Towards Improved Risk Bounds for Transductive Learning

Bowei Zhu[1,2,3] , **Shaojie Li**[1,2,3] , **Yong Liu**[1,2,3*]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
{bowei.zhu, lishaojie95, liuyonggsai}@ruc.edu.cn

Abstract

Transductive learning is a popular setting in statistic learning theory, reasoning from observed, specific training cases to specific test cases, which has been widely used in many fields such as graph neural networks and semi-supervised learning. Existing results provide fast rates of convergence based on the traditional local techniques, which need the surrogate function that upper bounds the uniform error within a localized region to be "sub-root". We derive new version of concentration inequality for empirical processes in transductive learning and apply generic chaining technique to relax the assumptions and gain tighter results in empirical risk minimization. Furthermore, we concentrate on the generalization of moment penalization algorithm. We design a novel estimator based on the second moment (variance) penalization and derive its learning rates, which is the first theoretical generalization analysis considering variance-based algorithms.

1 Introduction

Transductive learning is an important setting in statistic learning theory that has been widely used in many fields such as graph neural networks (GNNs) [Oono and Suzuki, 2020; Chien *et al.*, 2020; Esser *et al.*, 2021; Cong *et al.*, 2021; Tang and Liu, 2023b] and semi-supervised learning [Li *et al.*, 2021]. Different from sampled independent and identically distributed with replacement (i.i.d.), training examples in transductive learning settings are sampled independent and without replacement from a finite population and our goal is to reason from observed, specific training cases to specific test cases.

Theoretically, current issues regarding transduction learning remain understudied. Some works for transductive error bounds have been presented in transduction learning such as complexity-based bounds [Vapnik, 1982; Vapnik, 1999; El-Yaniv and Pechyony, 2009; Cortes and Mohri, 2006; Tolstikhin *et al.*, 2014], stability-based bounds [El-Yaniv and Pechyony, 2006; Cortes *et al.*, 2009], information theory [Tang and Liu, 2023a] and PAC-Bayesian bounds [Blum and Langford, 2003; Derbeko *et al.*, 2004; Bégin *et al.*, 2014]. There are some works [Blum and Langford, 2003; Cortes and Mohri, 2006] which consider the special case where the Bayes hypothesis has zero and is contained in the hypothesis class. But this assumption is clearly too restrictive in practice, where the Bayes hypothesis usually can not be assumed to be contained in the class. In fact, most results do not provide fast rates of convergence in the general transductive setting.

It is worth mentioning that [Tolstikhin *et al.*, 2014] provided the general fast rates of convergence in transduction learning based on the traditional local technique given by [Bartlett *et al.*, 2005]. However, this local technique requires the surrogate function (see Definition 4) ψ_m to be "sub-root", which might not be necessary. On the other hand, the Bernstein condition is also needed to derive the final results.

In this paper, we use functional technique to peeling the hypothesis space. Our novel peeling method does not rely on the "sub-root" assumption of the surrogate function or the Bernstein assumption of the loss function. We elaborate in detail that the results obtained by our peeling method will not be worse than those obtained by existing methods. Furthermore, under two common assumption hypothesis spaces: classes of polynomial growth and VC classes, we have obtained tighter bounds compared to the best results [Tolstikhin *et al.*, 2014] that have already been obtained. In addition, our results can also be bounded by empirical excess losses, which, to our knowledge, is the first estimating the risk bounds in transductive learning.

Finally, we employ this novel functional based peeling technique to design a moment-penalized based estimator that considering the variance information. To the best of our knowledge, generalization results for algorithms that consider variance information have not been discussed in transduction learning.

Our contribution can be summarized as follow:

1. We use novel functional based peeling technique to derive better uniform localized convergence upper bounds in transductive learning without "sub-root" assumption and Bernstein condition.

2. For non-parametric classes of polynomial growth and VC classes, our results for empirical risk minimization (ERM) exhibit the improvement relative to the previous re-

*Corresponding author.

sults [Tolstikhin *et al.*, 2014] in some cases. We also obtain the risk bounds from empirical data instead of population data.

3. We design the moment-penalized estimator in transductive learning and provide the generalization bounds for this variance-dependent algorithm, which have not bee discussed before.

2 Related Work

The concentration inequalities for the supremum of the standard empirical process for sampling with replacement have been well studied in the literature including Talagrand's inequality [Talagrand, 1996] and its versions due to [Bousquet, 2002a; Bousquet, 2002b] and Section 12 of [Boucheron *et al.*, 2013]. For transductive learning, we need a modified version on concentration inequalities for the supremum of the empirical process for sampling without replacement [Cortes and Vapnik, 1995; Tolstikhin *et al.*, 2014]. Some works for error bounds have been presented based on these inequalities such as [Vapnik, 1982; Vapnik, 1999; Blum and Langford, 2003; Derbeko *et al.*, 2004; El-Yaniv and Pechyony, 2009; Cortes and Mohri, 2006; Tolstikhin *et al.*, 2014]. The first general bound studied the binary loss functions, presented in [Vapnik, 1982], was implicit in the sense that the value of the bound was specified as an outcome of a computational procedure. [Blum and Langford, 2003; Derbeko *et al.*, 2004] developed several fast rates of PAC-Bayesian bounds which critically depends on the prior distribution over the hypothesis class. [Cortes and Mohri, 2006] considered a transductive regression with bounded squared loss and obtain a generalization error bound. [Tolstikhin *et al.*, 2014] provided the first general fast rates of convergence in transduction learning based on the traditional local technique [Bartlett *et al.*, 2005], which required the Bernstein condition and the surrogate function ψ_m to be "sub-root". To the best of our knowledge, there has not been achieved in previous literature for general optimal upper bounds relaxing the Bernstein condition and the "sub-root" surrogate function assumption. Generalization bounds for algorithms that consider variance information also have not been discussed in transductive learning.

Besides we have to mention that transductive bounds based on algorithmic stability have been studied for classification in [El-Yaniv and Pechyony, 2006], and for regression in [Cortes *et al.*, 2009]. However, both of them do not yield fast risk bounds.

