}\ell(W,Z_i) = \frac{m}{m+u}R_m(W,Z) + \frac{u}{m+u}R_u(W,Z)$ the error on Z. Denote by $\mathcal{D}(p,q) \triangleq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ the KL divergence between two Bernoulli distributions with success probability p and q. Then the $\mathcal{D}$-function introduced in [4] is expressed by $\mathcal{D}^*_\beta(p,q) \triangleq \mathcal{D}(p,q) + \frac{1-\beta}{\beta}\mathcal{D}(\frac{q-\beta p}{1-\beta}, q)$. By Theorem 5 and Theorem 6 in [4], for any $w \in \mathcal{W}$ we have

$$\mathbb{E}_Z\left[\exp\left\{m\mathcal{D}^*_\beta(R_m(w,Z), R_{m+u}(w,Z))\right\}\right] \le 3\log(m)\sqrt{\frac{mu}{m+u}}, \tag{32}$$

which implies that

$$\begin{aligned}
&\log\mathbb{E}_{W\otimes Z}\left[\exp\left\{m\mathcal{D}^*_\beta(R_m(W,Z), R_{m+u}(W,Z))\right\}\right]\\
=&\log\left(\int_w \mathbb{E}_Z\left[\exp\left\{m\mathcal{D}^*_\beta(R_m(w,Z), R_{m+u}(w,Z))\right\}\right]\,\mathrm{d}P_W(w)\right)\\
\le&\log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right).
\end{aligned} \tag{33}$$

By Donsker-Varadhan's variational formula we have

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(P_{Z,W}||P_{Z,W'})\\
\ge&\mathbb{E}_{Z,W}\left[m\mathcal{D}^*_\beta(R_m(W,Z), R_{m+u}(W,Z))\right] - \log\mathbb{E}_{W\otimes Z}\left[e^{m\mathcal{D}^*_\beta(R_m(W,Z),R_{m+u}(W,Z))}\right]\\
\ge&m\mathbb{E}_{Z,W}\left[\mathcal{D}^*_\beta(R_m(W,Z), R_{m+u}(W,Z))\right] - \log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right).
\end{aligned} \tag{34}$$

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

$$\begin{aligned}
&\mathbb{E}_{Z,W}\left[\mathcal{D}^*_\beta(R_m(W,Z), R_{m+u}(W,Z))\right]\\
=&\mathbb{E}_{Z,W}\left[\mathcal{D}(R_m(W,Z), R_{m+u}(W,Z))\right]\\
&+\frac{u}{m}\mathbb{E}_{Z,W}\left[\mathcal{D}\left(\frac{m+u}{u}R_{m+u}(W,Z) - \frac{m}{u}R_m(W,Z), R_{m+u}(W,Z)\right)\right]\\
\ge&2\mathbb{E}_{Z,W}\left[(R_m(W,Z) - R_{m+u}(W,Z))^2\right] + 2\frac{m}{u}\mathbb{E}_{Z,W}\left[(R_m(W,Z) - R_{m+u}(W,Z))^2\right]\\
=&2\frac{m+u}{u}\mathbb{E}_{Z,W}\left[\left(R_m(W,Z) - \frac{m}{m+u}R_m(W,Z) + \frac{u}{m+u}R_u(W,Z)\right)^2\right]\\
=&\frac{2u}{m+u}\mathbb{E}_{Z,W}\left[(R_m(W,Z) - R_u(W,Z))^2\right]\\
\ge&\frac{2u}{m+u}\left(\mathbb{E}_{Z,W}[R_m(W,Z) - R_u(W,Z)]\right)^2.
\end{aligned} \tag{35}$$

Combining Eq. (34) and Eq. (35) we obtain a degenerated bound compared with that provided in Theorem 1 with extra factors $\log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right)$:

$$|\mathbb{E}_{Z,W}[R_m(W,Z) - R_u(W,Z)]| \le \sqrt{\frac{1}{2}\left(\frac{1}{m}+\frac{1}{u}\right)\left[I(Z;W) + \log\left(3\log(m)\sqrt{\frac{mu}{m+u}}\right)\right]}. \tag{36}$$

7 Lemma

By the inequality $\frac{x}{x+1} \le \log(x+1)$, for any $t \ge 1$:

$$\frac{a_t}{\sum_{k=1}^t a_k} = \frac{a_t}{\sum_{k=1}^{t-1} a_k}\frac{\sum_{k=1}^{t-1} a_k}{\sum_{k=1}^t a_k} = \frac{a_t}{\sum_{k=1}^{t-1} a_k}\frac{1}{1+\frac{a_t}{\sum_{k=1}^{t-1} a_k}} \le \log\left(1 + \frac{a_t}{\sum_{k=1}^{t-1} a_k}\right). \tag{37}$$

Thus, we have

$$\begin{aligned}\sum_{t=1}^{T}\frac{a_t}{\epsilon+\sum_{k=1}^{t}a_k}\leq&\frac{a_1}{\epsilon+a_1}+\sum_{t=2}^{T}\frac{a_t}{\sum_{k=1}^{t}a_k}\\\leq&\frac{a_1}{\epsilon+a_1}+\sum_{t=2}^{T}\log\left(1+\frac{a_t}{\sum_{k=1}^{t-1}a_k}\right)\\=&\frac{a_1}{\epsilon+a_1}+\log\left(1+\frac{\sum_{t=1}^{T}a_t}{a_1}\right).\end{aligned}\tag{38}$$

