STARS

University of Central Florida
STARS

Electronic Theses and Dissertations

2017

Improved Multi-Task Learning Based on Local Rademacher Analysis

Niloofar Yousefi
University of Central Florida

 Part of the Industrial Engineering Commons
Find similar works at: https://stars.library.ucf.edu/etd
University of Central Florida Libraries http://library.ucf.edu

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation
Yousefi, Niloofar, "Improved Multi-Task Learning Based on Local Rademacher Analysis" (2017). *Electronic Theses and Dissertations*. 5544.
https://stars.library.ucf.edu/etd/5544

IMPROVED MULTI-TASK LEARNING BASED ON LOCAL RADEMACHER ANALYSIS

by

NILOOFAR YOUSEFI
B.S. Iran University of Science & Technology, 2008
M.S. University of Tehran, 2012

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctoral of Philosophy
in the Department of Industrial Engineering and Management Systems
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2017

Major Professor: Mansooreh Mollaghasemi

© 2017 Niloofar Yousefi

ABSTRACT

Considering a single prediction task at a time is the most commonly paradigm in machine learning practice. This methodology, however, ignores the potentially relevant information that might be available in other related tasks in the same domain. This becomes even more critical where facing the lack of a sufficient amount of data in a prediction task of an individual subject may lead to deteriorated generalization performance. In such cases, learning multiple related tasks together might offer a better performance by allowing tasks to leverage information from each other. Multi-Task Learning (MTL) is a machine learning framework, which learns multiple related tasks simultaneously to overcome data scarcity limitations of Single Task Learning (STL), and therefore, it results in an improved performance. Although MTL has been actively investigated by the machine learning community, there are only a few studies examining the theoretical justification of this learning framework. The focus of previous studies is on providing learning guarantees in the form of generalization error bounds. The study of generalization bounds is considered as an important problem in machine learning, and, more specifically, in statistical learning theory. This importance is twofold: (1) generalization bounds provide an upper-tail confidence interval for the true risk of a learning algorithm the latter of which cannot be precisely calculated due to its dependency to some unknown distribution P from which the data are drawn, (2) this type of bounds can also be employed as model selection tools, which lead to identifying more accurate learning models.

The generalization error bounds are typically expressed in terms of the empirical risk of the learning hypothesis along with a complexity measure of that hypothesis. Although different complexity measures can be used in deriving error bounds, Rademacher complexity has received considerable attention in recent years, due to its superiority to other complexity measures. In fact, Rademacher complexity can potentially lead to tighter error bounds compared to the ones obtained by other

complexity measures. However, one shortcoming of the general notion of Rademacher complexity is that it provides a global complexity estimate of the learning hypothesis space, which does not take into consideration the fact that learning algorithms, by design, select functions belonging to a more favorable subset of this space and, therefore, they yield better performing models than the worst case. To overcome the limitation of global Rademacher complexity, a more nuanced notion of Rademacher complexity, the so-called local Rademacher complexity, has been considered, which leads to sharper learning bounds, and as such, compared to its global counterpart, guarantees faster convergence rates in terms of number of samples. Also, considering the fact that locally-derived bounds are expected to be tighter than globally-derived ones, they can motivate better (more accurate) model selection algorithms.

While the previous MTL studies provide generalization bounds based on some other complexity measures, in this dissertation, we prove excess risk bounds for some popular kernel-based MTL hypothesis spaces based on the Local Rademacher Complexity (LRC) of those hypotheses. We show that these local bounds have faster convergence rates compared to the previous Global Rademacher Complexity (GRC)-based bounds. We then use our LRC-based MTL bounds to design a new kernel-based MTL model, which enjoys strong learning guarantees. Moreover, we develop an optimization algorithm to solve our new MTL formulation. Finally, we run simulations on experimental data that compare our MTL model to some classical Multi-Task Multiple Kernel Learning (MT-MKL) models designed based on the GRCs. Since the local Rademacher complexities are expected to be tighter than the global ones, our new model is also expected to exhibit better performance compared to the GRC-based models.

To my beloved parents

ACKNOWLEDGMENTS

In the past five years, I was very fortunate to receive support and help from many people, and I would like to take this opportunity to thank these individuals in writing.

First and foremost, I would like to thank my committee chair Prof. Mansooreh Mollaghasemi for giving me guidance and support. I am deeply grateful for having had the opportunity to be her Ph.D. student. I am especially thankful for the wonderful example she has provided as a successful woman, mentor, instructor and advisor.

I am also greatly indebted to my advisor Prof. Michael Georgiopoulos for creating the inspiring and creative research environment, in which I have performed my graduate studies. His ultimate willingness to invest himself into the guidance and support of young scientists and researchers, his ability to inspire his students and encourage them to stay engaged and focused on learning are unrivaled. His personality, patience, motivation, enthusiasm and dedication make him a great mentor, advisor, and more importantly a role model. I especially thank him for providing guidance at key moments during my research at the Machine Learning Lab (ML^2) at University of Central Florida.

My deep appreciation goes to my co-advisor Prof. Georgios Anagnostopoulos for his invaluable guidance and feedback on my research and for always being available to advise me. I would like to thank Dr. Anagnostopoulos for spending an enormous amount of time for our extensive and valuable discussions on various machine learning as well as mathematical problems. I would never forget the numerous discussions we had, which made my Ph.D productive and stimulating. This thesis owes its existence to his support and caring mentoring during various stages of my Ph.D. tenure.

I would also like to sincerely thank my other committee members, Dr. Luis Rabelo, Dr. Qipeng Zheng and Dr. Petros Xanthopoulos for devoting time to assess my research work, and providing valuable suggestions and comments through this process.

Moreover, I want to thank present and past members of the ML^2 for all the chats and good memories we shared during the stressful and difficult moments of our Ph.D endeavor.

Besides mentors, this dissertation is greatly influenced by the contribution of Prof. Marius Kloft who has given me tremendous help during the completion of this investigation. I would like to express my sincere appreciation and gratitude to Dr. Kloft for his invaluable and inspiring suggestions and sharing his vast knowledge and experience with me, which helped gaining a deeper understanding of statistical learning theory.

A very special thank goes to my parents for their unflagging love, faith and encouragement in all my pursuits. My greatest fortune is being blessed with a family that has been always supportive and encouraging even when it meant traveling thousand of miles to pursue life on a different continent. I appreciate their unconditional support that gives me freedom to explore the world.

Most importantly, I would like to thank my best friend, soul-mate, and husband, Ali who has been a true and great supporter. I would like to express my heartfelt appreciation for being by my side throughout my Ph.D. studies even during the most difficult times. His faith in me, quiet patience and unwavering love kept my spirits up and encouraged me to embark on this journey.

Finally, I acknowledge financial support from National Science Foundation (NSF) grant No. 1161228 (COMPASS Project), and No. 1200566 (AEGIS RET Program). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Introduction

While most traditional machine learning approaches focus on the learning of a single independent task at a time, Multi-Task Learning (MTL), in contrast, aims to training several related tasks together, with the hope of improving the overall performance of all tasks by allowing information sharing between them. More specially, when only a limited number of training samples per each task exists, MTL can benefit tasks by inducing a positive inductive bias in the learning process of multiple related tasks. Therefore, more effective training can be conducted in this way, which leads to improved generalization performance for each task compared to the "no transfer" scenario, where each task is learned in isolation.

Nowadays, MTL frameworks are routinely employed in a variety of settings. Some application domains include computer vision [1, 60, 79, 122, 137, 149], HIV therapy screening [15], collaborative filtering [22], age estimation from facial images [149], and sub-cellular location prediction [142] , Information retrieval [121, 22, 123], bioinformatics [15, 142, 90] and finance [52] just to name a few prominent ones.

The underlying assumption behind the MTL paradigm is based on tasks' relatedness. Therefore, the key concern of MTL is "how to capture tasks relatedness and integrate it into the learning formulation." In response to this question, several MTL approaches have been designed, which employ different strategies to capture task relatedness. Although, these models differ in how they model the relationship among tasks, they mostly formulate MTL as a regularized Empirical Risk Minimization (ERM) problem, in which the objective function is a composition of an over-the-tasks average error and a regularization term to encourage information sharing among tasks in

some capacity. More precisely, similar to many machine learning models, the regularized MTL formulation is typically given as

$$\min_{\boldsymbol{f}} \mathcal{L}(\boldsymbol{f}) + \lambda \Omega(\boldsymbol{f}) \tag{1.1}$$

where $\boldsymbol{f}$ is a vector-valued function that consists of the tasks' learning functions $(f_1, \ldots, f_T)$, $\mathcal{L}(\boldsymbol{f})$ is the averaged empirical loss over all tasks, and $\Omega(\boldsymbol{f})$ is the regularization term which is designed to enforce information sharing among tasks. Also, λ is the regularization parameter that allows for choosing the right trade-off between $\mathcal{L}(\boldsymbol{f})$ and $\Omega(\boldsymbol{f})$. There are many prior efforts which utilize this framework to model task relationships, among which we refer to [47, 155, 158, 129, 103, 102], just to name a few. It is worth pointing out that in some cases, instead of (or besides) the regularization term, some optimization constraints are also incorporated into MTL formulation (1.1) in favor of adding some other desired characteristics. A good example of this situation is where a clustering or grouping strategy is needed to be considered in order to allow different level of information sharing between different tasks. This goal can then be achieved by imposing clustering-type constraints into Problem (1.1). Also, in order to allow more flexibility and achieve better generalization performance, kernel-based regularizations have been proposed in the context of MTL [46]. Beside flexibility, simplicity and generality of kernels and their associated Reproducing Kernel Hilbert Space (RKHS)s, "availability of effective error bounds and stability analysis relative to perturbations of the data" [109] is another attractive feature of kernel-based regularizers. Interestingly, it can be shown that there is an equivalency (Ivanov-Tikhonov regularization equivalency) between (1.1) and optimization problem

$$\begin{aligned} &\min_{\boldsymbol{f}} \mathcal{L}(\boldsymbol{f}) \\ &\text{s.t. } \Omega(\boldsymbol{f}) \leq R, \end{aligned} \tag{1.2}$$

which can be efficiently used to identify the hypothesis space of the learning problem at hand. In more detail, given T learning tasks, the learning problem 1.2 seeks the vector-valued function $\boldsymbol{f} = (f_1, \ldots, f_T)$ from the hypothesis space $\mathcal{F} := \{\boldsymbol{f} = (f_1, \ldots, f_T) : \Omega(\boldsymbol{f}) \leq R\}$ such that the average empirical error $\mathcal{L}(\boldsymbol{f})$ is minimized.

Problem Statement

The study of generalization error bounds is important in machine learning problems, as they are uniform over the learning hypothesis space, that is, the bounds hold for any function f within the hypothesis space under consideration. These type of bounds provide upper-tail confidence intervals for the true risk. But, even more importantly, the same bounds can also be used as model selection tools where, among several alternatives, one can identify the model that most likely has the lowest risk. In particular, such a bound is usually based on an empirical measurement of its risk (error) and a measure of its complexity. Also, it is worth pointing out that

To be more concrete regarding the importance of the generalization bounds, recall that the main goal of any typical machine learning algorithm is to automate the process of learning a model based on some observations from a phenomenon in order to make good predictions in the future with the help of the learned model [21]. To make this more precise, consider a supervised learning paradigm (as we restrict ourselves to this case in this dissertation), and n training samples $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, which are identically and independently drawn from an unknown distribution P. A supervised learning algorithm then, aims to construct a function $f : \mathcal{X} \to \mathcal{Y}$, which depends on this observed data and generalizes well over any unseen future data (X, Y). Hence, the goal is to select a function $f \in \mathcal{F}$ with small risk or expected loss $\mathbb{E}_{(X,Y)\sim P}\left[\ell(f(X), Y)\right]$. However, the calculation of this quantity is impossible, as the distribution P is unknown. One common approach to estimate the true risk $\mathbb{E}_{(X,Y)\sim P}\left[\ell(f(X), Y)\right]$ is to relate this quantity by its

empirical counterpart along with a complexity measure of the function class $\mathcal{F}$. A typical form of a generalization bound is

$$\mathbb{E}_{(X,Y)\sim P}\left[\ell(f(X),Y)\right] \leq \frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i),Y_i) + h\left(\text{complexity of the function class } \mathcal{F}, n\right) \quad (1.3)$$

It is worth mentioning that different complexity measures such as Vapnik-Chevonenkis (VC) dimension, fat-shattering dimension or covering numbers can be used in deriving generalization bounds. However, in this dissertation, we use a more recent notion of complexity called Rademacher complexity, which, in turn, can be bounded by other complexity measures (such as covering numbers or VC dimension), and therefore improves existing bounds based on these other measures. More importantly, the generalization bounds based on empirical version of the Rademacher complexity are data dependent, meaning that they measure the complexity of the function class $\mathcal{F}$ based on the training samples. In other words, these data-dependent generalization bounds can be estimated based on finite samples and they are usually tighter than their distribution-dependent counterparts [111]. Also, data-dependent bounds (such as Rademacher-based bounds) are of more value as they can provide strong theoretical foundation in designing of new learning algorithms. As an example, for kernel-based hypotheses, the empirical Rademacher-based bounds are typically functions of the kernel matrix. This can lead to deriving kernel learning algorithms which, by considering a regularization on the trace of the kernels, benefits from strong learning guarantees. As an effective complexity measure, the Rademacher complexity was first proposed by [74], [9] and [107]. However, one shortcoming of the general notion of Rademacher complexity is that it provides the global estimation of the complexity of the function class $\mathcal{F}$. In other words, it does not take into consideration the fact that learning algorithms, typically, pick functions belonging to a more favorable subset of the function class, and they therefore yield better performance than the worst case. Recall that most learning algorithms tend to choose functions

inducing small empirical errors and also (hopefully) small generalization errors. Therefore, it is very likely that the function $\hat{f} \in \mathcal{F}$ minimizing the empirical risk $\frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$, lies in a neighborhood of the best function $f^* \in \mathcal{F}$ that minimizes the true risk $\mathbb{E}_{(X,Y)}\left[\ell(f(X), Y)\right]$. To overcome the limitation of *global* Rademacher complexity, a finer notion of Rademacher complexity, the so-called *local* Rademacher complexity, has been considered which leads to sharper learning bounds and as such also (compared to its global counterpart, i.e. GRC) guarantees a faster rate of convergence—the rate at which the empirical risk approaches the true risk—under some general conditions. Also, regarding the fact that local bounds are expected to be tighter than the global ones, they can motivate more efficient model selection algorithms.

The idea of LRC is to restrict the function class $\mathcal{F}$ to a mush smaller subset of it, by imposing a variance-type constraint on this class of functions. Since such a small class can also have smaller Rademacher complexity, they can lead to sharper bounds compared to GRC-based bounds.

Contributions

Although MTL has been actively investigated by the machine learning community, there are only a few studies examining the theoretical perspective of this learning framework. While these previous MTL studies provide generalization bounds based on some other complexity measures such as covering number and VC-dimension [11, 13, 2] or GRC [101, 102, 104, 105] of the learning hypothesis, we are not aware of any study taking advantage of the LRCs to derive error bounds for MTL. In this dissertation, we derive excess risk bounds for some popular kernel-based MTL hypothesis spaces based on the LRC of those hypotheses. It turns out that similar to the STL scenario, for kernel-based hypotheses, the data-dependent LRC-based MTL bounds are functions of the tail sum of the eigenvalues of the kernel matrices. Also, as expected (and it has been shown for STL [10]), we show that theses local bounds have faster convergence rate compared

to previously known GRC-based bounds. Furthermore, similar to what has been done in [38] for STL, we use our LRC-based MTL bounds to design a kernel-based MTL model which considers a constraint based on the tail sum of the eigenvalues of the kernels. Finally, we show that our new LRC-based MTL model consistently outperforms the traditional kernel learning algorithms, whose performances have been proven to be difficult to surpass in the past; such as uniform combination solution as well as convex combination of base kernels. Note that it can be shown that the latter case—with an ℓ_1 norm constraint on kernel parameter—corresponds to a model which is derived based on a GRC analysis. We show the superiority of our LRC model against this GRC-based algorithm, by performing a series of experiments.

Organization

The rest of this dissertation is organized as follows: In Chapter 2, we provide some background on statistical learning theory, and some of its concepts including supervised learning, generalization error bounds, global and local Rademacher complexity-based bounds. Finally, we introduce the general MTL setting and formulation at the end of Chapter 2. Chapter 3 presents a literature review associated with MTL models and generalization bounds. Also, the methodology of our study can be found in Chapter 4. Then, Chapter 5 details the derivation of the LRC-based generalization bound for MTL. Risk bounds are eventually found for several common MTL framework considering norm regularizers. A thorough analysis of the derived bound as well as an insightful comparison to the existing bounds is included which demonstrates the advantages of the LRC-based bounds. Additionally, due to the superiority of the derived LRC bounds, a new kernel-based MTL model along with an optimization algorithm are introduced in Chapter 6. The experimental evaluation of the work is given in Chapter 7. Chapter 8 provides summary and potential future directions of the work.

CHAPTER 2: BACKGROUND

With the ever-increasing amount of data, it is almost impossible for a human programmer or specialist to detect a meaningful pattern in data and translate it to some expertise or knowledge for the future use. For this reason, *machine leaning*, as an automated learning tool, has become a central part of human life over the past couple of decades. Machine learning refers to an automated process of detecting meaningful patterns from data, and it has applications in many real world problems where information extraction from large data sets is required. Ranking the relevant web pages given a submitted query into a search engine, filtering email messages by an anti-spam software, securing credit card transactions by a fraud detection software, detecting faces in digital photos, recognizing voice commands on smart phones, and accident preventing systems in cars are just some examples of machine learning applications in real world problems. Machine learning also appears in many other scientific guises such as bioinformatics, medicine, and astronomy. In this section, we provide some background on the main concepts underlying machine learning.

Automated Learning

Learning, of course, covers a wide range of processes which is difficult to define precisely. Consequently, machine learning has waded into several branches, each of which dealing with a different type of learning task. However, the common feature of all different types of machine learning models is that they automate the process of an inductive inference including, observing a phenomenon, building a model based on the observed phenomenon, and making predictions using the constructed model.

In this dissertation, we consider a special type of this learning process which is called supervised

learning, and primarily, we will be dealing with binary classification problems. In this framework, the phenomenon is defined as some instance-label pairs, where a label is either $+1$ or -1. A classification model is then constructed as a mapping function from the instances to the labels. This function is expected to make future predictions for unseen instances with as few mistakes as possible. Note that it is always possible to build a function that agrees very well with the observed (training) data. However such a model might exhibit a poor performance in predicting unseen future data. As example of this instance is the case where the training data are noisy. This phenomenon is referred to as *overfitting*, and it happens when the model can fit the training data too well. This type of models are usually too complex in the sense that they have too many free parameters to tune. Therefore, one way to avoid overfitting is to restrict the choice of the learning model to a set of predictors with less *complexity*. This set of predictors is called a *hypothesis class* and it is typically chosen in advance based on some specific assumptions or knowledge about the data. Another way to avoid overfitting is to add a penalty (to the learning process) for complicated hypothesis classes. This is usually referred to as *regularization* technique, and it is known as a very successful method in all machine learning problems. By using one of these (and usually the combination of both) techniques, we can expect that the learning model can be reasonably *generalized* from the observed data to future unseen instances. In other words, the model is expected to make future predictions with as small risk as possible. In the following, we discuss in more detail: "What is *regularization* technique?", "How to quantify the *complexity* of a hypothesis class?", or" How to measure the *generalization* of a model?". Before answering these fundamental questions, let us first to formally describe the supervised learning paradigm.

Formalization

Consider an input space $\mathcal{X}$ as the set of objects we want to label. Also, assume that the output space $\mathcal{Y}$ denotes the set of possible labels, which are chosen as $\{-1, +1\}$ for binary classification. We then assume that the training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are identically and independently drawn from an unknown distribution P defined on $\mathcal{X} \times \mathcal{Y}$. Now, given the training data, the objective of a learning algorithm is to choose a function $f : \mathcal{X} \to \mathcal{Y}$ among the functions in the hypothesis class $\mathcal{F}$, which generalizes well. In other words, this function should be chosen in such a way that the probability of error $P(f(X) \neq Y)$ is small for any unseen instance pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Therefore, the *true risk* of a predictor function f can be given as

$$R(f) := \mathbb{E}_{(X,Y)\sim P}\left[(f(X) \neq Y)\right] = \mathbb{E}\left[\mathbf{1}_{f(X)\neq Y}\right]. \tag{2.1}$$

Regularization

The objective of a learning algorithm is to choose a function $f \in \mathcal{F}$ that minimizes the risk $R(f)$. However, the true risk $R(f)$ cannot be calculated, due to its dependency to the unknown distribution P. But, we can quantify the consistency of the function f with the training data through an *empirical risk* defined as

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{f(X_i)\neq Y_i}. \tag{2.2}$$

which is commonly used as a criterion to choose a function f from the hypothesis space $\mathcal{F}$. This algorithm, which is called *Empirical Risk Minimization*, is based on the idea of choosing a predictor function $f \in \mathcal{F}$ which minimized (2.2). However, as mentioned earlier, this hypothesis class $\mathcal{F}$ is predefined based on some priori assumptions regarding the problem. Therefore, one may want

to enlarge $\mathcal{F}$ as much as possible to increase the chances of finding a good predictor f in $\mathcal{F}$. From the other side, enlarging $\mathcal{F}$ might increase the risk of overfitting. Therefore, a regularizer is usually imposed on $\mathcal{F}$ to prevent overfitting, while choosing a large class $\mathcal{F}$. Intuitively, the regularization function $\Omega(f)$ is a measure of the complexity of $\mathcal{F}$ that reflects some prior belief about the problem. Regularized empirical risk minimization algorithms solve the following problem

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{f(X_i) \neq Y_i} + \lambda \Omega(f) \tag{2.3}$$

which tries to balance between "better fit" and "less complexity" of $\mathcal{F}$.

Generalization Error Bounds

In a binary classification setting, given a predictor function $f : \mathcal{X} \to \mathbb{R}$, by convention, the point X is considered to be classified as class $+1$ if $f(X) > 0$, and class -1 if $f(X) < 0$, and it is considered misclassified otherwise. In other words, any instance (X_i, Y_i) is classified correctly by the predictor function f, only if $Y_i f(X_i) > 0$. Therefore, the risk associated to function f can be defined as

$$\mathbb{E}\left[\mathbf{1}_{Y f(X) \leq 0}\right]$$

where $\mathbf{1}_{Y f(X) \leq 0}$ is known as $0-1$ loss, and it is a non-convex, non-differentiable function. These characteristics make the optimization hard. For this reason, a convex or/and differentiable surrogate function ℓ, which upper-bounds this loss, is optimized. For ease of notation, let $Z_i := (X_i, Y_i)$ and $Z := (X, Y)$. Now we define the class of *loss* function $\mathcal{L}_{\mathcal{F}}$ associated with f as

$$\mathcal{L}_{\mathcal{F}} := \{\ell_f : (X, Y) \to \ell(f(X), Y), f \in \mathcal{F}\}$$

Also, for convenience, let us introduce the shorthand notations $P\ell_f := \mathbb{E}_{(X,Y)\sim P}\left[\ell(f(X), Y)\right]$ and $P_n\ell_f := \frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i), Y_i)$ as the true and empirical risks of f, respectively.

As noted earlier, the optimal goal is to characterize the true risk associated with the predictor function f. However, as it depends on the unknown probability P, it is impossible to estimate this quantity. One way to approximate the true risk $R(f)$ is to relate it to its empirical counterpart, and one prominent approach to do this is based on the theory of uniform convergence of empirical quantities to their mean (see e.g. [140]). In other words, this theory provides an upper-bound on the quantity

$$P\ell_f - P_n\ell_f \tag{2.4}$$

It it worth pointing that the bound on (2.4) is usually expressed as a function of the complexity of $\mathcal{F}$ along with a function of the number of samples n. An important aspect of the generalization bounds is that they are uniform over the learning hypothesis space $\mathcal{F}$, that is, these bounds hold for any function f which lies within the function class $\mathcal{F}$. Although, different complexity measures can be used in deriving generalization bounds, in this dissertation, we use a more recent notion of complexity called *Rademacher complexity*, which usually leads to tighter, high-quality bounds.

Relationship of Error Bounds to Empirical Processes

This section provides some insights on how the generalization error bounds can be obtained. Recall that in order to find the generalization bounds, one needs to obtain bounds on $Pf - P_nf$, in which the function f is chosen from a function space $\mathcal{F} := \{f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$. As mentioned before, this quantity is a random variable, and the randomness stems from (i) the unknown distribution P and (2) the function f which depends on how a learning algorithm processes the data to select this

function from $\mathcal{F}$ [20]. However, the latter randomness can be removed by considering a collection of random variables $Pf - P_n f$ indexed by the function set $\mathcal{F}$:

$$\{Pf - P_n f\}_{f \in \mathcal{F}}, \tag{2.5}$$

which is known as the *empirical process* in statistical learning theory. A more helpful quantity associated to an empirical process is

$$\sup_{f \in \mathcal{F}} (Pf - P_n f). \tag{2.6}$$

Note that a bound on (2.6) also acts a bound on (2.5). Since the supremum of the empirical process $Pf - P_n f$ in (2.6) is still random due to its dependency on the unknown distribution P, the bounds for this quantity takes the probabilistic form

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} (Pf - P_n f) \geq x\right] \leq \delta, \tag{2.7}$$

where $\delta > 0$. The expression in (2.7) is usually referred to as *concentration inequality* which provides a bound on the probability that a random variable Z differs from its expected value $\mathbb{E}Z$ by more than a certain amount. A number of methods have been proposed to derive these type of inequalities such as martingale methods [110, 106], decoupling methods [41], Talagrand's induction method [133, 134, 93, 115] and the so-called "entropy method" which is based on logarithmic Sobolev inequalities introduced in [80, 81, 16, 125, 100, 94, 17, 19, 19]. Being related to the tail bounds of empirical processes, we are interested to obtain probabilistic bounds for $\{Z - \mathbb{E}Z \geq x\}$, in which $Z := \sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are n independent random variables. One useful bound of this form is the so-called Hoeffding's tail inequality which is introduced bellow.

Theorem 1 (Hoeffding's inequality)**.** *Assume that $X_1, \ldots, X_n$ are n independent bounded random*

variables such that for each $i \in 1, \ldots, n$, X_i takes values in $[a_i, b_i]$. If $Z := \sum_{i=1}^{n}, X_i$, then for any $x > 0$, the following holds

$$\mathbb{P}\{Z - \mathbb{E}Z \geq x\} \leq \exp\left[\frac{-2x^2}{\sum_i^n (b_i - a_i)^2}\right]$$

and,

$$\mathbb{P}\{Z - \mathbb{E}Z \leq -x\} \leq \exp\left[\frac{-2x^2}{\sum_i^n (b_i - a_i)^2}\right]$$

The generalization of Hoeffding's inequality to functions of i.i.d. random variables is know as McDiarmid (or bounded differences) inequality.

Theorem 2 (McDiarmid's Inequality)**.** *For function $g : \mathcal{X}^n \to \mathbb{R}$, let $Z := g(X_1, \ldots, X_i, \ldots, X_n)$ and $Z_i := g(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$. Assume that for all $X_1, \ldots, X_n, X_1', \ldots, X_n' \in \mathcal{X}^n$, and for all $i \in \{1, \ldots, n\}$,*

$$|Z - Z_i| \leq c_i.$$

Then, for any $x > 0$,

$$\mathbb{P}\left\{|Z - EZ| > x\right\} \leq 2\exp\left[\frac{-2x^2}{\sum_{i=1}^n c_i^2}\right]$$

One limitation of Hoeffding type inequalities, however, is that they ignore the information about the variance of the X_is. Note that in many cases the variance of Z might be mush smaller than $\sum_{i=1}^n c_i^2$. For this reason, sharper bounds such as Bennett's and Bernstein's inequalities have been derived which provide improvements over Hoeffding's inequality.

Theorem 3 (Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n$ be n independent bounded random variables*

such that for all $i = \{1, \ldots, n\}$, $\mathbb{E}[X_i] = 0$, *and* $Var[X_i] := \sigma^2$. *If* $Z := \sum_{i=1}^{n} X_i$, *and there exist a constant* $c > 0$ *such that* $X_i \leq c$, *then for any* $x > 0$, *we have*

$$\mathbb{P}\{Z - \mathbb{E}Z \geq \sigma\sqrt{2nx} + \frac{cx}{3}\} \leq e^{-x}$$

The functional version of Bernstein's inequality applicable to function classes has been also proposed which is known as Talagrand's inequality. In the following, we introduce Bousquet's version of Talagrand's inequality presented in [19].

Theorem 4 (Talagrand's Concentration Inequality)**.** *Assume that for function class* $\mathcal{F} := \{f : \mathcal{X} \to \mathbb{R}\}$, *it holds that* $\mathbb{E}f(X_i) = 0, \forall i$, *and* $\sup_{\substack{f \in \mathcal{F}, \\ X \in \mathcal{X}}} f(X) \leq 1$. *Let*

$$Z := \sup_{f \in \mathcal{F}} \sum_{i}^{n} f(X_i),$$

and $\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i) \leq \sigma^2$ *with the real number* $\sigma > 0$. *Then, for any* $x > 0$, *we have*

$$\mathbb{P}\{Z - \mathbb{E}Z \geq \sqrt{2x\nu} + \frac{x}{3}\} \leq e^{-x}$$

where $\nu := n\sigma^2 + 2\mathbb{E}Z$.

Now based on these McDiarmid's and Talagrand's inequalities, we can derive such inequalities for the suprema of empirical processes. For this purpose, one can easily define $Z := \sup_{f \in \mathcal{F}}(Pf - P_n f)$ for which we obtain the following results.

Corollary 5 (McDiarmid's Inequality for the Suprema of Empirical Processes)**.** *Let the function*

class $\mathcal{F}$ map $X \in \mathcal{X}$ into $[0,1]$. McDiarmid inequality then gives, with probability at least $1-e^{-x}$,

$$\sup_{f\in\mathcal{F}} Pf - P_n f \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] + \sqrt{\frac{2x}{n}}$$

Corollary 6 (Talagrand's Inequality for the Suprema of Empirical Processes)**.** *Let $\mathcal{F}$ be a class of functions mapping $X \in \mathcal{X}$ into $[0,1]$. Assume that r is a positive real value for which $Var\left[f(X_i)\right] \leq r$ for all $f \in \mathcal{F}$. Then, for every $x > 0$, with probability at least $1-e^{-x}$,*

$$\sup_{f\in\mathcal{F}} Pf - P_n f \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] + \sqrt{\frac{2xr}{n}} + \frac{4x}{3n}$$

As you can see, the term $\mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right]$ is one of the main components of the bound in both inequalities above. Thanks to the symmetrization technique, this term can be also bounded based on the fact that for any functions f in $\mathcal{F}$, the expected deviation of the empirical mean $P_n f$ from its true one Pf, can be controlled by the Rademacher complexity of the function class $\mathcal{F}$.

Lemma 1 (Symmetrization Technique, Lemma A.5 in [10])**.** *For any function class $\mathcal{F}$*

$$\max\left\{\mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right], \mathbb{E}\left[\sup_{f\in\mathcal{F}} P_n f - Pf\right]\right\} \leq 2\mathfrak{R}(\mathcal{F}).$$

$\mathfrak{R}(\mathcal{F})$ is the so-called *Rademacher complexity* of the function class $\mathcal{F}$ which will be discussed in the next section.

Rademacher Complexity

Rademacher complexity quantifies how well the functions in a hypothesis class $\mathcal{G}$ can correlate with random noise, and therefore it measures the richness of the hypothesis set $\mathcal{G}$. In the following,

we provide the definition and some useful property of the Rademacher complexity which will be used in the future chapters.

Definition 7 (Rademacher Complexity)**.** *Given a hypothesis class $\mathcal{G} := \{g : \mathcal{Z} \to \mathbb{R}\}$ and a set of data $S := \{z_1, \dots z_n\}$ which are drawn identically and independently according to distribution P, the Empirical Rademacher Complexity of $\mathcal{G}$ is defined as*

$$\hat{\mathfrak{R}}(\mathcal{G}) := \mathbb{E}_\sigma \left\{ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right\}$$

where σ_is are independent uniformly-distributed $\{\pm 1\}$-valued random variables. Also, the Rademacher Complexity of $\mathcal{G}$ is defined as the expectation of the empirical Rademacher complexity over all samples of size n drawn according to P:

$$\mathfrak{R}(\mathcal{G}) := \mathbb{E}_{S \sim P^n} \left[\hat{\mathfrak{R}}(\mathcal{G}) \right].$$

Intuitively, for a *fixed* set S and a *fixed* Rademacher vector $\boldsymbol{\sigma} := \{\sigma_1, \dots, \sigma_n\}$, the supremum measures the maximum correlation between $g(z_i)$ and σ_i over all functions $g \in \mathcal{G}$. Therefore, by taking the expectations over the random vector $\boldsymbol{\sigma}$, the empirical Rademacher complexity measures how well, on average, the functions $g \in \mathcal{G}$ can be correlated with *random* noise over the *fixed* sample set S. Also, Rademacher complexity measures the expected noise-fitting ability of $\mathcal{G}$ over any *random* sample set of size n drawn according to P^n.

The following Talagrand lemma provides a useful property for the Rademacher Complexity. Using this property, the Rademacher complexity of a hypothesis space $\mathcal{F}$ after composition with a Lipschitz function ϕ can be upper-bounded in terms of the Rademacher complexity of the hypothesis set $\mathcal{F}$.

Lemma 2 (Talagrand's Contraction property [10])**.** *Let ϕ be a Lipschitz function with constant L,*

that is, $|\phi(x) - \phi(y)| \leq L|x - y|$*. Then for every function class* $\mathcal{F}$ *there holds*

$$\mathbb{E}_\sigma \mathfrak{R}(\phi \circ \mathcal{F}) \leq L \mathbb{E}_\sigma \mathfrak{R}(\mathcal{F}), \tag{2.8}$$

where $\phi \circ \mathcal{F} := \{\phi \circ f : f \in \mathcal{F}\}$ *and* $\circ$ *is the composition operator.*

In the following section we introduce some fundamental theorems providing generalization bound based on the Rademacher Complexity.

Rademacher Complexity-based Generalization Bounds

The following theorem, which is based on McDiarmid's inequality, serves as a general tool for providing generalization bounds based on Rademacher complexity.

Theorem 8 (Rademacher complexity-based generalization bound for general function class $\mathcal{F}$, Theorem 3.1 in [112])**.** *Let* $\mathcal{G}$ *be a family of functions mapping from* $\mathcal{Z}$ *to* $[0, 1]$*. Assume that* $S := \{z_1, \ldots, z_n\}$ *is a set of* n *samples which are drawn identically and independently according to the probability distribution* D*. Then for any* $g \in \mathcal{G}$ *and* $x > 0$*, the following holds with probability at least* $1 - e^{-x}$*,*

$$\mathbb{E}\left[g(z)\right] \leq \frac{1}{n}\sum_{i=1}^{n} g(z_i) + 2\mathfrak{R}(\mathcal{G}) + \sqrt{\frac{x}{2n}},$$
$$\mathbb{E}\left[g(z)\right] \leq \frac{1}{n}\sum_{i=1}^{n} g(z_i) + 2\hat{\mathfrak{R}}(\mathcal{G}) + \sqrt{\frac{9x}{2n}}$$

where $\mathfrak{R}(\mathcal{G})$ *and* $\hat{\mathfrak{R}}(\mathcal{G})$ *are Rademchar and empirical Rademacher complexities of* $\mathcal{G}$*.*

Based on Theorem 8 and Lemma 2, now we can introduce the following Theorem which provides a Rademacher-based bound for binary classification problem.

Theorem 9 (Rademacher complexity-based generalization bound for binary classification, Theorem 3.2 in [112]). *Let $\mathcal{L}_{\mathcal{F}} := \{\ell_f : (X, Y) \to \ell(f(X), Y), f \in \mathcal{F}\}$ be a class of loss functions with ranges in $[0, 1]$. Assume that the function class $\mathcal{F}$ is a set of functions $f : \mathcal{X} \to \mathbb{R}$, and $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is a set of n samples distributed identically and independently according to P. Also let the loss function ℓ be an L-Lipschitz, and upper-bound the $0 - 1$ loss function. Fix $L > 0$, then for any $f \in \mathcal{F}$ and $x > 0$, with probability at least $1 - e^{-x}$,*

$$\mathbb{E}\left[\ell(f(X), Y)\right] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) + 2L\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{x}{2n}},$$
$$\mathbb{E}\left[\ell(f(X), Y)\right] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) + 2L\hat{\mathfrak{R}}(\mathcal{F}) + \sqrt{\frac{9x}{2n}}$$

where $\mathfrak{R}(\mathcal{F})$ and $\hat{\mathfrak{R}}(\mathcal{F})$ are Rademchar and empirical Rademacher complexities of $\mathcal{F}$.

Note that the first term in the right-hand side of the above inequalities is the empirical risk of function f, and the second term is a measure of complexity of hypothesis class $\mathcal{F}$. Based on this observation, it is not hard to see that these error bounds can be used to design complexity-regularization algorithms, similar to (2.3), for model selection. These type of algorithm are usually of interest, as they minimize the upper bound on the true risk $\mathbb{E}\left[\ell(f(X), Y)\right]$, hence a better generalization performance is expected by utilizing them as learning algorithms. Also, it can be seen that the best error rate that can be achieved using global Rademacher complexity is at least of the order of $\mathcal{O}(1/\sqrt{n})$.

The derivations of these bounds are based on the application of McDiarmid's inequality which is the functional version of Hoeffding's inequality, and it does not use any information regarding the variance of the functions. For this reason, another type of functional inequality, namely Talagrand's inequality, has been used in the derivation of generalization bounds. Talagrand's inequality is based on Bernstein's concentration inequality and yields sharper bounds by incorporating additional data

on the variances of the functions into the derivations. Talagrand's inequality was first established in [132] and later improved by [82, 100, 125, 19]. The following section presents a generalization bound based on Talagrand's inequality, which requires a new definition of Rademacher complexity, namely the *local* Rademacher complexity.

Local Rademacher Complexity-based Generalization Bounds

local Rademacher complexity refers to the Rademacher complexity of a subset of the function class $\mathcal{F}$ which is determined by a variance constraint on the functions in that class.

Definition 10 (Local Rademacher Complexity (LRC))**.** *Given a hypothesis class $\mathcal{G} := \{g : \mathcal{Z} \to \mathbb{R}\}$, and a set of data $S := \{z_1, \dots z_n\}$ which are drawn identically and independently according to distribution P, the empirical local Rademacher complexity of the function class $\mathcal{G}$ at radius r is defined as*

$$\hat{\mathfrak{R}}(\mathcal{G}, r) := \mathbb{E}_\sigma \left\{ \sup_{\substack{g \in \mathcal{G}, \\ Pg^2 \leq r}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right\}$$

where σ_i's are independent random variables uniformly chosen from $\{\pm 1\}$. Also, the local Rademacher Complexity of $\mathcal{G}$ is defined as the expectation of the empirical local Rademacher complexity over all samples of size n drawn according to P:

$$\mathfrak{R}(\mathcal{G}, r) := \mathbb{E}_{S \sim P^n} \left[\hat{\mathfrak{R}}(\mathcal{G}, r) \right].$$

The reason for the definition of local Rademacher complexity is based on the fact that by incorporating the variance constraint better error rate for the bounds can be obtained. In other words, the key point in deriving fast rate bounds is that around the best function f^* (the function that mini-

mizes the true risk), the variance of the deviation between the empirical and true errors of functions can be controlled by a linear function of the expectation of this difference. Based on this observation, instead of considering the Rademacher complexity of the entire class, we can consider the Rademacher complexity of a subset of the class which is usually the intersection of the class with a ball centered at the best function f^* in the class [10]. Note that local Rademacher complexity is always smaller than its corresponding global one, as it considers a smaller subset of the class.