3 Preliminaries

3.1 Notations

In standard i.i.d. problem, we assume that a random sample $\mathbf{z}$ follows an unknown distribution $\mathbb{P}$ with the data support $\mathcal{Z}$. For each realization of $\mathbf{z}$, let $\ell(\cdot;\mathbf{z})$ be a real-valued loss function, defined over the hypothesis class $\mathcal{W}$. Given n i.i.d samples $\{\mathbf{z}_i\}_{i=1}^n$ drawn from $\mathbb{P}$ as training set. Then the population risk and the empirical risk are as follows:

$$\mathbb{P}\ell(\mathbf{w};\mathbf{z}) = \mathbb{E}_{\mathbf{z}}[\ell(\mathbf{w};\mathbf{z})], \quad \mathbb{P}_n\ell(\mathbf{w};\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{w};\mathbf{z}_i).$$

However, in transductive learning, the learner is provided with m labeled training points and u unlabeled test points. The objective of the learner is to obtain accurate predictions for the test points. Two different settings of transductive learning were given by [Vapnik, 1998]. One assumes that both the training and test sets are sampled i.i.d. from a same unknown distribution and the learner is provided with the labeled training and unlabeled test sets. Another assumes that the set consisting of N arbitrary input points without any other assumptions regrading its underlying source is given. In this paper, we study the second setting, as pointed out by [Vapnik, 1998], any upper generalization bound in the second setting can easily yield a bound for the first setting by just taking expectation.

Let's consider a finite set $\boldsymbol{Z}_N = (\boldsymbol{X}_N, \boldsymbol{Y}_N)$ containing N arbitrary input points. For each data point $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, we have $\mathbf{x}_i \in \boldsymbol{X}_N$ and its corresponding output $y_i \in \boldsymbol{Y}_N$ serves as the label. From this set, we uniformly sample $m < N$ elements $\boldsymbol{X}_m \subset \boldsymbol{X}_N$ without replacement, creating a dependency among the inputs within $\boldsymbol{X}_m$. Naturally, we also sample the corresponding outputs $\boldsymbol{Y}_m$ for the input examples in $\boldsymbol{X}_m$. The resulting training set is denoted as $\boldsymbol{Z}_m = (\boldsymbol{X}_m, \boldsymbol{Y}_m)$ and the test set is denoted as $\boldsymbol{Z}_u = (\boldsymbol{X}_u, \boldsymbol{Y}_u)$.

For any $\mathbf{w} \in \mathcal{W}$ and the loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ bounded by $[-B, B]$, the training error and the test error can be defined as

$$\hat{R}_m(\mathbf{w}) = \frac{1}{m}\sum_{\mathbf{z}\in\boldsymbol{Z}_m}\ell(\mathbf{w};\mathbf{z}), \quad R_u(\mathbf{w}) = \frac{1}{u}\sum_{\mathbf{z}\in\boldsymbol{Z}_u}\ell(\mathbf{w};\mathbf{z}).$$

For technical reasons that will become clear later, we also define the overall error with regard to both the union of the training and test sets as

$$R_N(\mathbf{w}) = \frac{1}{N}\sum_{\mathbf{z}\in\boldsymbol{Z}_N}\ell(\mathbf{w};\mathbf{z}).$$

Then, the main goal of the learner in transductive setting is to select proper parameters to minimizing the test error $R_u(\mathbf{w})$, which we will denote by $\mathbf{w}_u^*$. Since the labels of the test set examples are unknown, we can't compute $R_u(\mathbf{w})$ and need to estimate it based on the training sample $\boldsymbol{Z}_m$. A common choice is to replace the test error minimization by empirical risk minimization $\hat{\mathbf{w}}_m \in \arg\min_{\mathbf{w}\in\mathcal{W}}\hat{R}_m(\mathbf{w})$ and to use it as an approximation of $\mathbf{w}_u^*$. For $\mathbf{w} \in \mathcal{W}$ we define the excess risk in transductive learning:

$$\mathcal{E}_u(\mathbf{w}) = R_u(\mathbf{w}) - \inf_{\mathbf{w}'\in\mathcal{W}} R_u(\mathbf{w}') = R_u(\mathbf{w}) - R_u(\mathbf{w}_u^*).$$

From now on it will be convenient to introduce the following operators, mapping functions f (for example excess loss in Theorem 1) defined on $\mathbf{z}$ to $\mathbb{R}$:

$$E_N f = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{z}_i), \quad \mathbf{z}_i \in \boldsymbol{Z}_N,$$

$$\hat{E}_m f = \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{z}_j), \quad \mathbf{z}_j \in \boldsymbol{Z}_m.$$

3.2 Empirical Process Theory to Generalization

In this subsection, we introduce classical empirical process theory to construct surrogate function by upper bounding the "local Rademacher complexity". Here we give the definition of Rademacher complexity for completeness.

Definition 1 (Rademacher complexity [Wainwright, 2019])**.** *For a function class $\mathcal{F}$ that consists of mappings from $\mathcal{Z}$ to $\mathbb{R}$ and $f \in \mathcal{F}$, define*

$$\mathfrak{R}\mathcal{F} := \mathbb{E}_{\mathbf{z},v} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} v_i f(\mathbf{z}_i)$$

and

$$\mathfrak{R}_n\mathcal{F} := \mathbb{E}_{v} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} v_i f(\mathbf{z}_i),$$

as the Rademacher complexity and the empirical Rademacher complexity of $\mathcal{F}$, respectively, where $\{v_i\}_{i=1}^n$ are i.i.d. Rademacher variables and $\mathbb{P}(v_i = 1) = \mathbb{P}(v_i = -1) = \frac{1}{2}$.

Furthermore, we can apply Dudley's integral bound [Sridharan, 2010] to gain the upper bound on the local Rademacher complexity using the covering number of hypothsis class $\mathcal{F}$. Here we introduce the definition of covering number and Dudley's integral bound.

Definition 2 (Covering number [Wainwright, 2019])**.** *Assume $(\mathcal{M}, \mathrm{metr}(\cdot,\cdot))$ is a metric space, and $\mathcal{F} \subseteq \mathcal{M}$. The ε-convering number of the set $\mathcal{F}$ with respect to a metric $\mathrm{metr}(\cdot,\cdot)$ is the size of its smallest ε-net cover:*

$$\mathcal{N}(\varepsilon, \mathcal{F}, \mathrm{metr}) = \min\{m : \exists f_1, \ldots, f_m \in \mathcal{F} \text{ such that } \mathcal{F} \subseteq \cup_{j=1}^{m} \mathcal{B}(f_j, \varepsilon)\},$$

where $\mathcal{B}(f,\varepsilon) := \{\tilde{f} : \mathrm{metr}(\tilde{f}, f) \le \varepsilon\}$.

Assume that there is a function $\mathbf{w}_N^* \in \mathcal{W}$ satisfying $R_N(\mathbf{w}_N^*) = \inf_{\mathbf{w}\in\mathcal{W}} R_N(\mathbf{w})$. Define the excess loss class $\mathcal{F}^* = \{f : f(\mathbf{z}) = \ell(\mathbf{w};\mathbf{z}) - \ell(\mathbf{w}_N^*;\mathbf{z}), \mathbf{w} \in \mathcal{W}\}$.