8 Lemma

By definition, introducing integral variable $z\overset{\text{def}}{=}\phi^{-1}(t)$, we have

$$\begin{aligned}&I(X;Y|g(Z))\\=&h(X|\phi(Z))-h(X|Y,\phi(Z))\\=&-\int_t\int_x \mathrm{d}P_{X|\phi(Z)=t}(x)\,\mathrm{d}P_{\phi(Z)}(t)\log\,\mathrm{d}P_{X|\phi(Z)=t}(x)\\&+\int_t\int_y\int_x \mathrm{d}P_{X|Y=y,\phi(Z)=t}\,\mathrm{d}P_{Y|\phi(Z)=t}(y)\,\mathrm{d}P_{\phi(Z)}(t)\log\,\mathrm{d}P_{X|Y=y,\phi(Z)=t}(x)\\=&-\int_z\int_x \mathrm{d}P_{X|Z=z}(x)\,\mathrm{d}P_{Z}(z)\log\,\mathrm{d}P_{X|Z=z}(x)\\&+\int_z\int_y\int_x \mathrm{d}P_{X|Y=y,Z=z}\,\mathrm{d}P_{Y|Z=z}(y)\,\mathrm{d}P_{Z}(z)\log\,\mathrm{d}P_{X|Y=y,Z=z}(x)\\=&h(X|Z)-h(X|Y,Z)=I(X;Y|Z).\end{aligned}\tag{39}$$

Theorem 7. *Denote by w_T a realization of W_T returned by AdaGrad. Suppose $\ell(w,\mathbf{z})\leq B$ holds for any $w\in\mathcal{W},\mathbf{z}\in D$ and $\mathbb{E}_{Z,U_T}\left[R_u(w_T+U_T,Z)-R_u(w_T,Z)\right]\geq 0$ holds, then*

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

where $C_{m,u}\overset{def}{=}\frac{2B^2(m+u)\max(m,u)}{(m+u-1/2)(2\max(m,u)-1)}$.

9 Proof of Theorem 4

As a warm-up example, we will first prove this theorem for AdaNorm algorithm, a popular variant of AdaGrad used for analysis in previous studies. Denote by $g(W_t,Z)\overset{\text{def}}{=}\frac{1}{m}\sum_{i=1}^{m}\nabla\ell(W_t,Z_i)$ the gradient calculated on full training samples. We use $W^{[k]}\overset{\text{def}}{=}(W_0,\ldots,W_k)$ to denote the sequence of random variables. The update rule of AdaNorm is

$$\begin{aligned}\eta_{t-1}(W^{[t-1]})&=\frac{\eta}{\epsilon+\sqrt{\sum_{k=0}^{t-1}\|g(W_k,Z)\|^2}},\\W_t&=W_{t-1}-\eta_{t-1}(W^{[t-1]})g(W_{t-1},Z),\end{aligned}\tag{40}$$

where η and ϵ are two hyperparameters. Here η_{t-1} is a function of random variables $W^{[t-1]}$. Inspired by , we construct the following virtual process for $t\in[T]$:

$$\widetilde{W_0}=W_0,\widetilde{W_t}=\widetilde{W}_{t-1}-\eta_{t-1}(W^{[t-1]})g(W_{t-1},Z)+N_t,\tag{41}$$

where $N_t\overset{\text{def}}{=}\sigma_tN$. Here σ_t is a hyperparameter and N is a random variable following normal distribution, which is independent to $W^{[T]}$ and Z. For concise, we define $U_t\overset{\text{def}}{=}\sum_{k=1}^{t}N_k$. The transductive generalization gap is decomposed by

$$\begin{aligned}
&\mathbb{E}_{Z,W_T}\left[R_u(W_T,Z)-R_m(W_T,Z)\right]\\
=&\mathbb{E}_{Z,W_T,U_T}\left[R_m(\widetilde{W}_T,Z)-R_m(W_T,Z)\right]-\mathbb{E}_{Z,W_T,U_T}\left[R_u(\widetilde{W}_T,Z)-R_u(W_T,Z)\right]\\
&+\mathbb{E}_{Z,W_T,U_T}\left[R_u(\widetilde{W}_T,Z)-R_m(\widetilde{W}_T,Z)\right]\\
\leq&\mathbb{E}_{Z,W_T,U_T}\left[R_m(\widetilde{W}_T,Z)-R_m(W_T,Z)\right]+\sqrt{\frac{C_{m,u}}{2}\left(\frac{1}{m}+\frac{1}{u}\right)I(Z;\widetilde{W}_T)}.
\end{aligned}\tag{42}$$

Since

$$\mathbb{E}\left[R_u(W_T+U_T,Z)-R_u(W_T,Z)|W_T\right]\geq 0,\tag{43}$$

we have

$$\mathbb{E}_{Z,W_T,U_T}\left[R_u(\widetilde{W}_T,Z)-R_u(W_T,Z)\right]=\mathbb{E}_{W_T}\left[\mathbb{E}\left[R_u(W_T+U_T)-R_u(W_T)|W_T\right]\right]\geq 0.\tag{44}$$

Again, note that $\mathbb{E}\left[R_u(W_T+U_T,Z)-R_u(W_T,Z)|W_T\right]$ is a function of W_T. If we treat it as $f(W_T)$, we will have $f(W_T=w)=\mathbb{E}_{Z,U_T}\left[R_u(w+U_T,Z)-R_u(w,Z)\right]$. where we have used xx to obtain the last inequality. Now our last step is to provide a upper bound for $I(Z;\widetilde{W}_T)$. Following , the mutual information term is decomposed by

$$\begin{aligned}
&I(Z;\widetilde{W}_T)\\
=&I(Z;\widetilde{W}_{T-1}-\eta_{T-1}(W^{[T-1]})g(W_{T-1},Z)+U_T)\\
\leq&I(Z;\widetilde{W}_{T-1},-\eta_{T-1}(W^{[T-1]})g(W_{T-1},Z)+U_T)\\
=&I(Z;\widetilde{W}_{T-1})+I(Z;-\eta_{T-1}(W^{[T-1]})g(W_{T-1},Z)+U_T|\widetilde{W}_{T-1})
\end{aligned}$$

