Journal of Machine Learning Research 19 (2018) 1-47 Submitted 3/17; Revised 4/18; Published 8/18

Local Rademacher Complexity-based Learning Guarantees for Multi-Task Learning

Niloofar Yousefi NILOOFAR.YOUSEFI@UCF.EDU
Department of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816, USA

Yunwen Lei LEIYW@SUSTC.EDU.CN
Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, 518055, China

Marius Kloft KLOFT@CS.UNI-KL.DE
Department of Computer Science
Technische Universität Kaiserslautern
67653 Kaiserslautern, Germany

Mansooreh Mollaghasemi MANSOOREH.MOLLAGHASEMI@UCF.EDU
Department of Industrial Engineering & Management Systems
University of Central Florida
Orlando, FL 32816, USA

Georgios C. Anagnostopoulos GEORGIO@FIT.EDU
Department of Electrical and Computer Engineering
Florida Institute of Technology
Melbourne, FL 32901, USA

Editor: Massimiliano Pontil

Abstract

We show a Talagrand-type concentration inequality for Multi-Task Learning (MTL), with which we establish sharp excess risk bounds for MTL in terms of the Local Rademacher Complexity (LRC). We also give a new bound on the LRC for any norm regularized hypothesis classes, which applies not only to MTL, but also to the standard Single-Task Learning (STL) setting. By combining both results, one can easily derive fast-rate bounds on the excess risk for many prominent MTL methods, including—as we demonstrate—Schatten norm, group norm, and graph regularized MTL. The derived bounds reflect a relationship akin to a conservation law of asymptotic convergence rates. When compared to the rates obtained via a traditional, global Rademacher analysis, this very relationship allows for trading off slower rates with respect to the number of tasks for faster rates with respect to the number of available samples per task.

Keywords: Excess Risk Bounds, Local Rademacher Complexity, Multi-task Learning

1. Introduction

A commonly occurring problem, when applying machine learning in the sciences, 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. For example, in cancer genomics, tumor bioptic samples may be relatively scarce due to the limited number of cancer patients, when compared to samples of healthy individuals. Also, in neuroscience, electroencephalogram experiments are carried out on human subjects to record training data and typically involve only a few dozen subjects.

In such settings, when considering any type of prediction task per individual subject (for example, whether the subject is indeed suffering from a specific medical affliction or not), relying solely on the scarce data per individual most often leads to inadequate predictive performance. Such a direct approach completely ignores the advantages that might be gained, when considering intrinsic, strong similarities between subjects and, hence, tasks. For instance, in the area of genomics, different living organisms can be related to each other in terms of their evolutionary relationships as given by the tree of life. Taking into account such relationships may be instrumental in detecting genes of recently developed organisms, for which only a limited number of training data is available. While our discussion here has focused on the realm of biomedicine, similar limitations and opportunities to overcome them exist in other fields as well.

Transfer learning (Pan and Yang, 2010) and, in particular, Multi-Task Learning (MTL) (Caruana, 1997) leverage such underlying common links among a group of 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. An early example of such a learning paradigm is the neural network-based approach introduced by Caruana (1997), while more recent works consider convex MTL problems (Ando and Zhang, 2005; Evgeniou and Pontil, 2004; Argyriou et al., 2008a). At the core of each such MTL formulation lies a mechanism that encodes task relatedness into the learning problem (Evgeniou et al., 2005). Such relatedness mechanism can always be thought of as jointly constraining the tasks' hypothesis spaces, so that their geometry is mutually coupled, *e.g.*, via a block norm constraint (Yousefi et al., 2015). Thus, from a regularization perspective, the tasks mutually regularize their learning based on their inter-task relatedness. This process of information exchange during co-learning is often referred to as *information sharing*. With respect to learning theory results, the analysis of MTL goes back to the seminal work of Baxter (2000), which was followed up by the works of Ando and Zhang (2005); Maurer (2006a). Nowadays, MTL frameworks are routinely employed in a variety of settings. Some recent applications include computational genetics (Widmer et al., 2013), image segmentation (An et al., 2008), HIV therapy screening (Bickel et al., 2008), collaborative filtering (Cao et al., 2010), age estimation from facial images (Zhang and Yeung, 2010), and sub-cellular location prediction (Xu et al., 2011), just to name a few prominent ones.