Lemma 1 (Dudley's integral bound [Sridharan, 2010])**.** *Given $r > 0$ and class $\mathcal{F}$ that consists of functions defined on $\mathcal{Z}$,*

$$\mathfrak{R}_n\{f \in \mathcal{F} : \mathbb{P}_n[f^2] \le r\} \le \inf_{\varepsilon_0 > 0} \left\{ 4\varepsilon_0 + 12 \int_{\varepsilon_0}^{\sqrt{r}} \sqrt{\frac{\log \mathcal{N}(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n))}{n}} d\varepsilon \right\}.$$

4 Upper Bounds for Empirical Risk Minimization

In this section, we start with the novel functional based peeling method. We discussed this method and compare with traditional localized technique [Bartlett *et al.*, 2005; Tolstikhin *et al.*, 2014] in detail. Then we derive upper bounds for ERM in transductive learning and illustrate our fast rates in two common assumptions: non-parametric classes of polynomial growth and VC classes.

4.1 Functional based Peeling Method

In uniform localized convergence procedure, we firstly define a real valued function ψ_m that upper bounds the uniform error within a localized region $\{f \in \mathcal{F} : T(f) \le r\}$, $T : \mathcal{F} \to \mathbb{R}_+$ is a measurement functional.

Definition 3. *Let ψ_m be a function that maps $[0,+\infty)\times(0,1)$ to $(0,+\infty)$, which possibly depends on the observed samples $\{\mathbf{z}_i\}_{i=1}^m$. Assume ψ_m satisfies for arbitrary fixed δ, r, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}:T(f)\le r} (E_N - \hat{E}_m) f \le \psi_m(r;\delta).$$

Traditional localized technique provided by Corollary 5.3 of [Bartlett *et al.*, 2005] peels hypothesis space by a non-decreasing and nonnegative "sub-root" function such that $r \to \psi(r;\delta)/\sqrt{r}$ is also nonincreasing for $r > 0$. We relax this assumption by a meaningful surrogate function instead, which removes the "sub-root" condition.

Definition 4 (Meaningful Surrogate Function)**.** *If a function $\psi(r;\delta)$ defined over $[0,+\infty)\times(0,1)$ is non-decreasing, non-negative and bounded with respect to r for every fixed $\delta \in (0,1)$. This function is called a meaningful surrogate function.*

Note that the excess loss class in Theorem 3 is itself non-decreasing and non-negative, and the boundedness requirement can always be met by setting $\psi(r;\delta) = \psi(4B^2;\delta)$ for all $r \ge 4B^2$. Next, we give the main theorem.

Theorem 1. *For the excess loss class $\mathcal{F}^* = \{f : f(\mathbf{z}) = \ell(\mathbf{w};\mathbf{z}) - \ell(\mathbf{w}_N^*;\mathbf{z}), \mathbf{w} \in \mathcal{W}\}$, assume there is a meaningful surrogate function $\psi_m(r;\delta)$ that satisfies for all $\delta \in (0,1)$, $\lambda > 1$ and for all $r > 0$, with probability at least $1-\delta$*

$$\sup_{f\in\mathcal{F}^*:E_N f^2\le r} (E_N - \hat{E}_m) f \le \psi_m(r;\delta).$$

Then for any $\delta \in (0,1)$ and $r_0 \in (0, 4B^2)$, with probability at least $1-\delta$

$$(E_N - \hat{E}_m) f \le \psi_m\left(\max\left\{\lambda E_N f^2, r_0\right\}; \frac{\delta}{2\log_\lambda \frac{4B^2\lambda}{r_0}}\right).$$

Remark 1. The "cost" of this localized uniform convergence mainly from the additional $\log_\lambda \frac{4B^2\lambda}{r_0}$ term, which only appear in the form $\log\left(\frac{\delta}{2\log_\lambda \frac{4B^2\lambda}{r_0}}\right)$ in high-probability bounds, which is of a negiligible $O(\log\log n)$ order in general. The proof technique is motivated by [Xu and Zeevi, 2024], which peels with a variable functional rather than a fix value r^*.

Then, we compare this theorem with existing result in [Tolstikhin *et al.*, 2014] based on traditional local Rademacher complexities [Bartlett *et al.*, 2005].

Theorem 2 ([Bartlett *et al.*, 2005; Tolstikhin *et al.*, 2014])**.** *Assume that the loss function ℓ is bounded in the interval $[0,1]$ and there is a constant $B_e > 0$ such that for every $f \in \mathcal{F}^*$ we have $E_N f^2 \le B_e E_N f$. Assume that there is a **sub-root** function $\psi_m(r)$ such that*

$$\sup_{f\in\mathcal{F}^*;E_N f^2\le r} (E_N - \hat{E}_m) f \le \psi_m(r).$$

Let $\hat{R}_m^*$ be a fixed point of $\psi_m(r)$. Then for any $t > 0$ with probability at least $1 - \delta$, we have

$$(E_N - \hat{E}_m)f \leq \inf_{K>1} \frac{E_N f}{K} + 25\frac{K}{B_e}\lambda \hat{R}_m^*,$$

where $\lambda > 1$ is a constant to peeling the hypothesis space.

Remark 2. There are three issues we need to point out here. Firstly, Theorem 2 is a modified version of Theorem 11 in [Tolstikhin *et al.*, 2014]. which gave a direct result under the empirical risk minimization algorithm. However, their proof satisfies uniformly over the hypothesis space $\mathcal{F}^*$. For a more intuitive comparison between our results and theirs, here we give the modified results using their techniques uniformly over $\mathcal{F}^*$, the differences between [Tolstikhin *et al.*, 2014] and proof of Theorem 2 are given in Appendix.

Secondly, we want to point out that Theorem 2 assumes the bounded loss and the Bernstein condition. However, the bounded loss assumption can derive to the Bernstein condition. (Although it may be possible to assume the existence of smaller constants that satisfy the Bernstein condition on the bounded loss). Thus, in the following comparison, we will give the version of Theorem 1 that satisfies the Bernstein condition. Later in the rest of our paper, in order to highlight our core contribution, we will directly use the loss-bounded derivation of the constant that remove the Bernstein condition directly.

Finally, we notice that Theorem 11 in [Tolstikhin *et al.*, 2014] includes an extra term $O\left(B_e \log\left(\frac{1}{\delta}\right)\left(\frac{N}{m^2}\right)\right)$, which is involved from the concentration inequality for sampling without replacement (See Lemma 3). But Theorem 1 in our paper only focuses on the peeling technique and has not yet brought in the concentration inequality to highlight our main technique. When comparing with Theorem 1, we take out the "peeling technique" in Theorem 2 and focus on the improvement of the peeling technique.