Recursively repeating the above process, we obtain

$$I(Z;\widetilde{W}_T)\leq\sum_{t=1}^{T}I(-\eta_{t-1}(W^{[t-1]})g(W_{t-1},Z)+N_t;Z|\widetilde{W}_{t-1}).\tag{45}$$

Denote by $h(\cdot)$ the differential entropy, then

$$\begin{aligned}
&I(\Psi(V,X,y-U)+\sigma N;X|Y=y)\\
=&h(\Psi(V,X,y-U)+\sigma N|Y=y)-h(\Psi(V,X,y-U)+\sigma N|X,Y=y)
\end{aligned}\tag{46}$$

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

$$\begin{aligned}
&h(\Psi(V,X,y)+\sigma N|Y=y)\\
\leq&\frac{d}{2}\log\left(2\pi e\frac{\mathbb{E}\left[\|\Psi(V,X,y-U)+\sigma N\|^2|Y=y\right]}{d}\right)\\
=&\frac{d}{2}\log\left(2\pi e\frac{\mathbb{E}\left[\|\Psi(U,V,X,y-U)\|^2|Y=y\right]+\sigma^2\mathbb{E}\left[\|N\|^2\right]}{d}\right)\\
=&\frac{d}{2}\log\left(2\pi e\frac{\mathbb{E}\left[\|\Psi(V,X,y-U)\|^2|Y=y\right]+d\sigma^2}{d}\right).
\end{aligned}\tag{47}$$

For the second term, define $\Psi : \mathbb{R}^d \to \mathbb{R}^d$ as a function of random variables U, V, X, Y, then:

$$\begin{aligned}
&h(\Psi(V, y-U, X) + \sigma N | X, Y = y) \\
=& - \int_x \mathbb{E}_{U,V,N|X=x,Y=y} \left[\log P_{\Psi(V,y-U,X)+\sigma N|X=x,Y=y}\right] \, \mathrm{d}P_{X|Y=y}(x) \\
=& - \int_x \mathbb{E}_{U,V,N|X=x,Y=y} \left[\log P_{\Psi(V,y-U,X)+\sigma N|U,V,X=x,Y=y}\right] \, \mathrm{d}P_{X|Y=y}(x) \\
&- \int_x \mathbb{E}_{U,V,N|X=x,Y=y} \left[\log P_{U,V|X=x,Y=y}\right] \, \mathrm{d}P_{X|Y=y}(x) \\
=& - \mathbb{E}_{U,V,N,X|Y=y} \left[\log P_{\Psi(V,y-U,X)+\sigma N|U,V,X,Y=y}\right] \\
&- \int_x \mathbb{E}_{U,V|X=x,Y=y} \left[\log P_{U,V|X=x,Y=y}\right] \, \mathrm{d}P_{X|Y=y}(x) \\
=& h(\Psi(V, y-U, X) + \sigma N | U, V, X, Y = y) + \int_x h(U, V | X = x, Y = y) \, \mathrm{d}P_{X|Y=y}(x) \\
\geq& h(\Psi(V, y-U, X) + \sigma N | U, V, X, Y = y) \\
=& h(\sigma N) = \frac{d}{2} \log 2\pi e \sigma^2.
\end{aligned} \tag{48}$$

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

$$\Psi(V, y-U, X) = -\eta_{t-1}(W^{[t-2]}) g(\widetilde{W}_{t-1} - U_t, Z),$$

plugging Eqs. (47,48) into Eq. (46), we have

$$\begin{aligned}
&I(-\eta_{t-1}(W^{[t-1]}) g(W_{t-1}, Z) + N_t; Z | \widetilde{W}_{t-1}) \\
\leq& \mathbb{E}_{\widetilde{W}_{t-1}} \left[\frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E}\left[\eta_{t-1}^2 (W^{[t-2]}, \widetilde{W}_{t-1} - U_t, Z) \| g(\widetilde{W}_{t-1} - U_t, Z) \|^2 \right] + 1 \right) \middle| \widetilde{W}_{t-1} \right] \\
\leq& \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E}\left[\eta_{t-1}^2 (W^{[t-2]}, \widetilde{W}_{t-1} - U_t, Z) \| g(\widetilde{W}_{t-1} - U_t, Z) \|^2 \right] + 1 \right).
\end{aligned} \tag{49}$$