MTL learning guarantees are centered around notions of (global) Rademacher complexities, which were introduced to machine learning by Bartlett et al. (2002); Bartlett and Mendelson (2002); Koltchinskii and Panchenko (2000); Koltchinskii (2001); Koltchinskii and Panchenko (2002), and employed in the context of MTL by Maurer (2006a,b); Kakade et al. (2012); Maurer and Pontil (2013); Maurer (2016); Maurer et al. (2016). All these works are briefly surveyed in Sect. 1.3. It is worth noting that, if T denotes the number of tasks being co-learned and n denotes the number of available observations per task, then the fastest-converging error or excess risk bounds derived in these works are of the order $O(1/\sqrt{nT})$.

More recently, Koltchinskii (2006) and Bartlett et al. (2005) introduced a more nuanced variant of these complexities, 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, as shown by Bartlett et al. (2005), 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. This can be attributed to the fact that, unlike LRCs, 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. The end result of this distinction empowers a local analysis to provide less conservative and, hence, sharper bounds than the standard global analysis. 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 (Koltchinskii, 2010), multiple kernel learning (Kloft and Blanchard, 2011; Cortes et al., 2013), transductive learning (Tolstikhin et al., 2014), semi-supervised learning (Oneto et al., 2015) and bounds on the LRCs via covering numbers (Lei et al., 2015).

1.1 Our Contributions

Through a Talagrand-type concentration inequality adapted to the MTL case, this paper's main contribution is the derivation of 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 ensure us that the MTL hypothesis selected by a learning algorithm approaches the best-in-class solution as n increases beyond a certain threshold. We also prove a new bound on the LRC, which generally holds for hypothesis classes with any norm 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 Single-Task Learning (STL) setting. As a matter of fact, we demonstrate such results, in Sect. 4, for classes induced by graph-based, Schatten and group norm regularizers. Moreover, we prove matching lower bounds and, thus, show 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 from the typical GRC-based $O(1/\sqrt{n})$ bounds, 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, since T is typically considered fixed is such setting.

Fixing all other parameters when the number of samples per task n approaches infinity, our local bounds yield faster rates compared to their global counterparts. Also, it is observed that, if the number of tasks T and the radius R of the ball-norms can grow with n, there are cases wherein local analysis always improves over the global one. When our local bounds are compared to the ones in(Maurer and Pontil, 2013; Maurer, 2006b), which stem from a global analysis, one observes that our bounds yield faster, $O(1/T)$ and $O(1/n)$ convergence rates.

1.2 Organization

The paper is organized as follows: Sect. 2 lays the foundations for our analysis by considering a Talagrand-type concentration inequality suitable for deriving our bounds. Next, in Sect. 3, after suitably defining LRCs for MTL hypothesis spaces, we provide our LRC-based MTL excess risk bounds. Based on these bounds, we follow up this section with a local analysis of linear MTL frameworks, in which task-relatedness is presumed and enforced by imposing a norm constraint. More specifically, leveraging off Hölder's inequality, Sect. 4 presents generic upper bounds for the relevant LRC of any norm regularized hypothesis class. These results are subsequently specialized to the case of group norm, Schatten norm and graph regularized linear MTL. Sect. 5 supplies the corresponding excess risk bounds based on the LRC of the aforementioned hypothesis classes. The paper concludes with Sect. 6, which investigates the convergence rate of our LRC-based excess risk bounds for the previously mentioned hypothesis spaces. We also compare our local bounds with those obtained from a GRC-based analysis provided in Maurer and Pontil (2013); Maurer (2006b).