Before presenting the localized version of error bounds, first we introduce some concepts and definition which are used later in this section and also in the future chapters for the derivation of local generalization bounds.

Definition 11 (Sub-Root Function)**.** *A function $\psi : [0, \infty] \to [0, \infty]$ is sub-root if*

1. *ψ is non-negative,*
2. *ψ is non-decreasing,*
3. *$r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for $r > 0$.*

The following lemma is an immediate consequence of the above definition.

Lemma 3 (Lemma 3.2 [10])**.** *Assume that ψ is a sub-root function. Then one can show that ψ is continuous on $[0, \infty]$, and the equation $\psi(r) = r$ has a unique (non-zero) solution which is known as the fixed point of ψ and it is denotes by r^*. Moreover, for any $r > 0$, it holds that $r > \psi(r)$ if and only if $r^* \leq r$.*

The following provides another useful definition that will be needed in introducing the main result of this section.

Definition 12 (Star-Hull). *The star-hull of a function class $\mathcal{F}$ around the function f_0 is given as*

$$star(\mathcal{F}, f_0) := \{f_0 + \alpha(f - f_0) : f \in \mathcal{F}, \alpha \in [0, 1]\}$$

Now, we present a lemma from [21] which indicates that the local Rademacher complexity of the star-hull of any function class $\mathcal{F}$ can be considered as a sub-root function, and it has a unique fixed point. We will see later that this fixed point plays a key role in the local error bounds.

Lemma 4 (Lemma 6 in [21]). *For any function class $\mathcal{F}$, the local Rademacher complexity of its start-hull is a sub-root function.*

Now, we can state the main results of this section as the following theorem which is a consequence of Talagrand's inequality.

Theorem 13 (Local Rademacher complexity-based generalization bound for general function class $\mathcal{F}$, Theorem 3.3 in [10]). *Let $\mathcal{F}$ be a class of functions satisfying $\sup_x |f(x)| \leq b$. Let $\{X_i\}_{i=1}^n$ be a sequence of n random variables which are independently and identically distributed according to P. Assume that there exist a constant B and a function $V : \mathcal{F} \to \mathbb{R}^+$ such that for every $f \in \mathcal{F}$, it holds that $Pf^2 \leq V(f) \leq BPf$, where $Pf^2 := \mathbb{E}_{X\sim P}[(f(X))^2]$ and similarly $Pf := \mathbb{E}_{X\sim P} f(X)$. Let ψ be a sub-root function with the fixed point r^*. Suppose that*

$$B\mathfrak{R}(\mathcal{F}, r) \leq \psi(r), \forall r \geq r^*,$$

where $\mathfrak{R}(\mathcal{F}, r)$ is the LRC of the function class $\mathcal{F}$ defined as

$$\mathfrak{R}(\mathcal{F}, r) := \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{f\in\mathcal{F}, \\ Pf^2\leq r}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right\}$$

Then for any $f \in \mathcal{F}$, $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$Pf \leq \frac{K}{K-1} P_n f + \frac{704K}{B} r^* + \frac{(22b + 26BK)x}{n}.$$

where $P_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.

Also, if $\mathcal{F}$ is a convex class of functions, and for any $\alpha \in [0, 1]$, $V(\alpha f) \leq \alpha^2 V(f)$, then for any $f \in \mathcal{F}$, $K > 1$ and $x > 0$ the following inequality holds with probability at least $1 - e^{-x}$,

$$Pf \leq \frac{K}{K-1} P_n f + \frac{6K}{B} r^* + \frac{(22b + 5BK)x}{n}.$$

Now, it can be shown (as we see later in Section 5 of Chapter 5), considering additional conditions on the data distribution P or on the hypothesis set $\mathcal{F}$, can make an improvement on the (excess) risk bounds in term of the convergence rate. These assumptions are presented in the following.

Assumption 14. *Consider the loss function ℓ and the function class $\mathcal{F}$ which satisfy the following conditions*

1. *For every probability distribution P, there exists a function $f \in \mathcal{F}$, which satisfies $P\ell_{f^*} = \inf_{f \in \mathcal{F}} P\ell_f$.*
2. *There is a constant $B > 1$, such that for every $f \in \mathcal{F}$, we have $P(f - f^*) \leq BP(\ell_f - \ell_{f^*})$.*
3. *There exists a constant L, such that the loss function $\ell(\hat{Y}, Y)$ is L-Lipschitz in its first argument, that is for any $Y, \hat{Y}_1, \hat{Y}_2$,*

$$|\ell(\hat{Y}_1, Y) - \ell(\hat{Y}_2, Y)| \leq L|\hat{Y}_1 - \hat{Y}_2|.$$

According to [10], these assumptions are not too restrictive, as they hold for several commonly

used regularized ERM algorithms. With the help of Assumption 14 and Theorem 13, the following theorem presents an excess risk bound of the function class $\mathcal{F}$.

Theorem 15 (Distribution-dependent local Rademacher complexity-based excess risk bound for binary classification, Corollary 5.3 in [10])**.** *Assume that $\mathcal{F}$ is a class of functions satisfying $\sup_x |f(x)| \leq 1$. Also, let $(X_i, Y_i)_{i=1}^n$ be a sequence of n independent random variables distributed according to P. Suppose that Assumption 14 holds. Define $\mathcal{F}^* := \{f - f^*\}$, where f^* is the function satisfying $P\ell_{f^*} = \inf_{f\in\mathcal{F}} P\ell_f$. Also, let $\hat{f} \in \mathcal{F}$ be such that $P_n\ell_{\hat{f}} = \inf_{f\in\mathcal{F}} P_n\ell_f$. Assume that ψ is a sub-root function with the fixed point r^* such that $BL\mathfrak{R}(\mathcal{F}^*, r) \leq \psi(r)$, $\forall r \geq r^*$, where $\mathfrak{R}(\mathcal{F}^*, r)$ is the LRC of the function class $\mathcal{F}^*$, and it is defined as*

$$\mathfrak{R}(\mathcal{F}^*, r) := \mathbb{E}_{X,\sigma}\Big[\sup_{\substack{f\in\mathcal{F},\\ P(f-f^*)\leq r}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\Big].$$

Then for any $f \in \mathcal{F}$, $K > 1$, $x > 0$ and $\psi(r) \leq r$, with probability at least $1 - e^{-x}$,

$$P(\ell_{\hat{f}} - \ell_{f^*}) \leq \frac{704K}{B} r^* + \frac{(11L + 26BK)x}{n}.$$

where $P(\ell_{\hat{f}} - \ell_{f^*}) := \mathbb{E}_{X\sim P}\big[\ell(\hat{f}(X), Y) - \ell(f^*(X), Y)\big]$.

Also, if $\mathcal{F}$ is a convex class of functions, then for any $f \in \mathcal{F}$, $K > 1$ and $x > 0$, the following inequality holds with probability at least $1 - e^{-x}$,

$$P(\ell_{\hat{f}} - \ell_{f^*}) \leq \frac{6K}{B} r^* + \frac{(11L + 5BK)x}{n}. \tag{2.9}$$

The result of Theorem 15 uses a distribution-dependent measure of complexity of the function class $\mathcal{F}$. In other words, the sub-root function ψ in Theorem 15 is bounded in terms of the Rademacher averages that cannot be computed without knowing the probability distribution P. The next the-

orem, analogous to Corollary 5.4 in [10], presents a data-dependent version of (5.9) replacing the Rademacher complexity in Theorem 15 with its empirical counterpart. Indeed, this error bound can be directly computed from the data, without having a priori information of the distribution.

Theorem 16 (Data-dependent local Rademacher complexity-based excess risk bound for binary classification, Corollary 5.4 in [10])**.** *Assume that $\mathcal{F}$ is a convex class of functions satisfying $\sup_x |f(x)| \leq 1$. Also, let $(X_i, Y_i)_{i=1}^n$ be a sequence of n independent random variables distributed according to P. Suppose that Assumption 14 holds. Also, let $\hat{f} \in \mathcal{F}$ be such that $P_n \ell_{\hat{f}} = \inf_{f \in \mathcal{F}} P_n \ell_f$. Assume that $\hat{\psi}_n$ is a sub-root function with the fixed point $\hat{r}^*$. Define*

$$\hat{\psi}_n(r) := c_1 \hat{\mathfrak{R}}(\mathcal{F}^*, c_3 r) + \frac{c_2 x}{n}, \quad \hat{\mathfrak{R}}(\mathcal{F}^*, c_3 r) := \mathbb{E}_\sigma \Big[\sup_{\substack{f \in \mathcal{F}, \\ L^2 P_n (f - \hat{f})^2 \leq c_3 r}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \Big],$$

where $c_1 := 2L(B \vee 10L)$, $c_2 := 11L^2 + c_1$ and $c_3 := 2824 + 4B(11L + 27B)/c_2$. Then, for any $f \in \mathcal{F}$, $K > 1$ and $x > 0$, with probability at least $1 - 4e^{-x}$,

$$P(\ell_{\hat{f}} - \ell_{f^*}) \leq \frac{705K}{B} \hat{r}^* + \frac{(11L + 27BK)x}{n}. \tag{2.10}$$

It is worth mentioning that, the fact that data-dependent bounds of the form (2.10) can be computed from the data, does not imply that they are easy to compute. However, for several cases including binary classification and kernel classes they can be efficiently computed (see Section 6 of [10]). Another interesting application of data-dependent bounds can be considered in devising efficient learning algorithms using some criteria derived from these bounds, which potentially can lead to more accurate models.

Multi-Task Learning

MTL is a learning framework in which several multiple related task are jointly learned with the hope of achieving a better generalization performance compared to learning each task independently. As the key assumption in MTL is based on task relatedness, several MTL models have been proposed taking different approaches in capturing and modeling task's relations. However, the common feature of all these models is that they formulate the MTL problem using a regularized ERM framework in which the objective function is a composition of an over-the-tasks average loss function and a regularization term to encourage some sort of information sharing among tasks. More specifically, given T multiple learning tasks, each of which presented by a training set $\{(x_t^n, y_t^n)\}_{n=1}^{n_t}, t \in \mathbb{N}_T$, which is drawn from an unknown distribution $P_t(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, the objective is to learn a discriminative functions $f_t : \mathcal{X} \to \mathcal{Y}$ for each task $t \in \{1, \ldots, T\}$. Here $\mathcal{X}$ denotes the native space of samples for all tasks and $\mathcal{Y}$ is the corresponding output space. In general, assuming a linear model the predictions functions are given in the form of $f_t(\boldsymbol{x}) = \langle \boldsymbol{w}_t, \boldsymbol{x} \rangle + b_t, \forall t \in \{1, \ldots, T\}$. The extension of these to kernel-based models have been also proposed in the context of MTL [46] which allow more flexibility and can achieve better generalization performance. A kernel based MTL model usually considers the linear model $f_t(\boldsymbol{x}) := \langle \boldsymbol{w}_t, \boldsymbol{\phi}_t(\boldsymbol{x}) \rangle_{\mathcal{H}_t} + b_t$ for $t \in \{1, \ldots, T\}$, where $\boldsymbol{w}_t$ is the weight vector related to task t. Furthermore, the feature space $\mathcal{H}_t$, associated to task t, is induced with the feature mapping ϕ_t associated with the reproducing kernel function $k_t(x_t^i, x_t^j)$ for all $x_t^i, x_t^j \in \mathcal{X}$. The goal is then to learn the $\boldsymbol{w}_t$'s and b_t's jointly via the following regularized risk minimization problem:

$$\min_{\boldsymbol{f}} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \ell(f_t(x_t^i), y_t^i) + \lambda \Omega(\boldsymbol{f}) \tag{2.11}$$

where $\boldsymbol{f} := (f_1, \ldots, f_T)$ is a vector-valued function parametrized by $\boldsymbol{W} := (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T)$ and $\boldsymbol{b} := (b_1, \ldots, b_T)$. Also, $\Omega(\boldsymbol{f})$ is the *so-called* regularization term which is designed to en-

force some information sharing among tasks, $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell(f_t(x_t^i), y_t^i)$ is the averaged empirical loss over all tasks, and λ is the regularization parameter. This framework—known as *regularized MTL*—has be extensively employed in MTL literature among which, we can refer to [47, 155, 158, 129, 103, 102], just to name a few.

Interestingly, using the following proposition, it can be shown that Problem 2.11 can be converted to an equivalent optimization problem.

Proposition 17. *(Proposition 12 in [73], part (a)) Let $f, g : \mathcal{C} \mapsto \mathbb{R}$ be two functions with $\mathcal{C} \subseteq \mathcal{X}$. For any $\nu > 0$, there is a $\eta > 0$, such that the optimal solution of (2.12) is also optimal in (2.13)*

$$\min_{x\in\mathcal{C}} f(x) + \nu g(x) \tag{2.12}$$

$$\min_{x\in\mathcal{C}, g(x)\leq\eta} f(x) \tag{2.13}$$

Using Proposition 17, it is not hard to verify that Problem 2.11 is equivalent to

$$\begin{aligned}&\min_{\boldsymbol{f}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell(f_t(x_t^i), y_t^i)\\ &\text{s.t. } \Omega(\boldsymbol{f}) \leq R\end{aligned}$$

In other words, this learning problem seeks the vector-valued function $\boldsymbol{f} = (f_1, \ldots, f_T)$ from the hypothesis space $\mathcal{F} := \{\boldsymbol{f} = (f_1, \ldots, f_T) : \Omega(\boldsymbol{f}) \leq R\}$. Once the hypothesis space of a MTL model is given, one might be able to derive generalization bounds for the learning problem at hand. As mentioned earlier, generalization bounds are considered as important tools in understanding the performance of a learning model and they can reveal the potential capability of the model in a learning task. For this reason, many MTL efforts have been concentrated on this problem since it was first studied in [11] and later have been notably pursued in [102, 69, 105, 104, 103]. A seminal

work in this direction—presented in [101]—derived a MTL generalization bounds based on global Rademacher complexities. We introduce this result in the following theorem.

Theorem 18 (McDiarmiad-Type Inequality for MTL)**.** *Let $\mathcal{F}$ be a class of vector-valued functions $\{\boldsymbol{f} = (f_1, \ldots, f_T) : \mathcal{X} \mapsto \mathbb{R}^T\}$, that maps $\mathcal{X}$ into $[-b, b]^T$, and let $(X_t^i, Y_t^i)_{(i,t)=(1,1)}^{(n,T)}$ be a vector of nT independent random variables where for all fixed t, $(X_t^1, Y_t^1), \ldots, (X_t^n, Y_t^n)$ are identically distributed according to P_t. Also, assume $\ell : \mathbb{R} \to [0, 1]$ be an L Lipschitz loss function domination the $0-1$-loss function $\mathbf{1}_{(\infty,0]}(.)$. Let $\{\sigma_t^i\}_{t,i}$ be a sequence of independent Rademacher varietals. Then, for every $x > 0$ with probability at least $1 - e^{-x}$, the followings hold for any $\boldsymbol{f} \in \mathcal{F}$*

$$P\ell_{\boldsymbol{f}} \leq P_n\ell_{\boldsymbol{f}} + 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{Lbx}{nT}},$$

and,

$$P\ell_{\boldsymbol{f}} \leq P_n\ell_{\boldsymbol{f}} + 2\hat{\mathfrak{R}}(\mathcal{F}) + \sqrt{\frac{9Lbx}{nT}}$$

where the MTL Rademacher complexity $\mathfrak{R}(\mathcal{F})$, and its empirical counterpart $\hat{\mathfrak{R}}(\mathcal{F})$ are defined as following

$$\mathfrak{R}(\mathcal{F}) := \mathbb{E}_{X\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right\}, \qquad \hat{\mathfrak{R}}(\mathcal{F}) := \mathbb{E}_{\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right\}.$$

Also, the true error $P\ell_{\boldsymbol{f}}$ and the corresponding empirical loss $P_n\ell_{\boldsymbol{f}}$ are given as

$$P\ell_{\boldsymbol{f}} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(X_t,Y_t)\sim P_t}\ell(f_t(X_t), Y_t), \qquad P_n\ell_{\boldsymbol{f}} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell(f_t(X_t^i), Y_t^i).$$

The results of this theorem can be easily proved based on Theorem 16 and 17 in [101] which is based on McDiarmid inequality. It is worth pointing out that thus far, the best convergence

rate of the error bounds achieved by global Rademacher analysis—whether distribution- or data-dependent—in the context of MTL is of the order of $O(1/\sqrt{nT})$. One thing that can be concluded from these type of bounds is that, as the number of tasks T grows, the number of sample n per task can be significantly increased. This indicates that MTL can be specifically advantageous when the tasks lack a sufficient body of observed data. In recent years, these bounds and more specifically the data-dependent version of them are efficiently used to design MTL models with better generalization performance.

However, recall that McDiarmid-type inequities do not make use of the variances of the functions and consequently, the bounds derived based on global analysis (GRCs) ignore the fact that learning algorithms typically choose well-performing hypotheses that belong only to a subset of the entire hypothesis space under consideration. This motivated us to apply Talagrand's concentration inequality to derive generalization bounds for MTL that leads to a local analysis and the derivation of local Rademacher complexities. This local analysis provides less conservative and, hence, sharper bounds than when a global analysis is employed. To date, there have been only a few additional works attempting to reap the benefits of such local analysis in various contexts: active learning for binary classification tasks [75], multiple kernel learning [71, 33], transductive learning [136], semi-supervised learning [114] and bounds on the LRCs via covering numbers [84]. To the best of our knowledge, there has been no study that uses the notion of local Rademacher complexity in the context of MTL. Therefore, as a main contribution of this dissertation, we first derive sharp excess risk bounds for MTL in terms of distribution- and data-dependent LRC. Then, we propose a new MTL formulation using the criteria derived based on our data-dependent bound. At the end, we propose an efficient algorithm to solve our new MTL problem.

CHAPTER 3: LITERATURE REVIEW

Common traditional machine learning problems are often formulated as single task learning. These models, by definition, use only one task at a time, that is, they use previously collected labeled or unlabeled training data from one task to make future predictions about data of the same task. MTL, in contrast, is a machine learning paradigm that addresses the problem of jointly learning multiple related tasks together with the aim of improving the generalization performance of all tasks, by allowing them to share information among themselves. MTL has attracted a lot of attention over the past years since it was first introduced in [23]. The need of MTL may arise when only a few samples are available for the tasks. This is typically the case in many real world applications where gathering sufficient amount of training data for a reasonable prediction is expensive or even impossible. To illustrate the utility of MTL, it is worthwhile to compare it with human learning. People can intelligently speed up their learning process of new things while they apply knowledge and experience learned in the past. MTL studies can be categorized into experimental and theoretical studies; such studies are presented in the following sections.

Non-Theoretical Multi-Task Learning Studies

One of the most important and challenging problem in MTL is the assessment of tasks relatedness, which has motivated two main research questions in this field: "How can the information contained in multiple tasks be captured and incorporated into a MTL framework?" and "How can the knowledge be transferred between the tasks with different levels of similarities?". In response to these questions various novel MTL studies have been performed which we categorize in what follows.

How Does Information Sharing Occur among Tasks?

This line of research considers designing approaches that capture tasks similarities and incorporate them into a learning process. Studies in this area can be mainly categorized as outlined next:

Shared Features Learning

A commonly used approach in MTL is based on the idea that a common low-dimensional representation is shared across multiple related tasks. Inspired by this assumption, different methods have been proposed which aim at finding a good (shared) feature representation that captures tasks' underlying commonalities. It is worth pointing out that a typical approach is to employ the group Lasso-type regularizer on the tasks' weight matrix $\boldsymbol{W}$:

$$R(\boldsymbol{W}) := \|\boldsymbol{W}\|_{2,1} := \sum_{d=1}^{D} \|\boldsymbol{w}^d\|_2,$$

where $\boldsymbol{w}^d$ is the d-th row os the matrix $\boldsymbol{W}$. For example, in [5], a sparse representation shared across multiple tasks is learned by an $\ell_{2,1}$ regularizer which controls the number of learned features shared among tasks by imposing an ℓ_1-norm on the rows level of $\boldsymbol{W}$. Then an equivalent convex optimization problem is developed which jointly learns both the task functions and the features through two alternating steps. Their formulation is interestingly equivalent to the approach of employing a trace norm as a regularizer in [4, 67, 124, 43, 119] for MTL. Different from exixting trace norm-based MTL approaches, the authors in [58] proposed to use a capped trace norm regularizer to penalize only the singular values smaller than some threshold. Another study in [89] addresses the problem of joint feature selection across multiple tasks using an $\ell_{2,1}$-norm requlrizer which results to a convex , non-smooth optimization problem. They propose to reformulate

the problem as its equivalent smooth convex problem and use a Nesterov's algorithm to solve it. A more general MTL feature section model in [154] studies the problem of determining the most appropriate sparsity-enforcing norm by considering a family of $\ell_{1,q}$ norms for $1 \leq q \leq \infty$. They also provide a probabilistic interpretation of this general framework based on which they develop a probabilistic model using the non-informative Jeffreys prior. An expectation-maximization algorithm is then proposed to learn the models parameters including q. Similar works considering the sparsity-inducing norm requlrarizer on matrix $\boldsymbol{W}$ for common feature selection across tasks, include [92, 87, 146, 113, 147, 26].

Another widely utilized approach in MTL literature is to capture tasks' common features using a combination of two norm regularizers. In [64], the authors proposed a combination of $\ell_{\infty,1}$-norm along with a $\ell_{1,1}$-norm in the from of the following regularizer

$$R(\boldsymbol{W}) := \lambda_1 \|\boldsymbol{W}\|_{\infty,1} + \lambda_2 \|\boldsymbol{W}\|_{1,1}$$

where the first norm encourages sparsity at the row level of $\boldsymbol{W}$ for feature learning, and the second norm is added for element-wise sparsity in $\boldsymbol{W}$. In a similar attempt, the authors in [145] developed an online learning framework for Multi-Task (MT) feature selection by introducing norm regularizers $\ell_{1,1}$ and $\ell_{2,1}$. A similar study is proposed in [120] in which they used the combination of norm regularizers $\ell_{1,1}$ and $\ell_{2,1}$ for a regression problem, and they employed an iteratively reweighted least square algorithm to handle the optimization problem. Similarly, in [55], they proposed a MT Calibrated Multivariate Regression (CMR) model using a combination of $\ell_{1,2}$-norm and Frobenious norm in a form of a regularizer. In [55], unlike one similar prior study in [88], the authors showed that the dual problem of their proposed formulation is smooth which enables developing a fast optimization algorithm to solve the problem.

Task Relationship Learning

Another MTL paradigm assumes that closely related tasks should also share some parameters or prior distributions of hyper-parameters among each others. Several approaches have pursued shared parameter learning.

In [47], the authors proposed a regularized MTL framework in which all task vectors $\boldsymbol{w}_t$s are assumed to be the sum of two terms. In particular, for each task t, they assume that $\boldsymbol{w}_t = \boldsymbol{w}_0 + \boldsymbol{v}_t$, where $\boldsymbol{w}_0$ corresponds to the function common to all task, and $\boldsymbol{v}_t$ is designed to capture characteristics specific to each task. Then, they utilize the following regularization terms and estimate the common parameter $\boldsymbol{w}_0$ as well as the task-specific parameters $\boldsymbol{v}_t$'s, simultaneously.

$$\frac{\lambda_1}{T} \sum_{t=1}^{T} \|\boldsymbol{v}_t\|^2 + \lambda_2 \|\boldsymbol{w}_0\|^2 \tag{3.1}$$

They also showed that this regularization is equivalent to considering the closeness of a task's model parameters $\boldsymbol{w}_t$ to the average of these model parameters, $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t$, and this can be modeled by letting the regularizer to be

$$\rho_1 \sum_{t=1}^{T} \|\boldsymbol{w}_t\|^2 + \rho_2 \sum_{t=1}^{T} \|\boldsymbol{w}_t - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t\|^2 \tag{3.2}$$

Similar to [47], the authors in [155] proposed a MTL formulation which (beside the task parameters $\boldsymbol{w}_0$ and $\boldsymbol{v}_t$) finds an appropriate linear feature mapping $\phi_t(X) = \phi_t X$, which maps the tasks into a k-dimensional latent feature space where the learned task's hypotheses are similar. For this purpose, they considered regularizing the mapping function's complexity by adding the Frobenius norm penalty $\|\phi_t\|_F^2$ to (3.1) and optimizing it along with $\|\boldsymbol{w}_0\|^2$ and $\|\boldsymbol{v}_t\|^2$. Based on this idea,

they introduced the regularization term

$$\frac{\lambda_1}{T}\sum_{t=1}^{T}\|\boldsymbol{v}_t\|^2+\lambda_2\|\boldsymbol{w}_0\|^2+\frac{\lambda_3}{T}\sum_{t=1}^{T}\|\phi_t\|_F^2$$

where ϕ_t is a $k\times d$ matrix. In another effort similar to [47], the authors in [130] employed a method called Relative Margin Machine (RMM) to formulate a MTL problem which takes into account the tasks relatedness as well as the spread of the data. The idea behind RMM is to maximize the relative margin (rather than the absolute (classic) margin) of the data from a separating hyperplane to handle the presence of arbitrary affine transformations or data drifts in particular directions in the feature space.

Also, in [91], a weighted decomposition of $\boldsymbol{w}_t$ has been considered in the form of $\boldsymbol{w}_t=\boldsymbol{w}_0+\alpha_t\boldsymbol{v}_t$, where α_t (the weight of task t's bias) represents the divergence of task t from other tasks, and is learned along with $\boldsymbol{w}_0$ and $\boldsymbol{v}_t$ during the optimization process.

Instead of the decomposition on task's weight vectors $\boldsymbol{w}_t$'s, the authors in [2] assumed that for each task t, the linear predictor function $f_t(x)$ can be decomposed into two predictor functions as follows

$$f_t(x)=\boldsymbol{w}_t^T\boldsymbol{\phi}_t(x)+\boldsymbol{v}_t^T\boldsymbol{\Theta}\boldsymbol{\psi}_t(x)$$

where $\boldsymbol{\phi}$ is a known high-dimensional feature map and $\boldsymbol{\psi}$ corresponds to a low-dimensional feature map parameterized by an unknown matrix $\boldsymbol{\Theta}$. $\boldsymbol{w}$ and $\boldsymbol{v}$ are the weight vectors specific to each prediction problem and $\boldsymbol{\Theta}$ is designed to capture the common structure shared by all tasks. Their formulation results in a non-convex optimization problem for semi-supervised learning. They also investigated the computational complexity of the their proposed algorithm in terms of covering numbers. Following the approach in [2], the authors in [26] considered the problem of learning a shared structure from multiple related tasks using an improved alternating structure optimization.

They also showed that their non-convex formulation can be converted into a relaxed convex formulation, and they presented a theoretical condition under which the convex formulation finds a globally optimal solution to its non-convex counterpart.

In [158], a time series problem has been formulated as a MT regression problem in which the task (prediction at each time point) relatedness is captured through the regulirizer

$$\sum_{t=1}^{T-1} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2 \tag{3.3}$$

which ensures the smoothness between two regression models at successive time points.

Unlike the previous studies which make some prior assumptions about task relationships, there is another approach in which the task relationships are learned automatically during the training stage. One of the first efforts in this line of research is done by [46] in which the following regularizer is proposed to learn the task relation

$$\lambda_1 \sum_{t=1}^{T} \|\boldsymbol{w}_t\|^2 + \sum_{t=1}^{T} \sum_{s=1}^{T} \boldsymbol{A}_{st} \|\boldsymbol{w}_t - \boldsymbol{w}_s\|^2 \tag{3.4}$$

where the graph adjacency matrix $\boldsymbol{A} = \boldsymbol{A}_{st}$ represents the task similarities. Graph-based regularized MTL framework has been also studied in [50, 49, 144, 59, 129, 29, 30, 3].

In another attempt to explore task relationships, the model proposed in [150] formulates a MTL problem by introducing the regularizer

$$\frac{\lambda_1}{2}\text{tr}(\boldsymbol{W}\boldsymbol{W}^T) + \frac{\lambda_2}{2}\text{tr}(\boldsymbol{W}\boldsymbol{\Omega}^{-1}\boldsymbol{W}^T) \tag{3.5}$$

where the matrix $\boldsymbol{\Omega}$ represents the relationships among tasks and is is learned during the training phase. Similar approaches to this have also been studied in [25, 152, 153, 151]. Moreover, in

a slightly different approach [48], the first regularization term in (3.5) is replaced with the ℓ_1-norm $\lambda_1\|\boldsymbol{W}\|_1$ to enforce a sparse common subset of features among tasks. Moreover, the trace regularization term $\frac{\lambda_3}{2}\text{tr}(\boldsymbol{W}^T\boldsymbol{L}\boldsymbol{W})$ has been added to (3.5) to impose smoothness across features. Similar to [48], a MTL formulation is introduced in [53] which switches the regularization term $\frac{\lambda_3}{2}\text{tr}(\boldsymbol{W}^T\boldsymbol{L}\boldsymbol{W})$ with the sparse inducing regularization penalty $\lambda_3\|\boldsymbol{\Omega}^{-1}\|_1$ to impose a common sparse structure shared among tasks.

How Can the Knowledge Be Transferred Between the Tasks with Different Levels of Similarities?

One limitation of the previous methods is that they consider similar or equal contributions of all tasks to the joint learning process. However, this assumption might be easily violated in the existence of "outlier" tasks, which commonly occurs in many practical applications. Therefore, a major challenge in MTL is to enable tasks to selectively share information with only related tasks. To be more concrete, some tasks can have full, partial or no overlap with other tasks. In this case sharing information between unrelated tasks might be detrimental, in the sense that it may deteriorate each task's generalization performance. Inspired by this observation, another influential line of research in MTL considers designing models wherein transferring knowledge among tasks can benefit all tasks with different level of similarities. Equivalently, these approaches are interested in a situation when transferring information might not be beneficial to all tasks, or in the worst case it might even hurt the learning performance of a subset of tasks (or even all tasks). This situation, which is often referred to as negative transfer, attracted a lot of attention in the past decade.

Several models have been introduced that address this issue by exploiting the latent relationship among tasks using different approaches. For example, some methods [8, 143, 150, 153], utilize a probabilistic framework, where transferring information is based on a common prior among tasks.

These approaches are usually computationally expensive.

It is worth mentioning that Clustered Multi-Task Learning (CMTL) corresponds to another family of approaches wherein tasks can be clustered into several groups. Each group contains related tasks in terms of some notion of similarity. Based on the current literature, clustering strategies can be broadly categorized into two classes: task-level CMTL and feature-level CMTL.

Task-Level Clustered Multi-Task Learning

In task-level CMTL, it is assumed that the parameters of all tasks' models within each group are close to each other. For example, the approach in [46] assumes that the weight vectors of the tasks assigned to the same group are close to each other. More precisely, they considered a task clustering approach through the regularizer

$$\lambda_1 \sum_{k=1}^{c} \sum_{t=1}^{T} \rho_t^k \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_k\|^2 + \lambda_2 \|\bar{\boldsymbol{w}}_k\|^2$$

where c is the number of clusters and $\bar{\boldsymbol{w}}_k$ is the average parameter of the k-th group. This work has been later extended in [63] wherein the authors proposed the regularizer

$$\lambda_1 n \|\bar{\boldsymbol{w}}\|^2 + \lambda_2 \sum_{k=1}^{c} m_k \|\bar{\boldsymbol{w}}_k - \bar{\boldsymbol{w}}\|^2 + \lambda_3 \sum_{k=1}^{c} \sum_{t \in \mathcal{J}(k)} \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}_k\|^2$$

where the first term corresponds to the global penalty and measures how large the average weight vectors are, the second term quantifies the closeness of different clusters to each other, and the last term is a measure of within-cluster variance (quantifying the compactness of the clusters). Note that, these models are restricted in the sense that: (i) they are designed based on an unrealistic assumption, as similarity between tasks' models does not necessarily implies that a meaningful information sharing can happen between tasks, and (ii) for these methods, the group structure

(number of groups or basis tasks) is needed to be known a priori. A more flexible clustering approach has been proposed in [159] in which some representative tasks are identified and utilized to cluster all other tasks. Their proposed approach uses the regularizer

$$\lambda_1 \|\boldsymbol{W}\|_F^2 + \frac{\lambda_2}{2} \sum_{t=1}^{T} \sum_{s=1}^{T} \boldsymbol{Z}_{ts} \|\boldsymbol{w}_t - \boldsymbol{w}_s\|_2^2 + \frac{\lambda_3}{2} \|\boldsymbol{Z}\|_{0,q} \tag{3.6}$$

along with the constraints $\mathbf{0} \preceq \text{vec}\boldsymbol{Z} \preceq \mathbf{1}_{TT}$, and $\boldsymbol{Z}^T \mathbf{1}_T = \mathbf{1}_T$, where $\boldsymbol{Z} \in \mathbb{R}^{T \times T}$ is a matrix whose entries indicate the assignment of the tasks to the representative tasks. More specifically, $\boldsymbol{Z}_{ts}$ is a value between 0 and 1 which quantifies the probability that task t selects task s as its representative task. Also, letting $\mathcal{I}(x)$ be an indicator function (whose function value is one if $x \neq 0$ and is zero otherwise), the norm penalty $\|\boldsymbol{Z}\|_{0,q} := \sum_{t=1}^{T} \mathcal{I}\left(\|\boldsymbol{Z}(t,:)\|_q\right)$ determines the number of representative tasks by calculating the number of rows in $\boldsymbol{Z}$ whose ℓ_q-norm in non-zero. This approach is considered more flexible in the sense that it allows tasks to be assigned to multiple clusters and also it does not require a priori knowledge regarding the number of clusters.

Feature-Level Clustered Multi-Task Learning

The other clustering strategy in MTL is known as feature-level CMTL, which models task relatedness by learning shared features among the tasks within each group.

As an example, a decomposition technique has been used in [128] where each task parameter $\boldsymbol{w}_t$ is decomposed into two components $\boldsymbol{w}_t = \boldsymbol{c}_t + \boldsymbol{s}_t$, which correspond to the shared and task-specific features, respectively. They considered the regularizer

$$\lambda_c \|\boldsymbol{C}\|_* + \lambda_s \|\boldsymbol{S}\|_1 \tag{3.7}$$

where $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_T]$ and $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T]$, and the nuclear norm $\|.\|_*$ encourages a low

rank solution on matrix $\boldsymbol{C}$ for coupling the related tasks, and the ℓ_1-norm penalty characterizes the specific features by imposing the sparsity on matrix $\boldsymbol{S}$. Then the authors used an accelerated proximal gradient algorithm to solve the non-smooth optimization problem. A similar formulation has been proposed in [27], where instead of the ℓ_1-norm, an $\ell_{1,2}$-norm is employed to induce a group-sparse structure on the column level of matrix $\boldsymbol{S}$ for identifying the outlier tasks. They adopt an accelerated proximal method for solving their non-smooth optimization problem. A similar framework has been introduced in [28] which utilizes an ℓ_0-norm (the number of non-zero entries) on matrix $\boldsymbol{C}$ to capture the sparse discriminative features for each task along with a matrix rank constraint on matrix $\boldsymbol{S}$ to encourage a shared low-rank structure among multiple tasks. Then they applied an alternative convex relaxation technique to solve their non-convex optimization problem. This approach has been also followed in [54] by employing the $\ell_{1,2}$-norm on both matrices $\boldsymbol{C}$ and $\boldsymbol{S}$ to capture the shared features among tasks and discover the outlier tasks, respectively. Also, the combination of a $\ell_{1,2}$-norm on matrix $\boldsymbol{C}$ and a trace norm on matrix $\boldsymbol{S}$ has been proposed in [27] to identify the outlier tasks by imposing a group-sparse structure on matrix $\boldsymbol{C}$, and also to couple the related tasks by inducing a low-rank structure on matrix $\boldsymbol{S}$. Another similar approach, introduced in [141], also considers the decomposition scheme $\boldsymbol{W} = \boldsymbol{C} + \boldsymbol{S}$, where $\boldsymbol{C}$ reflects global similarities among tasks, and $\boldsymbol{S}$ captures the task-feature relationships. In this study, they considered the following regularization term

$$\lambda_1 \sum_{t=1}^{T} \|\boldsymbol{c}_t - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{c}_t\|_2^2 + \lambda_2 \|\boldsymbol{S}\|_K^2 \tag{3.8}$$

where, for some non-negative integer $k < \min(d, T)$ (corresponding to the number of clusters), $\|\boldsymbol{S}\|_K^2 := \sum_{i=k+1}^{\min(d,T)} \sigma_i^2(\boldsymbol{S})$, and $\sigma_1(\boldsymbol{S}) \geq \sigma_2(\boldsymbol{S}) \geq \ldots \geq \sigma_{\min(d,T)}(\boldsymbol{S})$ are the singular values of matrix $\boldsymbol{S}$. By including the $\|.\|_K$-norm regularizer, they assumed that "two tasks can be related only on a subset of features" [141]. This approach results in a non-convex formulation as well as a co-clustering structure capturing task-feature relationships.

Similarly, a feature-level clustering model has been introduced in [156], in which a shared feature representation for all tasks is sought through decomposing each task parameter into two parts: one responsible to capture the shared structure between tasks and the other to model the variations specific to each task. Based on the interactions among tasks and features, this model constructs different task clusters for different features through the regularization term

$$\lambda_1 \|\boldsymbol{C}\|_{clus} + \lambda_2 \|\boldsymbol{C}\|_F^2 + \|\boldsymbol{S}\|_F^2 \tag{3.9}$$

where for each feature d and each pair $(\boldsymbol{c}_t, \boldsymbol{c}_s)$, the penalty norm $\|\boldsymbol{C}\|_{clus} = \sum_{d=1}^{D} \sum_{t<s} |C_{dt} - C_{ds}|$ encourages C_{dt}, C_{ds} to be close to each other, leading to feature-specific task clusters. The study in [57] then extends this model by proposing a multi-level structure, which learns task groups in the context of MTL.

In another outlier detection strategy [76], the existence of k latent basis tasks is assumed and each task is presented as a linear combination of these basis tasks, that is $\boldsymbol{w}_t = \boldsymbol{L}\boldsymbol{s}_t$ or equivalently $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{S}$ where the columns of matrix $\boldsymbol{L} \in \mathbb{R}^{d \times k}$ represent the latent tasks, and matrix $\boldsymbol{S} \in \mathbb{R}^{k \times T}$ contains the weights of linear combination for each task. Then, they used the sparsity inducing ℓ_1-norm to enforce each task to be presented by only a few number of the latent tasks. Also, a Frobenius norm on matrix $\boldsymbol{L}$ is considered to regularize the predictor weights to avoid over-fitting.