Remark 3. Now we start to compare Theorem 1 with Theorem 2 given in [Tolstikhin *et al.*, 2014]. Overall, our results does not require the surrogate function ψ_m to be "sub-root". Despite weaker assumptions, our results are typically "tighter" than Theorem 2 under the same assumptions. Here we explain in detail.

On one hand, under the "sub-root" assumption, and taking the optimal choice of K, Theorem 2 can be rewrite as

$$(E_N - \hat{E}_m)f \leq 10\sqrt{\frac{\lambda \hat{R}_m^* E_N f}{B_e}}.$$

Under the same assumption $E_N f^2 \leq B_e E_N f$ and use Theorem 1 it is straightforward to have

$$\begin{aligned}&\psi_m\left(\lambda E_N f^2; \delta\right) \leq \psi_m(\lambda B_e E_N f; \delta)\\ \leq &\frac{\sqrt{\lambda B_e E_N f}}{\sqrt{\hat{R}_m^*}}\psi_m(\hat{R}_m^*; \delta) \leq \sqrt{\frac{\lambda \hat{R}_m^* E_N f}{B_e}}, \end{aligned} \tag{1}$$

where the first inequality applies the Bernstein condition. The second inequality holds because the surrogate function ψ_m is sub-root. The last inequality holds because $\hat{R}_m^*$ is the fixed point of $B_e\psi_m(r, \delta)$.

Then we can derive the conclusion that when $E_N f \leq \frac{\hat{R}_m^*}{B_e}$, $(E_N - \hat{E}_m)f$ is of order $\frac{\lambda \hat{R}_m^*}{B_e}$ both in Theorem 1 and Theorem 2. When $E_N f > \frac{\hat{R}_m^*}{B_e}$, $(E_N - \hat{E}_m)f$ is of order $\sqrt{\frac{\lambda \hat{R}_m^* E_N f}{B_e}}$ in Theorem 2. However our results in Theorem 1 is of order $\psi_m\left(\lambda E_N f^2; \delta\right)$, which is strictly improved because $\psi_m\left(\lambda E_N f^2; \delta\right) \leq \sqrt{\frac{\lambda \hat{R}_m^* E_N f}{B_e}}$ according to (1).

On the other hand, the removal of the "sub-root" requirement on ψ_m is also important. The "sub-root" inequality (the first inequality in (1)) becomes an equality when $\psi_m(r; \delta) = O\left(\sqrt{\frac{dr}{m}}\right)$ in the parametric case, where d is the parametric dimension. However, when the hypothesis space $\mathcal{F}$ is "rich", $\frac{\psi_m(r;\delta)}{\sqrt{r}}$ can be strictly decreasing but "sub-root" assumption only requires nonincreasing so that the "sub-root" assumption can become loose. For example, when $\mathcal{F}$ is a nonparametric class, we often have $\psi_m(r; \delta) = O\left(\sqrt{\frac{r^{1-\rho}}{n}}\right)$ for some $\rho \in (0, 1)$. Under this condition, the richer $\mathcal{F}$ is, the more loose the "sub-root" inequality is, which further lead to looser bounds. We will discuss the case in detail in Theorem 4.

4.2 Upper Bounds for Empirical Risk Minimization

In this subsection, we apply our results to the loss-dependent rates of empirical risk minimization (ERM) via a surrogate function ψ_m and its fixed point $\hat{R}_m^*$, and further give new results on two important families of classes: parametric classes of polynomial growth and VC classes. We denote the effective loss $\mathcal{L}^* = E_N\left[\ell(\mathbf{w}_N^*; \mathbf{z}) - \inf_{\mathbf{w}\in\mathcal{W}} \ell(\mathbf{w}; \mathbf{z})\right]$ on full dataset.

Theorem 3. *For the excess loss class $\mathcal{F}^* = \{f : f(\mathbf{z}) = \ell(\mathbf{w}; \mathbf{z}) - \ell(\mathbf{w}_N^*; \mathbf{z}), \mathbf{w} \in \mathcal{W}\}$, assume there is a meaningful surrogate function $\psi_m(r; \delta)$ that satisfies for all $\delta \in (0, 1)$ and for all $r > 0$, with probability at least $1 - \delta$*

$$\sup_{f\in\mathcal{F}: E_N f^2 \leq r} (E_N - \hat{E}_m)f \leq \psi_m(r; \delta).$$

Then $\hat{\mathbf{w}}_m \in \arg\min_{\mathcal{W}}\{\hat{R}_m(\mathbf{w})\}$ satisfies for any $\delta \in (0, 1)$ and $r_0 \in (0, 1)$, with probability at least $1 - \delta$

$$\begin{aligned}&R_N(\hat{\mathbf{w}}_m) - R_N(\mathbf{w}_N^*)\\ \leq &\max\left\{\psi_m\left(24B\mathcal{L}^*; \frac{\delta}{2\log_2 \frac{8B^2}{r_0}}\right), \frac{\hat{R}_m^*}{6B}, \frac{r_0}{48B}\right\},\end{aligned}$$

where $\hat{R}_m^*$ is the fixed point of $6B\psi_m\left(8r; \frac{\delta}{2\log_2 \frac{8B^2}{r_0}}\right)$.

Remark 4. Notice that the term r_0 can be selected very small. For example, we can set $r_0 = \frac{B^2}{m^4}$, which can make it much smaller than $\hat{R}_m^*$. ($\hat{R}_m^*$ can be calculated according to classical empirical process theory and Dudley's integral bound [Sridharan, 2010] and is at least of order

$O(B^2 \log \frac{1}{\delta}/m)$). Under this situation, $\log_\lambda \frac{4B^2\lambda}{r_0}$ is of order $\log m$, which only appear in the form $\log\left(\frac{2\log_\lambda \frac{4B^2\lambda}{r_0}}{\delta}\right)$ in the final result, which is of order $\log\log m$ and can be regarded as an absolute constant. Comparing with traditional localized Rademacher complexities [Bartlett *et al.*, 2005; Tolstikhin *et al.*, 2014], Theorem 3 can apply to broader settings. Since we bypass the "sub-root" assumption on ψ_m and adapt to the better parameter $\mathcal{L}^*$ instead of traditional $E_N f$.