1.3 Previous Related Works

An earlier work by Maurer (2006a), which considers linear MTL frameworks for binary classification, investigates the generalization guarantees based on Rademacher averages. In this framework, all tasks are pre-processed by a common bounded linear operator and operator norm constraints are used to control the complexity of the associated hypothesis spaces. The GRC-based error bounds derived are of order $O(1/\sqrt{nT})$. Another contemporary study (Maurer, 2006b) 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 benefited from renewed attention in recent years. Kakade et al. (2012) 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 GRC-based excess risk bound of order $O(1/\sqrt{nT})$. Moreover, Maurer and Pontil (2013) present a global Rademacher complexity analysis leading to excess risk bounds of order $O(\sqrt{\log(nT)/nT}$ for a trace norm regularized MTL model. Also, Maurer (2016) examines the bounding of (global) Gaussian complexities of function classes that result from considering composite maps, as is typical in several settings, including MTL. An application of the paper's results yields MTL risk bounds of order $O(1/\sqrt{nT})$. More recently, Maurer et al. (2016) presents excess risk bounds of order $O(1/\sqrt{nT})$ for both MTL and Learning-to-Learn (LTL) and reveals conditions, under which MTL is more beneficial over learning tasks independently.

Finally, although loosely related to our focus, we mention in passing a few works 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 by Thrun and Pratt (2012); Ben-David and Schuller (2003); Ben-David and Borbely (2008); Pentina and Lampert (2015) and Pentina and Ben-David (2015). Also, in the context of domain adaptation, similar considerations are examined by Mansour et al. (2009a,b,c); Cortes and Mohri (2011); Zhang et al. (2012); Mansour and Schain (2013) and Cortes and Mohri (2014).

1.4 Basic Assumptions & Notations

Consider T supervised learning tasks sampled from the same input-output space $\mathcal{X} \times \mathcal{Y}$. Each task t is represented by an independent random variable (X_t, Y_t) governed by a probability distribution μ_t. Also, the *i.i.d.* samples related to each task t are described by the sequence $(X_t^i, Y_t^i)_{i=1}^n$, drawn from μ_t.

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 function f we use $\mathbb{E} f(X, Y)$ and $\mathbb{E}_X f(X, Y)$ to denote the expectation with w.r.t. all the involved random variables and the conditional expectation w.r.t. the random variable X respectively. For any vector-valued function $\boldsymbol{f} = (f_1, \ldots, f_T)$, we introduce the following two notations:

$$P\boldsymbol{f} := \frac{1}{T}\sum_{t=1}^{T} P f_t, \qquad P_n \boldsymbol{f} := \frac{1}{T}\sum_{t=1}^{T} P_n f_t.$$

where $P f_t := \mathbb{E}[f_t(X_t)]$ and $P_n f_t := \frac{1}{n}\sum_{i=1}^n f_t(X_t^i)$. When well-defined, we denote the component-wise exponentiation of a vector $\boldsymbol{f}$ as $\boldsymbol{f}^\alpha = (f_1^\alpha, \ldots, f_T^\alpha), \forall \alpha \in \mathbb{R}$. For any loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ 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 $\ell_{f_t}((X_t, Y_t)) = \ell(f_t(X_t), Y_t)$.

Finally, in the subsequent material, we always assume the measurability of functions and suprema whenever necessary. Furthermore, operators on separable Hilbert spaces are assumed to be of trace class.

2. Talagrand-Type Inequality for Multi-Task Learning

Our derivation of LRC-based error bounds for MTL is founded on a Talagrand-type concentration inequality, which was adapted to the context of MTL and is presented next. It shows that the uniform deviation between the true and empirical means for 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 $\mathcal{F}$. A notable property of Theorem 1 is that the correlation among different components of $\boldsymbol{f}$, encoded by either the constraint on variances or the constraint imposed in the hypothesis space, is preserved. This last observation is congruent with the spirit of MTL. The proof of Theorem 1, which is deferred to Appendix A, is based on a so-called *Logarithmic Sobolev inequality* on log-moment generating functions.