Finally as a different strategy, in [70] the tasks are clustered into G different groups. For each group g, the group assignment matrix $\boldsymbol{Q}_g \in \mathbb{R}^{T \times T}$ is learned along with the model parameter $\boldsymbol{W}$ during the training stage. The regularization term is expressed as

$$\lambda \sum_{g}^{G} \text{tr}\left[\boldsymbol{W}\boldsymbol{Q}_g \left(\boldsymbol{W}\boldsymbol{Q}_g\right)^T\right]^{1/2} \tag{3.10}$$

with the constraint $\sum_g^G \boldsymbol{Q}_g = \boldsymbol{I}$, which ensures that each task belongs to only one group. In

this model, the tasks from different groups are learned independently; however, a shared feature representation is learned for the tasks within the same group. The formulation results in a a non-convex optimization problem. An alternating algorithm is then utilized to solve the problem, which converges to local optima, and suffers potentially from slow convergence. Interestingly, it can be shown [157] that an equivalent relationship exists between CMTL and alternating structure optimization, wherein the goal is to find a low-dimensional structure shared by all tasks.

One limitation of these methods is that they do not allow information sharing between the tasks from different groups. This might be restrictive in the sense that tasks in disjoint groups could still be inter-related, albeit weakly. Hence, assigning tasks into different groups may not take full advantage of MTL.

Theoretical Multi-Task Learning Studies

Most of the studies discussed above are only concerned with designing models that can capture the task relatedness in a meaningful way. However, one main concern regarding any machine learning problem (including MTL) is that if one can provide some generalization guarantee for the learning problem at hand. Despite the considerable success and application of MTL to different problems, only a few number of studies investigated the theoretical aspects and benefits of MTL since it was first advocated in [11]. However the generalization bounds derived in this work are based on the notion of VC dimension and depend on a complexity measure which can not be easily inferred given a particular setting and is often too loose to be of any use in practice. Therefore, more intuitive complexity measures such as Rademacher complexity, have been considered alternatively in modern MTL theoretical research. More specifically, Rademacher complexities can provide tighter generalization error bounds by incorporating the data distribution and learning samples in computing the complexity of the learning space. It is worth pointing out that Rademacher bounds

are always at least as good as the VC dimension bounds [68], which explains the popularity of the Rademacher bounds in recent works.

Among the latest approaches, investigating MTL generalization guarantees using Rademacher averages, the study in [101] considers linear MTL frameworks for binary classification. In these framework, data from all tasks are pre-processed by a common bounded linear operator. Moreover, operator norm constraints are used to control the complexity of the associated hypothesis spaces. Note that both distribution- and data-dependent error bounds derived based on GRC, the convergence rate is of order $O(1/\sqrt{nT})$. Another study, [102], provides bounds for the empirical and expected Rademacher complexities of linear transformation classes. Based on Hölder's inequality, GRC-based risk bounds of order $O(1/\sqrt{nT})$ are established for MTL hypothesis spaces with graph-based and L_{S_q}-Schatten norm regularizers, where $q \in \{2\} \cup [4, \infty]$.

The subject of MTL generalization guarantees experienced renewed attention in recent years. In [69], the authors take advantage of the strongly-convex nature of certain matrix-norm regularizers to easily obtain generalization bounds for a variety of machine learning problems. Part of their work is devoted to the realm of online and off-line MTL. In the latter case, which pertains to the focus of our work, the paper provides a distribution-dependent GRC-based excess risk bound of order $O(1/\sqrt{nT})$. Moreover, [104] presents a global Rademacher complexity analysis leading to both data and distribution-dependent excess risk bounds of order $O(\sqrt{\log(nT)/nT}$ for a trace norm regularized MTL model. Also, [103] examines the bounding of (global) Gaussian complexities of function classes that result from considering composite maps, as it is typical in MTL among other settings. An application of the paper's results yields MTL risk bounds of order $O(1/\sqrt{nT})$. More recently, [105] presents excess risk bounds of order $O(1/\sqrt{nT})$ for both MTL and Learning-to-Learn (LTL) settings and reveals conditions, under which MTL is more beneficial when compared to learning tasks independently.

Finally, due to being domains related to MTL, but, at the same time, less connected to the focus of this study, we only mention in passing a few efforts that pertain to generalization guarantees in the realm of life-long learning and domain adaptation. Generalization performance analysis in life-long learning has been investigated in [135, 13, 12, 117] and [116]. Also, in the context of domain adaptation, similar considerations are examined in [96, 98, 97, 34, 148, 99] and [35].

CHAPTER 4: METHODOLOGY

In this chapter, we generally present an overview of the methods we used to derive LRC-based excess risk bounds for MTL.

A Concentration Inequality Using The Entropy Method

Recall that the derivation of LRC-based error bounds makes use of Talagrand's inequality which by itself is the application of Bernstein's concentration inequality for function classes. As we pointed out earlier one elegant way to prove concentration inequalities is the so-called *entropy method* which is based on *logarithmic Sobolev* inequalities, and was first developed by [81, 80, 16, 100, 125, 19] to provide sharp concentration bounds for the suprema of empirical processes. These inequalities are usually considered as the exponential version of the well-known *Efron-Stein* inequality, and they are powerful tools, as they can provide general way to obtain results in many applications. As an example in [17], it has been shown that this method can be applied to retrieve the results of Talagrand's convex-distance inequality. In the following, we first present some notation which will be frequently used throughout the chapter.

Let $X_1, \ldots, X_n$ be n independent random variables taking values in a measurable space $\mathcal{X}$. Assume that $g : \mathcal{X}^n \to \mathbb{R}$ is a measurable function and define

$$Z := g(X_1, \ldots, X_n),$$

which is the quantity we are concerned about its concentration. Let $X_1', \ldots, X_n'$ denote an inde-

pendent copy of $X_1, \ldots, X_n$. Now, let

$$Z_i' := g(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n) \tag{4.1}$$

which is obtained by replacing the variable X_i with X_i'. Define the random variables V^+ and V^- by

$$V^+ := \sum_{i=1}^{n} \mathbb{E}'\left[\left(Z - Z_i'\right)_+^2\right]. \tag{4.2}$$

and

$$V^- := \sum_{i=1}^{n} \mathbb{E}'\left[\left(Z - Z_i'\right)_-^2\right].$$

where $(y)_+ = \max\{0, y\}$, and $(y)_- = \min\{0, y\}$. Also, $\mathbb{E}'[\cdot] := \mathbb{E}[\cdot|X]$ denotes the expectation only with respect to the variables $X_1', \ldots, X_n'$.

The first inequality we will present was proposed by [44], and later improved by [131].

Proposition 19 (Efron-Stein Inequality)**.** *Using the introduced notation above, we have*

$$\mathit{Var}(Z) \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n} \left(Z - Z_i'\right)^2\right]$$

Note that this inequality can be also written as

$$\mathrm{Var}(Z) \leq \mathbb{E}\left[V^+\right] = \mathbb{E}\left[V^-\right]. \tag{4.3}$$

Although this inequality can be very helpful in finding sharp bounds on the variance of complicated functions, it can not take into account the exponential nature of the tails. for this reason, the

authors in [17] showed that under some conditions the Efron-Stein inequality can be transformed to exponential bounds.

Theorem 20 (Exponential Version of Efron-stein Inequality, Theorem 2 in [17])**.** *Let* $\theta > 0$ *and* $\lambda \in (0, 1/\theta)$*. Then, the following inequalities hold*

$$\log \mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big) \leq \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E}\big[\exp\big(\frac{\lambda V^+}{\theta}\big)\big],$$

and

$$\log \mathbb{E}\big(e^{-\lambda(Z-\mathbb{E}Z)}\big) \leq \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E}\big[\exp\big(\frac{\lambda V^-}{\theta}\big)\big].$$

It is worth pointing out that we will be using only the first inequality in our derivations of Talagrand's inequality for MTL. However, for the sake of completeness, we provided all the results as they appeared in Theorem 2 of [17].

Recall that the ultimate goal is to derive a concentration bound for Z. In order to achieve this goal, the following lemma—presented in [20]— can be used to transfer the upper bound on the log-moment generating function $\log \mathbb{E}\big(e^{-\lambda(Z-\mathbb{E}Z)}\big)$, in Theorem 20, to a tail probability on Z.

Lemma 5 (Lemma 2.11 in [20])**.** *Let* Z *be a random variable,* $A, B > 0$ *be some constants. If for any* $\lambda \in (0, 1/B)$ *it holds*

$$\log \mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big) \leq \frac{A\lambda^2}{2\big(1-B\lambda\big)},$$

then for all $x \geq 0$*,*

$$P\big[Z \geq \mathbb{E}Z + \sqrt{2Ax} + Bx\big] \leq e^{-x}.$$

In order to take advantage of Lemma 5, one needs to make $\frac{\lambda\theta}{1-\lambda\theta}\log\mathbb{E}\big[\exp\big(\frac{\lambda V^+}{\theta}\big)\big]$ in the form of $\frac{A\lambda^2}{2\big(1-B\lambda\big)}$. The following results provide some helpful tools to achieve this goal.

We start by introducing a property which is satisfied by many important examples (including some interesting empirical processes), and can be efficiently used to bound the log-moment generating function of those empirical processes which satisfy this property.

Definition 21 (b-self bounding property, Section 3.3 in [18])**.** *A function $g : \mathcal{X}^n \to [0,\infty)$ is said to be b-self bounding ($b > 0$), if there exist functions $g_i : \mathcal{X}^{n-1} \to \mathbb{R}$, such that for all $X_1,\ldots,X_n \in \mathcal{X}$ and all $i \in \mathbb{N}_n$,*

$$0 \le g(X_1,\ldots,X_n) - g_i(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) \le b,$$

and,

$$\sum_{i=1}^{n}\big(g(X_1,\ldots,X_n) - g_i(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)\big) \le g(X_1,\ldots,X_n).$$

Theorem 22 (Theorem 6.12 in [18])**.** *Assume that $Z = g(X_1,\ldots,X_n)$ is a 1-self bounding function. Then for every $\lambda \in \mathbb{R}$,*

$$\log\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \le \phi(\lambda)\mathbb{E}Z, \tag{4.4}$$

where $\phi(\lambda) = e^\lambda - \lambda - 1$.

Corollary 23. *Assume that $Z = g(X_1,\ldots,X_n)$ is a b-self bounding function ($b > 0$). Then, for any $\lambda \in \mathbb{R}$ we have*

$$\log\mathbb{E}e^{\lambda Z} \le \frac{\big(e^{\lambda b}-1\big)}{b}\mathbb{E}Z.$$

Proof. Note that Eq. (4.4) can be rewritten as $\log\mathbb{E}[\exp(\lambda Z)] \le (e^\lambda - 1)\mathbb{E}[Z]$. The stated inequality follows immediately by rescaling Z to Z/b in the above inequality. □

A Talagrand-type Inequality for Suprema of Empirical Processes

The results presented above can be efficiently applied to the suprema of empirical processes which lead to the derivation of a Talagrand-type inequalities for these interesting quantities.

We start by providing an upper bound on the variance-type quantity V^+ for the suprema of an empirical process.

Theorem 24. *Let $X_1, \ldots, X_n$ be n independent variables which are identically distributed according to P. Assume that $\mathcal{F}$ is a countably infinite set of functions defined as $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ such that each function f in $\mathcal{F}$ satisfies $\mathbb{E}[f] = 0$. Let $Z := \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$, and*

$$Z_i' := \sup_{f\in\mathcal{F}} \big[\sum_{j=1}^n f(X_j) - f(X_i) + f(X_i') \big]$$

where $X_1', \ldots, X_n'$ denote an independent copy of $X_1, \ldots, X_n$. Also, define

$$W := \sup_{f\in\mathcal{F}} \sum_{i=1}^n \big[f(X_i)\big]^2$$

and

$$\Upsilon := \sup_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{E}\big[f(X_i)\big]^2.$$

Then the quantity V^+ defined in (4.2) *can be bounded as*

$$V^+ \leq W + \Upsilon.$$

Proof. Let $\hat{f}$ be such that $Z = \sum_{i=1}^{n} \hat{f}(X_i)$. It can be shown that for any $i \in 1, \ldots, n$,

$$Z - Z_i' \leq \hat{f}(X_i) - \hat{f}(X_i'),$$

and therefore,

$$(Z - Z_i')_+^2 \leq \left(\hat{f}(X_i) - \hat{f}(X_i')\right)^2.$$

Then, it follows from the assumption $\mathbb{E}'[f(X_i')] = 0$ that

$$\begin{aligned} V^+ &:= \sum_{i=1}^{n} \mathbb{E}'\left[\left(Z - Z_i'\right)_+^2\right] \\ &\leq \sum_{i=1}^{n} \mathbb{E}'\left[\left(\hat{f}(X_i) - \hat{f}(X_i')\right)^2\right] \\ &= \sum_{i=1}^{n} \left[\hat{f}(X_i)\right]^2 + \sum_{i=1}^{n} \mathbb{E}'\left[\hat{f}(X_i')\right]^2 \\ &\leq \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[f(X_i)\right]^2 + \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{E}\left[f(X_i)\right]^2 \\ &= W + \Upsilon \end{aligned}$$

□

Theorem 25. *Suppose that the conditions of Theorem 24 hold. Moreover, assume that the functions in $\mathcal{F}$ have ranges in $[-b, b]$ with $b > 0$. Let $W := \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[f(X_i)\right]^2$ with $\mathbb{E}W = \Sigma^2$. Then, for any $\lambda \in (0, 1/b)$*

$$\log \mathbb{E} e^{\lambda(W/b)} \leq \frac{\lambda}{b(1 - \lambda b)} \left[4b\mathfrak{R}(\mathcal{F}) + \Upsilon\right].$$

where Υ is defined as in Theorem 24, and the Rademacher complexity $\mathfrak{R}(\mathcal{F})$ is defined according

to Definition 7 as

$$\mathfrak{R}(\mathcal{F}) := \mathbb{E}_{X,\sigma} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(X_i)$$

with σ_i being n independent Rademacher variables.

Proof. Introduce

$$W_i := \sup_{f\in\mathcal{F}} \big[\sum_{j=1}^{n} [f(X_j)]^2 - [f(X_i)]^2 \big]$$

Let $\tilde{f}$ be the function such that $W = \sum_{i=1}^{n} [\tilde{f}(X_i)]^2$, and similarly $\tilde{f}^i$ be the function achieving the supremum in the definition of W_i. It can be shown that

$$\begin{aligned} W - W_i &= \sum_{i=1}^{n} [\tilde{f}(X_i)]^2 - \sup_{f\in\mathcal{F}} \big[\sum_{j=1}^{n} [f(X_j)]^2 - [f(X_i)]^2 \big] \\ &\le \sum_{i=1}^{n} [\tilde{f}(X_i)]^2 - \big[\sum_{j=1}^{n} [\tilde{f}(X_j)]^2 - [\tilde{f}(X_i)]^2 \big] \\ &= \big[\tilde{f}(X_i)\big]^2 \le b^2 \end{aligned}$$

Also, it can be easily shown that $W - W_i \ge 0$. From the other side, we have

$$\sum_{i=1}^{n} (W - W_i) \le \sum_{i=1}^{n} \big[\tilde{f}(X_i)\big]^2 \le \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \big[f(X_i)\big]^2 = W$$

Therefore, according to Definition 21, W/b is a b-self bounding function. Now applying Corollary 23 gives for any $\lambda \in (0, 1/b)$,

$$\log \mathbb{E} e^{\lambda(W/b)} \le \frac{e^{\lambda b} - 1}{b^2} \mathbb{E} W = \frac{e^{\lambda b} - 1}{b^2} \Sigma^2 \le \frac{\lambda \Sigma^2}{b(1 - \lambda b)}, \tag{4.5}$$

where the last inequality follows from $(e^x - 1)(1 - x) \leq x, \forall x \in [0, 1]$. Furthermore, the term Σ^2 can be bounded as

$$\begin{aligned}
\Sigma^2 &= \mathbb{E}_X \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \big[f(X_i)\big]^2 - \Upsilon + \Upsilon \\
&= \mathbb{E}_X \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \big[f(X_i)\big]^2 - \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{E}\big[f(X_i)\big]^2 + \Upsilon \\
&\leq \mathbb{E}_X \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \big[f(X_i)\big]^2 - \sum_{i=1}^{n} \mathbb{E}\big[f(X_i)\big]^2\right] + \Upsilon \\
&\leq 2\mathbb{E}_{X,\sigma} \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \sigma_i \big[f(X_i)\big]^2\right] + \Upsilon \\
&\leq 4b\mathbb{E}_{X,\sigma} \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} \sigma_i f(X_i)\right] + \Upsilon \\
&= 4b\mathfrak{R}(\mathcal{F}) + \Upsilon
\end{aligned}$$

where in the second last inequality we used Lemma 1—the standard symmetrization technique which relates the uniform deviation of an empirical average from its expectation to the Rademacher complexity— Also, the last inequality is the direct application of Lemma 2 with $\phi(x) = x^2$ with Lipschitz constant $2b$ on interval $[-b, b]$, and the last equality uses the definition of Rademacher complexity. Plugging the above inequality back into (4.5) completes the proof. □

Corollary 26. *Assume that the conditions of Theorem 25 hold. If* $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i)$*, and* W*,* Υ *and* $\mathfrak{R}(\mathcal{F})$ *are defined as in Theorem 25, then for any* $b > 0$ *and* $\lambda \in (0, 1/b)$*,*

$$\log \mathbb{E}\big(e^{\lambda(Z - \mathbb{E}Z)}\big) \leq \frac{\lambda^2}{(1 - 2\lambda b)} \left[4b\mathfrak{R}(\mathcal{F}) + 2\Upsilon\right]. \tag{4.6}$$

Proof. Note that by combining the results of Theorem 20 and Theorem 24, we can get for any

$\lambda \in (0, 1/b)$,

$$\log \mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big) \leq \frac{\lambda b}{1-\lambda b} \log \mathbb{E}\big[\exp\big(\frac{\lambda\big(W+\Upsilon\big)}{b}\big)\big].$$

which together with Theorem 25 gives

$$\begin{aligned}
\log \mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big) &\leq \frac{\lambda b}{1-\lambda b}\left[\frac{\lambda}{b(1-\lambda b)}\left[4b\mathfrak{R}(\mathcal{F})+\Upsilon\right]+\frac{\lambda\Upsilon}{b}\right]. \\
&\leq \frac{\lambda b}{1-\lambda b}\left(\frac{\lambda}{b(1-\lambda b)}\right)\left[4b\mathfrak{R}(\mathcal{F})+2\Upsilon\right] \\
&= \frac{\lambda^2}{(1-\lambda b)^2}\left[4b\mathfrak{R}(\mathcal{F})+2\Upsilon\right].
\end{aligned}$$

Also, regarding the fact that for any $\lambda \in (0, 1/2b)$, it holds that $(1-\lambda b)^2 \geq 1-2\lambda b \geq 0$, we have

$$\log \mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big) \leq \frac{\lambda^2}{(1-2\lambda b)}\left[4b\mathfrak{R}(\mathcal{F})+2\Upsilon\right].$$

□

Now, with the help of Lemma 5, we can convert the bound in (4.6) into a tail probability on $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i)$ as presented in the following.

Corollary 27. *If the conditions of Corollary 26 hold, then for any $x > 0$, with probablity at least $1 - e^{-x}$,*

$$Z \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{8x\Upsilon} + 4bx$$

where $\mathfrak{R}(\mathcal{F})$ and Υ are defined as in Theorem 25.

Proof. Note (4.6) together with Lemma 5, for $A = 2\left[4b\mathfrak{R}(\mathcal{F})+2\Upsilon\right]$ and $B = 2b$ gives with

probability at least $1 - e^{-x}$,

$$\begin{aligned} Z &\leq \mathbb{E}Z + \sqrt{4x\left[4b\mathfrak{R}(\mathcal{F}) + 2\Upsilon\right]} + 2bx \\ &\leq \mathbb{E}Z + 4\sqrt{bx\mathfrak{R}(\mathcal{F})} + \sqrt{8x\Upsilon} + 2bx \\ leq &\mathbb{E}Z + 2\mathfrak{R}(\mathcal{F}) + 2bx + \sqrt{8x\Upsilon} + 2bx \\ &\leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{8x\Upsilon} + 4bx \end{aligned}$$

where the second last inequality follows from $2\sqrt{uv} \leq u + v$, and the last step uses the symmetrization inequality $\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(X_i) \leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(X_i) = 2\mathfrak{R}(\mathcal{F})$ (Note that we assumed $\mathbb{E}f = 0$). □

CHAPTER 5: LOCAL RADEMACHER COMPLEXITY-BASED EXCESS RISK BOUNDS FOR MULTI-TASK LEARNING

Through a Talagrand-type concentration inequality adapted to the MTL case, one of the main contributions of this dissertation is the derivation of sharp bounds on the excess MTL risk in terms of the distribution- and data-dependent LRC. For a given number of tasks T, these bounds admit faster (asymptotic) convergence characteristics in the number of observations per task n, when compared to corresponding bounds hinging on the GRC. Thence, these faster rates allow for heightened confidence that the MTL hypothesis selected by a learning algorithm approaches the best-in-class solution as n increases beyond a certain threshold. We also derive a new bound on the LRC, which generally holds for hypothesis classes with any norm function or strongly convex regularizers. This bound readily facilitates the bounding of the LRC for a range of such regularizers (not only for MTL, but also for the standard i.i.d. setting), as we demonstrate for classes induced by graph-based, Schatten- and group-norm regularizers. Moreover, we prove matching lower bounds showing that, aside from constants, the LRC-based bounds are tight for the considered applications.

Our derived bounds reflect that one can trade off a slow convergence speed w.r.t. T for an improved convergence rate w.r.t. n. The latter one ranges, in the worst case, from the typical GRC-based bounds $O(1/\sqrt{n})$, all the way up to the fastest rate of order $O(1/n)$ by allowing the bound to depend less on T. Nevertheless, the premium in question becomes less relevant to MTL, in which T is typically considered as fixed.

In what follows, we use the following notational conventions: vectors and matrices are depicted in bold face. The superscript T, when applied to a vector/matrix, denotes the transpose of that quantity. We define $\mathbb{N}_T := \{1, \ldots, T\}$. For any random variables X, Y and functions f we use $\mathbb{E}f(X, Y)$ and $\mathbb{E}_X f(X, Y)$ to denote the expectation w.r.t. all the involved random variables

and the conditional expectation w.r.t. the random variable X. For any vector-valued function $\boldsymbol{f} = (f_1, \ldots, f_T)$, we introduce the following two notations:

$$P\boldsymbol{f} := \frac{1}{T}\sum_{t=1}^{T} Pf_t = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(f(X_t)), \qquad P_n\boldsymbol{f} := \frac{1}{T}\sum_{t=1}^{T} P_n f_t = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n} f(X_t^i).$$

We also denote $\boldsymbol{f}^\alpha = (f_1^\alpha, \ldots, f_T^\alpha), \forall \alpha \in \mathbb{R}$. For any loss function ℓ and any $\boldsymbol{f} = (f_1, \ldots, f_T)$ we define $\ell_{\boldsymbol{f}} = (\ell_{f_1}, \ldots, \ell_{f_T})$ where ℓ_{f_t} is the function defined by $\ell_{f_t}((X_t, Y_t)) = \ell(f_t(X_t), Y_t)$.

Talagrand-Type Inequality for Multi-Task Learning

The derivation of our LRC-based error bounds for MTL is founded on the following modified Talagrand's concentration inequality adapted to the context of MTL, showing that the uniform deviation between the true and empirical means in a vector-valued function class $\mathcal{F}$ can be dominated by the associated *multi-task Rademacher complexity* plus a term involving the variance of functions in $\mathcal{F}$. We defer the proof in Appendix A.

Theorem 28 (TALAGRAND-TYPE INEQUALITY FOR MTL)**.** *Let $\mathcal{F} = \{\boldsymbol{f} := (f_1, \ldots, f_T)\}$ be a class of vector-valued functions satisfying $\sup_{t,x} |f_t(x)| \leq b$. Let $X := (X_t^i)_{(t,i)=(1,1)}^{(T,N_t)}$ be a vector of $\sum_{t=1}^{T} N_t$ independent random variables where $X_t^1, \ldots, X_t^n, \forall t$ are identically distributed. Let $\{\sigma_t^i\}_{t,i}$ be a sequence of independent Rademacher variables. If $\frac{1}{T}\sup_{\boldsymbol{f}\in\mathcal{F}} \sum_{t=1}^{T} \mathbb{E}\left[f_t(X_t^1)\right]^2 \leq r$, then, for every $x > 0$, with probability at least $1 - e^{-x}$,*

$$\sup_{\boldsymbol{f}\in\mathcal{F}}(P\boldsymbol{f} - P_n\boldsymbol{f}) \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{8xr}{nT}} + \frac{12bx}{nT}, \tag{5.1}$$

where $n := \min_{t\in\mathbb{N}_T} N_t$, and the multi-task Rademacher complexity *of function class $\mathcal{F}$ is defined*

as

$$\mathfrak{R}(\mathcal{F}) := \mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}=(f_1,\ldots,f_T)\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\sigma_t^i f_t(X_t^i)\right\}.$$

Note that the same bound also holds for $\sup_{\boldsymbol{f}\in\mathcal{F}}(P_n\boldsymbol{f} - P\boldsymbol{f})$.

In Theorem 28, the data from different tasks assumed to be mutually independent, which is typical in the MTL setting [101]. To present the results in a clear way we always assume in the following that the available data for each task is the same, namely n.

Remark 29. *At this point, we would like to present the result of the above theorem for the special case* $T = 1$ *which corresponds to the traditional single task learning framework. It is very easy to verify that for* $T = 1$*, the bound in* (5.1) *can be written as*

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{8xr}{n}} + \frac{12bx}{n}, \tag{5.2}$$

where the function f *is chosen from an scalar-valued function class* $\mathcal{F}$*. This bound can be compared to the result of Theorem 2.1 of [10] (for* $\alpha = 1$*) which is presented as*

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{2xr}{n}} + \frac{8bx}{n} \tag{5.3}$$

Note that the difference between the constants in (5.2) *and* (5.3)*, is due to the fact that we failed in directly applying Bousquet's version of Talagrand inequality—similar to what has been done in [10] for scalar-valued functions—to the class of vector-valued functions. To make it more clear, let* Z *be defined as* (A.1) *with the jackknife replication* $Z_{s,j}$ *for which a lower bound* $Z''_{s,j}$ *can be found such that* $Z''_{s,j} \leq Z - Z_{s,j}$*. Then, in order to apply Theorem 2.5 of [20], one needs to show that the quantity* $\frac{1}{nT}\sum_{s=1}^{T}\sum_{j=1}^{n}\mathbb{E}_{s,j}[(Z''_{s,j})^2]$ *is bounded. This goal, ideally, can be achieved by including a constraint similar to* $\frac{1}{T}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{t=1}^{T}\mathbb{E}\left[f_t(X_t^1)\right]^2 \leq r$ *in Theorem 28. However,*

we could not come up with any obvious and meaningful way—appropriate for MTL—of defining this constraint to satisfy the boundedness condition $\frac{1}{nT}\sum_{s=1}^{T}\sum_{j=1}^{n}\mathbb{E}_{s,j}[(Z_{s,j}^{''})^2]$ *in terms of* r*. We would like emphasize that the key ingredient to the proof of Theorem 28 is the so-called* Logarithmic Sobolev *inequality—Theorem 20—which can be considered as the exponential version of* Efron-Stein *inequality.*

Excess MTL Risk Bounds Based on Local Rademacher Complexities

The cornerstone of Sect. 5's results is the presence of an upper bound of an empirical process's variance (the second term in the right-hand side of (5.1)). In this section, we consider the Rademacher averages associated with a smaller subset of the function class $\mathcal{F}$ and use them as a complexity term in the context of excess risk bounds. As pointed out in [10], these (local) averages are always smaller than the corresponding global Rademacher averages and allow for eventually deriving sharper generalization bounds. Herein, we exploit this very fact for MTL generalization guarantees.

Theorem 28 motivates us to extend the definition of classical LRC $\mathfrak{R}(\mathcal{F}^{sclr},r)$ for a scalar-valued function class $\mathcal{F}^{sclr}$ as

$$\mathfrak{R}(\mathcal{F}^{sclr},r):=\mathbb{E}_{X,\sigma}\big[\sup_{f\in\mathcal{F}^{sclr},V(f)\leq r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\big]$$

to the Multi-Task Local Rademacher Complexity (MT-LRC) using the following definition.

Definition 30 (MULTI-TASK LOCAL RADEMACHER COMPLEXITY)**.** *For a vector-valued function class* $\mathcal{F}$ *the* Local Rademacher Complexity $\mathfrak{R}(\mathcal{F},r)$ *and its empirical counterpart* $\hat{\mathfrak{R}}(\mathcal{F},r)$

are defined as

$$\begin{aligned}\mathfrak{R}(\mathcal{F}, r) &:= \mathbb{E}\Bigg[\sup_{\substack{\boldsymbol{f}=(f_1,\ldots,f_T)\in\mathcal{F}\\ V(\boldsymbol{f})\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\Bigg],\\ \hat{\mathfrak{R}}(\mathcal{F}, r) &:= \mathbb{E}_\sigma\Bigg[\sup_{\substack{\boldsymbol{f}=(f_1,\ldots,f_T)\in\mathcal{F}\\ V_n(\boldsymbol{f})\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\Bigg],\end{aligned} \tag{5.4}$$

where $V(\boldsymbol{f})$ and $V_n(\boldsymbol{f})$ are upper bounds on the variance and empirical variances of the functions in $\mathcal{F}$, respectively. This dissertation makes the choice $V(\boldsymbol{f}) = P\boldsymbol{f}^2$ and $V_n(\boldsymbol{f}) = P_n\boldsymbol{f}^2$ where

$$\begin{aligned}P\boldsymbol{f}^2 &:= \frac{1}{T}\sum_{t=1}^{T} Pf_t^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f_t(X_t)\right]^2,\\ P_n\boldsymbol{f}^2 &:= \frac{1}{T}\sum_{t=1}^{T} P_nf_t^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}(f_t(X_t^i))^2.\end{aligned}$$

Analogous to single task learning, the challenge in applying MT-LRC (5.4) to refine the existing learning rates is to find an optimal radius trading-off the size of the set $\{\boldsymbol{f} \in \mathcal{F} : V(\boldsymbol{f}) \le r\}$ and its complexity, which, as we show later, reduces to the calculation of the fixed-point of a sub-root function.

The definition of local Rademacher complexity is based on the fact that by incorporating the variance constraint, a better error rate for the bounds can be obtained. In other words, the key point in deriving fast rate bounds is that around the best function f^* (the function that minimizes the true risk), the variance of the deviation between the empirical and true errors of functions in the class is controlled by a linear function of the expectation of this difference. We will call a class with this property a *Bernstein* class, and we provide a definition of a vector-valued Berstein class $\mathcal{F}$ as following.

Definition 31 (VECTOR-VALUED BERNSTEIN CLASS)**.** *A vector-valued function class $\mathcal{F}$ is said to be a (β, B)-Bernstein class with respect to the probability measure P, if for every $0 < \beta \le 1$,*

$B \geq 1$ *and any* $\boldsymbol{f} \in \mathcal{F}$*, there exists a function* $V : \mathcal{F} \to \mathbb{R}^+$ *such that*

$$P\boldsymbol{f}^2 \leq V(\boldsymbol{f}) \leq BP\boldsymbol{f}^\beta. \tag{5.5}$$

It can be shown that the Bernstein condition (5.5) is not too restrictive and it holds, for example, for non-negative bounded functions with respect to any probability distribution [10]. Other examples include the class of excess risk functions $\mathcal{L}_\mathcal{F} := \{\ell_f - \ell_{f^*} : f \in \mathcal{F}\}$—with $f^* \in \mathcal{F}$ the minimizer of $P\ell_f$— when the function class $\mathcal{F}$ is convex and the loss function ℓ is strictly convex.

In this section, we show that under some mild assumptions on a vector-valued Bernstein class, the LRC-based excess risk bounds can be established for MTL. We assume that the loss function ℓ and the vector-valued hypothesis space $\mathcal{F}$ satisfy the following conditions:

Assumption 32.

1. *There is a function* $\boldsymbol{f}^* = (f_1^*, \ldots, f_T^*) \in \mathcal{F}$ *satisfying* $P\ell_{\boldsymbol{f}^*} = \inf_{\boldsymbol{f}\in\mathcal{F}} P\ell_{\boldsymbol{f}}$.
2. *There is constant* $B' \geq 1$*, such that for every* $\boldsymbol{f} \in \mathcal{F}$ *we have* $P(\boldsymbol{f} - \boldsymbol{f}^*)^2 \leq B'P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*})$.
3. *There exists a constant* L*, such that the loss function* ℓ *is* L*-Lipschitz in its first argument.*

As it has been pointed out in [10], there are many examples of regularized algorithms for which these conditions can be satisfied. More specifically, a uniform convexity condition on the loss function ℓ is usually sufficient to satisfy Assumption 32.2. As an example for which this assumption holds, [10] refereed to the quadratic loss function $\ell(f(X), Y) = (f(X) - Y)^2$ when the functions $f \in \mathcal{F}$ are uniformly bounded, More specifically, if for all $x \in \mathcal{X}$, $Y \in \mathcal{Y}$ and $f \in \mathcal{F}$, it holds that $|f(X) - Y| \in [0, 1]$, then it can be shown that the conditions of Assumption 32 are met with $L = 1$ and $B = 1$.

Now we can present the main result of this section showing that the excess error of MTL can be bounded by the fixed-point of a sub-root function dominating the MT-LRC. The proof of the results is provided in the Appendix B.

Theorem 33 (Distribution-dependent excess risk bound for MTL)**.** *Let* $\mathcal{F} := \{\boldsymbol{f} := (f_1, \ldots, f_T) : \forall t, f_t \in \mathbb{R}^{\mathcal{X}}\}$ *be a class of vector-valued functions* $\boldsymbol{f}$ *satisfying* $\sup_{t,x} |f_t(x)| \leq b$*. Also, Let* $X := (X_t^i, Y_t^i)_{(t,i)=(1,1)}^{(T,n)}$ *be a vector of* nT *independent random variables where for each task* t*,* $(X_t^1, Y_t^1) \ldots, (X_t^n, Y_t^n)$ *be identically distributed. Suppose that Assumption 32 holds. Define* $\mathcal{F}^* := \{\boldsymbol{f} - \boldsymbol{f}^*\}$*, where* $\boldsymbol{f}^*$ *is the function satisfying* $P\ell_{\boldsymbol{f}^*} = \inf_{\boldsymbol{f}\in\mathcal{F}} P\ell_{\boldsymbol{f}}$*. Let* $B := B'L^2$ *and* ψ *be a sub-root function with the fixed point* r^* *such that* $BL\mathfrak{R}(\mathcal{F}^*, r) \leq \psi(r), \forall r \geq r^*$*, where* $\mathfrak{R}(\mathcal{F}^*, r)$ *is the LRC of the functions class* $\mathcal{F}^*$*:*

$$\mathfrak{R}(\mathcal{F}^*, r) := \mathbb{E}_{X,\sigma}\Big[\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\leq r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\Big]. \tag{5.6}$$

Then, we have the following bounds in terms of the fixed point r^* *of* $\psi(r)$*:*

1. *For any* $\boldsymbol{f} \in \mathcal{F}$*,* $K > 1$ *and* $x > 0$*, with probability at least* $1 - e^{-x}$*,*

$$P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) \leq \frac{K}{K-1}P_n(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) + 800Kr^* + \frac{(48Lb + 28BK)Bx}{nT}. \tag{5.7}$$

2. *If the function class* $\mathcal{F}$ *is convex, then for any* $\boldsymbol{f} \in \mathcal{F}$*,* $K > 1$ *and* $x > 0$*, with probability at least* $1 - e^{-x}$*,*

$$P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) \leq \frac{K}{K-1}P_n(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) + 32Kr^* + \frac{(48Lb + 16BK)Bx}{nT}. \tag{5.8}$$

Corollary 34. *Let* $\hat{\boldsymbol{f}}$ *be any element of convex class* $\mathcal{F}$ *satisfying* $P_n\ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f}\in\mathcal{F}} P_n\ell_{\boldsymbol{f}}$*. Assume that the conditions of Theorem 33 hold. Then for any* $\boldsymbol{f} \in \mathcal{F}$*,* $x > 0$ *and* $r > \psi(r)$*, with probability*

at least $1 - e^{-x}$*,*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 32Kr^* + \frac{(48Lb + 16BK)Bx}{nT}. \tag{5.9}$$

Proof. The results follows by noticing that $P_n(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 0$. □

The next theorem, analogous to Corollary 5.4 in [10], presents a data-dependent version of (5.9) replacing the Rademacher complexity in Corollary 34 with its empirical counterpart. The proof of this Theorem, which repeats the same basic steps utilized in Theorem 5.4 in [10], can be found in Appendix B.

Theorem 35 (Data-dependent excess risk bound for MTL)**.** *Let* $\hat{\boldsymbol{f}}$ *be any element of convex class* $\mathcal{F}$ *satisfying* $P_n\ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f}\in\mathcal{F}} P_n\ell_{\boldsymbol{f}}$*. Assume that the condition of Theorem 33 hold. Define*

$$\hat{\psi}_n(r) = c_1\hat{\mathfrak{R}}(\mathcal{F}^*, c_3 r) + \frac{c_2 x}{nT}, \quad \hat{\mathfrak{R}}(\mathcal{F}^*, c_3 r) := \mathbb{E}_\sigma\Big[\sup_{\substack{\boldsymbol{f}\in\mathcal{F}, \\ L^2P_n(\boldsymbol{f}-\hat{\boldsymbol{f}})^2\leq c_3 r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\Big],$$

where $c_1 = 2L\max(B, 16Lb)$*,* $c_2 = 128L^2b^2 + 2bc_1$ *and* $c_3 = 4 + 128BK + 4B^2(48Lb + 16BK)/c_2$*. Then for any* $\boldsymbol{f} \in \mathcal{F}$*,* $K > 1$ *and* $x > 0$*, with probability at least* $1 - 4e^{-x}$*, we have*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 32K\hat{r}^* + \frac{(48Lb + 16BK)Bx}{nT},$$

where $\hat{r}^*$ *is the fixed point of the sub-root function* $\hat{\psi}_n(r)$*.*

An immediate consequence of the results of this section is that one can derive excess risk bounds for given regularized MTL hypothesis spaces. In the next section, by further bounding the fixed point r^* in Corollary 34 (and $\hat{r}^*$ in Theorem 35), we will derive distribution (and data)-dependent excess risk bounds for several commonly used norm-regularized MTL hypothesis spaces.

Local Rademacher Complexity Bounds for Norm Regularized MTL Models

This section presents very general MT-LRC bounds, based on the distribution-dependent excess risks established in Theorem 33, for hypothesis spaces defined by norm regularizers, which allows us to immediately derive, as specific application cases, LRC bounds for group-norm, Schatten-norm, and graph-regularized MTL models. It should be mentioned that similar data-dependent MT-LRC bounds are also available by a similar deduction process.