To illustrate the noticeable gaps between our results and previous works, we compare them on two important families of classes: non-parametric classes of polynomial growth and VC classes. To bound meaningful surrogate function in Definition 4, we need to build a connection between the supremum of the empirical process from sampling without replacement and Rademacher complexities using modified concentration inequalities. Here we provide the following lemma for transductive learning.

Lemma 2. *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ into $[-2B, 2B]$. Assume that there is some $r > 0$ such that for every $f \in \mathcal{F}$, $Var[f(\mathbf{z}_i)] \leq r$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\sup_{f\in\mathcal{F}}(E_N - \hat{E}_m)f \leq 2\mathfrak{R}_m\mathcal{F} + 2\sqrt{2\left(\frac{N}{m^2}\right) r \log\left(\frac{1}{\delta}\right)}.$$

Moreover, the same results hold for the quantity $\sup_{f\in\mathcal{F}}(\hat{E}_m - E_N)f$.

Remark 5. Lemma 2 uses classical empirical process theory to construct surrogate function by upper bounding the "local Rademacher complexity". Traditional concentration inequality for empirical process such as [Bartlett *et al.*, 2005] are derived from standard Talagrand's concentration inequality, where we set the random variable to be $\sup_f(\mathbb{P}-\mathbb{P}_n)f$ and use symmetrization inequality for Rademacher complexity to build a connection between $\mathbb{E}\sup_f(\mathbb{P}-\mathbb{P}_n)f$ and $\mathfrak{R}_n\mathcal{F}$. However, we can't directly apply Talagrand's concentration inequality for $\sup_f(E_N - \hat{E}_m)f$ because this samples are independent without replacement. We derive the concentration inequality for empirical process in transductive learning, which is motivated by [Bartlett *et al.*, 2005]. But we use the sub-gaussian type concentration for sampling without replacement in [Tolstikhin *et al.*, 2014] instead of Talagrand's concentration inequality [Bousquet, 2002a; Bousquet, 2002b].

Then, we derive the bounds with two families of classes.

Theorem 4 (Non-parametric classes of polynomial growth)**.** *Consider a loss class $\ell\circ\mathcal{W}$ with the metric entropy condition*

$$\log\mathcal{N}(\varepsilon, \ell\circ\mathcal{W}, d_{\ell\circ\mathcal{W}}) \leq O(\varepsilon^{-2\rho}),$$

under the conditions of Theorem 3, then for any $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, we have

$$R_N(\hat{\mathbf{w}}_m) - R_N(\mathbf{w}_N^*) \leq O\left(\max\left\{\sqrt{\frac{(B\mathcal{L}^*)^{1-\rho}}{m}}, \frac{B^{\frac{1-\rho}{1+\rho}}}{m^{\frac{1}{1+\rho}}}\right\}\right).$$

Furthermore, for any $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, we have

$$\mathcal{E}(\hat{\mathbf{w}}_m) \leq O\left(\frac{N}{u}\left(\max\left\{\sqrt{\frac{(B\mathcal{L}^*)^{1-\rho}}{m}}, \frac{B^{\frac{1-\rho}{1+\rho}}}{m^{\frac{1}{1+\rho}}}\right\}\right) + \frac{N}{m}\left(\max\left\{\sqrt{\frac{(B\mathcal{L}^*)^{1-\rho}}{u}}, \frac{B^{\frac{1-\rho}{1+\rho}}}{u^{\frac{1}{1+\rho}}}\right\}\right)\right).$$

Theorem 5 (VC classes)**.** *Consider a loss class $\ell\circ\mathcal{W}$ with the metric entropy condition*

$$\log\mathcal{N}(\varepsilon, \ell\circ\mathcal{W}, d_{\ell\circ\mathcal{W}}) \leq O\left(d\log\frac{1}{\varepsilon}\right),$$

under the conditions of Theorem 3, then for any $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, we have

$$R_N(\hat{\mathbf{w}}_m) - R_N(\mathbf{w}_N^*) \leq O\left(\max\left\{\sqrt{\frac{dB\mathcal{L}^*}{m}\log\frac{B}{3\mathcal{L}^*}}, \frac{Bd}{m}\log\frac{B}{3\mathcal{L}^*}, \frac{Bd\log m}{m}\right\}\right).$$

Furthermore, for any $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, we have

$$\mathcal{E}(\hat{\mathbf{w}}_m) \leq O\left(\frac{N}{u}\left(\max\left\{\sqrt{\frac{dB\mathcal{L}^*}{m}\log\frac{B}{3\mathcal{L}^*}}, \frac{Bd}{m}\log\frac{B}{3\mathcal{L}^*}, \frac{Bd\log m}{m}\right\}\right) + \frac{N}{m}\left(\max\left\{\sqrt{\frac{dB\mathcal{L}^*}{u}\log\frac{B}{3\mathcal{L}^*}}, \frac{Bd}{u}\log\frac{B}{3\mathcal{L}^*}, \frac{Bd\log u}{u}\right\}\right)\right).$$

Remark 6. Firstly, we compare Theorem 4, Theorem 5 with existing work [Tolstikhin *et al.*, 2014]. It is worth noting that we don't assume the Bernstein condition comparing with [Tolstikhin *et al.*, 2014]. In fact, we have already discussed in Remark 3 that under the assumption of the Bernstein condition, the order of our results is the same as [Tolstikhin *et al.*, 2014].

In fact, the technique of splitting the hypothesis space used in the proof of [Tolstikhin *et al.*, 2014] comes from the classical method [Bartlett *et al.*, 2005]. Using the same technique and removing the Bernstein condition, we can easily obtain the result [1]

$$\mathcal{E}(\hat{\mathbf{w}}_{m,\text{previous}}) \leq O\left(\frac{N}{u}\left(\max\left\{\sqrt{\frac{\mathcal{L}^*\hat{R}_m^*}{B}}, \frac{\hat{R}_m^*}{B}\right\}\right) + \frac{N}{m}\left(\max\left\{\sqrt{\frac{\mathcal{L}^* r_u^*}{B}}, \frac{r_u^*}{B}\right\}\right)\right).$$

Similarly, using Dudley's integral bound [Sridharan, 2010] and Lemma 2 and solving $r \leq O\left(B\psi_m(r;\delta)\right)$, we can derive the following two results.