Theorem 1 (TALAGRAND-TYPE INEQUALITY FOR MTL) *Let $\mathcal{F} = \{\boldsymbol{f} := (f_1, \ldots, f_T)\}$ be a class of vector-valued functions satisfying $\max_{t \in N_T} \sup_{x \in \mathcal{X}} |f_t(x)| \leq b$. Also, assume that $X := (X_t^i)_{(t,i)=(1,1)}^{(T,N_t)}$ is 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 $1 - e^{-x}$,*

$$\sup_{\boldsymbol{f} \in \mathcal{F}} (P\boldsymbol{f} - P_n\boldsymbol{f}) \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{8xr}{nT}} + \frac{12bx}{nT}, \tag{1}$$

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 1, the data from different tasks are assumed to be mutually independent, which is typically presumed in MTL (Maurer, 2006a). 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 2 *Note that Theorem 1 pertains to classes of uniformly bounded functions and is used in the sequel to bound the excess risk of multi-task learning function classes. However, using a new argument by Mendelson (2014), Theorem 1 can be extended beyond the case of classes of uniformly bounded loss functions. In particular, rather than adopting a concentration-based inequality, which is crucial to our approach here to bound the suprema of the resulting empirical processes, the approach in Mendelson (2014) relies on a "small ball" assumption. Such an assumption holds for functions with "well-behaved high-order moments" (e.g. heavy-tailed functions).*

Remark 3 *At this point, we present the result of the previous theorem for the special case of single task learning. It is very straightforward to verify that, for $T = 1$, the bound in* (1) *can be written as*

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{8xr}{n}} + \frac{12bx}{n}, \tag{2}$$

where the function f is chosen from a scalar-valued function class $\mathcal{F}$. This bound can be compared to the result of Theorem 2.1 of Bartlett et al. (2005), which for $\alpha = 1$ reads as

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{2xr}{n}} + \frac{8bx}{3n}. \tag{3}$$

Note that the difference between the constants in (2) *and* (3) *is due to the fact that we were unable to directly apply Bousquet's version of Talagrand's inequality (like it was done in Bartlett et al. (2005) for scalar-valued functions) to the class of vector-valued functions. To be more clear, let Z be defined as in* (A.2) *with the jackknife replication $Z_{s,j}$. We find a lower bound $Z^{''}_{s,j}$ such that $Z^{''}_{s,j} \leq Z - Z_{s,j}$. Then, in order to apply Theorem 2.5 of Bousquet (2002), 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 1. However, we found that—when dealing with MTL class of functions—it is not very straightforward to define such a constraint that satisfies 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. With that being said, the key ingredient to Theorem 1's proof is the so-called* Logarithmic Sobolev *inequality—Theorem A.1—which can be considered as the exponential version of* Efron-Stein*'s inequality.*

3. MTL Excess Risk Bounds based on Local Rademacher Complexities

At the heart of Theorem 1 lies a variance bound, which motivates us to consider Rademacher averages associated with a function sub-class enjoying small variances. As pointed out in Bartlett et al. (2005), 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.

Definition 4 (MULTI-TASK LOCAL RADEMACHER COMPLEXITY) *For a vector-valued function class $\mathcal{F} = \{\boldsymbol{f} = (f_1, \ldots, f_T)\}$, the* Multi-Task Local Rademacher Complexity (MT-LRC) $\mathfrak{R}(\mathcal{F}, r)$ *is defined as*

$$\mathfrak{R}(\mathcal{F}, r) := \mathbb{E}_{X,\sigma}\left[\sup_{\substack{\boldsymbol{f}=(f_1,\ldots,f_T)\in\mathcal{F}\\ V(\boldsymbol{f})\leq r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_t^i f_t(X_t^i)\right], \tag{4}$$

where $V(\boldsymbol{f})$ is an upper bound on the variance of the functions in $\mathcal{F}$.

For the case $T = 1$, it is clear that the MT-LRC reduces to the standard LRC for scalar-valued function classes. Analogous to single task learning, a challenge in using the MT-LRC to refine existing learning rates is to find an optimal radius trading-off the variance and the associated complexity, which, as we show later, reduces to the calculation of the fixed-point of a sub-root function.