Preliminaries

We consider linear MTL models where we associate to each task-wise function f_t a weight $\boldsymbol{w}_t \in \mathcal{H}$ by $f_t(X) = \langle \boldsymbol{w}_t, \phi(X) \rangle$. Here ϕ is a feature map associated to a Mercer kernel k satisfying $k(X, \tilde{X}) = \langle \phi(X), \phi(\tilde{X}) \rangle, \forall X, \tilde{X} \in \mathcal{X}$ and $\boldsymbol{w}_t$ belongs to the *reproducing kernel Hilbert space* $\mathcal{H}_K$ induced by k with inner product $\langle \cdot, \cdot \rangle$. We assume that the multi-task model $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T) \in \mathcal{H} \times \ldots \times \mathcal{H}$ is learned by a regularization scheme:

$$\min_{\boldsymbol{W}} \Omega(\boldsymbol{W}) + C \sum_{t=1}^{T} \sum_{i=1}^{n} \ell(\langle \boldsymbol{w}_t, \phi(X_t^i) \rangle_{\mathcal{H}}, Y_t^i), \tag{5.10}$$

where the regularizer $\Omega(\cdot)$ is used to enforce a priori information to avoid over-fitting. This regularization scheme amounts to performing *ERM* in the hypothesis space

$$\mathcal{F} := \left\{ X \mapsto [\langle \boldsymbol{w}_1, \phi(X_1) \rangle, \ldots, \langle \boldsymbol{w}_T, \phi(X_T) \rangle]^T : \Omega(\boldsymbol{D}^{1/2} \boldsymbol{W}) \leq R^2 \right\}, \tag{5.11}$$

where $\boldsymbol{D}$ is a given positive operator defined in $\mathcal{H}$. Note that the hypothesis spaces corresponding to group and Schatten norms can be recovered by taking $\boldsymbol{D} = \boldsymbol{I}$, and choosing their associated norms. More specifically, by choosing $\Omega(\boldsymbol{W}) = \frac{1}{2} \|\boldsymbol{W}\|_{2,q}^2$, we can retrieve an $L_{2,q}$-norm hy-

pothesis space in (5.11). Similarly, the choice $\Omega(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_{S_q}^2$ gives an L_{S_q}-Schatten norm hypothesis space in (5.11). Furthermore, the graph-regularized MTL [46, 102, 109] can be specialized by taking $\Omega(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{D}^{1/2}\boldsymbol{W}\|_F^2$, wherein $\|.\|_F$ is a Frobenius norm, and $\boldsymbol{D} = \boldsymbol{L} + \eta\boldsymbol{I}$ with $\boldsymbol{L}$ being a graph-Laplacian, and η being a regularization constant. On balance, all these MTL models can be considered as norm-regularized models. Also, for specific values of q, it can be shown that they are strongly convex.

General Bound on the Local Rademacher Complexity

Now, we can provide the main results of this section which give general LRC bounds for any general MTL hypothesis space of the form (5.11) in which $\Omega(\boldsymbol{W})$ is given as a strongly convex or a norm function of $\boldsymbol{W}$.

Theorem 36 (Distribution-dependent MT-LRC bounds by strong convexity)**.** *Let $\Omega(\boldsymbol{W})$ in* (5.10) *be μ-strongly convex with $\Omega^*(\mathbf{0}) = 0$ and $\|k\|_\infty \leq \mathcal{K} \leq \infty$. Let $X_t^1, \ldots, X_t^n$ be an i.i.d. sample drawn from P_t. Also, assume that for each task t, the eigenvalue-eigenvector decomposition of the Hilbert-Schmidt covariance operator J_t is given by $J_t := \mathbb{E}(\phi(X_t) \otimes \phi(X_t)) = \sum_{j=1}^{\infty} \lambda_t^j \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j$, where $(\boldsymbol{u}_t^j)_{j=1}^{\infty}$ forms an orthonormal basis of $\mathcal{H}$, and $(\lambda_t^j)_{j=1}^{\infty}$ are the corresponding eigenvalues, arranged in non-increasing order. Then for any given positive operator $\boldsymbol{D}$ on $\mathbb{R}^T$, any $r > 0$ and any non-negative integers $h_1, \ldots, h_T$:*

$$\mathfrak{R}(\mathcal{F}, r) \leq \min_{\{0 \leq h_t \leq \infty\}_{t=1}^T} \left\{ \sqrt{\frac{r \sum_{t=1}^T h_t}{nT}} + \frac{R}{T} \sqrt{\frac{2}{\mu} \mathbb{E}_{X,\sigma} \left\| \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\|_*^2} \right\}, \tag{5.12}$$

where $\boldsymbol{V} = \left(\sum_{j > h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^T$.

Proof. Note that with the help of LRC definition, we have for any function class $\mathcal{F}$,

$$
\begin{aligned}
\mathfrak{R}(\mathcal{F}, r) &= \frac{1}{nT}\mathbb{E}_{X,\sigma}\left\{\sup_{\substack{\boldsymbol{f}=(f_1,\ldots,f_T)\in\mathcal{F},\\ P\boldsymbol{f}^2\le r}}\sum_{i=1}^{n}\left\langle (\boldsymbol{w}_t)_{t=1}^{T}, \left(\sigma_t^i\phi(X_t^i)\right)_{t=1}^{T}\right\rangle\right\} \\
&= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ P\boldsymbol{f}^2\le r}}\left\langle (\boldsymbol{w}_t)_{t=1}^{T}, \left(\sum_{j=1}^{\infty}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\rangle\right\} \\
&\le \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{P\boldsymbol{f}^2\le r}\left\langle\left(\sum_{j=1}^{h_t}\sqrt{\lambda_t^j}\left\langle\boldsymbol{w}_t, \boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T},\right.\right. \\
&\qquad\qquad\left.\left.\left(\sum_{j=1}^{h_t}\sqrt{\lambda_t^j}^{-1}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\rangle\right\} \qquad (5.13) \\
&+ \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\langle (\boldsymbol{w}_t)_{t=1}^{T}, \left(\sum_{j>h_t}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\rangle\right\} \qquad (5.14) \\
&= A_1 + A_2.
\end{aligned}
$$

where in the last equality, we defined the term in (5.13) as A_1, and the term in (5.14) as A_2.

Step 1. Controlling A_1: Applying Cauchy-Schwartz (C.S.) inequality on A_1 yields the following

$$A_1 \leq \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{P\boldsymbol{f}^2\leq r}\left[\left(\sum_{t=1}^{T}\left\|\sum_{j=1}^{h_t}\sqrt{\lambda_t^j}\left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j\right\rangle \boldsymbol{u}_t^j\right\|^2\right)^{\frac{1}{2}}\right.\right.$$
$$\left.\left.\left(\sum_{t=1}^{T}\left\|\sum_{j=1}^{h_t}\sqrt{\lambda_t^j}^{-1}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle \boldsymbol{u}_t^j\right\|^2\right)^{\frac{1}{2}}\right]\right\}$$
$$= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{P\boldsymbol{f}^2\leq r}\left[\left(\sum_{t=1}^{T}\sum_{j=1}^{h_t}\lambda_t^j\left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j\right\rangle^2\right)^{\frac{1}{2}}\right.\right.$$
$$\left.\left.\left(\sum_{t=1}^{T}\sum_{j=1}^{h_t}\lambda_t^{j-1}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle^2\right)^{\frac{1}{2}}\right]\right\}.$$

With the help of Jensen's inequality and regarding the fact that $\mathbb{E}_{X,\sigma}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle^2 = \frac{\lambda_t^j}{n}$ and $P\boldsymbol{f}^2 \leq r$ implies $\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\lambda_t^j\left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j\right\rangle^2 \leq r$ (see Lemma 12 in the Appendix for the proof), we can further bound A_1 as

$$A_1 \leq \sqrt{\frac{r\sum_{t=1}^{T}h_t}{nT}}. \tag{5.15}$$

Step 2. Controlling A_2: We use strong convexity assumption on the regularizer in order to further bound the second term $A_2 = \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\langle \boldsymbol{D}^{1/2}\boldsymbol{W}, \boldsymbol{D}^{-1/2}\boldsymbol{V}\right\rangle\right\}$.

Let $\lambda > 0$. Applying (C.1) with $\boldsymbol{w} = \boldsymbol{D}^{1/2}\boldsymbol{W}$ and $\boldsymbol{v} = \lambda\boldsymbol{D}^{-1/2}\boldsymbol{V}$ gives

$$\left\langle \boldsymbol{D}^{1/2}\boldsymbol{W}, \lambda\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\rangle \leq \Omega(\boldsymbol{D}^{1/2}\boldsymbol{W}) + \left\langle \nabla\Omega^*\left(\boldsymbol{0}\right), \lambda\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\rangle + \frac{\lambda^2}{2\mu}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*^2.$$

Note that, regrading the definition of $\boldsymbol{V}$, we get $\mathbb{E}_\sigma \left\langle \triangledown \Omega^* \left(\mathbf{0} \right), \lambda \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\rangle = 0$, Therefore, taking supremum and expectation on both sides, dividing throughout by λ and T, and then optimizing over λ gives

$$\begin{aligned} A_2 &= \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{\boldsymbol{f} \in \mathcal{F}} \left\langle \boldsymbol{D}^{1/2} \boldsymbol{W}, \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\rangle \right\} \leq \min_{0<\lambda<\infty} \left\{ \frac{R^2}{\lambda T} + \frac{\lambda}{2\mu T} \mathbb{E}_{X,\sigma} \left\| \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\|_*^2 \right\} \\ &= \frac{R}{T} \sqrt{\frac{2}{\mu} \mathbb{E}_{X,\sigma} \left\| \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\|_*^2}. \end{aligned} \tag{5.16}$$

Combining (5.16) with (5.15) completes the proof. □

Remark 37. *Note that when considering a norm regularized space similar to* (5.11)*, more general result can be obtained with the help of Hölder inequality which hold for any norm regularizer* $\Omega(\boldsymbol{D}^{1/2}\boldsymbol{W})$ *and not necessarily strongly convex norms. More specifically, for any regularizer* $\Omega(\boldsymbol{W})$*, which is presented as a norm function* $\|.\|$ *of* $\boldsymbol{W}$*, we can derive a general LRC bound presented in the following theorem.*

Theorem 38 (Distribution-dependent MT-LRC bounds by Hölder inequality)**.** *Let the regularizer* $\Omega(\boldsymbol{W})$ *in* (5.10) *be given as a norm function in the form of* $\|.\|$*, where its dual conjugate is denoted by* $\|.\|_*$*. Let the kernels be uniformly bounded, that is* $\|k\|_\infty \leq \mathcal{K} \leq \infty$*, and* $X_t^1, \ldots, X_t^n$ *be an i.i.d. sample drawn from* P_t*. Also, assume that for each task* t*, the eigenvalue-eigenvector decomposition of the Hilbert-Schmidt covariance operator* J_t *is given by* $J_t := \mathbb{E}(\phi(X_t) \otimes \phi(X_t)) = \sum_{j=1}^\infty \lambda_t^j \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j$*, where* $(\boldsymbol{u}_t^j)_{j=1}^\infty$ *forms an orthonormal basis of* $\mathcal{H}$*, and* $(\lambda_t^j)_{j=1}^\infty$ *are the corresponding eigenvalues, arranged in non-increasing order. Then for any given positive operator* $\boldsymbol{D}$ *on* $\mathbb{R}^T$*, any* $r > 0$ *and any non-negative integers* $h_1, \ldots, h_T$*:*

$$\mathfrak{R}(\mathcal{F}, r) \leq \min_{0 \leq h_t \leq \infty} \left\{ \sqrt{\frac{r \sum_{t=1}^T h_t}{nT}} + \frac{\sqrt{2}R}{T} \mathbb{E}_{X,\sigma} \left\| \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\|_* \right\}, \tag{5.17}$$

where $\boldsymbol{V} = \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^T$

Proof. The proof of this theorem repeats the same steps as the proof of Theorem 36, except for controlling term A_2 in (5.14), in which the Hölder inequality can be efficiently used to further bound A_2 as following

$$\begin{aligned}
A_2 &= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\langle (\boldsymbol{w}_t)_{t=1}^T, \left(\sum_{j>h_t}\left\langle \frac{1}{n}\sum_{i=1}^n \sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle \boldsymbol{u}_t^j\right)_{t=1}^T\right\rangle\right\} \\
&= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\langle \boldsymbol{D}^{1/2}\boldsymbol{W}, \boldsymbol{D}^{-1/2}\boldsymbol{V}\right\rangle\right\} \\
&\overset{\text{Hölder}}{\leq} \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\|\boldsymbol{D}^{1/2}\boldsymbol{W}\right\|.\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*\right\} \\
&\leq \frac{\sqrt{2}R}{T}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*. \qquad (5.18)
\end{aligned}$$

□

Remark 39. *Notice that, obviously,* $\sqrt{2}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_* \leq \sqrt{\frac{2}{\mu}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*^2}$ *for any* $\mu \leq 1$*. Interestingly, for the cases considered in our study, it holds that* $\mu \leq 1$*. More specifically, from Theorem 3 and Theorem 12 in [69], it can be shown that* $R(\boldsymbol{W}) = 1/2\left\|\boldsymbol{W}\right\|_{2,q}^2$ *is* $\frac{1}{q^*}$*-strongly convex w.r.t. the group norm* $\|.\|_{2,q}$*. Similarly, using Theorem 10 in [69], it can be shown that the regularization function* $R(\boldsymbol{W}) = \frac{1}{2}\left\|\boldsymbol{W}\right\|_{S_q}^2$ *with* $q \in [1,2]$ *is* $(q-1)$*-strongly convex w.r.t. the* L_{S_q}*-Schatten norm* $\|.\|_{S_q}$*. Therefore, given the range of* q *in* $[1,2]$*, for which these two norms are strongly convex, it can be easily seen that* $\mu \leq \frac{1}{2}$ *and* $\mu \leq 1$ *for the group and Schatten-norm hypotheses, respectively. Therefore, for this cases, Hölder inequality yields slightly tighter bounds for MT-LRC.*

Remark 40. *It is worth mentioning that, when applied to the norm-regularized MTL models, the result of Theorem 38 could be more general than that of Theorem 36. More specially, for* $L_{2,q}$*-group and* L_{S_q}*-Schatten norm regularizers, Theorem 36 can only be applied to the special case of* $q \in [1,2]$*, for which these two norms are strongly convex. In contrast, Theorem 38 is applicable to any value of* q *for these two norms. For this reason and considering the fact that very similar*

results can be obtained from Theorem 36 and Theorem 38 (see Lemma 6 and Remark 41), we will use Theorem 38 in the sequel to find the LRC bounds of several norm regularized MTL models.

In what follows, we demonstrate the power of Theorem 38 by applying it to derive the LRC bounds for some popular MTL models, including group norm, Schatten norm and graph regularized MTL models extensively studied in the literature of MTL [102, 45, 7, 5, 3].

Group Norm Regularized MTL

We first consider a group norm regularized MTL capturing the inter-task relationships by the group norm regularizer $\frac{1}{2}\|\boldsymbol{W}\|_{2,q}^2 := \frac{1}{2}\big(\sum_{t=1}^T \|\boldsymbol{w}_t\|_2^q\big)^{2/q}$ [45, 126, 5, 92], for which the associated hypothesis space takes the form

$$\mathcal{F}_q := \left\{ X \mapsto [\langle \boldsymbol{w}_1, \phi(X_1)\rangle, \ldots, \langle \boldsymbol{w}_T, \phi(X_T)\rangle]^T : \frac{1}{2}\|\boldsymbol{W}\|_{2,q}^2 \leq R_{max}^2 \right\}. \tag{5.19}$$

Before presenting the result for the group-norm regularized MTL, we want to bring it into attention that A_1 does not depend on the $\boldsymbol{W}$-constraint in the hypothesis space, therefore the bound for A_1 is the same for all cases we consider in this study, despite the choice of the requlrizer. However, A_2 can be further bounded for different hypothesis spaces corresponding to different choice of regularization functions. In the following we start with a useful lemma which helps bounding A_2 for the group-norm hypothesis space (5.19). The proof of this Lemma, which is based on the application of the Khintchine (C.2) and Rosenthal (C.3) inequalities, is presented in Appendix C.

Lemma 6. *Assume that the kernels in* (5.10) *are uniformly bounded, that is* $\|k\|_\infty \leq \mathcal{K} \leq \infty$. *Then, for the group norm regularizer* $\frac{1}{2}\|\boldsymbol{W}\|_{2,q}^2$ *in* (5.19) *and for any* $1 \leq q \leq 2$, *the expectation*

$\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_{2,q^*}$ *for* $\boldsymbol{D}=\boldsymbol{I}$ *can be upper-bounded as*

$$\mathbb{E}_{X,\sigma}\left\|\left(\sum_{j>h_t}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\|_{2,q^*}\leq\frac{\sqrt{\mathcal{K}e}q^*T^{\frac{1}{q^*}}}{n}+\sqrt{\frac{eq^{*2}}{n}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}.$$

Remark 41. *Similarly as in Lemma 6, one can easily prove that*

$$\mathbb{E}_{X,\sigma}\left\|\left(\sum_{j>h_t}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\|_{2,q^*}^2\leq\frac{\mathcal{K}eq^{*2}T^{\frac{2}{q^*}}}{n^2}+\frac{eq^{*2}}{n}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}.\tag{5.20}$$

To see this, note that in the first step of the proof of Lemma 6 (see Appendix C), by replacing the outermost exponent $\frac{1}{q^*}$ *with* $\frac{2}{q^*}$*, and following the same procedure, one can verify* (5.20)*. Therefore, it can be concluded that very similar LRC bounds can be obtained via Theorem 36 and Theorem 38.*

Corollary 42. *Using Theorem 38, for any* $1\leq q\leq 2$*, the LRC of function class* $\mathcal{F}_q$ *in* (5.19) *can be bounded as*

$$\mathfrak{R}(\mathcal{F}_q,r)\leq\sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{q^*}},\frac{2eq^{*2}R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}+\frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}.\tag{5.21}$$

Proof Sketch: The proof of the corollary uses the result of Lemma 6 to upper bound A_2 for the group-norm hypothesis space (5.19) as,

$$A_2(\mathcal{F}_q)\leq\sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}+\frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}.\tag{5.22}$$

Now, combining (5.15) and (5.22) provides the bound on $\mathfrak{R}(\mathcal{F}_q, r)$ as

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{r\sum_{t=1}^{T} h_t}{nT}} + \sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}, \tag{5.23}$$

Then using inequalities shown below which hold for any $\alpha_1, \alpha_2 > 0$, any non-negative vectors $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^T$, any $0 \leq q \leq p \leq \infty$ and any $s \geq 1$,

$$(\star) \quad \sqrt{\alpha_1} + \sqrt{\alpha_2} \leq \sqrt{2(\alpha_1 + \alpha_2)} \tag{5.24}$$

$$(\star\star) \quad l_p - to - l_q: \quad \|\boldsymbol{a}_1\|_q = \langle \mathbf{1}, \boldsymbol{a}_1 \rangle^{\frac{1}{q}} \overset{\text{Hölder's}}{\leq} \left(\|\mathbf{1}\|_{(p/q)^*} \|\boldsymbol{a}_1^q\|_{(p/q)}\right)^{\frac{1}{q}} = T^{\frac{1}{q}-\frac{1}{p}} \|\boldsymbol{a}_1\|_p \tag{5.25}$$

$$(\star\star\star) \quad \|\boldsymbol{a}_1\|_s + \|\boldsymbol{a}_2\|_s \leq 2^{1-\frac{1}{s}} \|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_s \leq 2\|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_s, \tag{5.26}$$

we can obtain the desired result. See Appendix C for the detailed proof.

Remark 43. *Since the LRC bound above is not monotonic in q it is more reasonable to state the above bound in terms of $\kappa \geq q$; choosing $\kappa = q$ is not always the optimal choice. Trivially, for the group norm regularizer with any $\kappa \geq q$, it holds that $\|\boldsymbol{W}\|_{2,\kappa} \leq \|\boldsymbol{W}\|_{2,q}$ and therefore $\mathfrak{R}(\mathcal{F}_q, r) \leq \mathfrak{R}(\mathcal{F}_\kappa, r)$. Thus, we have the following bound on $\mathfrak{R}(\mathcal{F}_q, r)$ for any $\kappa \in [q, 2]$,*

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{\kappa^*}}, \frac{2e\kappa^{*2}R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{\kappa^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}\kappa^*T^{\frac{1}{\kappa^*}}}{nT}. \tag{5.27}$$

Remark 44 (Sparsity-inducing group-norm)**.** *Assuming a sparse representations shared across multiple tasks is a well-known presumption in MTL [7, 5] which leads to the use of group norm regularizer $\frac{1}{2}\|\boldsymbol{W}\|_{2,1}^2$. Notice that for any $\kappa \geq 1$, it holds that $\mathfrak{R}(\mathcal{F}_1, r) \leq \mathfrak{R}(\mathcal{F}_\kappa, r)$. Also, assuming an identical tail sum $\sum_{j\geq h}\lambda^j$ for all tasks, reduces the bound in* (5.27) *to the function $\kappa^* \mapsto \kappa^* T^{1/\kappa^*}$ in terms of κ. This function attains its minimum at $\kappa^* = \log T$. Thus, by choosing*

$\kappa^* = \log T$ *it is easy to show:*

$$\mathfrak{R}(\mathcal{F}_1, r) \leq \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty} \min\left(rT^{1-\frac{2}{\kappa^*}}, \frac{2e\kappa^{*2}R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{\kappa^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}\kappa^* T^{\frac{1}{\kappa^*}}}{nT}$$
$$\overset{(l_{\frac{\kappa^*}{2}}-to-l_\infty)}{\leq} \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty} \min\left(rT, \frac{2e^3(\log T)^2R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\infty}} + \frac{\sqrt{2\mathcal{K}}R_{max}e^{\frac{3}{2}}\log T}{nT}.$$

Remark 45 ($L_{2,q}$ Group-norm regularizer with $q \geq 2$)**.** *For any $q \geq 2$, Theorem 38 provides a LRC bound for the function class $\mathcal{F}_q$ in* (5.19) *as*

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty} \min\left(rT^{1-\frac{2}{q^*}}, \frac{2R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}, \tag{5.28}$$

where $q^* := \frac{q}{q-1}$.

Proof.

$$
\begin{aligned}
A_2(\mathcal{F}_q) &\overset{\text{Hölder's}}{\leq} \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}_q}\|\boldsymbol{W}\|_{2,q}\|\boldsymbol{V}\|_{2,q^*}\right\} \\
&\leq \frac{\sqrt{2}R_{max}}{T}\mathbb{E}_{X,\sigma}\left(\sum_{t=1}^{T}\left\|\sum_{j>h_t}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right\|^{q^*}\right)^{\frac{1}{q^*}} \\
&\overset{\text{Jensen's}}{\leq} \frac{\sqrt{2}R_{max}}{T}\left(\sum_{t=1}^{T}\left(\mathbb{E}_{X,\sigma}\left\|\sum_{j>h_t}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right\|^2\right)^{\frac{q^*}{2}}\right)^{\frac{1}{q^*}} \\
&= \frac{\sqrt{2}R_{max}}{T}\left(\sum_{t=1}^{T}\left(\sum_{j>h_t}\mathbb{E}_{X,\sigma}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle^2\right)^{\frac{q^*}{2}}\right)^{\frac{1}{q^*}} \\
&= \frac{\sqrt{2}R_{max}}{T}\left(\sum_{t=1}^{T}\left(\sum_{j>h_t}\frac{\lambda_t^j}{n}\right)^{\frac{q^*}{2}}\right)^{\frac{1}{q^*}} = \sqrt{\frac{2R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}
\end{aligned}
$$

By applying (5.24), (5.25) and (5.26), this last result together with the bound in (5.15) for A_1, yields the result. □

To investigate the tightness of the bound in (5.21), we derive the lower bound which holds for the LRC of $\mathcal{F}_q$ with any $q \geq 1$. The proof of the result can be found in Appendix C.

Theorem 46 (Lower bound)**.** *The following lower bound holds for the local Rademacher complexity of $\mathcal{F}_q$ in* (5.21) *with any $q \geq 1$. There is an absolute constant c so that $\forall t$, if $\lambda_t^1 \geq 1/(nR_{max}^2)$ then for all $r \geq \frac{1}{n}$ and $q \geq 1$,*

$$
\mathfrak{R}(\mathcal{F}_{q,R_{max},T},r) \geq \sqrt{\frac{c}{nT^{1-\frac{2}{q^*}}}\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{q^*}},\frac{R_{max}^2}{T}\lambda_1^j\right)}. \tag{5.29}
$$

A comparison between the lower bound in (5.29) and the upper bound in (5.21) can be clearly

illustrated by assuming identical eigenvalue tail sums $\sum_{j\geq\infty}\lambda_t^j$ for all tasks, for which the upper bound translates to

$$\mathfrak{R}(\mathcal{F}_{q,R_{max},T},r)\leq\sqrt{\frac{4}{nT^{1-\frac{2}{q^*}}}\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{q^*}},\frac{2eq^{*2}R_{max}^2}{T}\lambda_t^j\right)}+\frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}.$$

By comparing this to (5.29), we see that the lower bound matches the upper bound up to constants. The same analysis for MTL models with Schatten norm and graph regularizers yields similar results confirming that the LRC upper bounds that we have obtained are reasonably tight.

Remark 47. *It is worth pointing out that a matching lower bound on the local Rademacher complexity does not necessarily imply a tight bound on the expectation of an empirical minimizer. As it has been shown in Section 4 of [10], by direct analysis of the empirical minimizer, sharper bounds than the LRC-based bounds can be obtained. Consequently, based on Theorem 8 in [10], there might be cases in which the local Rademacher complexity bounds are constants, however $P\hat{f}$ is of some order depending on the number of samples n—O(1/n))—which decreases with n growing. As it has pointed out in that paper, under some mild conditions on the loss function ℓ, a similar argument also holds for the class of loss functions $\{\ell_f-\ell_{f^*}:f\in\mathcal{F}\}$.*

Schatten Norm Regularized MTL

In [7], the authors developed a spectral regularization framework for MTL where the L_{S_q}-Schatten norm $\frac{1}{2}\|\boldsymbol{W}\|_{S_q}^2:=\frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{W}^T\boldsymbol{W}\right)^{\frac{q}{2}}\right]^{\frac{2}{q}}$ is studied as a concrete example, corresponding to performing ERM in the following hypothesis space:

$$\mathcal{F}_{S_q}:=\left\{X\mapsto\left[\langle\boldsymbol{w}_1,\phi(X_1)\rangle,\ldots,\langle\boldsymbol{w}_T,\phi(X_T)\rangle\right]^T:\frac{1}{2}\|\boldsymbol{W}\|_{S_q}^2\leq R_{max}'^2\right\}. \tag{5.30}$$

Corollary 48. *For any* $1 \leq q \leq 2$ *in* (5.30)*, the LRC of function class* $\mathcal{F}_{S_q}$ *is bounded as*

$$\mathfrak{R}(\mathcal{F}_{S_q}, r) \leq \sqrt{\frac{4}{nT} \Big\| \Big(\sum_{j=1}^{\infty} \min\Big(r, \frac{2q^* R_{max}'^2}{T} \lambda_t^j\Big)\Big)_{t=1}^{T} \Big\|_1}.$$

The proof is provided in Appendix C.

Remark 49 (Sparsity-inducing Schatten-norm (trace norm))**.** *Trace-norm regularized MTL, corresponding to Schatten norm regularization with* $q = 1$ *[104, 119], imposes a low-rank structure on the spectrum of* $\boldsymbol{W}$ *and can also be interpreted as low dimensional subspace learning [6, 77, 70]. Note that for any* $q \geq 1$*, it holds that* $\mathfrak{R}(\mathcal{F}_{S_1}, r) \leq \mathfrak{R}(\mathcal{F}_{S_q}, r)$*. Therefore, choosing the optimal* $q^* = 1$*, we get*

$$\mathfrak{R}(\mathcal{F}_{S_1}, r) \leq \sqrt{\frac{4}{nT} \Big\| \Big(\sum_{j=1}^{\infty} \min\Big(r, \frac{2R_{max}'^2}{T} \lambda_t^j\Big)\Big)_{t=1}^{T} \Big\|_1}.$$

Remark 50 (L_{S_q} Schatten-norm regularizer with $q \geq 2$)**.** *For any* $q \geq 2$*, Theorem 38 provides a LRC bound for the function class* $\mathcal{F}_{S_q}$ *in* (5.30) *as*

$$\mathfrak{R}(\mathcal{F}_{S_q}, r) \leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min\left(r, \frac{2R_{max}'^2}{T} \lambda_t^j\right)\right)_{t=1}^{T} \right\|_1}. \tag{5.31}$$

Proof. Taking $q^* := \frac{q}{q-1}$, we first bound the expectation $\mathbb{E}_{X,\sigma} \|\boldsymbol{V}\|_{S_{q^*}}$. Take $\boldsymbol{U}_t^i$ as a matrix with T columns where the only non-zero column t of $\boldsymbol{U}_t^i$ is defined as $\sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j$. Based on the definition of $\boldsymbol{V} = \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^{T}$, we can then provide a bound for

this expectation as

$$\begin{aligned}
\mathbb{E}_{X,\sigma}\left\|\boldsymbol{V}\right\|_{S_{q^*}} &= \mathbb{E}_{X,\sigma}\left\|\left(\sum_{j>h_t}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\|_{S_{q^*}} \\
&= \mathbb{E}_{X,\sigma}\left\|\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i\boldsymbol{U}_t^i\right\|_{S_{q^*}} \\
&\overset{\text{Jensen}}{\leq}\left(\text{tr}\left(\sum_{t,s=1}^{T}\sum_{i,j=1}^{n}\mathbb{E}_{X,\sigma}\left(\sigma_t^i\sigma_s^j{\boldsymbol{U}_t^i}^T\boldsymbol{U}_s^j\right)\right)^{\frac{q^*}{2}}\right)^{\frac{1}{q^*}} \\
&= \left(\text{tr}\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}_X\left({\boldsymbol{U}_t^i}^T\boldsymbol{U}_t^i\right)\right)^{\frac{q^*}{2}}\right)^{\frac{1}{q^*}} \\
&= \left(\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j>h_t}\mathbb{E}_X\left\langle\frac{1}{n}\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle^2\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n}\sum_{t=1}^{T}\sum_{j>h_t}\lambda_t^j\right)^{\frac{1}{2}} = \sqrt{\frac{1}{n}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_1}
\end{aligned}$$

Note that replacing this into (5.17), and with the help of (5.24), (5.25) and (5.26), one can conclude the result. □

Graph Regularized MTL

The idea underlying graph regularized MTL is to force the models of related tasks to be close to each other, by penalizing the squared distance $\|\boldsymbol{w}_t-\boldsymbol{w}_s\|^2$ with different weights ω_{ts}. We consider the following MTL graph regularizer [102]

$$\Omega(\boldsymbol{W})=\frac{1}{2}\sum_{t=1}^{T}\sum_{s=1}^{T}\omega_{ts}\|\boldsymbol{w}_t-\boldsymbol{w}_s\|^2+\eta\sum_{t=1}^{T}\|\boldsymbol{w}_t\|^2=\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\boldsymbol{L}+\eta\boldsymbol{I}\right)_{ts}\left\langle\boldsymbol{w}_t,\boldsymbol{w}_s\right\rangle,$$

where $\boldsymbol{L}$ is the graph-Laplacian associated to a matrix of edge-weights ω_{ts}, $\boldsymbol{I}$ is the identity operator, and $\eta > 0$ is a regularization parameter. According to the identity $\sum_{t=1}^{T}\sum_{s=1}^{T}\big(\boldsymbol{L}+\eta\boldsymbol{I}\big)_{ts}\big\langle \boldsymbol{w}_t, \boldsymbol{w}_s\big\rangle = \|(\boldsymbol{L}+\eta\boldsymbol{I})^{1/2}\boldsymbol{W}\|_F^2$, the corresponding hypothesis space is:

$$\mathcal{F}_G := \Big\{X \mapsto [\Big\langle \boldsymbol{w}_1, \phi(X_1)\Big\rangle, \ldots, \Big\langle \boldsymbol{w}_T, \phi(X_T)\Big\rangle]^T : \frac{1}{2}\|\boldsymbol{D}^{1/2}\boldsymbol{W}\|_F^2 \leq R_{max}''^2\Big\}. \tag{5.32}$$

where we define $\boldsymbol{D} := \boldsymbol{L} + \eta\boldsymbol{I}$.

Corollary 51. *For any given positive definite matrix $\boldsymbol{D}$ in* (5.32)*, the LRC of $\mathcal{F}_G$ is bounded by*

$$\mathfrak{R}(\mathcal{F}_G, r) \leq \sqrt{\frac{4}{nT}\Big\|\Big(\sum_{j=1}^{\infty}\min\Big(r, \frac{2\boldsymbol{D}_{tt}^{-1}R_{max}''^2}{T}\lambda_t^j\Big)\Big)_{t=1}^{T}\Big\|_1}. \tag{5.33}$$

where $\big(\boldsymbol{D}_{tt}^{-1}\big)_{t=1}^{T}$ are the diagonal elements of $\boldsymbol{D}^{-1}$.

See Appendix C for the proof.

Excess Risk Bounds for Norm Regularized MTL Models

In this section we will provide the distribution and data-dependent excess risk bounds for the hypothesis spaces considered earlier. Note that the proofs are provided only for the hypothesis space $\mathcal{F}_q$ with $q \in [1, 2]$ in (5.19). However, in the cases involving the $L_{2,q}$-group norm with $q \geq 2$, as well as the L_{S_q}-Schatten and graph norms, the proofs can be obtained in a very similar way. More specifically, by using the LRC bounds of Remark 45, Corollary 48, Remark 50 and Corollary 51, one can follow the same steps of the proofs of this section to arrive at the results pertaining to these cases.

Theorem 52. (Distribution-dependent excess risk bound for a $L_{2,q}$ group-norm regularized MTL) *Assume that $\mathcal{F}_q$ in* (5.19) *is a convex class of functions with ranges in $[-b, b]$, and let the loss*

function ℓ of Problem (5.10) *be such that Assumption 32 is satisfied. Let $\hat{\boldsymbol{f}}$ be any element of $\mathcal{F}_q$ with $1 \leq q \leq 2$ which satisfies $P_n \ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f} \in \mathcal{F}_q} P_n \ell_{\boldsymbol{f}}$. Assume moreover that k is a positive semi-definite kernel on $\mathcal{X}$ such that $\|k\|_\infty \leq \mathcal{K} \leq \infty$. Denote by r^* the fixed point of $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$. Then, for any $\boldsymbol{f} \in \mathcal{F}$, $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$, the excess loss of function class $\mathcal{F}_q$ is bounded as*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 32Kr^* + \frac{(48Lb + 16BK)Bx}{nT}, \tag{5.34}$$

where for the fixed point r^* of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$, it holds that

$$r^* \leq \min_{0 \leq h_t \leq \infty} \frac{B^2 \sum_{t=1}^T h_t}{Tn} + 4BL \sqrt{\frac{2eq^{*2} R_{max}^2}{nT^2} \left\| \left(\sum_{j > h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{4\sqrt{2\mathcal{K}e} R_{max} q^* T^{\frac{1}{q^*}}}{nT}, \tag{5.35}$$

where $h_1, \ldots, h_T$ are arbitrary non-negative integers.

Proof. First notice that $\mathcal{F}_q$ is convex, thus it is star-shaped around any of its points. Hence according to Lemma 10, $\mathfrak{R}(\mathcal{F}_q, r)$ is a sub-root function. Moreover, because of the symmetry of σ_t^i and because $\mathcal{F}_q$ is convex and symmetric, it can be shown that $\mathfrak{R}(\mathcal{F}_q^*, r) \leq 2\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$, where $\mathfrak{R}(\mathcal{F}_q^*, r)$ is defined according to (5.6) for the class of functions $\mathcal{F}_q$. Therefore, it suffices to find the fixed point of $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$ by solving $\phi(r) = r$. For this purpose, we will use (5.23) as a bound for $\mathfrak{R}(\mathcal{F}_q, r)$, and solve $\sqrt{\alpha r} + \gamma = r$ (or equivalently $r^2 - (\alpha + 2\gamma)r + \gamma^2 = 0$) for r, where we define

$$\alpha = \frac{B^2 \sum_{t=1}^T h_t}{Tn}, \text{ and } \gamma = 2BL \sqrt{\frac{2eq^{*2} R_{max}^2}{nT^2} \left\| \left(\sum_{j > h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{2\sqrt{2\mathcal{K}e} R_{max} BLq^* T^{\frac{1}{q^*}}}{nT}. \tag{5.36}$$

It is not hard to verify that $r^* \leq \alpha + 2\gamma$. Substituting the definition of α and γ gives the result. $\square$

Remark 53. *If the conditions of Theorem 38 hold, then it can be shown that the following results hold for the fixed point of the considered hypothesis spaces in* (5.19)*,* (5.30) *and* (5.32)*.*

- *Group norm: For the fixed point r^* of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$ with any $1 \leq q \leq 2$ in* (5.19)*, it holds*

$$r^* \leq \min_{0 \leq h_t \leq \infty} \frac{B^2 \sum_{t=1}^{T} h_t}{Tn} + 4BL \sqrt{\frac{2eq^{*2} R_{max}^2}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}} + \frac{4\sqrt{2\mathcal{K}e} R_{max} q^* BLT^{\frac{1}{q^*}}}{nT}. \tag{5.37}$$

  Also, for any $q \geq 2$ in (5.19)*, it holds*

$$r^* \leq \min_{0 \leq h_t \leq \infty} \frac{B^2 \sum_{t=1}^{T} h_t}{Tn} + 4BL \sqrt{\frac{2R_{max}^2}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}}. \tag{5.38}$$

- *Schatten-norm: For the fixed point r^* of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_{S_q}, \frac{r}{4L^2})$ with any $1 \leq q \leq 2$ in* (5.30)*, it holds*

$$r^* \leq \min_{0 \leq h_t \leq \infty} \frac{B^2 \sum_{t=1}^{T} h_t}{Tn} + 4BL \sqrt{\frac{2q^* R_{max}'^2}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_1}. \tag{5.39}$$

  Also, for any $q \geq 2$ in (5.30)*, it holds*

$$r^* \leq \min_{0 \leq h_t \leq \infty} \frac{B^2 \sum_{t=1}^{T} h_t}{Tn} + 4BL \sqrt{\frac{2R_{max}'^2}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_1}. \tag{5.40}$$

- *Graph regularizer: For the fixed point r^* of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_G, \frac{r}{4L^2})$*

with any positive operator $\boldsymbol{D}$ *in* (5.32)*, it holds*

$$r^* \leq \min_{0\leq h_t\leq\infty} \frac{B^2\sum_{t=1}^T h_t}{Tn} + 4BL\sqrt{\frac{2R_{max}''^2}{nT^2}\left\|\left(\boldsymbol{D}_{tt}^{-1}\sum_{j>h_t}\lambda_t^j\right)_{t=1}^T\right\|_1}. \tag{5.41}$$

Regarding the fact that λ_t^js are decreasing with respect to j, we can assume $\exists d_t : \lambda_t^j \leq d_t j^{-\alpha_t}$ for some $\alpha_t > 1$. As examples, this assumption holds for finite rank kernels as well as convolution kernels. Thus, it can be shown that

$$\sum_{j>h_t}\lambda_t^j \leq d_t\sum_{j>h_t} j^{-\alpha_t} \leq d_t\int_{h_t}^{\infty} x^{-\alpha_t}dx = d_t\left[\frac{1}{1-\alpha_t}x^{1-\alpha_t}\right]_{h_t}^{\infty} = -\frac{d_t}{1-\alpha_t}h_t^{1-\alpha_t}. \tag{5.42}$$

Note that via $l_p - to - l_q$ conversion inequality in (5.25), for $p = 1$ and $q = \frac{q*}{2}$, we have

$$\frac{B^2\sum_{t=1}^T h_t}{Tn} \leq B\sqrt{\frac{B^2T\sum_{t=1}^T h_t^2}{n^2T^2}} \overset{(\star\star)}{\leq} B\sqrt{\frac{B^2T^{2-\frac{2}{q^*}}\left\|(h_t^2)_{t=1}^T\right\|_{\frac{q^*}{2}}}{n^2T^2}}.$$

Now, applying, (5.24) and (5.26), and inserting (5.42) into (5.35), it holds for group norm regularized MTL with $1 \leq q \leq 2$,

$$r^* \leq \min_{0\leq h_t\leq\infty} 2B\sqrt{\left\|\left(\frac{B^2T^{2-\frac{2}{q^*}}h_t^2}{n^2T^2} - \frac{32d_teq^{*2}R_{max}^2L^2}{nT^2(1-\alpha_t)}h_t^{1-\alpha_t}\right)_{t=1}^T\right\|_{\frac{q^*}{2}}} + \frac{4\sqrt{2\mathcal{K}e}R_{max}BLq^*T^{\frac{1}{q^*}}}{nT}. \tag{5.43}$$

Taking the partial derivative of the above bound with respect to h_t and setting it to zero yields the optimal h_t as

$$h_t = \left(16d_teq^{*2}R_{max}^2B^{-2}L^2T^{\frac{2}{q^*}-2}n\right)^{\frac{1}{1+\alpha_t}}.$$

Note that substituting the above for $\alpha := \min_{t\in\mathbb{N}_T} \alpha_t$ and $d = max_{t\in\mathbb{N}_T} d_t$ into (5.43), we can upper-bound the fixed point of r^* as

$$r^* \leq \frac{14B^2}{n}\sqrt{\frac{\alpha+1}{\alpha-1}}\left(dq^{*2}R_{max}^2 B^{-2}L^2 T^{\frac{2}{q^*}-2}n\right)^{\frac{1}{1+\alpha}} + \frac{10\sqrt{\mathcal{K}}R_{max}BLq^*T^{\frac{1}{q^*}}}{nT},$$

which implies that

$$r^* = O\left(\left(\frac{T^{1-\frac{1}{q^*}}}{q^*}\right)^{\frac{-2}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right).$$

It can be seen that the convergence rate can be as slow as $O\left(\frac{q^*T^{1/q^*}}{T\sqrt{n}}\right)$ (for small α, where at least one $\alpha_t \approx 1$), and as fast as $O(n^{-1})$ (when $\alpha_t \to \infty$, for all t). The bound obtained for the fixed point together with Theorem 52 provides a bound for the excess risk, which leads to the following remark.