Let $w^{[k]} \overset{\text{def}}{=} (w_1, \dots, w_k)$ and $\zeta(W^{[t-2]}, \widetilde{W}_{t-1} - N_t, Z) \overset{\text{def}}{=} \eta_{t-1}^2 (W^{[t-2]}, \widetilde{W}_{t-1} - N_t, Z) \| g(\widetilde{W}_{t-1} - N_t, Z) \|^2$, we have

$$\begin{aligned}
&\mathbb{E}_{W^{[t-2]}, \widetilde{W}_{t-1}, U_t, Z} \left[\zeta(W^{[t-2]}, \widetilde{W}_{t-1} - U_t, Z) \right] \\
=& \int_z \int_{w^{[t-2]}} \int_{\widetilde{w}_{t-1}} \int_u \zeta(w^{[t-2]}, \widetilde{w}_{t-1} - u, z) \, \mathrm{d}P_{W_{t-1}|Z,W_{t-2}}(\widetilde{w}_{t-1} - u) \, \mathrm{d}P_{N_t}(u) \, \mathrm{d}P_{W^{[t-2]}|Z}(w^{[t-2]}) \, \mathrm{d}P_Z(z) \\
=& \int_z \int_{w^{[t-2]}} \int_{\widetilde{w}_{t-1}} \int_{w_{t-1}} \zeta(w^{[t-2]}, w_{t-1}, z) \, \mathrm{d}P_{W_{t-1}|Z,W_{t-2}}(w_{t-1}) \, \mathrm{d}P_{N_t}(\widetilde{w}_{t-1} - w_{t-1}) \, \mathrm{d}P_{W^{[t-2]}|Z}(w^{[t-2]}) \, \mathrm{d}P_Z(z) \\
=& \int_z \int_{w^{[t-2]}} \int_{w_{t-1}} \zeta(w^{[t-2]}, w_{t-1}, z) \, \mathrm{d}P_{W_{t-1}|Z,W_{t-2}}(w_{t-1}) \, \mathrm{d}P_{W^{[t-2]}|Z}(w^{[t-2]}) \, \mathrm{d}P_Z(z) \\
=& \int_z \int_{w^{[t-1]}} \zeta(w^{[t-2]}, w_{t-1}, z) \, \mathrm{d}P_{W^{[t-1]},Z}(w^{[t-1]}, z) \\
=& \mathbb{E}_{W^{[t-1]},Z}[\zeta(W^{[t-1]}, Z)].
\end{aligned} \tag{50}$$

Plugging Eqs. (49, 50) into Eq. (45), we have

$$I(Z; \widetilde{W}_T) \leq \sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E}\left[\eta_{t-1}^2 (W^{[t-1]}, Z) \| g(W_{t-1}, Z) \|^2 \right] + 1 \right). \tag{51}$$

Combining Eq. (51) with Eq. (42), we have

$$\begin{aligned}
&\mathbb{E}_{Z,W_T}\left[R_u(W_T, Z) - R_m(W_T, Z)\right] \\
\leq& \sum_{t=1}^{T} \frac{d}{2} \log \left(\frac{1}{d\sigma_t^2} \mathbb{E}\left[\eta_{t-1}^2 (W^{[t-1]}, , Z) \| g(W_{t-1}, Z) \|^2 \right] + 1 \right) \\
&+ \mathbb{E}_{Z,W_T,U_T}\left[R_m(W_T + U_T, Z) - R_m(W_T, Z)\right]
\end{aligned} \tag{52}$$

Now let $\sigma_t = \sigma$ for $t \in [T]$, the second term in Eq. (52) can be further bounded by

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{d}{2} \log\left(\frac{1}{d\sigma^2}\mathbb{E}_{W^{[t-1]},Z}\left[\eta_{t-1}^2(W^{[t-2]}, W_{t-1}, Z)\|g(W_{t-1}, Z)\|^2\right] + 1\right) \\
\le&\frac{1}{2\sigma^2}\sum_{t=1}^{T}\mathbb{E}_{W^{[t-1]},Z}\left[\eta_{t-1}^2(W^{[t-1]}, Z)\|g(W_{t-1}, Z)\|^2\right] \\
\le&\frac{\eta^2}{2\sigma^2}\mathbb{E}_{W^{[T-1]},Z}\left[\sum_{t=1}^{T}\frac{\|g(W_{t-1}, Z)\|^2}{\sum_{k=0}^{t-1}\|g(W_k, Z)\|^2}\right] \\
\le&\frac{\eta^2}{2\sigma^2} + \frac{\eta^2}{2\sigma^2}\mathbb{E}_{W^{[T-1]},Z}\left[\log\left(1 + \sum_{t=1}^{T}\|g(W_{t-1}, Z)\|^2\right)\right],
\end{aligned}$$

where the second inequality is due to

$$\eta_{t-1}^2(W^{[t-1]}, Z) = \left(\frac{\eta}{\epsilon + \sqrt{\sum_{k=0}^{t-1}\|g(W_k, Z)\|^2}}\right)^2 \le \frac{\eta^2}{\sum_{k=0}^{t-1}\|g(W_k, Z)\|^2}.$$

The last inequality is obtained by Lemma 4.2 from ali kavis. Now we show how to modify the above process to match the case for AdaGrad algorithm. Recall that the update rule of Adagrad is

$$\begin{aligned}
\eta_{t-1}(W^{[t-1]}) &= \frac{\eta}{\epsilon + \sqrt{\sum_{k=1}^{t-1} g(W_k, Z) \odot g(W_k, Z)}}, \\
W_t &= W_{t-1} - \eta_{t-1}(W^{[t-1]}) \odot g(W_{t-1}, Z),
\end{aligned} \tag{53}$$