Definition 5 (SUB-ROOT FUNCTION) *A function $\psi : [0, \infty] \to [0, \infty]$ is sub-root if and only if it is non-decreasing and the function $r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for $r > 0$.*

Lemma 6 (Lemma 3.2 Bartlett et al. (2005)) *If ψ is a sub-root function, then it is continuous on $[0, \infty]$, and the equation $\psi(r) = r$ has a unique (non-zero) solution r^*, which is known as the fixed point of ψ. Moreover, for any $r > 0$, it holds that $r > \psi(r)$ if and only if $r^* \leq r$.*

Intuitively, the model sought for by learning algorithms would hopefully attain a small generalization error and enjoy a small variance, when there is a relationship between risks and variances. The concept of local Rademacher complexity allows us to focus on identifying such models.

Definition 7 (VECTOR-VALUED BERNSTEIN CLASS) *Let $0 < \beta \leq 1$ and $B > 0$. A vector-valued function class $\mathcal{F}$ is said to be a (β, B)-Bernstein class with respect to the probability measure P if there exists a function $V : \mathcal{F} \to \mathbb{R}^+$ such that*

$$P\boldsymbol{f}^2 \leq V(\boldsymbol{f}) \leq B(P\boldsymbol{f})^\beta, \quad \forall \boldsymbol{f} \in \mathcal{F}. \tag{5}$$

It can be shown that the Bernstein condition (5) is not too restrictive and it holds, for example, for non-negative bounded functions with respect to any probability distribution as shown in (Bartlett et al., 2004). Other examples include the class of excess risk functions $\mathcal{L}_{\mathcal{F}} := \{\ell_f - \ell_{f^*} : f \in \mathcal{F}\}$—with $f^* \in \mathcal{F}$ being 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, LRC-based excess risk bounds can be established for MTL. We will assume that the loss function ℓ and the vector-valued hypothesis space $\mathcal{F}$ satisfy the following conditions:

Assumption 8

1. *There is a function* $\boldsymbol{f}^* = (f_1^*, \dots, f_T^*) \in \mathcal{F}$ *satisfying* $P\ell_{\boldsymbol{f}^*} = \inf_{\boldsymbol{f}\in\mathcal{F}} P\ell_{\boldsymbol{f}}$.
2. *There is a constant* $B' \geq 1$ *and* $0 < \beta \leq 1$, *such that for every* $\boldsymbol{f} \in \mathcal{F}$ *we have* $P(\boldsymbol{f} - \boldsymbol{f}^*)^2 \leq B'\big(P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*})\big)^{\beta}$.
3. *There is a constant* L, *such that the loss function* ℓ *is* L*-Lipschitz in its first argument.*

As pointed out in Bartlett et al. (2005), many regularized algorithms satisfy these conditions. More specifically, a uniform convexity condition on the loss function ℓ is usually sufficient to satisfy Assumption 8.2. A typical example is the quadratic loss function $\ell(f(X), Y) = (f(X) - Y)^2$. More specifically, if $|f(X) - Y| \leq 1$ for any $f \in \mathcal{F}, x \in \mathcal{X}$ and $Y \in \mathcal{Y}$, then it can be shown that the conditions of Assumption 8 are met with $L = 1$ and $B = 1$.

We now present the main result of this section showing that the excess risk 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 Appendix B.

Theorem 9 (Excess risk bound for MTL) *Let* $\mathcal{F} := \{\boldsymbol{f} := (f_1, \dots, f_T)\}$ *be a class of vector-valued functions* $\boldsymbol{f}$ *satisfying* $\max_{t\in\mathbb{N}_T} \sup_{x\in\mathcal{X}} |f_t(x)| \leq b$. *Assume that* $X := (X_t^i, Y_t^i)_{(t,i)=(1,1)}^{(T,n)}$ *is a vector of* nT *independent random variables, where for each task t, the samples* $(X_t^1, Y_t^1) \dots,$ (X_t^n, Y_t^n) *are identically distributed. Suppose that Assumption 8 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* $B := \max(B'L^2, 1)$ *and* ψ *be a sub-root function with 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* $\mathcal{F}^*$*:*

$$\mathfrak{R}(\mathcal{F}^*, r) := \mathbb{E}_{X,\sigma}\Big[\sup_{\substack{\boldsymbol{f}\in\mathcal{F}, \\ L^2P(\boldsymbol{f}-\boldsymbol{f}^*)^2\leq r}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_t^i f_t(X_t^i)\Big]. \tag{6}$$