Remark 54 (Excess risk bounds for selected norm regularized MTL problems)**.** *Assume that $\mathcal{F}_q$, $\mathcal{F}_{S_q}$ and $\mathcal{F}_G$ are convex classes of functions with ranges in $[-b, b]$, and let the loss function ℓ of Problem* (5.10) *be such that Assumption 32 are satisfied. Assume moreover that k is a positive semidefinite kernel on $\mathcal{X}$ such that $\|k\|_\infty \leq \mathcal{K} \leq \infty$. Also, denote $\alpha := \min_{t\in\mathbb{N}_T} \alpha_t$ and $d = max_{t\in\mathbb{N}_T} d_t$.*

- *Group norm: Assume that $\hat{\boldsymbol{f}}$ satisfies $P_n\ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f}\in\mathcal{F}_q} P_n\ell_{\boldsymbol{f}}$, and r^* is the fixed point of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$ with $1 \leq q \leq 2$ in* (5.19)*. Then, for any $\boldsymbol{f} \in \mathcal{F}_q$, $K > 1$ and $x > 0$, it holds with probability at least $1 - e^{-x}$,*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq \min_{\kappa\in[q,2]} 448K\sqrt{\frac{\alpha+1}{\alpha-1}}\left(d\kappa^{*2}R_{max}^2L^2\right)^{\frac{1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}}\left(T^{\frac{2}{\kappa}}\right)^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} + \frac{320\sqrt{\mathcal{K}}R_{max}BKL\kappa^*T^{\frac{1}{\kappa^*}}}{nT} + \frac{(48Lb+16BK)Bx}{nT}. \quad (5.44)$$

Also, for any $K > 1$, $x > 0$ and $\boldsymbol{f} \in \mathcal{F}_q$ with $q \geq 2$ in (5.19)*, it holds with probability at least $1 - e^{-x}$,*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq \min_{q \in [2,\infty]} 256K\sqrt{\frac{\alpha+1}{\alpha-1}} \left(dR_{max}^2 L^2\right)^{\frac{1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} \left(T^{\frac{2}{q}}\right)^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb + 16BK)Bx}{nT}. \tag{5.45}$$

- *Schatten-norm: Assume that $\hat{\boldsymbol{f}}$ satisfies $P_n \ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f} \in \mathcal{F}_{S_q}} P_n \ell_{\boldsymbol{f}}$, and r^* is the fixed point of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_{S_q}, \frac{r}{4L^2})$ with $1 \leq q \leq 2$ in* (5.30)*. Then, for any $\boldsymbol{f} \in \mathcal{F}_{S_q}$, $K > 1$, $x > 0$, it holds with probability at least $1 - e^{-x}$,*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq \min_{q \in [1,2]} 256K\sqrt{\frac{\alpha+1}{\alpha-1}} \left(dq^* R_{max}'^2 L^2\right)^{\frac{1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb + 16BK)Bx}{nT}. \tag{5.46}$$

Also, for any $K > 1$, $x > 0$ and $\boldsymbol{f} \in \mathcal{F}_{S_q}$ with $q \geq 2$ in (5.30)*, it holds with probability at least $1 - e^{-x}$,*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 256K\sqrt{\frac{\alpha+1}{\alpha-1}} \left(dR_{max}'^2 L^2\right)^{\frac{1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb + 16BK)Bx}{nT}. \tag{5.47}$$

- *Graph regularizer: Assume that $\hat{\boldsymbol{f}}$ satisfies $P_n \ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f} \in \mathcal{F}_G} P_n \ell_{\boldsymbol{f}}$, and r^* is the fixed point of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_G, \frac{r}{4L^2})$ with any positive operator $\boldsymbol{D}$ in* (5.32)*. Then, for any $\boldsymbol{f} \in \mathcal{F}_G$, $K > 1$ and $x > 0$, it holds with probability at least $1 - e^{-x}$,*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 256K\sqrt{\frac{\alpha+1}{\alpha-1}} \left(dR_{max}''^2 L^2 \boldsymbol{D}_{max}^{-1}\right)^{\frac{1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb + 16BK)Bx}{nT}. \tag{5.48}$$

where $\boldsymbol{D}_{max}^{-1} := \max_{t\in\mathbb{N}_T} \boldsymbol{D}_{tt}^{-1}$.

Corollary 55. (Data-dependent excess risk bound for a MTL problem with a $L_{2,q}$ group-norm regularizer) *Assume the convex class $\mathcal{F}_q$ in* (5.19) *has ranges in $[-b, b]$, and let the loss function ℓ in Problem* (5.10) *be such that Assumption 32 are satisfied. Let $\hat{\boldsymbol{f}}$ be any element of $\mathcal{F}_q$ with $1 \leq q \leq 2$ which satisfies $P_n\ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f}\in\mathcal{F}_q} P_n\ell_{\boldsymbol{f}}$. Assume moreover that k is a positive semidefinite kernel on $\mathcal{X}$ such that $\|k\|_\infty \leq \mathcal{K} \leq \infty$. Let $\boldsymbol{K}_t$ be the $n \times n$ normalized Gram matrix (or kernel matrix) of task t with entries $(\boldsymbol{K}_t)_{ij} := \frac{1}{n}k(X_t^i, X_t^j) = \frac{1}{n}\left\langle \hat{\phi}(X_t^i), \hat{\phi}(X_t^j)\right\rangle$. Let $\hat{\lambda}_t^1, \dots \hat{\lambda}_t^n$ be the ordered eigenvalues of matrix $\boldsymbol{K}_t$, and $\hat{r}^*$ be the fixed point of*

$$\hat{\psi}_n(r) = c_1\hat{\mathfrak{R}}(\mathcal{F}_q^*, c_3 r) + \frac{c_2 x}{nT},$$

where $c_1 = 2L\max(B, 16Lb)$, $c_2 = 128L^2b^2 + 2bc_1$ *and* $c_3 = 4 + 128BK + 4B^2(48Lb + 16BK)/c_2$, *and*

$$\hat{\mathfrak{R}}(\mathcal{F}_q^*, c_3 r) := \mathbb{E}_\sigma\left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F}_q,\\ L^2P_n\left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)^2\leq c_3 r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\,\middle|\, \left\{x_t^i\right\}_{t\in\mathbb{N}_T, i\in\mathbb{N}_n}\right]. \tag{5.49}$$

Then, for any $\boldsymbol{f} \in \mathcal{F}$, $K > 1$ and $x > 0$, with probability at least $1 - 4e^{-x}$ the excess loss of function class $\mathcal{F}_q$ is bounded as

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 32K\hat{r}^* + \frac{(48Lb + 16BK)Bx}{nT}, \tag{5.50}$$

where for the fixed point $\hat{r}^*$ of the empirical local Rademacher complexity $\hat{\psi}_n(r)$, it holds

$$\hat{r}^* \leq \frac{c_1^2 c_3 \sum_{t=1}^{T}\hat{h}_t}{nTL^2} + 4\sqrt{\frac{2c_1^2 q^{*2} R_{max}^2}{nT^2}\left\|\left(\sum_{j>\hat{h}_t}^{n}\hat{\lambda}_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{2c_2 x}{nT},$$

where $\hat{h}_1, \ldots, \hat{h}_T$ are arbitrary non-negative integers, and $(\hat{\lambda}_t^j)_{j=1}^n$ are eigenvalues of the normalized Gram matrix $\boldsymbol{K}$ obtained from kernel function k.

Proof. The proof of the result is provided in Appendix D. □

Discussion

Global vs. Local Rademacher Complexity Bounds

This section is devoted to compare the excess risk bounds based on local Rademacher complexity to those of the global ones.

First, note that to obtain the GRC-based bounds, we apply Theorem 16 of [101], as we consider the same setting and assumptions for tasks' distributions as considered in this work. This theorem presents an MTL bound based on the notion of GRC.

Theorem 56 (MTL excess risk bound based on GRC; Theorem 16 of [101])**.** *Let the vector-valued function class $\mathcal{F}$ be defined as $\mathcal{F} := \{\boldsymbol{f} = (f_1, \ldots, f_T) : \mathcal{X} \mapsto [-b, b]^T\}$. Assume that $X = (X_i^t)_{(i,t)=(1,1)}^{(n,T)}$ is a vector of independent random variables where for all fixed t, $X_1^t, \ldots, X_n^t$ are identically distributed according to P_t. Let the loss function ℓ be L-Lipschitz in its first argument. Then for any $\boldsymbol{f} \in \mathcal{F}$ and $x > 0$, with probability at least $1 - e^{-x}$,*

$$P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) \leq P_n(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) + 2L\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{2Lbx}{nT}}. \tag{5.51}$$

Proof. As it has been shown in [101], the proof of this theorem is based on using McDiarmid's inequality for Z defined in Theorem 28, and noticing that for the function class $\mathcal{F}$ with values in $[-b, b]$, it holds that $|Z - Z_{s,j}| \leq 2b/nT$. □

It can be observed that, in order to obtain the excess risk bound in the above theorem, one has to bound the GRC term $\mathfrak{R}(\mathcal{F})$ in (5.51). Therefore, we first upper-bound the GRC of different hypothesis spaces considered in the previous sections.

Theorem 57 (Distribution-dependent GRC bounds)**.** *Assume that the conditions of Theorem 36 hold. Then, the following results hold for the GRC of the hypothesis spaces in* (5.19)*,* (5.30) *and* (5.32)*, respectively.*

- *Group-norm regularizer: For any* $1 \leq q \leq 2$ *in* (5.19)*, the GRC of the function class* $\mathcal{F}_q$ *can be bounded as*

$$\forall \kappa \in [q, 2]: \quad \mathfrak{R}(\mathcal{F}_q) \leq \sqrt{\frac{2e\kappa^{*2} R_{max}^2}{nT^2} \left\| (\boldsymbol{tr}\left(J_t\right))_{t=1}^T \right\|_{\frac{\kappa^*}{2}}} + \frac{\sqrt{2\mathcal{K}e} R_{max} \kappa^* T^{\frac{1}{\kappa^*}}}{nT}. \tag{5.52}$$

  Also, for any $q \geq 2$ *in* (5.19)*, the GRC of the function class* $\mathcal{F}_q$ *can be bounded as*

$$\mathfrak{R}(\mathcal{F}_q) \leq \sqrt{\frac{2R_{max}^2}{nT^2} \left\| (\boldsymbol{tr}\left(J_t\right))_{t=1}^T \right\|_{\frac{q^*}{2}}}. \tag{5.53}$$

- *Schatten-norm regularizer: For any* $1 \leq q \leq 2$ *in* (5.30)*, the GRC of the function class* $\mathcal{F}_{S_q}$ *can be bounded as*

$$\mathfrak{R}(\mathcal{F}_{S_q}) \leq \sqrt{\frac{2q^* R_{max}'^2}{nT^2} \left\| (\boldsymbol{tr}\left(J_t\right))_{t=1}^T \right\|_1}. \tag{5.54}$$

  Also, for any $q \geq 2$ *in* (5.30)*, the GRC of the function class* $\mathcal{F}_{S_q}$ *can be bounded as*

$$\mathfrak{R}(\mathcal{F}_{S_q}) \leq \sqrt{\frac{2R_{max}'^2}{nT^2} \left\| (\boldsymbol{tr}\left(J_t\right))_{t=1}^T \right\|_1}. \tag{5.55}$$

- *Graph regularizer: For any positive operator* $\boldsymbol{D}$ *in* (5.32)*, the GRC of the function class* $\mathcal{F}_G$

can be bounded as

$$\mathfrak{R}(\mathcal{F}_G) \leq \sqrt{\frac{2R_{max}'^2}{nT^2} \left\| \left(\boldsymbol{D}_{tt}^{-1} \boldsymbol{tr}(J_t) \right)_{t=1}^{T} \right\|_1}. \tag{5.56}$$

where for the covariance operator $J_t = \mathbb{E}(\phi(X_t) \otimes \phi(X_t)) = \sum_{j=1}^{\infty} \lambda_t^j \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j$*, the* trace $\boldsymbol{tr}(J_t)$ *is defined as*

$$\boldsymbol{tr}(J_t) := \sum_j \left\langle J_t \boldsymbol{u}_t^j, \boldsymbol{u}_t^j \right\rangle = \sum_{j=1}^{\infty} \lambda_t^j.$$

Proof. The proof of the results can be found in Appendix D. □

Notice that, assuming a unique bound for the traces of all tasks' kernels, the bound in (5.52) is determined by $O\left(\frac{q^* T^{\frac{1}{q^*}}}{T\sqrt{n}}\right)$. Also, taking $q^* = \log T$, we obtain a bound of order $O\left(\frac{\log T}{T\sqrt{n}}\right)$. We can also remark that, when the kernel traces are bounded, the bounds in (5.53), (5.54), (5.55) and (5.56) are of the order of $O\left(\frac{1}{\sqrt{nT}}\right)$.

Note that for the purpose of comparison, we concentrate only on the parameters R, n, T, q^* and α and assume all the other parameters are fixed and hidden in the big-O notation. Also, for the sake of simplicity, we assume that the eigenvalues of all tasks satisfy $\lambda_t^j \leq dj^{-\alpha}$ (with $\alpha > 1$). Note that from Theorem 56, it follows that a bound on the global Rademacher complexity provides also a bound on the excess risk. This together with Theorem 57, gives the GRC-based excess risk bounds

of the following forms, for any $\boldsymbol{f} \in \mathcal{F}$ (note that $q \geq 1$)

$$\begin{aligned}
&\text{Group norm:} && \text{(a)} \quad \forall \kappa \in [q,2], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}^2 \kappa^{*2})^{\frac{1}{2}} \left(T^{\frac{2}{\kappa}} \right)^{-\frac{1}{2}} n^{-\frac{1}{2}} \right). \\
&&& \text{(b)} \quad \forall q \in [2,\infty], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}^2)^{\frac{1}{2}} \left(T^{\frac{2}{q}} \right)^{-\frac{1}{2}} n^{-\frac{1}{2}} \right). \\
&\text{Schatten-norm:} && \text{(c)} \quad \forall q \in [1,2], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}'^2 q^*)^{\frac{1}{2}} T^{-\frac{1}{2}} n^{-\frac{1}{2}} \right). \\
&&& \text{(d)} \quad \forall q \in [2,\infty], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}'^2)^{\frac{1}{2}} T^{-\frac{1}{2}} n^{-\frac{1}{2}} \right). \\
&\text{Graph regularizer:} && \text{(e)} && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}''^2)^{\frac{1}{2}} T^{-\frac{1}{2}} n^{-\frac{1}{2}} \right). \qquad (5.57)
\end{aligned}$$

which can be compared to their LRC-based counterparts as following

$$\begin{aligned}
&\text{Group norm:} && \text{(a)} \quad \forall \kappa \in [q,2], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}^2 \kappa^{*2})^{\frac{1}{1+\alpha}} \left(T^{\frac{2}{\kappa}} \right)^{-\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \right). \\
&&& \text{(b)} \quad \forall q \in [2,\infty], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}^2)^{\frac{1}{1+\alpha}} \left(T^{\frac{2}{q}} \right)^{-\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \right). \\
&\text{Schatten-norm:} && \text{(c)} \quad \forall q \in [1,2], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}'^2 q^*)^{\frac{1}{1+\alpha}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \right). \\
&&& \text{(d)} \quad \forall q \in [2,\infty], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}'^2)^{\frac{1}{1+\alpha}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \right). \\
&\text{Graph regularizer:} && \text{(e)} && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R_{max}''^2)^{\frac{1}{1+\alpha}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \right). \qquad (5.58)
\end{aligned}$$

It can be seen that holding all the parameters fixed when n approaches to infinity, the local bounds yield faster rates, since $\alpha > 1$. However, when T grows to infinity, the convergence rate of the local bounds could be only as good as those obtained by the global analysis.

A close appraisal of the results in (5.57) and (5.58) points to a conservation of asymptotic rates between n and T, when all other remaining quantities are held fixed. This phenomenon is more apparent for the Schatten norm and graph-based regularization cases. It can be seen that, for both the global and local analysis results, the rates (exponents) of n and T sum up to -1. In the local analysis case, the trade-off is determined by the value of α, which can facilitate faster n-rates

and compromise with slower T-rates. A similar trade-off is witnessed in the case of group norm regularization, but this time between n and $T^{2/\kappa}$, instead of T, due to the specific characteristic of the group norm.

As mentioned earlier in Remark 43, the bounds for the class of group norm regularizer for $1 \leq q \leq 2$ is not monotonic in q; they are minimized for $q^* = \log T$. Therefore, we split our analysis for this case as follows:

1. First, we consider $q^* \geq \log T$, which leads to the optimal choice $\kappa^* = q^*$, and taking the minimum of the global and local bounds gives

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq O\Big(\min\Big\{(R_{max}q^*)(T^{\frac{2}{q}})^{-\frac{1}{2}}n^{-\frac{1}{2}}, (R_{max}q^*)^{\frac{2}{1+\alpha}}(T^{\frac{2}{q}})^{-\frac{1}{1+\alpha}}n^{\frac{-\alpha}{1+\alpha}}\Big\}\Big). \tag{5.59}$$

   It is worth mentioning that, for any value of $\alpha > 1$, if the number of tasks T as well as the radius R_{max} of the $L_{2,q}$ ball can grow with n, the local bound improves over the global one whenever $\frac{T^{1/q}}{R_{max}} = O(\sqrt{n})$.

2. Secondly, assume that $q^* \leq \log T$, in which case the best choice is $\kappa^* = \log T$. Then, the excess risk bound reads

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq O\left(\min\left\{\left(\frac{R_{max}\log T}{T}\right)n^{-1/2}, \left(\frac{R_{max}\log T}{T}\right)^{\frac{2}{1+\alpha}}n^{\frac{-\alpha}{1+\alpha}}\right\}\right), \tag{5.60}$$

   and the local analysis improves over the global one, when $\frac{T}{R_{max}\log T} = O(\sqrt{n})$.

Also, a similar analysis for Schatten norm and graph regularized hypothesis spaces shows that the local analysis is beneficial over the global one, whenever the number of tasks T and the radius R can grow, such that $\frac{\sqrt{T}}{R} = O(\sqrt{n})$.

Comparisons to Related Works

Also, it would be interesting to compare our (global and local) results for the trace norm regularized MTL with the GRC-based excess risk bound provided in [104] wherein they apply a trace norm regularizer to capture the tasks' relatedness. It is worth mentioning that they consider a slightly different hypothesis space for $\boldsymbol{W}$, which in our notation reads as

$$\mathcal{F}_{S_1} := \left\{ \boldsymbol{W} : \frac{1}{2} \|\boldsymbol{W}\|_{S_1}^2 \leq T R_{max}'^2 \right\}. \tag{5.61}$$

It is based on the premise that, assuming a common vector $\boldsymbol{w}$ for all tasks, the regularizer should not be a function of number of tasks [104]. Given the task-averaged covariance operator $C := 1/T \sum_{t=1}^{T} J_t$, the excess risk bound in [104] reads as

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 2\sqrt{2} L R'_{max} \left(\sqrt{\frac{\|C\|_\infty}{n}} + 5\sqrt{\frac{\ln(nT)+1}{nT}} \right) + \sqrt{\frac{bLx}{nT}}.$$

where loss function ℓ is L-Lipschitz and $\mathcal{F}$ has ranges in $[-b, b]$. One can easily verify that the trace norm is a Schatten norm with $q = 1$. Note that for any $q \geq 1$ it holds that $\mathcal{F}_{S_1} \subseteq \mathcal{F}_{S_q}$, which implies $\mathfrak{R}(\mathcal{F}_{S_1}) \leq \mathfrak{R}(\mathcal{F}_{S_q})$. This fact, in conjunction with Theorem 57 and Theorem 56 (applied to the class of excess loss functions) yields a GRC-based excess risk bound. Therefore, considering the trace norm hypothesis space (5.61) and the optimal value of $q^* = 2$, translates our global and local bounds to the following

1. GRC-based excess risk bound:

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 4 L R'_{max} \sqrt{\frac{\left\| \left(\mathbf{tr}\left(J_t\right) \right)_{t=1}^{T} \right\|_1}{nT}} + \sqrt{\frac{bLx}{nT}}.$$

2. LRC-based excess risk bound ($\forall \alpha > 1$):

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 256K\sqrt{\frac{\alpha+1}{\alpha-1}}\left(2dR'^2_{max}L^2\right)^{\frac{1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb+16BK)Bx}{nT}. \quad (5.62)$$

Now, assume that each operator J_t is of rank M and denote its maximum eigenvalue by λ_t^{max}. If $\lambda_{max} := \max_{t\in\mathbb{N}_T}\{\lambda_t^{max}\}$, then it is easy to verify that $\mathbf{tr}(J_t) \leq M\lambda_t^{max}$ and $\|C\|_\infty \leq \lambda_{max}$, which leads to the following GRC-based bounds

$$\text{Ours:} \quad P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 4LR'_{max}\sqrt{\frac{M\lambda_{max}}{n}} + \sqrt{\frac{bLx}{nT}}, \quad (5.63)$$

$$\text{[104]:} \; P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 2\sqrt{2}LR'_{max}\left(\sqrt{\frac{\lambda_{max}}{n}} + 5\sqrt{\frac{\ln(nT)+1}{nT}}\right) + \sqrt{\frac{bLx}{nT}}. \quad (5.64)$$

One can observe that as $n \to \infty$, in all cases the bound vanishes. However, it does so at a rate of $n^{-\alpha/1+\alpha}$ for our local bound in (5.62), at a slower rate of $1/\sqrt{n}$ for our global bound in (5.63), and at the slowest rate of $\sqrt{\ln n/n}$ for the one in (5.64).

We remark that, as $T \to \infty$, all bounds converge to a non-zero limit: our local bound in (5.62) at a fast rate of $1/T$, the one in (5.63) at a slower rate of $\sqrt{1/T}$, and the bound in (5.64) at a the slowest rate of $\sqrt{\ln T/T}$.

Another interesting comparison can be performed between our bounds and the one introduced in [102] for a graph regularized MTL. For this purpose we consider the following hypothesis space similar to [102]

$$\mathcal{F}_G = \left\{\boldsymbol{W} : \frac{1}{2}\left\|\boldsymbol{D}^{1/2}\boldsymbol{W}\right\|_F^2 \leq TR''^2_{max}\right\}. \quad (5.65)$$

A bound on the empirical GRC of the aforementioned hypothesis space has been provided in [102].

However, similar to the proof of Corollary 51, we can easily convert it to a distribution dependent GRC bound which matches our global bound in (5.56) (for the defined hypothesis space (5.65)) and in our notation reads as

$$\mathfrak{R}\left(\mathcal{F}_G\right) \leq \sqrt{\frac{2R''^2_{max}}{nT}\left\|\left(\boldsymbol{D}_{tt}^{-1}\mathbf{tr}(J_t)\right)_{t=1}^{T}\right\|_1}.$$

Now, with $\boldsymbol{D} := \boldsymbol{L} + \eta\boldsymbol{I}$ (where $\boldsymbol{L}$ is the graph-Laplacian, $\boldsymbol{I}$ is the identity operator, and $\eta > 0$ is a regularization parameter) and the assumption that the J_ts are of rank M, it can be shown that

$$\left\|\left(\boldsymbol{D}_{tt}^{-1}\mathbf{tr}(J_t)\right)_{t=1}^{T}\right\|_1 = \sum_{t=1}^{T}\boldsymbol{D}_{tt}^{-1}\mathbf{tr}(J_t) \leq M\lambda_{max}\left(\sum_{t=1}^{T}\boldsymbol{D}_{tt}^{-1}\right) = M\lambda_{max}\mathbf{tr}\left(\boldsymbol{D}^{-1}\right) =$$
$$= M\lambda_{max}\mathbf{tr}\left(\boldsymbol{L}+\eta\boldsymbol{I}\right)^{-1} = M\lambda_{max}\left(\sum_{t=1}^{T}\frac{1}{\delta_t+\eta}+\frac{1}{\eta}\right) \leq M\lambda_{max}\left(\frac{T}{\delta_{min}+\eta}+\frac{1}{\eta}\right).$$

where λ_{max} is defined as before. Also, we define $(\delta_t)_{t=1}^{T}$ as the eigenvalues of Laplacian matrix $\boldsymbol{L}$ with $\delta_{min} := min_{t\in\mathbb{N}_T}\delta_t$. Therefore, the matching GRC-based excess risk bounds can be obtained as

$$\text{Ours \& [102]}: \qquad P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq \frac{2LR''_{max}}{\sqrt{n}}\sqrt{2M\lambda_{max}\left(\frac{1}{\delta_{min}}+\frac{1}{T\eta}\right)} + \sqrt{\frac{bLx}{nT}} \tag{5.66}$$

Also, from Remark 54, the LRC-based bound is given as

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 256K\sqrt{\frac{\alpha+1}{\alpha-1}}\left(dR''^2_{max}L^2\boldsymbol{D}^{-1}_{max}\right)^{\frac{1}{1+\alpha}}B^{\frac{2\alpha}{\alpha+1}}n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb+16BK)Bx}{nT}. \tag{5.67}$$

The above results show that when $n \to \infty$, both GRC and LRC bounds approach zero, albeit, the global bound with a rate of $\sqrt{1/n}$, and the local one with a faster rate of $n^{-\alpha/\alpha+1}$, since $\alpha > 1$. Also, as $T \to \infty$, both bounds approach non-zero limits. However, the global bound does so at a

rate of $\sqrt{1/T}$ and the local one at a faster rate of $1/T$.

CHAPTER 6: A NEW MULTI-TASK LEARNING MODEL USING LOCAL RADEMACHER COMPLEXITY

As we showed in the previous chapter, the local Rademacher complexity of kernel-based MT hypotheses can be both upper- and lower-bounded in terms of the tail sum of the eigenvalues of the kernel matrix. Motivated by this observation, in this chapter, we introduce a new family of MTL hypothesis based on convex combination of base kernels, in which the tail sum of kernels' eigenvalues is constrained. Furthermore, we extend the LRC-based kernel learning algorithm in [33] to a MTL setting. As shown in [33]—for the single task learning case—, it turns out that our algorithm for MTL also leads to a convex optimization problem, which can be solved using existing kernel learning algorithms.

Motivation and Analysis

As pointed out earlier in Chapter 1, a commonly utilized information sharing strategy for MTL is to use a (partially) common feature mapping ϕ to map the data from all tasks to a (partially) shared feature space $\mathcal{H}$. Such an MTL approach, not only allows information sharing across tasks, but also enjoys the non-linearity that is brought by the feature mapping ϕ. However, when applying kernel-based models, it is crucial to carefully choose the kernel function, as using inappropriate kernel functions may lead to deteriorated generalization performance. A widely adapted strategy for kernel selection is to learn a convex combination of some base kernels [71, 78], which combined with MTL, results in the MT-MKL approach. Such a method conically combines M pre-selected basis kernel functions $k_1, \cdots, k_M$, with the combination coefficients $\boldsymbol{\theta} := [\theta_1, \cdots, \theta_M]$, which are learned during the training stage in a pre-defined feasible region. For example, a widely used and theoretically well studied feasible region is given by the L_p-norm constraint [71]: $\Psi(\boldsymbol{\theta}) := \{\boldsymbol{\theta} :$

$\boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta}\|_p \leq 1\}$. As such, each task features a common kernel function $k(.,.) := \sum_{m=1}^{M} \theta_m k_m$. One such MT-MKL model is proposed in [127]. Besides, a more general MT-MKL approach with conically combined multiple objective functions and L_p-norm Multiple Kernel Learning (MKL) constraint is introduced in [86], and further extended and theoretically studied in [85]. A MT-MKL model that allows both feature and kernel selection is proposed in [65] and extended in [66].

In this section, we consider a MT-MKL approach where T inter-related tasks are learned via a standard MKL scheme using a prescribed collection of RKHSs $\{\mathcal{H}_m\}_{m=1}^{M}$, such that each $\mathcal{H}_m$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_m}$ and that has an associated feature mapping $\phi_m : \mathcal{X} \to \mathcal{H}_m$. The associated reproducing kernel $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is such that $k_m(x_1, x_2) = \langle \phi_m(x_1), \phi_m(x_2) \rangle_{\mathcal{H}_m}$ for all $x_1, x_2 \in \mathcal{X}$. It can be shown that this consideration implies an equivalent RKHS where the feature space of task t is served by $\mathcal{H}_t := \bigoplus_{m=1}^{M} \sqrt{\theta_t^m} \mathcal{H}_m$ with induced feature mapping $\phi_t := [\sqrt{\theta_t^1}\phi_1{}' \cdots \sqrt{\theta_t^M}\phi_M{}']'$ and endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{t,\boldsymbol{\theta}}} = \sum_{m=1}^{M} \langle \cdot, \cdot \rangle_{\mathcal{H}_m}$. Moreover, the reproducing kernel function for this feature space is given as $k_{t,\boldsymbol{\theta}}(x_t^i, x_t^j) = \sum_{m=1}^{M} \theta_t^m k_m(x_t^i, x_t^j)$ for all $x_t^i, x_t^j \in \mathcal{X}$. Note that each $\phi_m : \mathcal{X} \mapsto \mathcal{H}_m$ is selected before training. Finally, we consider the following MT-MKL hypothesis:

$$\mathcal{F} := \left\{ \boldsymbol{f} := (f_1, \ldots, f_T) : \forall t, \boldsymbol{w}_t \in \mathcal{H}_t, \sum_{t=1}^{T} \|\boldsymbol{w}_t\|_{\mathcal{H}_t}^2 \leq R, \boldsymbol{\theta} \succeq \mathbf{0} \right\}, \tag{6.1}$$

where, for each task t, $f_t := \langle \boldsymbol{w}_t, \phi_t(\boldsymbol{x}) \rangle + b_t$, where $b_t \in \mathbb{R}$ and $\boldsymbol{w}_t := (\boldsymbol{w}_t^1, \ldots, \boldsymbol{w}_t^M)$. Also, $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$ with $\boldsymbol{\theta}_t := (\theta_t^1, \ldots, \theta_t^M)$. It is worth pointing out that the (global) Rademacher complexity of the hypothesis space $\mathcal{F}$ can be upper-bounded in terms of the trace of the combined kernel $k_{t,\boldsymbol{\theta}}$ (see Theorem 57 in Chapter 5 for more details). Similar results have been shown in the context of single task learning, which inspired most kernel learning algorithms to restrict their hypothesis to the constraint $\mathbf{tr}(k_{\boldsymbol{\theta}}) \leq 1$, where $k_{\boldsymbol{\theta}}$ is a non-negative linear combination of M base kernels with the combination coefficient $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_M)$.

As demonstrated in the previous chapter, however, the LRC-based MTL-bounds are determined by the tail sum of the eigenvalues of the kernel, instead of its trace, *i.e.* the sum of all its eigenvalues. Also, regarding the fact that LRCs can potentially lead to tighter learning bounds compared to their GRC counterparts [33], this motivates us to consider an LRC-based constraint to restrict the hypothesis class $\mathcal{F}$ as

$$\mathcal{F}' := \left\{ \boldsymbol{f} := (f_1, \ldots, f_T) \in \mathcal{F} : \sum_{t=1}^{T} \sum_{j>h_t} \lambda_t^j(k_{t,\boldsymbol{\theta}}) \leq 1 \right\}, \tag{6.2}$$

where the non-negative integers $h_1, \ldots, h_T$ are free parameters used to control the tail sum. Note that the hypothesis class $\mathcal{F}'$ is not convex, since the tail sum of the eigenvalues can be written as the difference between the trace and the head sum of the eigenvalues, which are both convex functions. Therefore, following the approach in [33], we work with a more convenient hypothesis space which defines a convex set. For each task t and each kernel m, denoting $\tilde{\theta}_t^m := \theta_t^m / \|\boldsymbol{\theta}_t\|_1$, one can show

$$\begin{aligned} \sum_{t=1}^{T} \sum_{j>h_t} \lambda_t^j(k_{t,\boldsymbol{\theta}}) &= \sum_{t=1}^{T} \sum_{j>h_t} \lambda_t^j \left(\sum_{m=1}^{M} \tilde{\theta_t^m} \|\boldsymbol{\theta}_t\|_1 k_m \right) \geq \sum_{t=1}^{T} \sum_{m=1}^{M} \tilde{\theta_t^m} \sum_{j>h_t} \lambda_t^j \left(\|\boldsymbol{\theta}_t\|_1 k_m \right) \\ &= \sum_{t=1}^{T} \sum_{m=1}^{M} \theta_t^m \sum_{j>h_t} \lambda_t^j(k_m) \end{aligned} \tag{6.3}$$

where the inequality is due to the concavity of the function $\sum_{j>h_t} \lambda_t^j(.)$. Now, we alternatively consider the following hypothesis class

$$\mathcal{F}'' := \left\{ \boldsymbol{f} := (f_1, \ldots, f_T) \in \mathcal{F} : \sum_{t=1}^{T} \sum_{m=1}^{M} \theta_t^m \sum_{j>h_t} \lambda_t^j(k_m) \leq 1 \right\}, \tag{6.4}$$

which is convex (and therefor, more convenient to work with), since it is just the restriction of the convex class $\mathcal{F}$ with the linear constraint $\sum_{t=1}^{T} \sum_{m=1}^{M} \theta_t^m \sum_{j>h_t} \lambda_t^j(k_m) \leq 1$. Moreover, $\mathcal{F}''$ is a

richer class compared to $\mathcal{F}'$, as one can easily show that $\mathcal{F}' \subseteq \mathcal{F}''$ (see Propositions 4 and 5 in [33]).

A New Convex Formulation for MTL

In this section, we present our new MTL model corresponding to the hypothesis $\mathcal{F}''$. First note that for each task t, if one normalizes the base kernels k_ms as $\tilde{k}_t^m := \frac{k_m}{\sum_{j>h_t} \lambda_t^j(k_m)}$, then the hypothesis class $\mathcal{F}''$ can be equivalently written as

$$\mathcal{F}'' = \left\{ \boldsymbol{f} := (f_1, \ldots, f_T) : \forall t, \boldsymbol{w}_t \in \tilde{\mathcal{H}}_t, \sum_{t=1}^{T} \|\boldsymbol{w}_t\|_{\tilde{\mathcal{H}}_t}^2 \leq R, \boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta}\|_{1,1} \leq 1 \right\}, \tag{6.5}$$

where $\|\boldsymbol{\theta}\|_{1,1} := \sum_{t=1}^{T} \sum_{m=1}^{M} \theta_t^m$, and $\tilde{\mathcal{H}}_t$ is the feature space of task t induced by the normalized kernels $\tilde{k}_t^m$s. At this point, we would like to remark that, in practice, one needs to replace the kernels k_ms by their empirical counterparts, that is, the kernel matrices $\boldsymbol{K}_m$s, and consequently, considers the eigenvalues of the kernel matrices.

Here, we consider a linear MTL model involving T tasks. We assume that each task is presented by a training set $\{(x_t^i, y_t^i)\}_{i=1}^n$ sampled from $\mathcal{X} \times \mathcal{Y}$ based on some probability distribution $P_t(x, y)$, where $\mathcal{X}$ denotes an arbitrary set that serves as the native space of samples for all tasks, and $\mathcal{Y}$ represents the output space corresponding to the labels. Without loss of generality, we assume that the same number n of labeled samples are available for learning each task. Now, utilizing a regularized empirical risk minimization framework, we consider solving the following optimization

problem

$$\min_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta}} \frac{1}{2}\sum_{t=1}^{T}\|\boldsymbol{w}_t\|^2_{\tilde{\mathcal{H}}_t} + C\sum_{t=1}^{T}\sum_{i=1}^{n}\ell\left(f_t(x_t^i), y_t^i\right)$$
$$\textit{s.t. } \boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta}\|_{1,1} \leq 1. \tag{6.6}$$

where $\boldsymbol{b} := (b_1, \ldots, b_T)$. Note that one just requires to normalize the kernel matrices $\boldsymbol{K}_t^m$s as $\tilde{\boldsymbol{K}}_t^M := \frac{\boldsymbol{K}_m}{\sum_{j>h_t}\lambda_t^j(\boldsymbol{K}_m)}$, and then use any of the existing ℓ_1-norm MKL solvers in order to solve the above optimization problem. Finally, as pointed out earlier, the eigenvalue-tail sum of the kernel $\boldsymbol{K}_m$, for each task t, can be computed as $\sum_{j>h_t}\lambda_t^j(\boldsymbol{K}_m) = \mathbf{tr}(\boldsymbol{K_m}) - \sum_{j=1}^{h_t}\lambda_t^j(\boldsymbol{K}_m)$ with a computational complexity of the order $O(h_t n^2)$.