For non-parametric classes of polynomial growth:

[1] This classical result is different from [Tolstikhin *et al.*, 2014] because we don't use the Bernstein condition.

$$\begin{aligned}&\mathcal{E}(\hat{\mathbf{w}}_{m,\text{previous}})\\ \leq& O\left(\frac{N}{u}\left(\max\left\{\sqrt{\frac{\mathcal{L}^* B^{\frac{1-\rho}{1+\rho}}}{m^{\frac{1}{1+\rho}}}}, \frac{B^{\frac{1-\rho}{1+\rho}}}{m^{\frac{1}{1+\rho}}}\right\}\right)\right.\\ &\left.+\frac{N}{m}\left(\max\left\{\sqrt{\frac{\mathcal{L}^* B^{\frac{1-\rho}{1+\rho}}}{u^{\frac{1}{1+\rho}}}}, \frac{B^{\frac{1-\rho}{1+\rho}}}{u^{\frac{1}{1+\rho}}}\right\}\right)\right).\end{aligned}$$

For VC classes:

$$\begin{aligned}&\mathcal{E}(\hat{\mathbf{w}}_{m,\text{previous}})\\ \leq& O\left(\frac{N}{u}\left(\max\left\{\sqrt{\frac{dB\mathcal{L}^*\log m}{m}}, \frac{Bd\log m}{m}\right\}\right)\right.\\ &\left.+\frac{N}{m}\left(\max\left\{\sqrt{\frac{dB\mathcal{L}^*\log u}{u}}, \frac{Bd\log u}{u}\right\}\right)\right).\end{aligned}$$

Here, $A \asymp B$ means there exist two positive constants c_1, c_2 such that $c_1 A \leq B \leq c_2 A$. Then we will start the discussions.

For non-parametric classes of polynomial growth, we consider the following two cases:

When $B \asymp 1$ and $\mathcal{L}^* \asymp m^{-a} + u^{-a}$, where a is a positive constant. This case depicts a bounded loss with a small order for $\mathcal{L}^*$ to obtain a tight bound. In this case, we can easily derive that when $0 \leq a \leq \frac{1}{1+\rho}$ performs better than traditional result since our result is of order $(\mathcal{L}^*)^{\frac{1-\rho}{2}}$ and traditional result is of order $(\mathcal{L}^*)^{\frac{1}{2}}$ w.r.t. $\mathcal{L}^*$.

When $B \asymp m^b + u^b$ and $\mathcal{L}^* \ll B^2$ where b is a positive constant. This case depicts that the worst-case boundedness parameter is considered to scale with m so that we want to reduce the dependence on B. In this case, when $B^{\frac{2}{1+\rho}} \leq \mathcal{L}^* \ll B^2$, our result also gains an improvement relative to the previous result. And the larger ρ is, the more improvement our results are. For example, when ρ is almost 1, and $m \asymp u$, our improvement can be as large as $O(m^{\frac{1}{4}})$.

For VC classes, we discuss the case that when $\mathcal{L}^* \geq \left(\frac{B}{(\log m + \log u)^a}\right)$, where a is a positive constant. We find that our results is $\log\log m$ (or $\log\log u$) term instead of $\log m$ (or $\log u$), which is strictly tighter than previous results.

Next, we notice that we don't know the population "effective loss" $\mathcal{L}^*$ in practice. We denote the empirical "effective loss" $\widehat{\mathcal{L}^*} = \hat{E}_m\left[\ell(\mathbf{w}_m^*; \mathbf{z}) - \inf_{\mathbf{w}\in\mathcal{W}} \ell(\mathbf{w}; \mathbf{z})\right]$ in this part and build a connection between the optimal upper bounds without the knowledge of $\mathcal{L}^*$.

Theorem 6 (Estimating loss-dependent rate from data)**.** *Note that term $\mathcal{L}^*$ is defined as $E_N\left[\ell(\mathbf{w}_N^*; \mathbf{z}) - \inf_{\mathbf{w}\in\mathcal{W}} \ell(\mathbf{w}; \mathbf{z})\right]$ and denote $\widehat{\mathcal{L}^*} = \hat{E}_m\left[\ell(\mathbf{w}_m^*; \mathbf{z}) - \inf_{\mathbf{w}\in\mathcal{W}} \ell(\mathbf{w}; \mathbf{z})\right]$. Under the conditions of Theorem 3, then for any fixed $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&R_N(\hat{\mathbf{w}}_m) - R_N(\mathbf{w}_N^*)\\ \leq& \max\left\{\psi_m\left(cB\widehat{\mathcal{L}^*}; \frac{\delta}{2\log_2 m + 6}\right), \frac{c\hat{R}_m^*}{m}, \frac{cB\log\frac{2}{\delta}}{m}\right\},\end{aligned}$$

where c is an absolute constant.

Remark 7. The term $\frac{B\log\frac{2}{\delta}}{m}$ in Theorem 6 is negligible, because $\hat{R}_m^*$ is at least of order $\frac{B^2\log\frac{1}{\delta}}{m}$ for most practical applications. This order is unavoidable in traditional "local Rademacher complexity" analysis and two-sided concentration inequalities. This generalization error bound shows that without knowledge of $\mathcal{L}^*$, one can estimate the order of our loss-dependent rate by using $\widehat{\mathcal{L}^*} = \hat{E}_m\left[\ell(\mathbf{w}_m^*; \mathbf{z}) - \inf_{\mathbf{w}\in\mathcal{W}} \ell(\mathbf{w}; \mathbf{z})\right]$ as a proxy. Despite replacing $\mathcal{L}^*$ by $\widehat{\mathcal{L}^*}$, other quantities in the bound remain unchanged in order.

5 Upper Bounds for Moment Penalization

The risk bounds provided in Section 4.2 consider the parameter $\mathcal{L}^*$ within their ψ_m function (or $E_N f$ using traditional localized peeling techniques), which may still be much larger than the optimal variance $\mathcal{V}^* := \mathrm{Var}[\ell(\mathbf{w}_N^*; \mathbf{z})]$. An example is given in [Namkoong and Duchi, 2017] in i.i.d. problems where $\mathcal{V}^* = 0$ and the optimal rate is at most $O\left(\frac{\log m}{m}\right)$, while the excess risk bound of ERM is proved to be slower than $O\left(\frac{1}{\sqrt{m}}\right)$.

We follow the path of penalizing empirical second moment in standard minimization settings [Namkoong and Duchi, 2017; Xu and Zeevi, 2024; Foster and Syrgkanis, 2023] to design an estimator that achieves the bias-variance trade-off for transductive learning. In order to adapt to $\mathcal{V}^*$, we use a sample-splitting two-stage estimation procedure which is inspired by the prior work in standard i.i.d minimization settings [Xu and Zeevi, 2024; Foster and Syrgkanis, 2023]. Without loss of generality, we assume the size of the whole dataset is $2N$ and the training dataset is $2m$ and split the training dataset into the primary dataset S and the auxiliary dataset S', both of which are of size m for training dataset and u for test dataset. We denote E_m the empirical distribution of the primary dataset and $E_{S'}$ the empirical distribution of the auxiliary dataset. Then we define the two-stage sample-splitting moment-penalized estimator.