Note that here $\eta_{t-1}(W^{[t-1]})$ is a vector that has the same dimension as $g(W_t, Z)$, and $\odot$ denotes Hadamard dot. To keep the consistent notation, we do not use bold. Let $g^k(W_t, Z)$ be the k-th element of $g^k(W_t, Z)$ (so as $\eta_{t-1}^k(W^{[t-1]})$), by the same technique, we have

$$\begin{aligned}
&I(-\eta_{t-1}(W^{[t-1]}) \odot g(W_{t-1}, Z) + N_t; Z|\widetilde{W}_{t-1}) \\
\le&\mathbb{E}_{\widetilde{W}_{t-1}}\left[\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\|\eta_{t-1}(W^{[t-2]}, \widetilde{W}_{t-1} - N_t, Z) \odot g(\widetilde{W}_{t-1} - N_t, Z)\|^2\right] + 1\right)\middle|\widetilde{W}_{t-1}\right] \\
\le&\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\sum_{j=1}^{d}\left(\eta_{t-1}^j(W^{[t-2]}, \widetilde{W}_{t-1} - N_t, Z)g^j(\widetilde{W}_{t-1} - N_t, Z)\right)^2\right] + 1\right) \\
=&\frac{d}{2}\log\left(\frac{1}{d\sigma_t^2}\mathbb{E}\left[\sum_{j=1}^{d}\left(\eta_{t-1}^j(W^{[t-1]}, Z)g^j(W_{t-1}, Z)\right)^2\right] + 1\right).
\end{aligned} \tag{54}$$

The rest steps are similarly:

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{d}{2}\log\left(\frac{1}{d\sigma^2}\mathbb{E}\left[\sum_{j=1}^{d}\left(\eta_{t-1}^j(W^{[t-2]}, W_{t-1}, Z)g^j(W_{t-1}, Z)\right)^2\right] + 1\right) \\
\le&\frac{1}{2\sigma^2}\sum_{t=1}^{T}\mathbb{E}_{W^{[t-1]},Z}\left[\sum_{j=1}^{d}\left(\eta_{t-1}^j(W^{[t-1]}, Z)g^j(W_{t-1}, Z)\right)^2\right] \\
\le&\frac{\eta^2}{2\sigma^2}\mathbb{E}_{W^{[T-1]},Z}\left[\sum_{t=1}^{T}\sum_{j=1}^{d}\frac{\left(g^j(W_{t-1}, Z)\right)^2}{\epsilon^2 + \sum_{k=0}^{t-1}\left(g^j(W_k, Z)\right)^2}\right].
\end{aligned}$$

Lastly, we discuss the standard RMSprop algorithm. RMSprop-Grad:

$$\begin{aligned}
\eta_{t-1} &= \frac{\eta}{\epsilon + \sqrt{\alpha(\beta, t)\sum_{k=0}^{t-1}\beta^{t-1-k}\|g(W_k, Z)\|^2}}, \\
W_t &= W_{t-1} - \eta_{t-1}(W^{[t-1]})g(W_{t-1}, Z),
\end{aligned} \tag{55}$$

where $\alpha(\beta, t) \overset{\text{def}}{=} (1-\beta)/(1-\beta^t)$. RMSprop:

$$\begin{aligned}\eta_{t-1} &= \frac{\eta}{\epsilon + \sqrt{\alpha(\beta, t) \sum_{k=0}^{t-1} \beta^{t-1-k} g(W_k, Z) \odot g(W_k, Z)}}, \\ W_t &= W_{t-1} - \eta_{t-1}(W^{[t-1]}) g(W_{t-1}, Z),\end{aligned} \tag{56}$$

By the same technique, we have

$$\begin{aligned}&\sum_{t=1}^{T} \frac{d}{2} \log\left(\frac{1}{d\sigma^2} \mathbb{E}_{W^{[t-1]}, Z}\left[\eta_{t-1}^2(W^{[t-2]}, W_{t-1}, Z)\|g(W_{t-1}, Z)\|^2\right] + 1\right) \\ \leq& \frac{(1-\beta^T)\eta^2}{2(1-\beta)\sigma^2} + \frac{(1-\beta^T)\eta^2}{2(1-\beta)\sigma^2} \mathbb{E}_{W^{[T-1]}, Z}\left[\log\left(1 + \sum_{t=1}^{T} \beta^{T-t} \|g(W_{t-1}, Z)\|^2\right)\right],\end{aligned}$$

Also,

$$\begin{aligned}&\sum_{t=1}^{T} \frac{d}{2} \log\left(\frac{1}{d\sigma^2} \mathbb{E}\left[\sum_{j=1}^{d} \left(\eta_{t-1}^j(W^{[t-2]}, W_{t-1}, Z) g^j(W_{t-1}, Z)\right)^2\right] + 1\right) \\ \leq& \frac{\eta^2}{2(1-\beta)\sigma^2} \mathbb{E}_{W^{[T-1]}, Z}\left[\sum_{t=1}^{T} \sum_{j=1}^{d} \frac{(1-\beta^t)\left(g^j(W_{t-1}, Z)\right)^2}{\epsilon^2 + \sum_{k=0}^{t-1} \left(g^j(W_k, Z)\right)^2}\right].\end{aligned}$$

10 Corollary

Now we discuss how to extend the proof to the case that momentum technique is adopted. Denote by $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ and $\Psi : \mathbb{R}^d \to \mathbb{R}^d$ to function appied on W and (W, Z), respectively, a general form of the update rule is

$$W_t = \Phi(W_{t-1}) - \eta_{t-1}(W^{[t-1]}) \Psi(W_{t-1}, Z). \tag{57}$$

One can recover the update rule of AdaGrad, AdaNorm by letting $\Phi(W_{t-1}) = W_{t-1}$ and $\Psi(W_{t-1}, Z) = g(W_{t-1}, Z)$. Following previous work, we construct the following virtual process for $t \in [T]$:

$$\widetilde{W_0} = W_0, \widetilde{W_t} = \Phi(\widetilde{W_{t-1}}) - \eta_{t-1}(W^{[t-1]}) \Psi(W_{t-1}, Z) + N_t. \tag{58}$$