In this dissertation and for our experiments (presented in the next chapter), we consider T inter-related Support Vector Machine (SVM) problems. Therefore, (6.6) can be equivalently expressed as

$$\begin{aligned}\min_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{t=1}^{T}\|w_t\|^2_{\mathcal{H}_t} + C\sum_{t=1}^{T}\sum_{i=1}^{n}\xi_t^i \\ \textit{s.t.} \quad & \forall i,\ y_t^i\left(\left\langle w_t, \phi_t\left(x_t^i\right)\right\rangle_{\mathcal{H}_t} + b_t\right) \geq 1 - \xi_t^i,\ \xi_t^i \geq 0 \\ & \boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta}\|_{1,1} \leq 1.\end{aligned} \tag{6.7}$$

where $\boldsymbol{\xi} := (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T)$. Note that a very similar formulation can be derived for regression using algorithms such as SVR at this stage. Thus, the algorithm that we present in the following can be easily extended to regression problems by considering the ϵ-insensitive loss function used in the SVR context [56].

Algorithm

One can easily verify that the primal-dual form of (6.7) with respect to $\{\boldsymbol{\theta}\}$ and $\{\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\xi}\}$ can be given as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\alpha}} \quad & \sum_{t=1}^{T} \boldsymbol{\alpha}_t' \mathbf{1}_n - \frac{1}{2} \sum_{t=1}^{T} \boldsymbol{\alpha}_t' \boldsymbol{Y}_t \boldsymbol{K}_t \boldsymbol{Y}_t \boldsymbol{\alpha}_t \\ \textit{s.t.} \quad & \forall t,\ \mathbf{0} \preceq \boldsymbol{\alpha}_t \preceq C\mathbf{1},\ \boldsymbol{y}_t' \boldsymbol{\alpha}_t = 0 \\ & \boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta}\|_{1,1} \leq 1. \end{aligned} \tag{6.8}$$

where $\boldsymbol{\alpha} := (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_1)$ with $\boldsymbol{\alpha}_t$ being the Lagrangian dual variable for the minimization problem w.r.t.$\{\boldsymbol{w}_t, \boldsymbol{b}_t, \boldsymbol{\xi}_t\}$. A block coordinate descent framework, also known as the *non-linear Gauss-Seidel* method, applied to decompose Problem 6.8 into two subproblems. A convergence proof of this method can be found in [14], p. 268-269. The first subproblem is given as the following maximization problem with respect to $\boldsymbol{\alpha}$,

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{t=1}^{T} \boldsymbol{\alpha}_t' \mathbf{1}_n - \frac{1}{2} \sum_{t=1}^{T} \boldsymbol{\alpha}_t' \boldsymbol{Y}_t \boldsymbol{K}_t \boldsymbol{Y}_t \boldsymbol{\alpha}_t \\ \textit{s.t.} \quad & \forall t,\ \mathbf{0} \preceq \boldsymbol{\alpha}_t \preceq C\mathbf{1},\ \boldsymbol{y}_t' \boldsymbol{\alpha}_t = 0 \end{aligned} \tag{6.9}$$

the above problem can be efficiently solved via `LIBSVM` [24]. In the second block, we consider T independent minimization problems each of which with respect to the task-specific parameter $\boldsymbol{\theta}_t$ as

$$\begin{aligned} \min_{\boldsymbol{\theta}_t} \ & \boldsymbol{\theta}_t' \boldsymbol{q}_t \\ \textit{s.t. } & \boldsymbol{\theta}_t \succeq \mathbf{0}, \|\boldsymbol{\theta}_t\|_1 \leq 1. \end{aligned} \tag{6.10}$$

where $\boldsymbol{q}_t := (q_t^1, \ldots, q_t^M)$ and $q_t^m := -\frac{1}{2}\boldsymbol{\alpha}_t' \boldsymbol{Y}_t \boldsymbol{K}_t^m \boldsymbol{Y}_t \boldsymbol{\alpha}_t$. Note that a closed form solution can be found for the linear optimization problem (6.10) as

$$\left(\theta_t^j\right)_{j=1}^M = \begin{cases} 1 & \text{if } j = \arg\min_i q_t^i, \text{ and } q_t^j < 0 \\ 0 & \text{otherwise.} \end{cases}$$

CHAPTER 7: EXPERIMENTS

In this chapter, we present the results of our experiments with the algorithm we introduced in Chapter 6. We compare the performance of our model in (6.7) with several kernel learning algorithms on both classification and regression problems. In particular, we compare our model, denoted by `LRC-conv`, with the following algorithms:

1. Uniform combination (`unif`): This model is the most straightforward MKL algorithm whose performance has been difficult to surpass in the past [36, 31]. In this model, the kernel parameters are all fixed and set equal to $1/M$, where M is the number of base kernels.

2. ℓ_1-regularized combination (`l1-com`): We also evaluate the performance of our model against the classical ℓ_1-norm MKL [78]. In this model, an ℓ_1-norm constraint is imposed on kernel parameters θ_ms, and they are learned during the training process. It is worth pointing out that if one normalizes the base kernels such that they all have a trace of one, then the ℓ_1-norm MKL corresponds to the learning of a non-negative linear combination of kernels with a linear constraint in the form of $\mathbf{tr}(\boldsymbol{K}_{\boldsymbol{\theta}}) \leq \Lambda$ on the trace of the combined kernel.

3. ℓ_2-regularized combination (`l2-com`): This algorithm optimizes the kernel parameter θ_ms, by imposing an ℓ_2-norm constraint on them. We consider this model for our experiments, as it has been shown [37] that it can achieve significant improvement over the ℓ_1-norm MKL in some cases, specially when dealing with large number of base kernels.

4. Finally, an Independent Task Learning (`ITL`) model is used as a baseline, wherein each task is trained in isolation via a traditional single-task MKL strategy. The average performance over all tasks is taken to gauge the effectiveness of this method versus alternatives.

Experimental Setting

For all experiments, we have utilized 1 Linear, 1 Polynomial of degree 2, and a combination of Gaussian kernels of the form $\boldsymbol{K}_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2/\sigma\right)$, with varying spread parameter $\sigma \in \{2^{\sigma_0}, 2^{\sigma_0+1} \ldots, 2^{1-\sigma_1}, 2^{\sigma_1}\}$. An exception is the case of SARCOS dataset, for which, we only used Gaussian kernels after noticing that adding linear kernels can lead to degraded performance in all algorithms. Note that similar to the approach of base kernel selection in [39], the values of σ_0 and σ_1 for different datasets are chosen in a way so that the base kernels are sufficiently different. However, for the sake of a fair comparison, we use the same set of kernels across different methods. All kernel functions are normalized as $k(\boldsymbol{x}_1, \boldsymbol{x}_2) \leftarrow k(\boldsymbol{x}_1, \boldsymbol{x}_2)/\sqrt{k(\boldsymbol{x}_1, \boldsymbol{x}_1)k(\boldsymbol{x}_2, \boldsymbol{x}_2)}$. Note that, in order to accentuate the need for MTL, we intentionally keep the training set size small as only 10% of the samples we use for each experiment. The rest of the data is split in equal sizes for validation and testing. Also, the SVM regularization parameter C is chosen over the set $\{2^{-13}, \ldots, 2^{13}\}$, and the optimal model parameters h_ts for our `LRC-conv` model are determined via cross-validation over the set $\{2^0, \ldots 2^5\}$ in all experiments.

Benchmark Datasets

We evaluate the performance of our method on the following datasets:

Letter Recognition: This dataset is a collection of handwritten words which are collected by Rob Kassel at MIT spoken Language System Group. This dataset consists of eight tasks discriminating between the letters: 'c' vs. 'e', 'g' vs. 'y', 'm' vs. 'n', 'a' vs. 'g', 'i' vs. 'j', 'a' vs. 'o', 'f' vs. 't' and 'h' vs. 'n'. Each letter is represented by 8 by 16 pixel image, which forms a 128 dimensional feature vector. 200 samples are randomly chosen for each letter. However, we made an exception for the letter j, for which only 189 samples were available. Note that some features,

such as curvature of lines, number of strokes or contiguity of characters, are universal between different letters, while some other features are specific to each letter. Such adaptive characteristics of different letters, makes the MTL a suitable framework for letter recognition problem.

Landmine Detection dataset consist of 29 binary classification tasks collected from various landmine fields. Each data sample is represented by a 9-dimensional feature vector extracted from radar images and is associated to a binary class label y. The feature vectors correspond to regions of landmine fields and include four moment-based features, three correlation-based features, one energy ratio feature, and one spatial variance feature. The objective is to identify whether there is a landmine or not based on a region's features. An information sharing in this example can occur, since some region features such as surface reflection coefficient, energy-ratio and spatial variance are common/similar in different geographical landmine fields.

Spam Detection dataset was obtained from ECML PAKDD 2006 Discovery challenge for the spam detection problem. For our experiments, we used the Task B dataset, which contains labeled training data (emails) from inboxes of 15 different users. The goal is to construct a binary classifier for each user, detecting spam $(+)$ emails from the non-spam $(-)$ ones. Each email is represented by the term frequencies of the words resulting in $150K$ features from which we chose the 1000 most frequent ones. Since some aspects of relevant or junk emails are usually common across users (*e.g.* some spams are universal to all users), the problem of learning spam detection models for a set of users can be cast as a MTL problem.

Sentiment Analysis dataset contains Amazon product reviews on different domains, such as books, dvd ans so on. We choose 24 domains (corresponding to 24 tasks), and 100 reviews per domain. Each domain is considered as a binary classification task with reviews labeled $(+)$ when rating > 3, and labeled $(-)$ when rating < 3. Note that the reviews with rating $= 3$ are excluded in this experiment as such sentiments were ambiguous and hard to predict, even with very large

amount of data. Our features were defined using the 4000 most frequent bigrams, yielding a dictionary of size 76. Note that usually the same set of words are used to describe"good" or "bad" items in all domains. This commonality can be efficiently captured using an MTL framework.

SARCOS Inverse Kinematics dataset is generated from an inverse dynamics prediction system of a seven degrees-of-freedom (DOF) SARCOS anthropomorphic robot arm. This dataset consists of 28 dimensions: the first 21 dimensions are considered as features (including 7 joint positions, 7 joint velocities and 7 joint accelerations), and the last 7 dimensions, corresponding to 7 joint torques, are used as outputs. Therefore, there are 7 tasks and the inputs are shared among all the tasks. for each 21-dimensional observation, the goal is then to predict 7 joint torques for the seven DOF. This dataset involves 48933 observations from which we randomly sampled 2000 examples for our experiments. An MTL model is expected to show a promising performance in this problem, as all tasks (joints) shares the same set of features including joint positions, velocities and accelerations.

Short-term Electricity Load Forecasting dataset which was released for the Global Energy Forecasting Competition (GEFComp2012). This dataset contains hourly-load history of a US utility in 20 different zones from January 1st, 2004 to December 31, 2008. The goal is to predict the 1-hour-ahead electricity load of these 20 zones. For this purpose, we considered predictors consisting of a delay vector of 8 lagged hourly loads along with the calendar information including years, seasons, months, weekdays and holidays. Note that we normalized the data to unify the units of different features. Finally, we randomly sampled 2000 (non-sequential) examples per each task for our experiments. It is not hard to verify that the consumption patterns of some users in different zones might share some similarities, which might be properly captured using an MTL model.

Experimental Report

In this section, we report the results of our experiments with `LRC-conv`, and compare it with other kernel-based methods that we mentioned earlier in this chapter. For the statistical analysis of our experiments, we used the approach in [42], which has been widely applied by the machine learning community to compare multiple methods over multiple datasets. This is a common circumstance, specifically when the general performance of a method (and not its performance on a particular problem) is required to be assessed. In particular, we use Friedman's test [51] which can be considered as the non-parametric equivalent of the repeated-measures ANOVA. It is noteworthy that the *repeated-measures ANOVA* method is usually considered as the common statistical method for comparing between more than two related sample means. However, the assumptions of this method are most likely violated while studying the performance of machine learning algorithms. As an example, one assumption made by ANOVA is that the data come from normal distributions, although no guarantee exists for normality of performance distributions over a set of problems. The more important assumption of ANOVA is sphericity which refers to the situation where the variances of the differences between all possible pairs of groups are equal, *i.e.* the level of dependence between pairs of groups is roughly equal. However, this condition cannot be taken for granted when analyzing machine learning algorithms on multiple datasets. The violation of this assumption can have a great effect on the post-hoc test, and thus, it is not a suitable test for comparing learning algorithms.

For the reasons mentioned above, here we use Friedman's test to perform a statistical comparison between the methods that we considered in our study. Using this test, all algorithms for each dataset are ranked separately with the best performing algorithm being ranked first, and the second best ranked second and so on. Note that the average rank is assigned in the case of ties. The null hypothesis is that all methods are indistinguishable in terms of performance. Once the null hypoth-

esis is rejected, a post-hoc test is usually employed for pairwise comparisons between methods. The post-hoc tests are usually designed for situation in which one has already performed an F-test consisting of three or more means, but an additional exploration is needed to provide specific informations on which mean(s) are significantly different from others.

Suppose that we are provided with K algorithms and N datasets, where r_n^k denotes the rank of the k^{th} method on the nth dataset. Assuming that the average rank for each algorithm is obtained as $R_k := \frac{1}{N}\sum_{n=1}^{N} r_k^n$, then Friedman's test compares the average rank of the algorithms under the null hypothesis of the equivalency of all algorithms (all R_ks are equal). The statistic for this test is given as

$$\chi_F^2 := \frac{12N}{K(K+1)}\left(\sum_{k=1}^{K} R_k^2 - \frac{K(K+1)^2}{4}\right),$$

which is χ_F^2 with degrees of freedom $K-1$. A more advanced statistic (a modified version of Friedman test's statistic) was derived later by [62] as

$$F_F := \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}$$

which is F-distributed with $(K-1,(K-1)(N-1))$ degrees of freedom.

Now, when the null hypothesis is rejected, a post-hoc test can be further used to compare the general performance of the methods among themselves. More specifically, if one is interested to compare one—most likely a newly proposed—model against alternatives, a simple method such as Holm's test [61] can be further used. Defining $SE := \sqrt{\frac{K(K+1)}{6N}}$, this test is a sequentially step-down procedure with the test statistic

$$z := \frac{R_k - R_l}{SE}$$

to compare the k^{th} method against the l^{th} one. In more detail, Holm's test requires one to find the

p-values for all pairwise comparisons and then sort them from the most to the least significant one, so that $p_1 \leq p_2 \leq \ldots \leq p_{K-1}$. Now, considering the significance level α, if p_1 is smaller than $\alpha/(K-1)$, then the hypothesis corresponding to p_1 will be rejected and one proceeds with the next comparison, that is, p_2 to $\alpha/(K-2)$. We continue this process until a certain null hypothesis cannot be rejected in which case all remaining hypotheses are retained as well.

Table 7.1 reports the average performance over all task, and 20 runs of randomly sampled training sets. We used the accuracy percentage for classification problems, and the Mean Square Error (MSE) for the regression ones. The rank of each model is indicated by a superscript next to the performance of that model on the relevant data set, while the superscript next to the name of each model reflects its average rank over all data sets.

Table 7.1: Experimental comparison between `LRC-conv` and four other methods on six benchmark datasets. The superscript next to each model indicates its rank. The best performing algorithm gets rank of 1.

	Classification Accuracy				Regression MSE	
	Landmine	Letter	Spam	Sentiment	SARCOS	Load$^{\times 10^{-4}}$
σ	1, 16	1, 14	3, 18	1, 3	8, 16	3, 18
`ITL`$^{(4.5)}$	$59.13^{(2)}$	$83.75^{(5)}$	$84.12^{(5)}$	$57.83^{(5)}$	$19.95^{(5)}$	$46.17^{(5)}$
`unif`$^{(3.1)}$	$59.01^{(3)}$	$84.62^{(4)}$	$85.15^{(3)}$	$59.70^{(3)}$	$15.39^{(2)}$	$45.34^{(4)}$
`l1-com`$^{(3)}$	$58.63^{(5)}$	$85.94^{(2)}$	$85.17^{(2)}$	$59.86^{(2)}$	$19.75^{(4)}$	$44.13^{(3)}$
`l2-com`$^{(3.33)}$	$58.76^{(4)}$	$85.63^{(3)}$	$84.68^{(4)}$	$59.52^{(4)}$	$17.03^{(3)}$	$43.77^{(2)}$
`LRC-conv`$^{(1)}$	$61.86^{(1)}$	$89.58^{(1)}$	$88.34^{(1)}$	$61.32^{(1)}$	$11.19^{(1)}$	$41.40^{(1)}$

The results show that the LRC-based kernel learning approach leads to consistent improvements over all other kernel-based methods in all experiments. This is even more evident in two datasets *Landmine* and *Sentiment*, which are more difficult tasks as the classification accuracies are below 60%. Also, we observe that in 5 out of 6 datasets, ITL has the worst performance among all other

methods. This observation is consistent with the fact that MTL, in general, may improve over ITL, specially when a meaningful relationship exists between all tasks, and only a limited number of training samples per each task is available.

We used Friedman's and Holm's post-hoc tests for our statistical analysis at the significance level $\alpha = 0.05$. Applying the Friedman's test, on can easily verify that all algorithms are not equivalent. Therefore, we employ Holm's post-hoc test as following for our further analysis.

Table 7.2: Comparison of our `LRC-conv` method against the other methods with Holm's test

k	Methods	$z = \frac{R_k - R_{\text{LRC}}}{SE}$	p value	Adjusted α
1	`ITL`	3.83	0.0001	0.01
2	`l2-com`	2.55	0.0106	0.017
3	`l1-com`	2.37	0.0171	0.025
4	`unif`	2.19	0.0285	0.05

As shown in Table 7.2, the advantage of our model is statistically significant compared to all other methods. In particular, one can observe that in all experiments, our model outperforms all other kernel learning algorithms, since their corresponding p-values are all smaller than the adjusted α values obtained by Holm's post-hoc test. With this observation, one can verify the benefit of learning kernels based on our LRC approach, compared to other kernel learning methods (considered here) including ℓ_1-combination MKL, which can be viewed as learning a trace-constrained. Note that the latter case corresponds to a kernel learning algorithm based on a GRC analysis. This may justify the fact that using a LRC-based MTL hypothesis space (similar to (6.5)) can improve over the traditional MT-MKL algorithms such as uniform combination and ℓ_2-norm, as well as the GRC-based ℓ_1-norm MKL methods.

CHAPTER 8: CONCLUSION AND FUTURE DIRECTIONS

With the ever increasing amount of data being collected in many domains and applications everyday, an urgent need arises for tools to extract the patterns in large and complex datasets and translate them into meaningful information buried in the data. Machine learning strives to address this problem by building algorithms, which observe a phenomenon, construct a model and then use it for future predictions. A very common example of this automated learning process deals with supervised learning scenario in which the training data are given as input-output pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. For each pair, the variable x is related to y. This relationship is encapsulated by an unbeknown distribution $P(x, y)$, and can be modeled by a predictor function f which is supposed to generalize well in the future. That is to say, the mapping function f—from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$—should be constructed in such a way that it minimizes the probability of error when comparing the response $f(x)$ to y for a pair (x, y) drawn independently from P. Mathematically speaking, given the performance measure $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the expected loss $R(f) := \mathbb{E}_{(x,y)\sim P}\ell(f(x), y)$—associated to f on the pair (x, y)—should be as small as possible. Note that $R(f)$ cannot be evaluated, as it depends on the unknown distribution P. However, its empirical counterpart, known as empirical loss, can be calculated as $\hat{R}(f) := 1/n \sum_{i=1}^{n} \ell(f(x_i), y_i)$. Note that the empirical loss cannot tell anything about the generalization error on unseen new data. Therefore, the important question here is that *"is there a way to create a learning algorithm with a guarantee that the deviation between $R(f)$ and $\hat{R}(f)$ is small for a large sample size n?"* It is worth mentioning that the practical successes of machine learning algorithms largely rely on these guarantees which also ensure the quality of the solution produced.

The machine learning theory tries to address this question with the help of concentration inequalities, which quantify the gap between empirical averages and their expected values. These inequalities provide probabilistic upper-bounds on the deviation of empirical and expected errors in

the form of generalization error bounds. Beside characterizing the predictive ability of a learning model, generalization error bounds can be used in designing new learning algorithms, which potentially lead to more accurate prediction models as they enjoy strong learning guarantees. Furthermore, minimizing generalization bounds can be considered as an alternative to model selection, which tries to identify the "right" complexity of the learning space to prevent overfitting.

With all these being said, one commonly occurring problem when applying machine learning to solve application problems, is the lack of a sufficient amount of training data to attain acceptable performance results; either obtaining such data may be very costly or they may be unavailable due to technological limitations. In such situations, relying solely on the scarce data per individual prediction task most often leads to inadequate predictive performance. MTL leverages the underlying common links among a group of related tasks, while respecting the tasks individual idiosyncrasies to the extent warranted. This is achieved by phrasing the learning process as a joint, mutually dependent learning problem which has been shown to be beneficial compared to learning each task in isolation, as typically done in practice. Nowadays, MTL frameworks are routinely employed in a variety of settings. Some recent applications include computational genetics, image segmentation, HIV therapy screening, collaborative filtering, age estimation from facial images, and sub-cellular location prediction, just to name a few prominent ones.

A commonly utilized information sharing strategy for MTL is to use a (partially) common feature mapping ϕ to map the data from all tasks to a (partially) shared feature space $\mathcal{H}$. Such a method, named kernel-based MTL, not only allows information sharing across tasks, but also enjoys the non-linearity that is brought by the feature mapping ϕ. Although, the practical aspect of kernel-based MTL has been widely studied by machine learning community, the theoretical research of such settings is limited to a few studies.

The theoretical analysis of MTL in the existing work mostly investigates the generalization learn-

ing guarantees in the form of error bounds in terms of the (global) Rademacher complexities. To formally recapitulate the essence of these efforts, let T denote the number of tasks being co-learned and n denote the number of available observations per task. Then, in terms of convergence rates in the number of samples and tasks, respectively, the fastest-converging error or excess risk bounds derived in these efforts—whether distribution- or data-dependent—are of the order $O(\sqrt{1/nT})$.

In this dissertation, we investigate MTL generalization guarantees based on a more nuanced notion of complexity, termed Local Rademacher Complexity (LRC), as opposed to the original Global Rademacher Complexity (GRC). This new, modified function class complexity measure is attention-worthy, since an LRC-based analysis is capable of producing more rapidly-converging excess risk bounds ("fast rates"), when compared to the ones obtained via a GRC analysis. Note that the convergence rate is considered as an important factor in analyzing error bounds associated to a learning problem. This importance can be understood from the definition of the convergence rate; the rate at which the empirical risk of the learning problem approaches its true counterpart. As one of the main contributions of this dissertation, through a Talagrand-type concentration inequality adapted to the MTL case, we derived sharp bounds on the MTL excess risk in terms of the distribution- and data-dependent LRC. For a given number of tasks T, these bounds admit faster (asymptotic) convergence characteristics in the number of observations per task n, when compared to corresponding bounds hinging on the GRC. Hence, these faster rates allow us to increase the confidence that the MTL hypothesis selected by a learning algorithm approaches the best-in-class solution as n increases beyond a certain threshold. Our derived bounds reflect that one can trade off a slow convergence speed w.r.t. T for an improved convergence rate w.r.t. n. The latter one ranges, in the worst case, from the typical GRC-based bounds $O(\sqrt{1/n})$, all the way up to the fastest rate of order $O(1/n)$ by allowing the bound to depend less on T. Nevertheless, the premium in question becomes less relevant to MTL, in which T is typically considered as fixed.

Also, we show that our LRC-based bounds can be both upper- and lower-bounded in terms of

the tails sum of the eigenvalues of the tasks' kernel matrices. This motivated us to design a new kernel-based MTL formulation based on a non-negative combination of kernels with an LRC-based constraint on the tail sum of the eigenvalues of the kernel matrices. The resulting optimization problem is convex and can be efficiently solved using existing kernel learning algorithms. Finally, via a series of experiments, we show that our new MTL model consistently outperforms traditional MT kernel learning approaches, which have shown promising performances in the past.

To summarize the contributions of this work, we prove the first Talagrand's concentration inequality for vector-valued MTL function classes. Note that the existing Talagrand's bounds are applicable only to scalar-valued function classes associated to STL scenario. Therefore, the initial step of deriving Talagrand's bound needs to be revisited such that it can be applied to vector-valued function spaces, appropriate for MTL. This step itself, relies on proving a Logorithmic Sobolev inequality, which is considered as the exponential version of Efron-Stein inequality; a powerful tool in bounding the variance of general functions of independent random variables. Based on our revised Talagrand's concentration inequality, we then develop a new excess risk bound for MTL, which improves over the existing bounds in the context of MTL. Our risk bound achieves a faster convergence rate compared to McDiarmid-based bounds. Furthermore, in order to derive a more informative form of the improved bound, we consider the application of our risk bound to a number of common kernel-based MTL spaces. It turns out that the bounds for all cases that we considered in this study are functions of the tail-sum of the eigenvalues of tasks' kernels. Based on this new piece of information, we finally propose a new MTL model which outperforms traditional MTL approaches, including a model designed based on a GRC analysis.

While MTL mainly focuses on improving the generalization performance of all available tasks, there exists another transfer learning scenario, which focuses on doing well only on a new target task, by exploiting the knowledge from past experiences. This learning paradigm is known as Learning-To-Learn (LTL), and it transfers knowledge from some previously learned tasks to im-

prove learning of a new task. Similar to MTL, though, the assumption is that the tasks share some common properties, which can be utilized to enhance the learning of future tasks. The success in this line of research can have major impact in real world applications, as it can help building machines which learn from experience to perform a new task. A direction for future investigation, then, is to extend our MTL analysis in this dissertation to the case of LTL. Regarding the fact that there are only a few efforts investigating the theoretical aspects of this problem, an improved local analysis (similar to what has been done for MTL here) might lead to designing more powerful LTL frameworks.

Beside LTL, our study here can be extended to other learning frameworks such as deep Neural Networks, which have been proven to achieve remarkable performance in many areas such as natural language processing, speech recognition, social network filtering, image captioning, machine translation and bioinformatics. Despite the success of deep networks in solving machine learning problems, there exist a lack of a theoretical analysis in designing of the network architecture and the training process. Filling this theoretical gap can remedy some difficulties of training deep networks. For instance, tight generalization bounds in this context can guide the design of new learning algorithms in which one does not need to pre-specify the number of layers of the network. Instead the network architecture is leaned in an adaptive fashion based on the complexity of the learning problem at hand. Such an adaptive structural learning has been studied based on a global analysis in [32], which can be extended to local analysis similar to what has been done in this dissertation.

APPENDIX A: PROOFS OF THE RESULTS I

Proofs of the results for "Talagrand-Type Inequality for Multi-Task Learning"

Proof of Theorem 28

Before laying out the details, we first provide a sketch of the proof. Defining

$$Z := \sup_{\boldsymbol{f}\in\mathcal{F}} \Big[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)]\Big], \tag{A.1}$$

we first apply Theorem 20 to control the log-moment generating function $\log\mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big)$. From Theorem 20, we know that the main component to control $\log\mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big)$ is the variance-type quantity $V^+ = \sum_{s=1}^{T}\sum_{j=1}^{N_s}\mathbb{E}'\big[\big(Z - Z'_{s,j}\big)_+^2\big]$. In the next step, we show that V^+ can also be bounded in terms of two other quantities denoted by W and Υ. Applying Theorem 20 for a specific value of θ, then gives a bound for $\log\mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big)$ in terms of $\log\mathbb{E}[e^{\frac{\lambda}{b'}(W+\Upsilon)}]$. We then turn to controlling W and Υ, respectively. Our idea to tackle W is to show that it is a self-bounding function, according to which we can apply Corollary 23 to control $\log\mathbb{E}[e^{\frac{\lambda W}{b'}}]$. The term Υ is closely related to the constraint imposed on the variance of functions in $\mathcal{F}$, and can be easily upper bounded in terms of r. We finally apply Lemma 5 to transfer the upper bound on the log-moment generating function $\log\mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big)$ to the tail probability on Z. To clarify the process we divide the proof into four main steps.

Step 1. Controlling the log-moment generating function of Z with the random variable W and variance Υ. Let $X' := (X_t'^i)_{(t,i)=(1,1)}^{(T,N_t)}$ be an independent copy of $X := (X_t^i)_{(t,i)=(1,1)}^{(T,N_t)}$. Define the quantity

$$Z'_{s,j} := \sup_{\boldsymbol{f}\in\mathcal{F}} \Big[\frac{1}{TN_s}\big[\mathbb{E}' f_s(X'^j_s) - f_s(X'^j_s)\big] - \frac{1}{TN_s}\big[\mathbb{E} f_s(X^j_s) - f_s(X^j_s)\big] + \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}[\mathbb{E} f_t(X^i_t) - f_t(X^i_t)]\Big], \quad \text{(A.2)}$$

where $Z'_{s,j}$ is obtained from Z by replacing the variable X^j_s with X'^j_s. Let $\hat{\boldsymbol{f}} := (\hat{f}_1, \dots \hat{f}_T)$ be such that $Z = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\big[\mathbb{E}\hat{f}_t(X^i_t) - \hat{f}_t(X^i_t)\big]$, and introduce

$$W := \sup_{\boldsymbol{f}\in\mathcal{F}} \Big[\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}[\mathbb{E} f_t(X^i_t) - f_t(X^i_t)]^2\Big],$$
$$\Upsilon := \sup_{\boldsymbol{f}\in\mathcal{F}} \Big[\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}\mathbb{E}[\mathbb{E} f_t(X^i_t) - f_t(X^i_t)]^2\Big].$$

It can be shown that for any $j \in \mathbb{N}_n$ and any $s \in \mathbb{N}_T$:

$$Z - Z'_{s,j} \le \frac{1}{TN_s}\big[\mathbb{E}\hat{f}_s(X^j_s) - \hat{f}_s(X^j_s)\big] - \frac{1}{TN_s}\big[\mathbb{E}'\hat{f}_s(X'^j_s) - \hat{f}_s(X'^j_s)\big]$$

and therefore

$$(Z - Z'_{s,j})_+^2 \le \frac{1}{T^2N_s^2}\big([\mathbb{E}\hat{f}_s(X^j_s) - \hat{f}_s(X^j_s)] - [\mathbb{E}'\hat{f}_s(X'^j_s) - \hat{f}_s(X'^j_s)]\big)^2.$$

Then, it follows from the identity $\mathbb{E}'[\mathbb{E}'\hat{f}_s(X_s'^j) - \hat{f}_s(X_s'^j)] = 0$ that

$$\begin{aligned}
\sum_{s=1}^{T}\sum_{j=1}^{N_s}\mathbb{E}'\big[\big(Z - Z'_{s,j}\big)_+^2\big] &\le \sum_{s=1}^{T}\sum_{j=1}^{N_s}\frac{1}{T^2N_s^2}\mathbb{E}'\Big[\Big([\mathbb{E}\hat{f}_s(X_s^j) - \hat{f}_s(X_s^j)] - [\mathbb{E}'\hat{f}_s(X_s'^j) - \hat{f}_s(X_s'^j)]\Big)^2\Big] \\
&= \sum_{s=1}^{T}\sum_{j=1}^{N_s}\frac{1}{T^2N_s^2}[\mathbb{E}\hat{f}_s(X_s^j) - \hat{f}_s(X_s^j)]^2 + \sum_{s=1}^{T}\sum_{j=1}^{N_s}\frac{1}{T^2N_s^2}\mathbb{E}'[\mathbb{E}'\hat{f}_s(X_s'^j) - \hat{f}_s(X_s'^j)]^2 \\
&\le \sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{s=1}^{T}\sum_{j=1}^{N_s}\frac{1}{T^2N_s^2}[\mathbb{E}f_s(X_s^j) - f_s(X_s^j)]^2 + \sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{s=1}^{T}\sum_{j=1}^{N_s}\frac{1}{T^2N_s^2}\mathbb{E}[\mathbb{E}f_s(X_s^j) - f_s(X_s^j)]^2 \\
&= W + \Upsilon.
\end{aligned}$$

Introduce $b' := \frac{2b}{nT}$. Applying Theorem 20 and the above bound on $\sum_{s=1}^{T}\sum_{j=1}^{N_s}\mathbb{E}'\big[\big(Z - Z'_{s,j}\big)_+^2\big]$ then gives the following bound on the log-moment generating function of Z:

$$\log\mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big) \le \frac{\lambda b'}{1-\lambda b'}\log\mathbb{E}e^{\frac{\lambda}{b'}(W+\sigma^2)}, \quad \forall\lambda\in(0,1/b'). \tag{A.3}$$

Step 2. Controlling the log-moment generating function of W. We now upper bound the log-moment generating function of W by showing that it is a self-bounding function. For any $s\in\mathbb{N}_T, j\in\mathbb{N}_{N_s}$, introduce

$$W_{s,j} := \sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)]^2 - \frac{1}{T^2N_s^2}[\mathbb{E}f_s(X_s^j) - f_s(X_s^j)]^2\Big].$$

Note that $W_{s,j}$ is a function of $\{X_t^i, t\in\mathbb{N}_T, i\in\mathbb{N}_t\}\backslash\{X_s^j\}$. Letting $\tilde{\boldsymbol{f}} := (\tilde{f}_1,\ldots,\tilde{f}_T)$ be the function achieving the supremum in the definition of W, it can be checked that (note that $b' = \frac{2b}{nT}$)

$$T^2[W - W_{s,j}] \le \frac{1}{N_s^2}[\mathbb{E}\tilde{f}_s(X_s^j) - \tilde{f}_s(X_s^j)]^2 \le \frac{4b^2}{n^2} = T^2b'^2. \tag{A.4}$$

Similarly, if $\tilde{\boldsymbol{f}}^{s,j} := (\tilde{f}_1^{s,j}\ldots,\tilde{f}_T^{s,j})$ is the function achieving the supremum in the definition of

$W_{s,j}$, then one can derive the following inequality

$$T^2[W - W_{s,j}] \geq \frac{1}{N_s^2}[\mathbb{E}\tilde{f}_s^{s,j}(X_s^j) - \tilde{f}_s^{s,j}(X_s^j)]^2 \geq 0.$$

Also, it can be shown that

$$\begin{aligned}\sum_{s=1}^{T}\sum_{i=1}^{N_s} W - W_{s,j} &\leq \frac{1}{T^2}\sum_{s=1}^{T}\frac{1}{N_s^2}\sum_{i=1}^{N_s}[\mathbb{E}\tilde{f}_s(X_s^j) - \tilde{f}_s(X_s^j)]^2 \\ &= \sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)]^2\Big]. \end{aligned} \tag{A.5}$$

Therefore (according to Definition 21), W/b' is a b'-self bounding function. Applying Corollary 23 then gives the following inequality for any $\lambda \in (0, 1/b')$:

$$\log \mathbb{E}e^{\lambda(W/b')} \leq \frac{(e^{\lambda b'} - 1)}{b'^2}\mathbb{E}W = \frac{(e^{\lambda b'} - 1)}{b'^2}\Sigma^2 \leq \frac{\lambda\Sigma^2}{b'(1 - \lambda b')}, \tag{A.6}$$

where we introduce $\Sigma^2 := \mathbb{E}W$ and the last step uses the inequality $(e^x - 1)(1 - x) \leq x, \forall x \in [0, 1]$. Furthermore, the term Σ^2 can be controlled as follows: (here (σ_t^i) is a sequence of independent Rademacher variables, independent of X_t^i):

$$\begin{aligned}\Sigma^2 &\leq \frac{1}{T^2}\mathbb{E}_X \sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}\big[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)\big]^2 - \sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}\mathbb{E}\big[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)\big]^2\Big] + \Upsilon \\ &\leq 2\mathbb{E}_{X,\sigma}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}\sigma_t^i\big[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)\big]^2\Big] + \Upsilon \\ &\leq 8b\mathbb{E}_{X,\sigma}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}\sigma_t^i\big[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)\big]\Big] + \Upsilon \\ &\leq \frac{16b\mathfrak{R}(\mathcal{F})}{nT} + \Upsilon,\end{aligned}$$

where the first inequality follows from the definition of W and Υ, and the second inequality follows

from the standard symmetrization technique used to related Rademacher complexity to uniform deviation of empirical averages from their expectation [10]. The third inequality comes from a direct application of Lemma 2 with $\phi(x) = x^2$ (with Lipschitz constant $4b$ on $[-2b, 2b]$), and the last inequality uses Jensen's inequality together with the definition of $\mathfrak{R}(\mathcal{F})$ and the fact that $\frac{1}{N_t^2} \leq \frac{1}{nN_t}$. Plugging the previous inequality on Σ^2 back into (A.6) gives

$$\log \mathbb{E}e^{\lambda(W/b')} \leq \frac{\lambda}{b'(1-\lambda b')}\Big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT} + \Upsilon\Big], \quad \forall \lambda \in (0, 1/b'). \tag{A.7}$$

Step 3. Controlling the term Υ. Note that Υ can be upper bounded as

$$\begin{aligned}
\Upsilon :&= \sup_{\boldsymbol{f}\in\mathcal{F}} \Big[\frac{1}{T^2}\sum_{s=1}^{T}\frac{1}{N_s^2}\sum_{j=1}^{N_s}\mathbb{E}[\mathbb{E}f_s(X_s^j) - f_s(X_s^j)]^2\Big] \\
&\leq \frac{1}{nT^2}\sup_{\boldsymbol{f}\in\mathcal{F}} \Big[\sum_{s=1}^{T}\mathbb{E}[\mathbb{E}f_s(X_s^1) - f_s(X_s^1)]^2\Big] \\
&\leq \frac{1}{nT^2}\sup_{\boldsymbol{f}\in\mathcal{F}} \Big[\sum_{s=1}^{T}\mathbb{E}[f_s(X_s^1)]^2\Big] \\
&\leq \frac{r}{nT}.
\end{aligned} \tag{A.8}$$

where the last inequality follows from the assumption $\frac{1}{T}\sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\sum_{s=1}^{T}\mathbb{E}[f_s(X_s^1)]^2\Big] \leq r$ of the theorem.