Definition 5 (Two-stage Sample-splitting Moment-penalized Estimator)**.** *We use a sample-splitting two-stage estimation procedure.*

- *At the first stage, we derive a preliminary estimate of $\mathcal{L}_0^* := E\ell(\mathbf{w}_N^*; \mathbf{z})$ via the "auxiliary" data set S', which we refer to as $\widehat{\mathcal{L}_0^*}$.*
- *At the second stage, we perform regularized empirical risk minimization on the "primal" data set S, which considering the moment penalization. Consider the excess loss class $\mathcal{F}^* = \{f : f(\mathbf{z}) = \ell(\mathbf{w}; \mathbf{z}) - \ell(\mathbf{w}_N^*; \mathbf{z}), \mathbf{w} \in \mathcal{W}\}$. Let $\psi_m(r; \delta)$ be a meaningful surrogate function that satisfies $\forall\delta \in (0,1)$ and $\forall r > 0$, with probability at least $1-\delta$,*

$$\begin{aligned}&2\mathfrak{R}_m\{f \in \mathcal{F} : \hat{E}_m f^2 \leq 2r\} + \sqrt{\frac{2r\log\left(\frac{8}{\delta}\right)}{m}}\\ &+\frac{9\log\left(\frac{8}{\delta}\right)}{m} \leq \psi_m(r; \delta).\end{aligned}$$

Then give a fixed $\delta \in (0,1)$, let the moment-penalized estimator $\hat{\mathbf{w}}_{\mathrm{MP}}$ be

$$\hat{\mathbf{w}}_{\mathrm{MP}} \in \arg\min_{\mathbf{w}\in\mathcal{W}} \left\{ \hat{E}_m \ell(\mathbf{w};\mathbf{z}) + \psi_m\left(16\hat{E}_m[(\ell(\mathbf{w};\mathbf{z}) - \widehat{\mathcal{L}_0^*})^2]; \frac{\delta}{2\log_2 m + 5}\right)\right\}.$$

Using the estimator provided in Definition 5, we can derive the following variance-dependent rate.

Theorem 7. *Given arbitrary preliminary estimate $\widehat{\mathcal{L}_0^*} \in [-B, B]$, the generalization error of the moment-penalized estimator $\hat{\mathbf{w}}_{\mathrm{MP}}$ in Definition 5 is bounded by*

$$R_N(\hat{\mathbf{w}}_{\mathrm{MP}}) - R_N(\mathbf{w}_N^*) \le 2\psi_m\left(c_0\left[\max\left\{\mathcal{V}^*, \hat{R}_m^*, (\widehat{\mathcal{L}_0^*} - \mathcal{L}_0^*)^2\right\}\right]; \frac{\delta}{2\log_2 m + 5}\right),$$

with probability at least $1-\delta$, where c_0 is an absolute constant and the term $\hat{R}_m^*$ is the fixed point of $16B\psi_m\left(r; \frac{\delta}{2\log_2 m+5}\right)$.

Remark 8. Notice that we don't need the assumption that ψ_m is "sub-root" function in Theorem 7 and we can easily find that the first stage estimation error $(\widehat{\mathcal{L}_0^*} - \mathcal{L}_0^*)^2$ can be omitted if $(\widehat{\mathcal{L}_0^*} - \mathcal{L}_0^*)^2 \le O(\hat{R}_m^*)$.

Further more, to bound the first stage estimation error $(\widehat{\mathcal{L}_0^*} - \mathcal{L}_0^*)^2$, we add the "sub-root" assumption in Theorem 8.

Theorem 8. *Given arbitrary preliminary estimate $\widehat{\mathcal{L}_0^*} \in [-B, B]$, the generalization error of the moment-penalized estimator $\hat{\mathbf{w}}_{\mathrm{MP}}$ in Definition 5 is bounded by*

$$R_N(\hat{\mathbf{w}}_{\mathrm{MP}}) - R_N(\mathbf{w}_N^*) \le \max\left\{2\psi_m\left(c_1\mathcal{V}^*; \frac{\delta}{2\log_2 m + 5}\right), \frac{c_1\hat{R}_m^*}{8B}\right\},$$

with probability at least $1-\delta$, where c_0 is an absolute constant and the term $\hat{R}_m^*$ is the fixed point of $16B\psi_m(r; \frac{\delta}{2\log_2 m+5})$.

Remark 9. Similarly, $\frac{\delta}{2\log_2 m+5}$ only appear in the form $\log\log\left(\frac{2\log_2 m+5}{\delta}\right)$ in the final result, which is of order $\log\log m$ and can be regarded as an absolute constant for all practical purposes. We have to emphasize that the "sub-root" assumption is only used to bound the first-stage estimation error $(\widehat{\mathcal{L}_0^*} - \mathcal{L}_0^*)^2$ defined in Definition 5. We can also apply the Dudley's integral bound and derive the results for non-parametric classes of polynomial growth and VC classes.

Theorem 9 (Non-parametric classes of polynomial growth)**.** *Consider a loss class $\ell \circ \mathcal{W}$ with the metric entropy condition*

$$\log \mathcal{N}(\varepsilon, \ell\circ\mathcal{W}, d_{\ell\circ\mathcal{W}}) \le O(\varepsilon^{-2\rho}),$$

under the conditions of Theorem 8, then for any fixed $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, we have

$$\begin{aligned} &R_N(\widehat{\mathbf{w}}_{\mathrm{MP}}) - R_N(\mathbf{w}_u^*) \\ \le& O\left(\frac{N}{u}\left(\max\left\{\sqrt{\frac{(\mathcal{V}^*)^{1-\rho}}{m}}, \frac{B^{\frac{1-\rho}{1+\rho}}}{m^{\frac{1}{1+\rho}}}\right\}\right)\right. \\ &\left. + \frac{N}{m}\left(\max\left\{\sqrt{\frac{(\mathcal{V}^*)^{1-\rho}}{u}}, \frac{B^{\frac{1-\rho}{1+\rho}}}{u^{\frac{1}{1+\rho}}}\right\}\right)\right). \end{aligned}$$