The key step in the rest proof is to providing a upper bound for $I(Z; \widetilde{W_T})$:

$$\begin{aligned}I(Z; \widetilde{W_T}) =& I(Z; \Phi(\widetilde{W_{T-1}}) - \eta_{T-1}(W^{[T-1]}) \Psi(W_{T-1}, Z) + N_t) \\ \leq& I(Z; \Phi(\widetilde{W_{T-1}}), -\eta_{T-1}(W^{[T-1]}) \Psi(W_{T-1}, Z) + N_t) \\ =& I(Z; \Phi(\widetilde{W_{T-1}})) + I(Z; -\eta_{T-1}(W^{[T-1]}) \Psi(W_{T-1}, Z) + N_t | \Phi(\widetilde{W_{T-1}})) \\ \leq& I(Z; \widetilde{W_{T-1}}) + I(Z; -\eta_{T-1}(W^{[T-1]}) \Psi(W_{T-1}, Z) + N_t | \widetilde{W_{T-1}}).\end{aligned} \tag{59}$$

By recursively applying the above inequality, we have

$$I(Z; \widetilde{W_T}) \leq \sum_{t=1}^{T} I(-\eta_{t-1}(W^{[t-1]}) \Psi(W_{t-1}, Z) + N_t; Z | \widetilde{W}_{t-1}). \tag{60}$$

Repeating the process in Th xx and replacing $g(W, Z)$ with $\Psi(W, Z)$, we obtain

$$\begin{aligned}&I(-\eta_{t-1}(W^{[t-1]}) \Psi(W_{t-1}, Z) + N_t; Z | \widetilde{W}_{t-1}) \\ \leq& \frac{d}{2} \log\left(\frac{1}{d\sigma_t^2} \mathbb{E}\left[\eta_{t-1}^2(W^{[t-1]}, Z)\|\Psi(W_{t-1}, Z)\|^2\right] + 1\right).\end{aligned} \tag{61}$$

Define

$$\begin{aligned}\begin{bmatrix} W_t \\ V_t \end{bmatrix} &= \begin{bmatrix} W_{t-1} - \frac{1}{1-\beta^t} V_t \\ \beta V_{t-1} + (1-\beta)\eta_{t-1}(W^{[t-1]}, Z) g(W_{t-1}, Z) \end{bmatrix} \\ &= \begin{bmatrix} W_{t-1} - \frac{\beta}{1-\beta^t} V_{t-1} - \frac{1-\beta}{1-\beta^t} \eta_{t-1}(W^{[t-1]}, Z) g(W_{t-1}, Z) \\ \beta V_{t-1} + (1-\beta)\eta_{t-1}(W^{[t-1]}, Z) g(W_{t-1}, Z) \end{bmatrix} \\ &= \begin{bmatrix} I & -\frac{\beta}{1-\beta^t} I \\ 0 & \beta I \end{bmatrix} \begin{bmatrix} W_{t-1} \\ V_{t-1} \end{bmatrix} + \eta_{t-1}(W^{[t-1]}, Z) \begin{bmatrix} \frac{\beta-1}{1-\beta^t} g(W_{t-1}, Z) \\ (1-\beta) g(W_{t-1}, Z) \end{bmatrix}.\end{aligned} \tag{62}$$

where $\beta \in (0,1)$ is the hyper-parameter. We can treat $\Theta \overset{\text{def}}{=} (W_t, V_t)$ as a single variable and define

$$\Phi(\Theta) \overset{\text{def}}{=} \begin{bmatrix} I & -\frac{\beta}{1-\beta^t} I \\ 0 & \beta I \end{bmatrix} \Theta, \Psi(\Theta) \overset{\text{def}}{=} \begin{bmatrix} \frac{\beta-1}{1-\beta^t} g(W_{t-1}, Z) \\ (1-\beta) g(W_{t-1}, Z) \end{bmatrix}. \tag{63}$$

Since $\begin{vmatrix} I & -\frac{\beta}{1-\beta^t} I \\ 0 & \beta I \end{vmatrix} = \beta > 0$, we conclude that Φ is a one-to-one mapping.

11 Proof of Theorem 5

In this part, we discuss the super-sample setting. We firstly consider the case $m = u = \frac{n}{2}$. To this end, we have to introduce some extra notations. We introduce m random variables $S_1, \ldots, S_m$, where each S_i takes value in $\{0,1\}$ with equal probability, namely $\mathbb{P}\{S_i = 0\} = \mathbb{P}\{S_i = 1\} = \frac{1}{2}$. For a given instances set $D = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, let $\widetilde{Z} = (\widetilde{z}_1, \ldots, \widetilde{z}_m)$ be the sequence drawn without replacement from D, where each $\widetilde{z}_i$ is an instance pair. That is to say, each time we sample a pair instance from D without replacement. Taking $m = 2$ for example, a possible $\widetilde{Z}$ could be $\widetilde{Z} = ((\mathbf{z}_2, \mathbf{z}_3), (\mathbf{z}_1, \mathbf{z}_4))$. Here we *do not* consider the order within pairs. When $m = 2$, all possible cases are as follows:

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

here we use set and tuple to indicate whether we consider order or not. Define $\widetilde{z}_i \overset{\text{def}}{=} (\widetilde{z}_{i,0}, \widetilde{z}_{i,1})$, then the transductive generalization risk is given by

$$\mathcal{E}(w, \widetilde{z}, S) \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \left[\ell(w, \widetilde{z}_{i,S_i}) - \ell(w, \widetilde{z}_{i,1-S_i}) \right]. \tag{64}$$