Step 4. Transferring from the bound on log-moment generating function of Z to tail probabilities. Plugging the bound on $\log \mathbb{E}e^{\lambda W/b'}$ given in (A.7) and the bound on Υ given in (A.8) back into (A.3) immediately yields the following inequality on the log-moment generating function of

Z for any $\lambda \in (0, 1/2b')$:

$$\begin{aligned}\log \mathbb{E}[e^{\lambda(Z-\mathbb{E}Z)}] &\leq \frac{\lambda b'}{1-\lambda b'}\Big[\frac{\lambda}{b'(1-\lambda b')}\big[16(nT)^{-1}b\mathfrak{R}(\mathcal{F})+\Upsilon\big]+\frac{\lambda\Upsilon}{b'}\Big] \\ &\leq \frac{\lambda b'}{1-\lambda b'}\frac{\lambda}{b'(1-\lambda b')}\Big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT}+2\Upsilon\Big] \\ &\leq \frac{2\lambda^2}{2(1-2\lambda b')}\Big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT}+\frac{2r}{nT}\Big],\end{aligned} \tag{A.9}$$

where the second inequality uses $(1-\lambda b')^2 \geq 1-2\lambda b' > 0$ since $\lambda \in (0, 1/2b')$. That is, the conditions of Lemma 5 hold and we can apply it (with $A = 2\big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT}+\frac{2r}{nT}\big]$ and $B = 2b'$) to get the following inequality with probability at least $1-e^{-x}$ (note that $b' = \frac{2b}{nT}$):

$$\begin{aligned}Z &\leq \mathbb{E}[Z]+\sqrt{4x\Big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT}+\frac{2r}{nT}\Big]}+2b'x \\ &\leq \mathbb{E}[Z]+8\sqrt{\frac{bx\mathfrak{R}(\mathcal{F})}{nT}}+\sqrt{\frac{8xr}{nT}}+\frac{4bx}{nT} \\ &\leq \mathbb{E}[Z]+2\mathfrak{R}(\mathcal{F})+\frac{8bx}{nT}+\sqrt{\frac{8xr}{nT}}+\frac{4bx}{nT} \\ &\leq 4\mathfrak{R}(\mathcal{F})+\sqrt{\frac{8xr}{nT}}+\frac{12bx}{nT},\end{aligned}$$

where the third inequality follows from $2\sqrt{uv} \leq u+v$, and the last step uses the following inequality due to the symmetrization technique (here the ghost sample X' is an *i.i.d.* copy of the

initial sample X)

$$\begin{aligned}
\mathbb{E}Z &= \mathbb{E}_X\Big[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{T}\mathbb{E}_{X'}\Big[\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\big(f_t\big(X_t'^i\big)-f_t\big(X_t^i\big)\big)\Big]\Big] \\
&\le \mathbb{E}_{X,X'}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\big(f_t\big(X_t'^i\big)-f_t\big(X_t^i\big)\big)\Big] \\
&= \mathbb{E}_{X,X',\sigma}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\sigma_t^i\big(f_t\big(X_t'^i\big)-f_t\big(X_t^i\big)\big)\Big] \\
&\le 2\mathfrak{R}(\mathcal{F}).
\end{aligned}$$

Note that the second identity holds since for any σ_t^i, the random variable $f_t(X_t'^i) - f_t(X_t^i)$ has the same distribution as $\sigma_t^i(f_t(X_t'^i) - f_t(X_t^i))$.

APPENDIX B: PROOFS OF THE RESULTS II

Proofs of the results for "Excess MTL Risk Bounds based on Local Rademacher Complexities"

Theorem 59 is at the core of proving Theorem 33 in Sect. 5. We first present some useful lemmas.

Lemma 7 (Young's inequality)**.** *If $p, q > 0$ with $p^{-1} + q^{-1} = 1$, then we have*

$$\frac{x^p}{p} + \frac{y^q}{q} \geq xy, \quad \forall x, y \geq 0. \tag{B.1}$$

Lemma 8. *Let $c_1, c_2 > 0$ and $s > q > 0$. Then the equation $x^s - c_1 x^q - c_2 = 0$ has a unique positive solution x_0 satisfying*

$$x_0 \leq \left[c_1^{\frac{s}{s-q}} + \frac{s c_2}{s-q} \right]^{\frac{1}{s}}.$$

Furthermore, for any $x \geq x_0$, we have $x^s \geq c_1 x^q + c_2$.

Proof. Denote $p(x) := x^s - c_1 x^q - c_2$. The uniqueness of a positive solution for the equation $p(x) = 0$ is shown in Lemma 7.2 in [40]. Let x_0 be this unique positive solution. Then, it follows from Young's inequality (B.1) that

$$x_0^s = c_1 x_0^q + c_2 \leq \frac{x_0^{q \cdot \frac{s}{q}}}{\frac{s}{q}} + \frac{c_1^{\frac{s}{s-q}}}{\frac{s}{s-q}} + c_2 = \frac{q}{s} x_0^s + \frac{s-q}{s} c_1^{\frac{s}{s-q}} + c_2,$$

from which we have $x_0^s \leq c_1^{\frac{s}{s-q}} + \frac{s c_2}{s-q}$. The inequality $p(x) \geq 0$ for any $x \geq x_0$ follows immediately from $p(x_0) = 0$, $\lim_{x \to \infty} p(x) = \infty$ and the uniqueness of zero points for the equation $p(x) = 0$. □

Also, we will need the following lemma for the second step of the proof of Theorem 59.

Lemma 9. *Let $K > 1, r > 0$. Assume that $\mathcal{F} = \{\boldsymbol{f} := (f_1, \ldots, f_T) : \forall t, f_t \in \mathbb{R}^{\mathcal{X}}\}$ is a*

vector-valued (β, B)-Bernstein class of functions. Also, let the rescaled version of $\mathcal{F}$ be defined as

$$\mathcal{F}_r := \left\{ \boldsymbol{f}' = \left(f'_1, \ldots, f'_T\right) : f'_t := \frac{r f_t}{\max\left(r, V(\boldsymbol{f})\right)}, \boldsymbol{f} = (f_t, \ldots, f_T) \in \mathcal{F} \right\}.$$

If $V_r^+ := \sup_{\boldsymbol{f}' \in \mathcal{F}_r} [P\boldsymbol{f}' - P_n \boldsymbol{f}'] \leq \frac{r^{\frac{1}{\beta}}}{BK}$, then

$$\forall \boldsymbol{f} \in \mathcal{F} \qquad P\boldsymbol{f} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}}}{K}. \tag{B.2}$$

Proof. We prove (B.2) by considering two cases. Let $\boldsymbol{f}$ be any element in $\mathcal{F}$. If $V(\boldsymbol{f}) \leq r$, then $\boldsymbol{f}' = \boldsymbol{f}$ and the inequality $V_r^+ \leq \frac{r^{\frac{1}{\beta}}}{BK}$ translates to

$$P\boldsymbol{f} \leq P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}}}{BK} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}}}{K}. \tag{B.3}$$

If $V(\boldsymbol{f}) \geq r$, then $\boldsymbol{f}' = r\boldsymbol{f}/V(\boldsymbol{f})$ and the inequality $V_r^+ \leq \frac{r^{\frac{1}{\beta}}}{BK}$ translates to

$$\begin{aligned}
P\boldsymbol{f} &\leq P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}-1} V(\boldsymbol{f})}{BK} \leq P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}-1} P\boldsymbol{f}^\beta}{K} \\
&= P_n \boldsymbol{f} + \frac{1}{TK} \sum_{t=1}^{T} \mathbb{E}\left[r^{\frac{1}{\beta}-1} f_t^\beta \right] \\
&\overset{\text{(B.1)}}{\leq} P_n \boldsymbol{f} + \frac{1}{TK} \sum_{t=1}^{T} \mathbb{E}\left[\frac{[f_t^\beta]^{\frac{1}{\beta}}}{\frac{1}{\beta}} + \frac{[r^{\frac{1}{\beta}-1}]^{\frac{1}{1-\beta}}}{\frac{1}{1-\beta}} \right] \\
&= P_n \boldsymbol{f} + \frac{\beta}{K} P\boldsymbol{f} + \frac{(1-\beta) r^{\frac{1}{\beta}}}{K},
\end{aligned}$$

where we have used the Bernstein's condition $V(\boldsymbol{f}) \leq BP\boldsymbol{f}^\beta$. The above inequality can be

equivalently written as

$$P\boldsymbol{f} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + \frac{1-\beta}{K-\beta} r^{\frac{1}{\beta}} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}}}{K}. \tag{B.4}$$

Eq. (B.2) follows by combining (B.3) and (B.4) together. □

The following provides another useful definition that will be needed in introducing the result of Theorem 59.

Definition 58 (Star-Hull)**.** *The star-hull of a function class $\mathcal{F}$ around the function f_0 is defined as*

$$star(\mathcal{F}, f_0) := \{f_0 + \alpha(f - f_0) : f \in \mathcal{F}, \alpha \in [0,1]\}$$

Now, we present a lemma from [10] which indicates that the local Rademacher complexity of the star-hull of any function class $\mathcal{F}$ is a sub-root function, and it has a unique fixed point.

Lemma 10 (Lemma 3.4 in [10])**.** *For any function class $\mathcal{F}$, the local Rademacher complexity of its start-hull is a sub-root function.*

Theorem 59 (Distribution-dependent bound for MTL)**.** *Let $\mathcal{F} = \{\boldsymbol{f} := (f_1, \ldots, f_T) : \forall t, f_t \in \mathbb{R}^{\mathcal{X}}\}$ be a class of vector-valued functions satisfying $\sup_{t,x} |f_t(x)| \leq b$. Let $X := (X_t^i, Y_t^i)_{(t,i)=(1,1)}^{(T,n)}$ be a vector of nT independent random variables where $(X_t^1, Y_t^1), \ldots, (X_t^n, Y_t^n), \forall t \in \mathbb{N}_T$ are identically distributed. Assume that $\mathcal{F}$ is a (β, B)-Bernstein class of vector-valued functions. Let ψ be a sub-root function with the fixed point r^*. Suppose that*

$$B\mathfrak{R}(\mathcal{F}, r) \leq \psi(r), \quad \forall r \geq r^*,$$

where $\mathfrak{R}(\mathcal{F}, r) := \mathbb{E}\Big[\sup_{\boldsymbol{f} \in \mathcal{F}, V(\boldsymbol{f}) \leq r} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_t^i f_t(X_t^i)\Big]$ is the LRC of the function class $\mathcal{F}$. Then,

1. *For any $K > 1$, and $x > 0$, with probability at least $1 - e^{-x}$, every $\boldsymbol{f} \in \mathcal{F}$ satisfies*

$$P\boldsymbol{f} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} \max\left((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}}\right) + \left(\frac{2^{\beta+3} B^2 K^\beta x}{nT}\right)^{\frac{1}{2-\beta}} + \frac{24Bbx}{(2-\beta)nT}. \tag{B.5}$$

2. *If $\mathcal{F}$ is <u>convex</u> and $V(\alpha f) \leq \alpha^2 V(f)$ for any $\alpha \in [0,1], f \in \mathcal{F}$, then for any $K > 1$, and $x > 0$, the following inequality holds with probability at least $1 - e^{-x}$ for every $\boldsymbol{f} \in \mathcal{F}$*

$$P\boldsymbol{f} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + (2K)^{\frac{\beta}{2-\beta}} 4^{\frac{2}{2-\beta}} \max\left((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}}\right) + \left(\frac{2^{\beta+3} B^2 K^\beta x}{nT}\right)^{\frac{1}{2-\beta}} + \frac{24Bbx}{(2-\beta)nT}. \tag{B.6}$$

Proof. Similar to Lemma 9, define for the vector-valued function class $\mathcal{F}$,

$$\mathcal{F}_r := \left\{ \boldsymbol{f}' = (f'_1, \ldots, f'_T) : f'_t := \frac{r f_t}{\max(r, V(\boldsymbol{f}))}, \boldsymbol{f} = (f_t, \ldots, f_T) \in \mathcal{F} \right\}.$$

The proof can be broken down in two steps. The first step applies Theorem 28 and the seminal peeling technique [138, 139] to establish an inequality on the uniform deviation over the function class $\mathcal{F}_r$. The second step then uses the Bernstein assumption $V(\boldsymbol{f}) \leq BP\boldsymbol{f}^\beta$ to convert this inequality stated for $\mathcal{F}_r$ to a uniform deviation inequality for $\mathcal{F}$.

Step 1. Controlling uniform deviations for $\mathcal{F}_r$. To apply Theorem 28 to $\mathcal{F}_r$, we need to control the variances and uniform bounds for elements in $\mathcal{F}_r$. We first show $P\boldsymbol{f}'^2 \leq r, \forall \boldsymbol{f}' \in \mathcal{F}_r$. Indeed, for any $\boldsymbol{f} \in \mathcal{F}$ with $V(\boldsymbol{f}) \leq r$, the definition of $\mathcal{F}_r$ implies $f'_t = f_t$ and, hence, $P\boldsymbol{f}'^2 = P\boldsymbol{f}^2 \leq V(\boldsymbol{f}) \leq r$. Otherwise, if $V(\boldsymbol{f}) \geq r$, then $f'_t = r f_t / V(\boldsymbol{f})$ and we get

$$P\boldsymbol{f}'^2 = \frac{1}{T}\sum_{t=1}^T P f'^2_t = \frac{r^2}{\left[V(\boldsymbol{f})\right]^2}\left(\frac{1}{T}\sum_{t=1}^T P f_t^2\right) \leq \frac{r^2}{\left[V(\boldsymbol{f})\right]^2} V(\boldsymbol{f}) \leq r.$$

Therefore, $\frac{1}{T}\sup_{\boldsymbol{f}'\in\mathcal{F}_r}\sum_{t=1}^{T}\mathbb{E}[f_t'(X_t)]^2 \leq r$. Also, since functions in $\mathcal{F}$ admit a range of $[-b,b]$ and since $0 \leq r/\max(r,V(\boldsymbol{f})) \leq 1$, the inequality $\sup_{t,x}|f_t'(x)| \leq b$ holds for any $\boldsymbol{f}' \in \mathcal{F}_r$. Applying Theorem 28 to the function class $\mathcal{F}_r$ then yields the following inequality with probability at least $1-e^{-x}, \forall x > 0$

$$\sup_{\boldsymbol{f}'\in\mathcal{F}_r}[P\boldsymbol{f}' - P_n\boldsymbol{f}'] \leq 4\mathfrak{R}(\mathcal{F}_r) + \sqrt{\frac{8xr}{nT}} + \frac{12bx}{nT}. \tag{B.7}$$

It remains to control the Rademacher complexity of $\mathcal{F}_r$. Denote $\mathcal{F}(u,v) := \left\{\boldsymbol{f} \in \mathcal{F} : u \leq V(\boldsymbol{f}) \leq v\right\}, \forall 0 \leq u \leq v$, and introduce the notation

$$\mathfrak{R}_n\boldsymbol{f}' := \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t'(X_t^i), \qquad \mathfrak{R}_n(\mathcal{F}_r) := \sup_{\boldsymbol{f}'\in\mathcal{F}_r}\left[\mathfrak{R}_n\boldsymbol{f}'\right].$$

Note that $\mathfrak{R}(\mathcal{F}_r) = \mathbb{E}\mathfrak{R}_n(\mathcal{F}_r)$. Our assumption implies $V(\boldsymbol{f}) \leq BP\boldsymbol{f}^\beta \leq Bb^\beta, \forall \boldsymbol{f} \in \mathcal{F}$. Fix $\lambda > 1$ and define k to be the smallest integer such that $r\lambda^{k+1} \geq Bb^\beta$. Then, it follows from the union bound inequality

$$\mathfrak{R}(\mathcal{G}_1 \cup \mathcal{G}_2) \leq \mathfrak{R}(\mathcal{G}_1) + \mathfrak{R}(\mathcal{G}_2) \tag{B.8}$$

that

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_r) &= \mathbb{E}\Big[\sup_{\boldsymbol{f}'\in\mathcal{F}_r} \mathfrak{R}_n \boldsymbol{f}'\Big] = \mathbb{E}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \frac{r}{\max(r, V(\boldsymbol{f}))}\sigma_t^i f_t(X_t^i)\Big] \\
&\overset{\text{(B.8)}}{\leq} \mathbb{E}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}(0,r)} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_t^i f_t(X_t^i)\Big] + \mathbb{E}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}(r,Bb^\beta)} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \frac{r}{V(\boldsymbol{f})}\sigma_t^i f_t(X_t^i)\Big] \\
&\overset{\text{(B.8)}}{\leq} \mathbb{E}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}(0,r)} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_t^i f_t(X_t^i)\Big] + \sum_{j=0}^{k}\lambda^{-j}\mathbb{E}\Big[\sup_{\boldsymbol{f}\in\mathcal{F}(r\lambda^j, r\lambda^{j+1})} \mathfrak{R}_n \boldsymbol{f}\Big] \\
&\leq \mathfrak{R}(\mathcal{F}, r) + \sum_{j=0}^{k}\lambda^{-j}\mathfrak{R}\big(\mathcal{F}, r\lambda^{j+1}\big) \\
&\leq \frac{\psi(r)}{B} + \frac{1}{B}\sum_{j=0}^{k}\lambda^{-j}\psi(r\lambda^{j+1}).
\end{aligned}$$

The sub-root property of ψ implies that for any $\xi \geq 1$, $\psi(\xi r) \leq \xi^{\frac{1}{2}}\psi(r)$, and hence

$$\mathfrak{R}(\mathcal{F}_r) \leq \frac{\psi(r)}{B}\Big(1 + \sqrt{\lambda}\sum_{j=0}^{k}\lambda^{-j/2}\Big) \leq \frac{\psi(r)}{B}\Big(1 + \frac{\lambda}{\sqrt{\lambda}-1}\Big).$$

Taking the choice $\lambda = 4$ in the above inequality implies that $\mathfrak{R}(\mathcal{F}_r) \leq 5\psi(r)/B$, which, together with the inequality $\psi(r) \leq \sqrt{r/r^*}\psi(r^*) = \sqrt{rr^*}, \forall r \geq r^*$, gives

$$\mathfrak{R}(\mathcal{F}_r) \leq \frac{5}{B}\sqrt{rr^*}, \quad \forall r \geq r^*.$$

Combining (B.7) and the above inequality together, for any $r \geq r^*$ and $x > 0$, we derive the following inequality with probability at least $1 - e^{-x}$,

$$\sup_{\boldsymbol{f}'\in\mathcal{F}_r}[P\boldsymbol{f}' - P_n\boldsymbol{f}'] \leq \frac{20}{B}\sqrt{rr^*} + \sqrt{\frac{8xr}{nT}} + \frac{12bx}{nT}. \tag{B.9}$$

Step 2. Transferring uniform deviations for $\mathcal{F}_r$ to uniform deviations for $\mathcal{F}$. Setting $A =$

$20\sqrt{r^*}/B + \sqrt{8x/nT}$ and $C = 12bx/nT$, the upper bound (B.9) can be written as $A\sqrt{r} + C$, that is, $\sup_{\boldsymbol{f}' \in \mathcal{F}_r}[P\boldsymbol{f}' - P_n\boldsymbol{f}'] \leq A\sqrt{r}+C$. Now, according to Lemma 9, if $\sup_{\boldsymbol{f}' \in \mathcal{F}_r}[P\boldsymbol{f}' - P_n\boldsymbol{f}'] \leq \frac{r^{\frac{1}{\beta}}}{BK}$, then for any $\boldsymbol{f} \in \mathcal{F}$,

$$P\boldsymbol{f} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}}}{K}.$$

Therefore, in order to use the result of Lemma 9, we let $A\sqrt{r} + C = r^{\frac{1}{\beta}}/(BK)$. Assume r_0 is the unique positive solution of the equation $A\sqrt{r} + C = r^{\frac{1}{\beta}}/(BK)$, which can be written as

$$r^{\frac{1}{\beta}} - ABKr^{\frac{1}{2}} - BKC = 0.$$

Lemma 8 then implies

$$\begin{aligned} r_0^{\frac{1}{\beta}} &\leq (ABK)^{\frac{2}{2-\beta}} + \frac{2BKC}{2-\beta} \\ &\leq (BK)^{\frac{2}{2-\beta}} 2^{\frac{\beta}{2-\beta}} \Big[(20B^{-1})^{\frac{2}{2-\beta}} (r^*)^{\frac{1}{2-\beta}} + \Big(\frac{8x}{nT}\Big)^{\frac{1}{2-\beta}} \Big] + \frac{24BKbx}{(2-\beta)nT}, \end{aligned} \tag{B.10}$$

where we have used the inequality $(x+y)^p \leq 2^{p-1}(x^p + y^p)$ for any $x, y \geq 0, p \geq 1$. If $r^* \leq r_0$, we can take $r = r_0$ in (B.9) to derive $V_{r_0}^+ \leq A\sqrt{r_0} + C = r_0^{\frac{1}{\beta}}/(BK)$, which, coupled with (B.10) and Lemma 9, then implies

$$P\boldsymbol{f} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} (r^*)^{\frac{1}{2-\beta}} + \Big(\frac{2^{\beta+3} B^2 K^\beta x}{nT}\Big)^{\frac{1}{2-\beta}} + \frac{24Bbx}{(2-\beta)nT}, \tag{B.11}$$

If $r^* > r_0$, Lemma 8 implies $A\sqrt{r^*} + C \leq (r^*)^{\frac{1}{\beta}}/(BK)$. We now take $r = r^*$ in (B.9) to derive $V_{r^*}^+ \leq A\sqrt{r^*} + C \leq (r^*)^{\frac{1}{\beta}}/(BK)$, from which, via Lemma 9, we get that

$$P\boldsymbol{f} \leq \frac{K}{K-\beta} P_n \boldsymbol{f} + \frac{r_*^{\frac{1}{\beta}}}{K}. \tag{B.12}$$

The stated inequality (B.5) follows immediately by combining (B.11) and (B.12) together.

The proof of the second part follows from the fact that $\mathcal{F}_r \subseteq \{\boldsymbol{f} \in star(\mathcal{F}, 0) : V(\boldsymbol{f}) \leq r\}$, where $star(\mathcal{F}, f_0)$ is defined according to Definition 58. Also, since any convex class $\mathcal{F}$ is star-shaped around any of its points, we have $\mathcal{F}_r \subseteq \{\boldsymbol{f} \in \mathcal{F} : V(\boldsymbol{f}) \leq r\}$. Therefore, $\mathfrak{R}(\mathcal{F}_r)$ in (B.7) can be bounded as $\mathfrak{R}(\mathcal{F}_r) \leq \mathcal{R}(\mathcal{F}, r) \leq \psi(r)/B$. The rest proof of (B.6) is analogous to that of the first part and is omitted for brevity. □

Proof of Theorem 33

Note that the proof of this theorem relies on the results of Theorem 59. Introduce the following class of excess loss functions

$$\mathcal{H}^*_{\mathcal{F}} := \{h_{\boldsymbol{f}} = (h_{f_1}, \ldots, h_{f_T}), h_{f_t} : (X_t, Y_t) \mapsto \ell(f_t(X_t), Y_t) - \ell(f^*_t(X_t), Y_t), \boldsymbol{f} \in \mathcal{F}\}. \quad \text{(B.13)}$$

It can be shown that $\sup_{t,x} |h_{f_t}(x, y)| = \sup_{t,x} |\ell(f_t(x), y) - \ell(f^*_t(x), y)| \leq L \sup_{t,x} |f_t(x) - f^*_t(x)| \leq 2Lb$. Also, Assumption 32 implies

$$P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*})^2 \leq L^2 P(\boldsymbol{f} - \boldsymbol{f}^*)^2 \leq B' L^2 P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}), \quad \forall h_{\boldsymbol{f}} \in \mathcal{H}^*_{\mathcal{F}},$$

By taking $B = B' L^2$, we have for all $h_{\boldsymbol{f}} \in \mathcal{H}^*_{\mathcal{F}}$,

$$V(h_{\boldsymbol{f}}) := P h^2_{\boldsymbol{f}} \leq L^2 P(\boldsymbol{f} - \boldsymbol{f}^*)^2 \leq BP(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) = BP h_{\boldsymbol{f}}.$$

which implies that $\mathcal{H}^*_{\mathcal{F}}$ is a $(1, B)$-Bernstein class of vector-valued functions. Also, note that one can verify

$$\begin{aligned} B\mathfrak{R}(\mathcal{H}^*_{\mathcal{F}}, r) &= B\mathbb{E}_{X,\sigma}\left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ V(h_{\boldsymbol{f}})\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i h_{f_t}(X_t^i, Y_t^i)\right] \\ &= B\mathbb{E}_{X,\sigma}\left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ V(h_{\boldsymbol{f}})\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i \ell_{f_t}(X_t^i, Y_t^i)\right] \\ &\le BL\mathfrak{R}(\mathcal{F}^*, r) \le \psi(r), \end{aligned}$$

where the second last inequality is due to Talagrand's Lemma [83]. Applying Theorem 59 (which is the extension of Theorem 3.3 of [10] to MTL function classes) to the function class $\mathcal{H}^*_{\mathcal{F}}$ completes the proof.

The following lemma, as a consequence of Corollary 2.2 in [10], is essential in proving Theorem 35.

Lemma 11. *Assume that the functions in vector-valued function class* $\mathcal{F} = \{\boldsymbol{f} = (f_1, \ldots, f_T)\}$ *satisfy* $\sup_{t,x} |f_t(x)| \le b$ *with* $b > 0$. *For every* $x > 0$, *if* r *satisfies*

$$r \ge 32L^2 b\mathbb{E}_{\sigma,X}\left\{\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right\} + \frac{128L^2b^2x}{nT},$$

then, with probability at least $1 - e^{-x}$,

$$\left\{\boldsymbol{f}\in\mathcal{F} : L^2P\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2 \le r\right\} \subset \left\{\boldsymbol{f}\in\mathcal{F} : L^2P_n\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2 \le 2r\right\}.$$

Proof. First, define

$$\mathcal{F}_r^* := \left\{ \boldsymbol{f}' = (f_1', \dots, f_T') : \forall t, f_t' = (f_t - f_t^*)^2, \boldsymbol{f} = (f_1, \dots, f_T) \in \mathcal{F}, L^2 P(\boldsymbol{f} - \boldsymbol{f}^*)^2 \le r \right\}.$$

Note that for all $t \in \mathbb{N}_T$, $(f_t - f_t^*)^2 \in [0, 4b^2]$. Also, for any function in $\mathcal{F}_r^*$, it holds that

$$P\boldsymbol{f}'^2 = \frac{1}{T}\sum_{t=1}^{T} P f_t'^2 = \frac{1}{T}\sum_{t=1}^{T} P\left(f_t - f_t^*\right)^4 \le \frac{4b^2}{T}\sum_{t=1}^{T} P\left(f_t - f_t^*\right)^2 = 4b^2 P\left(\boldsymbol{f} - \boldsymbol{f}^*\right)^2 \le \frac{4b^2 r}{L^2}.$$

Therefore, by Theorem 28, with probability at least $1 - e^{-x}$, every $\boldsymbol{f}' \in \mathcal{F}_r^*$ satisfies

$$P_n \boldsymbol{f}' \le P\boldsymbol{f}' + 4\mathfrak{R}(\mathcal{F}_r^*) + \sqrt{\frac{32b^2 x r}{nTL^2}} + \frac{48b^2 x}{nT}, \tag{B.14}$$

where

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_r^*) &= \mathbb{E}_{\sigma,X}\left\{ \sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_t^i (f_t(X_t^i) - f_t^*(X_t^i))^2 \right\} \\
&\le 4b\mathbb{E}_{\sigma,X}\left\{ \sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_t^i f_t(X_t^i) \right\}.
\end{aligned} \tag{B.15}$$

The last inequality follows from the facts that $g(x) = x^2$ is $4b$-Lipschitz on $[-2b, 2b]$ and $\boldsymbol{f}^*$ is fixed. This together with (B.14), gives

$$\begin{aligned}
P_n\boldsymbol{f}' &\le P\boldsymbol{f}' + 16b\mathbb{E}_{\sigma,X}\left\{ \sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_t^i f_t(X_t^i) \right\} + \sqrt{\frac{32b^2xr}{nTL^2}} + \frac{48b^2x}{nT} \\
&\le \frac{r}{L^2} + 16b\mathbb{E}_{\sigma,X}\left\{ \sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_t^i f_t(X_t^i) \right\} + \frac{r}{2L^2} + \frac{64b^2x}{nT}.
\end{aligned} \tag{B.16}$$

Multiplying both sides by L^2 completes the proof. $\square$

Proof of Theorem 35

With $c_1 = 2L\max(B, 16Lb)$ and $c_2 = 128L^2b^2 + 2bc_1$, define the function $\psi(r)$ as

$$\psi(r) = \frac{c_1}{2}\mathbb{E}\left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\le r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right] + \frac{(c_2-2bc_1)x}{nT}. \tag{B.17}$$

Since $\mathcal{F}$ is convex, it is star-shaped around any of its points, thus using Lemma 3.4 in [10] it can be shown that $\psi(r)$ defined in (B.17) is a sub-root function. With the help of Corollary 34 and Assumption 32, we have with probability at least $1-e^{-x}$

$$L^2P\left(\hat{\boldsymbol{f}}-\boldsymbol{f}^*\right)^2 \le BP\left(\ell_{\hat{\boldsymbol{f}}}-\ell_{\boldsymbol{f}^*}\right) \le 32BKr + \frac{(48Lb+16BK)B^2x}{nT}. \tag{B.18}$$

where $B := B'L^2$. Denote the right hand side of the last inequality by s. Since $s \ge r \ge r^*$, then by the property of sub-root functions it holds that $s \ge \psi(s)$ which together with (B.17), gives

$$s \ge 32L^2b\mathbb{E}\left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\le s}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right] + \frac{128L^2b^2x}{nT}.$$

Applying Lemma 11, we have with probability at least $1-e^{-x}$,

$$\left\{\boldsymbol{f}\in\mathcal{F}, L^2P\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2 \le s\right\} \subset \left\{\boldsymbol{f}\in\mathcal{F}, L^2P_n\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2 \le 2s\right\}.$$

Combining this with (B.18), gives with probability at least $1-2e^{-x}$,

$$L^2P_n\left(\hat{\boldsymbol{f}}-\boldsymbol{f}^*\right)^2 \leq 2\left(32BKr+\frac{(48Lb+16BK)B^2x}{nT}\right)$$
$$\leq 2\left(32BK+\frac{(48Lb+16BK)B^2}{c_2}\right)r=cr. \tag{B.19}$$

where $c:=2(32BK+(48Lb+16BK)B^2/c_2)$ and in the last inequality we used the fact that $r\geq\psi(r)\geq c_2x/nT$. Applying the triangle inequality, if (B.19) holds, then for any $\boldsymbol{f}\in\mathcal{F}$, we have

$$L^2P_n\left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)^2 \leq \left(\sqrt{L^2P_n\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2}+\sqrt{L^2P_n\left(\boldsymbol{f}^*-\hat{\boldsymbol{f}}\right)^2}\right)^2$$
$$\leq \left(\sqrt{L^2P_n\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2}+\sqrt{cr}\right)^2. \tag{B.20}$$

Now, applying Lemma 11 for $r\geq\psi(r)$, implies that with probability at least $1-3e^{-x}$,

$$\left\{\boldsymbol{f}\in\mathcal{F}, L^2P\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2\leq r\right\}\subset\left\{\boldsymbol{f}\in\mathcal{F}, L^2P_n\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2\leq 2r\right\},$$

which coupled with (B.20), implies that with probability at least $1-3e^{-x}$,

$$\left\{\boldsymbol{f}\in\mathcal{F}, L^2P\left(\boldsymbol{f}-\boldsymbol{f}^*\right)^2\leq r\right\}\subset\left\{\boldsymbol{f}\in\mathcal{F}, L^2P_n\left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)^2\leq\left(\sqrt{2}+\sqrt{c}\right)^2r\right\}.$$

Also, with the help of Lemma A.4 in [10], it can be shown that with probability at least $1-e^{-x}$,

$$\mathbb{E}\left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\leq r}}\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right]\leq 2\mathbb{E}_\sigma\left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F},\\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\leq r}}\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right]+\frac{4bx}{nT}.$$

Thus, we will have with probability at least $1 - 4e^{-x}$,

$$\begin{aligned}
\psi(r) &\leq c_1 \mathbb{E}_\sigma \left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F}, \\ L^2 P(\boldsymbol{f}-\boldsymbol{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_t^i f_t(X_t^i) \middle| \left\{x_t^i\right\}_{t\in\mathbb{N}_T, i\in\mathbb{N}_n} \right] + \frac{c_2 x}{nT} \\
&\leq c_1 \mathbb{E}_\sigma \left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F}, \\ L^2 P_n(\boldsymbol{f}-\hat{\boldsymbol{f}})^2 \leq \left(\sqrt{2}+\sqrt{c}\right)^2 r}} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_t^i f_t(X_t^i) \middle| \left\{x_t^i\right\}_{t\in\mathbb{N}_T, i\in\mathbb{N}_n} \right] + \frac{c_2 x}{nT} \\
&\leq c_1 \mathbb{E}_\sigma \left[\sup_{\substack{\boldsymbol{f}\in\mathcal{F}, \\ L^2 P_n(\boldsymbol{f}-\hat{\boldsymbol{f}})^2 \leq (4+2c) r}} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_t^i f_t(X_t^i) \middle| \left\{x_t^i\right\}_{t\in\mathbb{N}_T, i\in\mathbb{N}_n} \right] + \frac{c_2 x}{nT} \\
&\leq \hat{\psi}(r). \qquad \text{(B.21)}
\end{aligned}$$

Setting $r = r^*$ and applying Lemma 4.3 of [10], gives $r^* \leq \hat{r}^*$ which together with (B.18) yields the result.

APPENDIX C: PROOFS OF THE RESULTS III

Proofs of the results for "Local Rademacher Complexity Bounds for MTL models with Strongly Convex Regularizers"

In the following, we would like to provide some basic notions of convex analysis which are helpful in understanding the results of Sect. 5.

Definition 60 (STRONG CONVEXITY)**.** *A function $R : \mathcal{X} \mapsto \mathbb{R}$ is μ-strong convex w.r.t. a norm $\|.\|$ if and only if $\forall x, y \in \mathcal{X}$ and $\forall \alpha \in (0, 1)$, we have*

$$R(\alpha x + (1-\alpha)y) \leq \alpha R(x) + (1-\alpha)R(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2.$$

Definition 61 (STRONG SMOOTHNESS)**.** *A function $R^* : \mathcal{X} \mapsto \mathbb{R}$ is $\frac{1}{\mu}$-strong smooth w.r.t. a norm $\|.\|_*$ if and only if R^* is everywhere differentiable and $\forall x, y \in \mathcal{X}$, we have*

$$R^*(x+y) \leq R^*(x) + \langle \nabla R^*(x), y \rangle + \frac{1}{2\mu}\|y\|_*^2.$$

Property 62 (Theorem 3 in [69]: strong convexity/strong smoothness duality)**.** *A function R is μ-strongly convex w.r.t. the norm $\|.\|$ if and only if its Fenchel conjugate R^* is $\frac{1}{\mu}$-strongly smooth w.r.t. the dual norm $\|.\|_*$. The Fenchel conjugate R^* is defined as*

$$R^*(\boldsymbol{w}) := \sup_{\boldsymbol{v}} \{\langle \boldsymbol{w}, \boldsymbol{v} \rangle - R(\boldsymbol{v})\}.$$

Property 63 (FENCHEL-YOUNG INEQUALITY)**.** *The definition of Fenchel dual implies that for any strong convex function R,*

$$\forall \boldsymbol{w}, \boldsymbol{v} \in S,\ \langle \boldsymbol{w}, \boldsymbol{v} \rangle \leq R(\boldsymbol{w}) + R^*(\boldsymbol{v}).$$

Combining this with the strong duality property of R^* gives the following

$$\langle \boldsymbol{w}, \boldsymbol{v} \rangle - R(\boldsymbol{w}) \leq R^*(\boldsymbol{v}) \leq R^*(\mathbf{0}) + \langle \triangledown R^*(\mathbf{0}), \boldsymbol{v} \rangle + \frac{1}{2\mu} \|\boldsymbol{v}\|_*^2 . \tag{C.1}$$

Lemma 12. *Assume that the conditions of Theorem 36 hold. Then, for ever $\boldsymbol{f} \in \mathcal{F}_q$,*

(a) *$P\boldsymbol{f}^2 \leq r$ implies $1/T \sum_{t=1}^{T} \sum_{j=1}^{\infty} \lambda_t^j \left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j \right\rangle^2 \leq r$.*

(b) $\mathbb{E}_{X,\sigma} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 = \frac{\lambda_t^j}{n}$.

Proof.

Part (**a**)

$$\begin{aligned}
P\boldsymbol{f}^2 =& \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left(\left\langle \boldsymbol{w}_t, \phi(X_t^i) \right\rangle \right)^2 \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left(\left\langle \boldsymbol{w}_t \otimes \boldsymbol{w}_t, \phi(X_t^i) \otimes \phi(X_t^i) \right\rangle \right) \\
=& \frac{1}{T} \sum_{t=1}^{T} \left\langle \boldsymbol{w}_t \otimes \boldsymbol{w}_t, \mathbb{E}_X \left(\phi(X_t^i) \otimes \phi(X_t^i) \right) \right\rangle = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{\infty} \lambda_t^j \left\langle \boldsymbol{w}_t \otimes \boldsymbol{w}_t, \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j \right\rangle \\
=& \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{\infty} \lambda_t^j \left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j \right\rangle \left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j \right\rangle = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{\infty} \lambda_t^j \left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j \right\rangle^2 \leq r.
\end{aligned}$$

Part (**b**)

$$\begin{aligned}
&\mathbb{E}_{X,\sigma} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 = \frac{1}{n^2} \mathbb{E}_{X,\sigma} \sum_{i,k=1}^{n} \sigma_t^i \sigma_t^k \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \left\langle \phi(X_t^k), \boldsymbol{u}_t^j \right\rangle \\
&\overset{\boldsymbol{\sigma_t} \text{i.i.d.}}{=} \frac{1}{n^2} \mathbb{E}_X \left(\sum_{i=1}^{n} \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right) = \frac{1}{n} \left\langle \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_X \left(\phi(X_t^i) \otimes \phi(X_t^i) \right), \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j \right\rangle \\
&= \frac{1}{n} \sum_{l=1}^{\infty} \lambda_t^l \left\langle \boldsymbol{u}_t^l \otimes \boldsymbol{u}_t^l, \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j \right\rangle = \frac{\lambda_t^j}{n}.
\end{aligned}$$

□

The following lemmas are used in the proof of the LRC bound for the $L_{2,q}$-group norm regularized MTL in Corollary 42.

Lemma 13 (Khintchine-Kahane Inequality [118])**.** *Let $\mathcal{H}$ be an inner-product space with induced norm $\|\cdot\|_{\mathcal{H}}$, $v_1, \ldots, v_M \in \mathcal{H}$ and $\sigma_1, \ldots, \sigma_n$ i.i.d. Rademacher random variables. Then, for any $p \geq 1$, we have that*

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^{n} \sigma_i v_i \right\|_{\mathcal{H}}^{p} \leq \left(c \sum_{i=1}^{n} \|v_i\|_{\mathcal{H}}^2 \right)^{\frac{p}{2}}. \tag{C.2}$$

where $c := \max\{1, p-1\}$. The inequality also holds for p in place of c.