Theorem 10 (VC classes)**.** *Consider a loss class $\ell \circ \mathcal{W}$ with the metric entropy condition*

$$\log \mathcal{N}(\varepsilon, \ell\circ\mathcal{W}, d_{\ell\circ\mathcal{W}}) \le \left(d\log\frac{1}{\varepsilon}\right),$$

under the conditions of Theorem 8, then for any fixed $\delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$, we have

$$\begin{aligned} &R_N(\widehat{\mathbf{w}}_{\mathrm{MP}}) - R_N(\mathbf{w}_u^*) \\ \le& O\left(\frac{N}{u}\left(\max\left\{\sqrt{\frac{dB^2\mathcal{V}^*}{m}\log\frac{B^2}{3\mathcal{V}^*}}, \frac{Bd}{m}\log\frac{B^2}{3\mathcal{V}^*}, \frac{Bd\log m}{m}\right\}\right)\right. \\ &\left. + \frac{N}{m}\left(\max\left\{\sqrt{\frac{dB^2\mathcal{V}^*}{u}\log\frac{B^2}{3\mathcal{V}^*}}, \frac{Bd}{u}\log\frac{B^2}{3\mathcal{V}^*}, \frac{Bd\log u}{u}\right\}\right)\right). \end{aligned}$$

Remark 10. The results for moment penalized estimator is similar to the ERM algorithm. And to the best of our knowledge, our results are the only generalization result in transductive learning that consider variance information. Similarly, we can derive the bound with the variance-dependent rate from data.

Theorem 11 (Estimating Variance-dependent Bounds from Data)**.** *Consider the empirical centered second moment*

$$\widehat{\mathcal{V}^*} := \hat{E}_m\left[\ell(\hat{\mathbf{w}}_{\mathrm{NMP}};\mathbf{z}) - \widehat{\mathcal{L}_0^*})^2\right],$$

where $\widehat{\mathcal{L}_0^*} \in [-B, B]$ is the preliminary estimate of $\mathcal{L}^*$ obtained in the first-stage, ψ_m is defined in Definition 5, For any fixed $\delta \in (0,1)$, by performing the moment-penalized estimator in Definition 5, with probability at least $1-\frac{\delta}{2}$,

$$\mathcal{E}(\hat{\mathbf{w}}_{\mathrm{MP}}) \le \max\left\{4\psi_m\left(16\widehat{\mathcal{V}^*}; \frac{\delta}{2\log_2 m + 5}\right), \frac{\hat{R}_m^*}{8B}\right\},$$

where $\hat{R}_m^*$ is the fixed point of $8\psi_m\left(r; \frac{\delta}{2\log_2 m+5}\right)$.

Remark 11. One should view Theorem 11 as a relaxation of the original variance-dependent rate in Theorem 8. We also notice that the "sub-root" assumption in Theorem 11 is not needed here as we do not discuss the precision of $\widehat{\mathcal{L}_0^*}$.

6 Conclusion

In this paper, we develop a novel functional based peeling technique to derive better uniform localized convergence upper bounds in transductive learning without "sub-root" assumption for functions that upper bound the uniform error within a localized region. Our method can obtain tighter risk bounds comparing with existing work [Tolstikhin *et al.*, 2014] for ERM. Furthermore, we design a novel estimator based on the second moment penalization and derive its generalization bounds, which are the first results in transductive learning.

Ethical Statement

There are no ethical issues.

Acknowledgments

This research was supported by National Natural Science Foundation of China (No.62476277), National Key Research and Development Program of China(NO. 2024YFE0203200), CCF-ALIMAMA TECH Kangaroo Fund(No.CCF-ALIMAMA OF 2024008), and 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.

References

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

[Bégin *et al.*, 2014] Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. Pac-bayesian theory for transductive learning. In *Artificial Intelligence and Statistics*, pages 105–113. PMLR, 2014.

[Blum and Langford, 2003] Avrim Blum and John Langford. Pac-mdl bounds. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 344–357. Springer, 2003.

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

[Bousquet, 2002a] Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[Bousquet, 2002b] Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

[Chien *et al.*, 2020] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. *arXiv preprint arXiv:2006.07988*, 2020.

[Cong *et al.*, 2021] Weilin Cong, Morteza Ramezani, and Mehrdad Mahdavi. On provable benefits of depth in training graph convolutional networks. *Advances in Neural Information Processing Systems*, 34:9936–9949, 2021.

[Cortes and Mohri, 2006] Corinna Cortes and Mehryar Mohri. On transductive regression. *Advances in neural information processing systems*, 19, 2006.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[Cortes *et al.*, 2009] Corinna Cortes, Mehryar Mohri, Dmitry Pechyony, and Ashish Rastogi. Stability analysis and learning bounds for transductive regression algorithms. *arXiv preprint arXiv:0904.0814*, 2009.

[Derbeko *et al.*, 2004] 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.

[El-Yaniv and Pechyony, 2006] Ran El-Yaniv and Dmitry Pechyony. Stable transductive learning. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, pages 35–49. Springer, 2006.

[El-Yaniv and Pechyony, 2009] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

[Esser *et al.*, 2021] Pascal Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. *Advances in Neural Information Processing Systems*, 34:27043–27056, 2021.

[Foster and Syrgkanis, 2023] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

[Koltchinskii and Panchenko, 2000] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

[Li *et al.*, 2021] Chen Li, Xutan Peng, Hao Peng, Jianxin Li, and Lihong Wang. Textgtl: Graph-based transductive learning for semi-supervised text classification via structure-sensitive interpolation. In *IJCAI*, pages 2680–2686, 2021.

[Meir and Zhang, 2003] Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

[Namkoong and Duchi, 2017] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30, 2017.

[Oono and Suzuki, 2020] Kenta Oono and Taiji Suzuki. Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks. *Advances in Neural Information Processing Systems*, 33:18917–18930, 2020.

[Sridharan, 2010] Karthik Sridharan. Note on refined dudley integral covering number bound. https://www.cs.cornell.edu/ sridharan/dudley.pdf, 2010.

[Talagrand, 1996] Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.

[Tang and Liu, 2023a] Huayi Tang and Yong Liu. Information-theoretic generalization bounds for transductive learning and its applications. *arXiv preprint arXiv:2311.04561*, 2023.

[Tang and Liu, 2023b] Huayi Tang and Yong Liu. Towards understanding the generalization of graph neural networks. *arXiv preprint arXiv:2305.08048*, 2023.

[Tolstikhin *et al.*, 2014] Ilya Tolstikhin, Gilles Blanchard, and Marius Kloft. Localized complexities for transductive learning. In *Conference on Learning Theory*, pages 857–884. PMLR, 2014.

[Vapnik, 1982] Vladimir Vapnik. Estimation of dependences based on empirical data, 1982.

[Vapnik, 1998] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

[Vapnik, 1999] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[Wainwright, 2019] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[Xu and Zeevi, 2024] Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *Mathematics of Operations Research*, 2024.