One can verify that

$$\mathbb{E}_{\widetilde{Z},S} \left[\mathcal{E}(w, \widetilde{Z}, S) \right] = \mathbb{E}_Z \left[R_u(w, Z) - R_m(w, Z) \right]. \tag{65}$$

This means that, by firstly sample $\widetilde{Z}$ and then sampling S, constructing the training and test set as $D_{\text{train}} = \{\widetilde{z}_{i,S_i}\}_{i=1}^m$ and $D_{\text{test}} = \{\widetilde{z}_{i,1-S_i}\}_{i=1}^m$, which is equivalent to sampling from Z. Denote by S' the independent copy of S. Suppose that $\ell(\cdot, \mathbf{z}_i) \in [0,1]$ for $i \in [n]$, we have that $\ell(w, \widetilde{z}_{i,S_i}) - \ell(w, \widetilde{z}_{i,1-S_i})$ is a variable bounded between $[-1,1]$. By Hoeffding Lemma,

$$\begin{aligned} &\mathbb{E}_{S'} \left[\exp \left\{ \lambda \mathcal{E}(w, \widetilde{z}, S') \right\} \right] \\ =&\mathbb{E}_{S'} \left[\exp \left\{ \frac{\lambda}{m} \sum_{i=1}^{m} \ell(w, \widetilde{z}_{i,S'_i}) - \ell(w, \widetilde{z}_{i,1-S'_i}) \right\} \right] \leq \exp \left\{ \frac{\lambda^2}{2m} \right\}. \end{aligned} \tag{66}$$

Therefore

$$\begin{aligned} &\log \mathbb{E}_{S', W | \widetilde{Z} = \widetilde{z}} \left[\exp \left\{ \lambda \mathcal{E}(W, \widetilde{z}, S') \right\} \right] \\ =&\log \left(\int_w \mathbb{E}_{S'} \left[\exp \left\{ \lambda \mathcal{E}(w, \widetilde{z}, S') \right\} \right] \mathrm{d}P_{W | \widetilde{Z} = \widetilde{z}}(w) \right) \leq \frac{\lambda^2}{2m}. \end{aligned} \tag{67}$$

where we have used the fact that $P_{S', W | \widetilde{Z} = \widetilde{z}} = P_{W | \widetilde{Z} = \widetilde{z}} P_{S'}$, due to S' is independent from both $\widetilde{Z}$ and W. By Donsker-Varadhan's variational formula, for any $\lambda \in \mathbb{R}$:

$$\begin{aligned} &I(S; W | \widetilde{Z} = \widetilde{z}) = \mathrm{D}_{\mathrm{KL}}(P_{S, W | \widetilde{Z} = \widetilde{z}} || P_{S', W | \widetilde{Z} = \widetilde{z}}) \\ \geq& \mathbb{E}_{S, W | \widetilde{Z} = \widetilde{z}} \left[\lambda \mathcal{E}(W, \widetilde{z}, S) \right] - \log \mathbb{E}_{S', W | \widetilde{Z} = \widetilde{z}} \left[\exp \left\{ \lambda \mathcal{E}(W, z, S') \right\} \right] \\ \geq& \lambda \mathbb{E}_{S, W | \widetilde{Z} = \widetilde{z}} \left[\mathcal{E}(W, \widetilde{z}, S) \right] - \frac{\lambda^2}{2m}, \end{aligned} \tag{68}$$

which implies that

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

Taking expectation on both side, we have obtain

$$\mathbb{E}_{S,W}\left[R_u(W,Z)-R_m(W,Z)\right] = \mathbb{E}_{\widetilde{Z},S,W}\left[\mathcal{E}(W,\widetilde{Z},S)\right] \le \mathbb{E}_{\widetilde{Z}}\sqrt{\frac{1}{2m}I(S;W|\widetilde{Z})}. \tag{69}$$

For the second part, note that $\mathbb{E}_{S'}\left[\exp\left\{-\lambda\mathcal{E}(w,\widetilde{z},S')\right\}\right] \le \exp\{\lambda^2B^2/2m\}$. By the same technique in the Proof of Theorem 1, we have

$$\mathbb{E}_{S,W|\widetilde{Z}=\widetilde{z}}\left[\mathcal{E}^2(W,\widetilde{z},S)\right] \le \frac{4B^2}{m}(I(S;W|\widetilde{Z}=\widetilde{z})+\log 3). \tag{70}$$

Taking expectation on both side, we have obtain

$$\mathbb{E}_{S,W}\left[(R_u(W,Z)-R_m(W,Z))^2\right] = \mathbb{E}_{\widetilde{Z},S,W}\left[\mathcal{E}^2(W,\widetilde{Z},S)\right] \le \frac{4B^2}{m}(I(S;W|\widetilde{Z})+\log 3). \tag{71}$$

Now we discuss how to extend this case to more general setting. Since m and u are symmetric, without loss of generality we assume that $m<u$, and assume that there exist $k\in\mathbb{N}_+$ such that $u=km$. Note that the case we have discussed is exactly an instance in this setting where $k=1$. We also introduce super-transductive samples $\widetilde{Z}=(\widetilde{Z}_1,\dots,\widetilde{Z}_m)$ by sampling a k-tuple each without replacement from D, where each $\widetilde{z}_i=(\widetilde{z}_{i,0},\dots,\widetilde{z}_{i,k})$. Similarly, we introduce the random variable S_i takes value in $\{0,\dots,k\}$ with equal probability. The transductive generalization gap is given by

$$\mathcal{E}(w,\widetilde{z},S) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m}\ell(w,\widetilde{z}_{i,S_i}) - \frac{1}{km}\sum_{i=1}^{m}\sum_{j=0}^{k}\mathbb{I}\{j\ne S_i\}\ell(w,\widetilde{z}_{i,j}). \tag{72}$$