Lemma 14 (Rosenthal-Young Inequality; Lemma 3 of [72])**.** *Let the independent non-negative random variables $X_1, \ldots, X_n$ satisfy $X_i \leq B < +\infty$ almost surely for all $i = 1, \ldots, n$. If $q \geq \frac{1}{2}$, $c_q := (2qe)^q$, then it holds*

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^q \leq C_q \left[\left(\frac{B}{n} \right)^q + \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} X_i \right)^q \right]. \tag{C.3}$$

Proof of Lemma 6

For the group norm regularizer $\|\boldsymbol{W}\|_{2,q}$, we can further bound the expectation term in (5.16) for $\boldsymbol{D} = \boldsymbol{I}$ as follows

$$
\begin{aligned}
\mathbb{E} &:= \mathbb{E}_{X,\sigma} \left\| \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^{T} \right\|_{2,q^*} \\
&= \mathbb{E}_{X,\sigma} \left(\sum_{t=1}^{T} \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right\|^{q^*} \right)^{\frac{1}{q^*}} \\
&\overset{\text{Jensen}}{\leq} \mathbb{E}_X \left(\sum_{t=1}^{T} \mathbb{E}_\sigma \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right\|^{q^*} \right)^{\frac{1}{q^*}} \\
&\overset{\text{(C.2)}}{\leq} \mathbb{E}_X \left(\sum_{t=1}^{T} \left(q^* \sum_{i=1}^{n} \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right\|^2 \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} \\
&= \sqrt{\frac{q^*}{n}} \mathbb{E}_X \left(\sum_{t=1}^{T} \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^{n} \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} \\
&\overset{\text{Jensen}}{\leq} \sqrt{\frac{q^*}{n}} \left(\sum_{t=1}^{T} \mathbb{E}_X \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^{n} \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} \qquad \text{(C.4)}
\end{aligned}
$$

Note that for $q \leq 2$, it holds that $q^*/2 \geq 1$. Therefore, we cannot employ Jensen's inequality to move the expectation operator inside the inner term, and instead we need to apply the Rosenthal-

Young (R+Y) inequality (see Lemma 14 in the Appendix), which yields

$$\mathbb{E} \overset{\text{R+Y}}{\leq} \sqrt{\frac{q^*}{n}} \left(\sum_{t=1}^{T} (eq^*)^{\frac{q^*}{2}} \left(\left(\frac{\mathcal{K}}{n} \right)^{\frac{q^*}{2}} + \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_X \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right) \right)^{\frac{1}{q^*}}$$
$$= \sqrt{\frac{q^*}{n}} \left(\sum_{t=1}^{T} (eq^*)^{\frac{q^*}{2}} \left(\left(\frac{\mathcal{K}}{n} \right)^{\frac{q^*}{2}} + \left(\sum_{j>h_t} \lambda_t^j \right)^{\frac{q^*}{2}} \right) \right)^{\frac{1}{q^*}}. \tag{C.5}$$

The last quantity can be further bounded using the sub-additivity of $\sqrt[q^*]{\cdot}$ and $\sqrt{\cdot}$ respectively in $(\dagger\dagger)$ and $(\dagger)$ below,

$$\mathbb{E} \overset{(\dagger)}{\leq} q^* \sqrt{\frac{e}{n}} \left[\left(T \left(\frac{\mathcal{K}}{n} \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} + \left(\sum_{t=1}^{T} \left(\sum_{j>h_t} \lambda_t^j \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} \right]$$
$$\overset{(\dagger\dagger)}{\leq} q^* \sqrt{\frac{e}{n}} \left[T^{\frac{1}{q^*}} \sqrt{\frac{\mathcal{K}}{n}} + \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}^{\frac{1}{2}} \right]$$
$$= \frac{\sqrt{\mathcal{K}e} q^* T^{\frac{1}{q^*}}}{n} + \sqrt{\frac{eq^{*2}}{n} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}}. \tag{C.6}$$

Proof of Corollary 42

Substituting the result of Lemma 6 into (5.18) gives,

$$A_2(\mathcal{F}_q) \leq \sqrt{\frac{2eq^{*2} R_{max}^2}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e} R_{max} q^* T^{\frac{1}{q^*}}}{nT}. \tag{C.7}$$

Now, combining (5.15) and (C.7) provides the bound on $\mathfrak{R}(\mathcal{F}_q, r)$ as

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_q, r) &\leq \sqrt{\frac{r\sum_{t=1}^{T} h_t}{nT}} + \sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT} \qquad \text{(C.8)}\\
&\overset{(\star)}{\leq} \sqrt{\frac{2}{nT}\left(r\sum_{t=1}^{T} h_t + \frac{2eq^{*2}R_{max}^2}{T}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}\right)} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}\\
&\overset{(\star\star)}{\leq} \sqrt{\frac{2}{nT}\left(rT^{1-\frac{2}{q^*}}\left\|(h_t)_{t=1}^{T}\right\|_{\frac{q^*}{2}} + \frac{2eq^{*2}R_{max}^2}{T}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}\right)} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}\\
&\overset{(\star\star\star)}{\leq} \sqrt{\frac{4}{nT}\left\|\left(rT^{1-\frac{2}{q^*}}h_t + \frac{2eq^{*2}R_{max}^2}{T}\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}.
\end{aligned}$$

where in $(\star)$, $(\star\star)$ and $(\star\star\star)$ we applied following inequalities receptively, according which for all non-negative numbers α_1 and α_2, and non-negative vectors $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^T$ with $0 \leq q \leq p \leq \infty$ and $s \geq 1$ it holds

$$\begin{aligned}
&(\star)\sqrt{\alpha_1} + \sqrt{\alpha_2} \leq \sqrt{2(\alpha_1+\alpha_2)}\\
&(\star\star)\quad l_p - to - l_q: \quad \|\boldsymbol{a}_1\|_q = \langle \mathbf{1}, \boldsymbol{a}_1\rangle^{\frac{1}{q}} \overset{\text{Hölder}}{\leq} \left(\|\mathbf{1}\|_{(p/q)^*}\left\|\boldsymbol{a}_1^q\right\|_{(p/q)}\right)^{\frac{1}{q}} = T^{\frac{1}{q}-\frac{1}{p}}\|\boldsymbol{a}_1\|_p\\
&(\star\star\star)\quad \|\boldsymbol{a}_1\|_s + \|\boldsymbol{a}_2\|_s \leq 2^{1-\frac{1}{s}}\|\boldsymbol{a}_1+\boldsymbol{a}_2\|_s \leq 2\|\boldsymbol{a}_1+\boldsymbol{a}_2\|_s.
\end{aligned}$$

Since inequality $(\star\star\star)$ holds for all non-negative h_t, it follows

$$\begin{aligned}\mathfrak{R}(\mathcal{F}_q, r) &\leq \sqrt{\frac{4}{nT}\left\|\left(\min_{h_t\geq 0} rT^{1-\frac{2}{q^*}}h_t + \frac{2eq^{*2}R_{max}^2}{T}\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT} \\ &\leq \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{q^*}}, \frac{2eq^{*2}R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}.\end{aligned}$$

Proof of Theorem 46

$$
\begin{aligned}
\mathfrak{R}(\mathcal{F}_{q,R,T}, r) &= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{P\boldsymbol{f}^2\le r,\\ \|\boldsymbol{W}\|_{2,q}^2\le 2R^2}} \sum_{t=1}^{T}\left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i)\right\rangle\right\} \\
&= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{1/T\sum_{t=1}^{T}\mathbb{E}\langle \boldsymbol{w}_t,\phi(X_t)\rangle^2\le r,\\ \|\boldsymbol{W}\|_{2,q}^2\le 2R^2}} \sum_{t=1}^{T}\left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i)\right\rangle\right\} \\
&\ge \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{\forall t\ \mathbb{E}_X\langle \boldsymbol{w}_t,\phi(X_t)\rangle^2\le r,\\ \|\boldsymbol{W}\|_{2,q}^2\le 2R^2,\\ \|\boldsymbol{w}_1\|_2=\ldots=\|\boldsymbol{w}_t\|_2}} \sum_{t=1}^{T}\left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i)\right\rangle\right\} \\
&= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{\forall t\ \mathbb{E}_X\langle \boldsymbol{w}_t,\phi(X_t)\rangle^2\le r,\\ \forall t\ \|\boldsymbol{w}_t\|_2^2\le 2R^2T^{-\frac{2}{q}}}} \sum_{t=1}^{T}\left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i)\right\rangle\right\} \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{\forall t\ \mathbb{E}_X\langle \boldsymbol{w}_t,\phi(X_t)\rangle^2\le r,\\ \forall t\ \|\boldsymbol{w}_t\|_2^2\le 2R^2T^{-\frac{2}{q}}}} \left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i)\right\rangle\right\} \\
&= \mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{\mathbb{E}_X\langle \boldsymbol{w}_1,\phi(X_1)\rangle^2\le r,\\ \|\boldsymbol{w}_1\|_2^2\le 2R^2T^{-\frac{2}{q}}}} \left\langle \boldsymbol{w}_1, \frac{1}{n}\sum_{i=1}^{n}\sigma_1^i\phi(X_1^i)\right\rangle\right\} \\
&= \mathfrak{R}(\mathcal{F}_{1,RT^{-\frac{1}{q}},1}, r).
\end{aligned}
$$

According to [108], it can be shown that there is a constant c such that if $\lambda_t^1 \ge \frac{1}{nR_{max}^2}$, then for all $r \ge \frac{1}{n}$ it holds $\mathfrak{R}(\mathcal{F}_{1,RT^{-\frac{1}{q}},1}, r) \ge \sqrt{\frac{c}{n}\sum_{j=1}^{\infty}\min\left(r, R^2T^{-\frac{2}{q}}\lambda_1^j\right)}$, which with some algebra manipulations gives the desired result.

The following lemma is used in the proof of the LRC bounds for the L_{S_q}-Schatten norm regularized MTL in Corollary 48.

Lemma 15 (Non-commutative Khintchine's inequality [95])**.** *Let $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ be a set of arbitrary $m \times n$ matrices, and let $\sigma_1, \ldots, \sigma_n$ be a sequence of independent Bernoulli random variables. Then for all $p \geq 2$,*

$$\left[\mathbb{E}_\sigma \left\| \sum_{i=1}^{n} \sigma_i \boldsymbol{Q}_i \right\|_{S_p}^{p}\right]^{1/p} \leq p^{1/2} \max\left\{ \left\| \left(\sum_{i=1}^{n} \boldsymbol{Q}_i^T \boldsymbol{Q}_i \right)^{1/2} \right\|_{S_p}, \left\| \left(\sum_{i=1}^{n} \boldsymbol{Q}_i \boldsymbol{Q}_i^T \right)^{1/2} \right\|_{S_p} \right\}. \quad \text{(C.9)}$$

Proof of Corollary 48

In order to find an LRC bound for a L_{S_q}-Schatten norm regularized hypothesis space (5.30), one just needs to bound the expectation term in (5.12). Define $\boldsymbol{U}_t^i$ as a matrix with T columns, whose only non-zero t^{th} column equals $\sum_{j>h_t} \left\langle \frac{1}{n}\phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j$. Also, note that for the Schatten norm

regularized hypothesis space (5.30), it holds that $\boldsymbol{D} = \boldsymbol{I}$. Therefore, we will have

$$
\begin{aligned}
\mathbb{E}_{X,\sigma} \left\| \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\|_* &= \mathbb{E}_{X,\sigma} \left\| \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^{T} \right\|_{S_{q^*}} \\
&= \mathbb{E}_{X,\sigma} \left\| \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_t^i \boldsymbol{U}_t^i \right\|_{S_{q^*}} \overset{\text{Jensen}}{\leq} \mathbb{E}_X \left\{ \mathbb{E}_\sigma \left\| \sum_{t=1}^{T} \sum_{i=1}^{n} \sigma_t^i \boldsymbol{U}_t^i \right\|_{S_{q^*}}^{q^*} \right\}^{\frac{1}{q^*}} \\
&\overset{\text{(C.9)}}{\leq} \sqrt{q^*} \mathbb{E}_X \max \left\{ \left\| \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \left(\boldsymbol{U}_t^i \right)^T \boldsymbol{U}_t^i \right)^{1/2} \right\|_{S_{q^*}}, \left\| \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \boldsymbol{U}_t^i \left(\boldsymbol{U}_t^i \right)^T \right)^{1/2} \right\|_{S_{q^*}} \right\} \\
&\overset{(\dagger\dagger\dagger)}{=} \sqrt{q^*} \mathbb{E}_X \left\| \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \left(\boldsymbol{U}_t^i \right)^T \boldsymbol{U}_t^i \right)^{1/2} \right\|_{S_{q^*}} = \sqrt{q^*} \mathbb{E}_X \left(\mathbf{tr} \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \left(\boldsymbol{U}_t^i \right)^T \boldsymbol{U}_t^i \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} \\
&= \sqrt{q^*} \mathbb{E}_X \left(\left(\sum_{t=1}^{T} \sum_{i=1}^{n} \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right\|^2 \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} \\
&= \sqrt{q^*} \mathbb{E}_X \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right\|^2 \right)^{\frac{1}{2}} \\
&= \frac{\sqrt{q^*}}{n} \mathbb{E}_X \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{j>h_t} \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{1}{2}} \\
&\overset{\text{Jensen}}{\leq} \frac{\sqrt{q^*}}{n} \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{j>h_t} \lambda_t^j \right)^{\frac{1}{2}} = \sqrt{\frac{q^*}{n} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_1}.
\end{aligned} \tag{C.10}
$$

where in $(\dagger\dagger\dagger)$, we assumed that the first term in the max argument is the largest one.

Proof of Corollary 51

Similar to the proof of Corollary 48, for the graph regularized hypothesis space (5.32), one can bound the expectation term in (5.12) as

$$
\begin{aligned}
\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_{*} &= \mathbb{E}_{X,\sigma}\left[\operatorname{tr}\left(\boldsymbol{V}^{T}\boldsymbol{D}^{-1}\boldsymbol{V}\right)\right]^{\frac{1}{2}} \\
&\overset{\text{Jensen}}{\leq} \mathbb{E}_{X}\left(\frac{1}{n^{2}}\sum_{t,s=1}^{T,T}\sum_{i,l=1}^{n,n}\sum_{j>h_t}\sum_{k>h_s}\boldsymbol{D}_{st}^{-1}\mathbb{E}_{\sigma}\left(\sigma_{t}^{i}\sigma_{s}^{l}\right)\left\langle\phi(X_{t}^{i}),\boldsymbol{u}_{t}^{j}\right\rangle\left\langle\phi(X_{s}^{l}),\boldsymbol{u}_{s}^{k}\right\rangle\left\langle\boldsymbol{u}_{t}^{j},\boldsymbol{u}_{s}^{k}\right\rangle\right)^{\frac{1}{2}} \\
&= \mathbb{E}_{X}\left(\frac{1}{n}\sum_{t=1}^{T}\boldsymbol{D}_{tt}^{-1}\sum_{j>h_t}\frac{1}{n}\sum_{i=1}^{n}\left\langle\phi(X_{t}^{i}),\boldsymbol{u}_{t}^{j}\right\rangle^{2}\right)^{\frac{1}{2}} \\
&\overset{\text{Jensen}}{\leq} \left(\frac{1}{n}\sum_{t=1}^{T}\boldsymbol{D}_{tt}^{-1}\sum_{j>h_t}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X}\left\langle\phi(X_{t}^{i}),\boldsymbol{u}_{t}^{j}\right\rangle^{2}\right)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{T}\sum_{j>h_t}\boldsymbol{D}_{tt}^{-1}\lambda_{t}^{j}\right)^{\frac{1}{2}} = \sqrt{\frac{1}{n}\left\|\left(\boldsymbol{D}_{tt}^{-1}\sum_{j>h_t}\lambda_{t}^{j}\right)_{t=1}^{T}\right\|_{1}}.
\end{aligned}
\tag{C.11}
$$

APPENDIX D: PROOFS OF THE RESULTS IV

Proof of the results for "Excess Risk Bounds for MTL models with Strongly Convex Regularizers"

Proof of Corollary 55

First notice that $\hat{\mathfrak{R}}(\mathcal{F}_q^*, c_3 r) \leq 2\hat{\mathfrak{R}}(\mathcal{F}_q, \frac{c_3 r}{4L^2})$. Assume that $(\hat{\boldsymbol{u}}_t^j)_{j\geq 1}$ is an orthonormal basis of $\mathcal{H}_K$ of matrix $\boldsymbol{K}_t$. Then similar to the proof of Theorem 38 it can be shown that

$$\begin{aligned}
\hat{\mathfrak{R}}(\mathcal{F}_q, r) &\leq \frac{1}{T}\mathbb{E}_\sigma \left\{ \sup_{P_n \boldsymbol{f}^2 \leq r} \left[\left(\sum_{t=1}^{T} \sum_{j=1}^{\hat{h}_t} \hat{\lambda}_t^j \left\langle \boldsymbol{w}_t, \hat{\boldsymbol{u}}_t^j \right\rangle^2 \right)^{\frac{1}{2}} \right.\right. \\
&\qquad \left.\left. \left(\sum_{t=1}^{T} \sum_{j=1}^{\hat{h}_t} \hat{\lambda}_t^{j^{-1}} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \hat{\phi}(X_t^i), \hat{\boldsymbol{u}}_t^j \right\rangle^2 \right)^{\frac{1}{2}} \right] \right\} \\
&\quad + \frac{\sqrt{2}R}{T} \mathbb{E}_\sigma \left\| \boldsymbol{D}^{-1/2} \hat{\boldsymbol{V}} \right\|_{2,q^*} \\
&\leq \sqrt{\frac{r \sum_{t=1}^{T} \hat{h}_t}{nT}} + \frac{\sqrt{2}R}{T} \mathbb{E}_\sigma \left\| \left(\sum_{j>\hat{h}_t}^{n} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \hat{\phi}(X_t^i), \hat{\boldsymbol{u}}_t^j \right\rangle \hat{\boldsymbol{u}}_t^j \right)_{t=1}^{T} \right\|_{2,q^*}
\end{aligned}$$

where the last inequality is obtained by replacing $\hat{\boldsymbol{V}} = \left(\sum_{j>\hat{h}_t}^{n} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \hat{\phi}(X_t^i), \hat{\boldsymbol{u}}_t^j \right\rangle \hat{\boldsymbol{u}}_t^j \right)_{t=1}^{T}$ and $\boldsymbol{D} = \boldsymbol{I}$, and regarding the fact that $\mathbb{E}_\sigma \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \hat{\phi}(X_t^i), \hat{\boldsymbol{u}}_t^j \right\rangle^2 = \frac{\hat{\lambda}_t^j}{n}$ and $P_n \boldsymbol{f}^2 \leq r$ implies $\frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{n} \hat{\lambda}_t^j \left\langle \boldsymbol{w}_t, \hat{\boldsymbol{u}}_t^j \right\rangle^2 \leq r$.

Now, similar to the proof of Lemma 6, it can be shown that

$$\mathbb{E}_\sigma \left\| \left(\sum_{j>\hat{h}_t} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \hat{\phi}(X_t^i), \hat{\boldsymbol{u}}_t^j \right\rangle \hat{\boldsymbol{u}}_t^j \right)_{t=1}^{T} \right\|_{2,q^*} \leq \sqrt{\frac{q^{*2}}{n} \left\| \left(\sum_{j>\hat{h}_t} \hat{\lambda}_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}}.$$

Note that, for the empirical LRC, the expectation is taken only with respect to the Radamacher

variables $(\sigma_t^i)_{(t,i=1)}^{(T,n)}$. Therefore, we get

$$\hat{\mathfrak{R}}(\mathcal{F}_q, \frac{c_3 r}{4L^2}) \leq \sqrt{\frac{c_3 r \sum_{t=1}^{T} \hat{h}_t}{4nTL^2}} + \sqrt{\frac{2q^{*2}R_{max}^2}{nT^2} \left\| \left(\sum_{j>\hat{h}_t}^{n} \hat{\lambda}_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}},$$

which implies,

$$\begin{aligned}\hat{\psi}_n(r) &\leq 2c_1 \left(\sqrt{\frac{c_3 r \sum_{t=1}^{T} \hat{h}_t}{4nTL^2}} + \sqrt{\frac{2q^{*2}R_{max}^2}{nT^2} \left\| \left(\sum_{j>\hat{h}_t}^{n} \hat{\lambda}_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}} \right) + \frac{c_2 x}{nT} \\ &= \sqrt{\frac{c_1^2 c_3 r \sum_{t=1}^{T} \hat{h}_t}{nTL^2}} + \sqrt{\frac{8c_1^2 q^{*2}R_{max}^2}{nT^2} \left\| \left(\sum_{j>\hat{h}_t}^{n} \hat{\lambda}_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}} + \frac{c_2 x}{nT}.\end{aligned}$$

Denote the right hand side by $\hat{\psi}_n^{ub}(r)$. Solving the fixed point equation $\hat{\psi}_n^{ub}(r) = \sqrt{\alpha r} + \gamma = r$ for

$$\alpha = \frac{c_1^2 c_3 \sum_{t=1}^{T} \hat{h}_t}{nTL^2}, \qquad \gamma = \sqrt{\frac{8c_1^2 q^{*2}R_{max}^2}{nT^2} \left\| \left(\sum_{j>\hat{h}_t}^{n} \hat{\lambda}_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}} + \frac{c_2 x}{nT}, \tag{D.1}$$

gives $\hat{r}^* \leq \alpha + 2\gamma$. Substituting α and γ completes the proof.

Proof of the results in Sect. 5: "Discussion"

Proof of Theorem 57

Note that regarding the definition of A_2 in (5.14), the global rademacher complexity for each case can be obtained by replacing the tail-sum $\sum_{j>h_t} \lambda_t^j$ in the bound of its corresponding $A_2(\mathcal{F})$ by $\sum_{j=1}^{\infty} \lambda_t^j = \mathbf{tr}(J_t)$. Indeed, similar to the proof of Lemma 6, it can be shown that for the group

norm with $q \in [1, 2]$,

$$\begin{aligned}\mathfrak{R}(\mathcal{F}_q) &= \mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}=(f_1,\ldots,f_T)\in\mathcal{F}_q} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right\} \\ &\leq \frac{\sqrt{2}R}{T}\mathbb{E}_{X,\sigma}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i)\right)_{t=1}^{T}\right\|_{2,q^*}.\end{aligned}$$

Also, one can verify the following

$$\begin{aligned}\mathbb{E}_{X,\sigma}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i)\right)_{t=1}^{T}\right\|_{2,q^*} &= \mathbb{E}_{X,\sigma}\left\|\left(\sum_{j=1}^{\infty}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\|_{2,q^*} \\ &\leq \sqrt{\frac{q^{*2}\mathcal{K}eT^{\frac{2}{q^*}}}{n^2}} + \sqrt{\frac{eq^{*2}}{n}\left\|\left(\sum_{j=1}^{\infty}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} \\ &= \frac{\sqrt{\mathcal{K}e}q^*T^{\frac{1}{q^*}}}{n} + \sqrt{\frac{eq^{*2}}{n}\left\|\left(\operatorname{tr}\left(J_t\right)\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}. \qquad \text{(D.2)}\end{aligned}$$

where the inequality is obtained in a similar way as in Lemma 6. The GRC bounds for the other cases can be easily derived in a very similar manner.

LIST OF REFERENCES

[1] Qi An, Chunping Wang, Ivo Shterev, Eric Wang, Lawrence Carin, and David B Dunson. Hierarchical kernel stick-breaking process for multi-task image analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 17–24. ACM, 2008.

[2] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[3] Andreas Argyriou, Stéphan Clémençon, and Ruocong Zhang. Learning the graph of relations among multiple tasks. *ICML 2014 workshop on New Learning Frameworks and Models for Big Data*, 2013.

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, pages 41–48, 2007.

[5] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[6] Andreas Argyriou, Andreas Maurer, and Massimiliano Pontil. An algorithm for transfer learning in a heterogeneous environment. In *Machine Learning and Knowledge Discovery in Databases*, pages 71–85. Springer, 2008.

[7] Andreas Argyriou, Massimiliano Pontil, Yiming Ying, and Charles A Micchelli. A spectral regularization framework for multi-task structure learning. In *Advances in neural information processing systems*, pages 25–32, 2007.

[8] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, 4:83–99, 2003.

[9] Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

[10] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

[11] Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12.149-198, 2000.

[12] Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.

[13] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

[14] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, second edition edition, 1999.

[15] Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the 25th international conference on Machine learning*, pages 56–63. ACM, 2008.

[16] Sergey Bobkov and Michel Ledoux. Poincarés inequalities and talagrands concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107(3):383–400, 1997.

[17] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *Annals of Probability*, pages 1583–1614, 2003.

[18] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

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

[20] Olivier Bousquet. Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. 2002.

[21] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[22] Bin Cao, Nathan N Liu, and Qiang Yang. Transfer learning for collective link prediction in multiple heterogenous domains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 159–166, 2010.

[23] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[24] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[25] Anveshi Charuvaka and Huzefa Rangwala. Convex multi-task relationship learning using hinge loss. In *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, pages 63–70. IEEE, 2014.

[26] Jianhui Chen, Lei Tang, Jun Liu, and Jieping Ye. A convex formulation for learning a shared predictive structure from multiple tasks. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1025–1038, 2013.

[27] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50. ACM, 2011.

[28] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Low-rank and sparse multi-task learning. In *Low-Rank and Sparse Modeling for Visual Analysis*, pages 151–180. Springer, 2014.

[29] Xi Chen, Seyoung Kim, Qihang Lin, Jaime G Carbonell, and Eric P Xing. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*, 2010.

[30] Carlo Ciliberto, Lorenzo Rosasco, and Silvia Villa. Learning multiple visual tasks while discovering their structure. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 131–139, 2015.

[31] Corinna Cortes. Can learning kernels help performance. *Invited talk at International Conference on Machine Learning (ICML)*, 2009.

[32] Corinna Cortes, Xavi Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Adanet: Adaptive structural learning of artificial neural networks. *arXiv preprint arXiv:1607.01097*, 2016.

[33] Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2760–2768. Curran Associates, Inc., 2013.

[34] Corinna Cortes and Mehryar Mohri. *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, chapter Domain Adaptation in Regression, pages 308–323. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[35] Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103 – 126, 2014. Algorithmic Learning Theory.

[36] Corinna Cortes, Mehryar Mohri, and Afshin Rostami. Tutorial: Learning kernels. *ICML*, 2011.

[37] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L 2 regularization for learning kernels. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press, 2009.

[38] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 247–254, 2010.

[39] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828, 2012.

[40] Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

[41] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

[42] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[43] Miroslav Dudık, Zaıd Harchaoui, and Jérôme Malick. Lifted coordinate descent for learning with trace-norm regularization.

[44] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[45] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.

[46] Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. In *Journal of Machine Learning Research*, pages 615–637, 2005.

[47] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

[48] Hongliang Fei and Jun Huan. Structured feature selection and task relationship inference for multi-task learning. *Knowledge and information systems*, 35(2):345–364, 2013.

[49] Sergey Feldman, Maya Gupta, and Bela Frigyik. Multi-task averaging. In *Advances in Neural Information Processing Systems*, pages 1169–1177, 2012.

[50] Rémi Flamary, Alain Rakotomamonjy, and Gilles Gasso. Learning constrained task similarities in graph-regularized multi-task learning. *Regularization, Optimization, Kernels, and Support Vector Machines*, page 103, 2014.

[51] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[52] Joumana Ghosn and Yoshua Bengio. Multi-task learning for stock selection. *Advances in Neural Information Processing Systems*, pages 946–952, 1997.

[53] André R Gonçalves, Puja Das, Soumyadeep Chatterjee, Vidyashankar Sivakumar, Fernando J Von Zuben, and Arindam Banerjee. Multi-task sparse structure learning. In *Pro-*

ceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 451–460. ACM, 2014.

[54] Pinghua Gong, Jieping Ye, and Changshui Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 895–903. ACM, 2012.

[55] Pinghua Gong, Jiayu Zhou, Wei Fan, and Jieping Ye. Efficient multi-task feature learning with calibration. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 761–770. ACM, 2014.

[56] Steve R Gunn et al. Support vector machines for classification and regression. 1998.

[57] Lei Han and Yu Zhang. Learning multi-level task groups in multi-task learning. *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[58] Lei Han and Yu Zhang. Multi-stage multi-task learning with reduced rank. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[59] Jingrui He and Rick Lawrence. A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 25–32, 2011.

[60] Bernd Heisele, Thomas Serre, Massimiliano Pontil, Thomas Vetter, and Tomaso Poggio. Categorization by learning and combining object parts. In *Advances in neural information processing systems*, pages 1239–1245, 2001.

[61] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[62] Ronald L Iman and James M Davenport. Approximations of the critical region of the friedman statistic. *Communications in Statistics-Theory and Methods*, 9(6):571–595, 1980.

[63] Laurent Jacob, Jean-philippe Vert, and Francis R Bach. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pages 745–752, 2009.

[64] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.

[65] Tony Jebara. Multi-task feature and kernel selection for svms. In *Proceedings of the twenty-first international conference on Machine learning*, page 55. ACM, 2004.

[66] Tony Jebara. Multitask sparsity via maximum entropy discrimination. *Journal of Machine Learning Research*, 12:75–110, 2011.

[67] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*, pages 457–464. ACM, 2009.

[68] Matti Kääriäinen. Generalization error bounds using unlabeled data. In *International Conference on Computational Learning Theory*, pages 127–142. Springer, 2005.

[69] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13(1):1865–1890, 2012.

[70] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011.

[71] Marius Kloft and Gilles Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Wein-

berger, editors, *Advances in Neural Information Processing Systems 24*, pages 2438–2446. Curran Associates, Inc., 2011.

[72] Marius Kloft and Gilles Blanchard. On the convergence rate of lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 13(1):2465–2502, 2012.

[73] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

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

[75] Vladimir Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *J. Mach. Learn. Res.*, 11:2457–2485, December 2010.

[76] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

[77] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

[78] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[79] Àgata Lapedriza, David Masip, and Jordi Vitrià. On the use of independent tasks for face recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–6. IEEE, 2008.

[80] Michel Ledoux. Isoperimetry and gaussian analysis. In *Lectures on probability theory and statistics*, pages 165–294. Springer, 1996.

[81] Michel Ledoux. On talagrand's deviation inequalities for product measures. *ESAIM: Probability and statistics*, 1:63–87, 1997.

[82] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

[83] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[84] Yunwen Lei, Lixin Ding, and Yingzhou Bi. Local rademacher complexity bounds based on covering numbers. *arXiv:1510.01463 [cs.AI]*, 2015.

[85] Cong Li, Michael Georgiopoulos, and Georgios C Anagnostopoulos. Conic multi-task classification. In *Machine Learning and Knowledge Discovery in Databases*, pages 193–208. Springer, 2014.

[86] Cong Li, Michael Georgiopoulos, and Georgios C. Anagnostopoulos. Pareto-path multitask multiple kernel learning. *Neural Networks and Learning Systems, IEEE Transactions on*, 26(1):51–61, Jan 2015.

[87] Han Liu, Mark Palatucci, and Jian Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 649–656. ACM, 2009.

[88] Han Liu, Lie Wang, and Tuo Zhao. Multivariate regression with calibration. In *Advances in neural information processing systems*, pages 127–135, 2014.

[89] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 339–348. AUAI Press, 2009.

[90] Qi Liu, Qian Xu, Vincent W Zheng, Hong Xue, Zhiwei Cao, and Qiang Yang. Multi-task learning for cross-platform sirna efficacy prediction: an in-silico study. *BMC bioinformatics*, 11(1):181, 2010.

[91] Yintao Liu, Anqi Wu, Dong Guo, Ke-Thia Yao, and Cauligi S Raghavendra. Weighted task regularization for multitask learning. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 399–406. IEEE, 2013.

[92] Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

[93] Malwina J Luczak and Colin McDiarmid. Concentration for locally acting permutations. *Discrete Mathematics*, 265(1):159–171, 2003.

[94] Gabor Lugosi and Pascal Massart. A sharp concentration inequality with applications. 1999.

[95] F. Lust-Piquard. Khintchine inequalities in cp ($1 < p < \infty$). *COMPTES RENDUS DE L ACADEMIE DES SCIENCES SERIE I-MATHEMATIQUE*, 303(7):289–292, 1986.

[96] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*. Omnipress, June 2009.

[97] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1041–1048. Curran Associates, Inc., 2009.

[98] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rÉnyi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in*

Artificial Intelligence, UAI '09, pages 367–374, Arlington, Virginia, United States, 2009. AUAI Press.

[99] Yishay Mansour and Mariano Schain. Robust domain adaptation. *Annals of Mathematics and Artificial Intelligence*, 71(4):365–380, 2013.

[100] Pascal Massart. About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, pages 863–884, 2000.

[101] Andreas Maurer. Bounds for linear multi-task learning. *The Journal of Machine Learning Research*, 7:117–139, 2006.

[102] Andreas Maurer. The rademacher complexity of linear transformation classes. In *Learning Theory*, pages 65–78. Springer, 2006.

[103] Andreas Maurer. *Algorithmic Learning Theory: 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014. Proceedings*, chapter A Chain Rule for the Expected Suprema of Gaussian Processes, pages 245–259. Springer International Publishing, Cham, 2014.

[104] Andreas Maurer and Massimiliano Pontil. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, volume 30, pages 55–76, 2013.

[105] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *arXiv preprint arXiv:1505.06279*, 2015.

[106] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[107] Shahar Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE transactions on Information Theory*, 48(1):251–263, 2002.

[108] Shahar Mendelson. On the performance of kernel classes. *The Journal of Machine Learning Research*, 4:759–771, 2003.

[109] Charles A Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6(Jul):1099–1125, 2005.

[110] Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer, 2009.

[111] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pages 1097–1104, 2009.

[112] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[113] Guillaume Obozinski, Ben Taskar, and Michael I Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[114] Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115 – 125, 2015.

[115] Dmitriy Panchenko et al. A note on talagrands concentration inequality. *Elect. Comm. in Probab*, 6:55–65, 2001.

[116] Anastasia Pentina and Shai Ben-David. Multi-task and lifelong learning of kernels. In *Algorithmic Learning Theory*, pages 194–208. Springer, 2015.

[117] Anastasia Pentina and Christoph H Lampert. Lifelong learning with non-iid tasks. In *Advances in Neural Information Processing Systems*, pages 1540–1548, 2015.

[118] G Peshkir and Albert Nikolaevich Shiryaev. The khintchine inequalities and martingale expanding sphere of their action. *Russian Mathematical Surveys*, 50(5):849–904, 1995.

[119] Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010.

[120] Jian Pu, Yu-Gang Jiang, Jun Wang, and Xiangyang Xue. Multiple task learning using iteratively reweighted least square.

[121] Yuting Qi, Dehong Liu, David Dunson, and Lawrence Carin. Multi-task compressive sensing with dirichlet process priors. In *Proceedings of the 25th international conference on Machine learning*, pages 768–775. ACM, 2008.

[122] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Transfer learning for image classification with sparse prototype representations. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[123] Rajat Raina, Andrew Y Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 713–720. ACM, 2006.

[124] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[125] Emmanuel Rio. Une inégalité de bennett pour les maxima de processus empiriques. In *Annales de l'IHP Probabilités et statistiques*, volume 38, pages 1053–1057, 2002.

[126] Bernardino Romera-Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. Exploiting unrelated tasks in multi-task learning. In *International Conference on Artificial Intelligence and Statistics*, pages 951–959, 2012.

[127] Wojciech Samek, Alexander Binder, and Motoaki Kawanabe. Multi-task learning via non-sparse multiple kernel learning. In *Computer Analysis of Images and Patterns*, pages 335–342. Springer, 2011.

[128] Li Shen, Gang Sun, Zhouchen Lin, Qingming Huang, and Enhua Wu. Adaptive sharing for image classification. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2183–2190. AAAI Press, 2015.

[129] M Signoretto, R Langone, M Pontil, and J Suykens. Graph based regularization for multi-linear multitask learning. 2014.

[130] Yunyan Song and Wenxin Zhu. Multi-task support vector machine for data classification. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(7):341–350, 2016.

[131] J Michael Steele. An efron-stein inequality for nonsymmetric statistics. *The Annals of Statistics*, pages 753–758, 1986.

[132] Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

[133] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.

[134] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

[135] S Thrun. Learning to learn: Introduction. In *In Learning To Learn*, 1996.

[136] I. Tolstikhin, G. Blanchard, and M. Kloft. Localized complexities for transductive learning. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pages 857–884. JMLR, 2014.

[137] Antonio Torralba, Kevin P Murphy, and William T Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–762. IEEE, 2004.

[138] Sara Van De Geer. A new approach to least-squares estimation, with applications. *The Annals of Statistics*, pages 587–602, 1987.

[139] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

[140] Vladimir Naumovich Vapnik and Samuel Kotz. *Estimation of dependences based on empirical data*, volume 40. Springer-Verlag New York, 1982.

[141] Linli Xu, Aiqing Huang, Jianhui Chen, and Enhong Chen. Exploiting task-feature co-clusters in multi-task learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[142] Qian Xu, Sinno Jialin Pan, Hannah Hong Xue, and Qiang Yang. Multitask learning for protein subcellular location prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(3):748–759, 2011.

[143] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research*, 8:35–63, 2007.

[144] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Oswald Lanz, and Nicu Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *Proceedings of the IEEE international conference on computer vision*, pages 1177–1184, 2013.

[145] Haiqin Yang, Irwin King, and Michael R Lyu. Online learning for multi-task feature selection. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1693–1696. ACM, 2010.

[146] Xiaolin Yang, Seyoung Kim, and Eric P Xing. Heterogeneous multitask learning with joint sparsity constraints. In *Advances in neural information processing systems*, pages 2151–2159, 2009.

[147] Xiao-Tong Yuan, Xiaobai Liu, and Shuicheng Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing*, 21(10):4349–4360, 2012.

[148] Chao Zhang, Lei Zhang, and Jieping Ye. Generalization bounds for domain adaptation. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3320–3328. Curran Associates, Inc., 2012.

[149] Yu Zhang and Dit-Yan Yeung. Multi-task warped gaussian process for personalized age estimation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2622–2629. IEEE, 2010.

[150] Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*, 2012.

[151] Yu Zhang and Dit-Yan Yeung. Multi-task boosting by exploiting task relationships. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 697–710. Springer, 2012.

[152] Yu Zhang and Dit-Yan Yeung. Multilabel relationship learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(2):7, 2013.

[153] Yu Zhang and Dit-Yan Yeung. A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):12, 2014.

[154] Yu Zhang, Dit-Yan Yeung, and Qian Xu. Probabilistic multi-task feature selection. In *Advances in neural information processing systems*, pages 2559–2567, 2010.

[155] Vincent Wenchen Zheng, Sinno Jialin Pan, Qiang Yang, and Jeffrey Junfeng Pan. Transferring multi-device localization models using latent multi-task learning. In *AAAI*, volume 8, pages 1427–1432, 2008.

[156] Wenliang Zhong and James Kwok. Convex multitask learning with flexible task clusters. *arXiv preprint arXiv:1206.4601*, 2012.

[157] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems*, pages 702–710, 2011.

[158] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822. ACM, 2011.

[159] Qiang Zhou and Qi Zhao. Flexible clustered multi-task learning by learning representative tasks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):266–278, 2016.