By the same technique, one can find that

$$\mathbb{E}_{S,W}\left[R_u(W,Z)-R_m(W,Z)\right] = \mathbb{E}_{\widetilde{Z},S,W}\left[\mathcal{E}(W,\widetilde{Z},S)\right] \le \mathbb{E}_{\widetilde{Z}}\sqrt{\frac{1}{2m}I(S;W|\widetilde{Z})}. \tag{73}$$

It seems that both cases has the same results, which is due to the value of n is different. By replacing m with $\frac{n}{k+1}$ when n is fixed, the final result is directly correlated with k:

$$\mathbb{E}_{S,W}\left[R_u(W,Z)-R_m(W,Z)\right] = \mathbb{E}_{\widetilde{Z},S,W}\left[\mathcal{E}(W,\widetilde{Z},S)\right] \le \mathbb{E}_{\widetilde{Z}}\sqrt{\frac{k+1}{2n}I(S;W|\widetilde{Z})}.$$

One can find that this bound increase with the increase of k. To guarantee a meaningful result, k should be required to satisfied $k=o(n)$.

12 Corollary

We show that combining the proposed transductive supersamples and techniques introduced in previous work, we can obtain more sharper bounds. For concise, we only discuss the cases that $m=u$. Proofs of more general cases that $u=km$ is similar, thus we directly present the results and omit the proof. First, we define the prediction matrix $F\in\mathbb{R}^{m\times 2}$, where the i-th row is given by $F_{i,:}\overset{\text{def}}{=}(f(\widetilde{z}_{i,0};W),f(\widetilde{z}_{i,1};W))$. Denote by $g(F_i,S_i)\overset{\text{def}}{=}\ell(F_{i,S_i},\widetilde{y}_{i,S_i})-\ell(F_{i,1-S_i},\widetilde{y}_{i,1-S_i})$ the function of F_i,S_i and S_i' the independent copy of S_i. Assume that $|f(\mathbf{z};W)|\le B$ for $\mathbf{z}\in D$, then we have

$$\mathbb{E}_{S_i'}\left[\exp\left\{\lambda g(F_i,S_i')\right\}\right] \le \exp\left\{\frac{\lambda^2B^2}{8m}\right\}. \tag{74}$$

By Donsker-Varadhan's variational formula:

$$\begin{aligned} I(F_i;S_i) &\ge \lambda\mathbb{E}_{F_i,S_i|\widetilde{Z}=\widetilde{z}}\left[g(F,S_i)\right] - \log\mathbb{E}_{F_i,S_i'|\widetilde{Z}=\widetilde{z}}\left[\exp\left\{\lambda g(F_i,S_i')\right\}\right] \\ &\ge \lambda\mathbb{E}_{F_i,S_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i,S_i)\right] - \frac{\lambda^2B^2}{8m}. \end{aligned} \tag{75}$$

Then we have

$$\mathbb{E}_{S,W|\widetilde{Z}=\widetilde{z}}\left[\ell(W,\widetilde{z}_{i,S_i}) - \ell(W,\widetilde{z}_{i,1-S_i})\right] = \mathbb{E}_{F_i,S_i|\widetilde{Z}=\widetilde{z}}\left[g(F_i,S_i)\right] \le \sqrt{\frac{2I(F_i;S_i|\widetilde{Z}=\widetilde{z})}{m}}, \tag{76}$$

which implies that

$$\begin{aligned}
&\mathbb{E}_{S,W}\left[R_u(W,Z) - R_m(W,Z)\right] = \mathbb{E}_{\widetilde{Z},S,W}\left[\mathcal{E}(W,\widetilde{Z},S)\right] \\
=&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z},S,W}\left[\ell(W,\widetilde{Z}_{i,S_i}) - \ell(W,\widetilde{Z}_{i,S_i})\right] \le \frac{B}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{\frac{I(\ell(F_i;S_i|\widetilde{Z})}{2m}}.
\end{aligned} \tag{77}$$

Next, we define the loss matrix $L \in \{0,1\}^{m\times 2}$, where the i-th row is $L_{i,:} \overset{\text{def}}{=} (\ell(W,\widetilde{z}_{i,0}),\ell(W,\widetilde{z}_{i,1}))$. Denote by $g(L_i,S_i) = L_{i,S_i} - L_{i,1-S_i}$, by the same technique we have

$$\mathbb{E}_{S,W}\left[R_u(W,Z) - R_m(W,Z)\right] \le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{\frac{I(L_i;S_i|\widetilde{Z})}{2m}}. \tag{78}$$

Finally, we define the loss gap vector $\delta \in \{-1,1,0\}^m$, where the i-th entry is $\delta_i \overset{\text{def}}{=} \ell(W,\widetilde{z}_{i,0}) - \ell(W,\widetilde{z}_{i,1})$. Denote by $g(\delta_i,S_i) \overset{\text{def}}{=} (-1)^{S_i}\delta_i$, by the same technique we have

$$\mathbb{E}_{S,W}\left[R_u(W,Z) - R_m(W,Z)\right] \le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\widetilde{Z}}\sqrt{\frac{I(\delta_i;S_i|\widetilde{Z})}{2m}}. \tag{79}$$