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

$$P(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) \leq \frac{K}{K-\beta} P_n(\ell_{\boldsymbol{f}} - \ell_{\boldsymbol{f}^*}) + (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} \max\Big((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}}\Big) + \Big(\frac{2^{\beta+3}B^2K^{\beta}x}{nT}\Big)^{\frac{1}{2-\beta}} + \frac{48LBbx}{(2-\beta)nT}. \tag{7}$$

The following corollary is direct by noting that $P_n(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 0$.

Corollary 10 *Let* $\hat{\boldsymbol{f}}$ *be any element of function 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 9 hold. Then for any* $x > 0$ *and* $r > \psi(r)$, *with probability at least* $1 - e^{-x}$,

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} \max\Big((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}}\Big) + \Big(\frac{2^{\beta+3}B^2K^{\beta}x}{nT}\Big)^{\frac{1}{2-\beta}} + \frac{48LBbx}{(2-\beta)nT}. \tag{8}$$

An immediate consequence of this section's results is that one can derive excess risk bounds for given regularized MTL hypothesis spaces. In the sequel, we will derive excess risk bounds for several commonly used norm regularized MTL hypothesis spaces by further bounding the fixed point r^* appearing in Corollary 10.

4. Local Rademacher Complexity Bounds for Norm Regularized MTL Models

This section presents very general MT-LRC bounds 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.

4.1 Preliminaries

We consider linear MTL models, where we associate to each task a functional $f_t(X) := \langle \boldsymbol{w}_t, \phi(X) \rangle$. Here, $\boldsymbol{w}_t$ belongs to a *Reproducing Kernel Hilbert Space (RKHS)* $\mathcal{H}$, equipped with an inner product $\langle ., . \rangle$ and an induced norm $\|.\| := \sqrt{\langle ., . \rangle}$. Also, $\phi : \mathcal{X} \to \mathcal{H}$ is a feature map associated to $\mathcal{H}$'s reproducing kernel k satisfying $k(X, \tilde{X}) = \langle \phi(X), \phi(\tilde{X}) \rangle, \forall X, \tilde{X} \in \mathcal{X}$. 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 using the regularized cost function:

$$\min_{\boldsymbol{W}} \Omega \left(\boldsymbol{D}^{1/2} \boldsymbol{W} \right) + C \sum_{t=1}^{T} \sum_{i=1}^{n} \ell(\left\langle \boldsymbol{w}_t, \phi(X_t^i) \right\rangle, Y_t^i), \tag{9}$$

where the regularizer $\Omega(\cdot)$ may be used to reflect a priori information. This regularization scheme amounts to performing *Empirical Risk Minimization (ERM)* using the hypothesis space

$$\mathcal{F} := \left\{ X \mapsto \left[\langle \boldsymbol{w}_1, \phi(X_1) \rangle, \ldots, \langle \boldsymbol{w}_T, \phi(X_T) \rangle \right]^T : \Omega(\boldsymbol{D}^{1/2} \boldsymbol{W}) \leq R^2 \right\}, \tag{10}$$

where $\boldsymbol{D}$ is a given positive operator defined on $\mathcal{H}$. Note that the hypothesis spaces corresponding to the group and Schatten norms can be recovered by setting $\boldsymbol{D} = \boldsymbol{I}$ and by using their corresponding norms. More specifically, by choosing $\Omega(\boldsymbol{W}) = \frac{1}{2} \|\boldsymbol{W}\|_{2,q}^2$, one obtains an $L_{2,q}$-group norm hypothesis space in (10). Similarly, the choice $\Omega(\boldsymbol{W}) = \frac{1}{2} \|\boldsymbol{W}\|_{S_q}^2$ gives an L_{S_q}-Schatten norm hypothesis space in (10). Furthermore, the graph regularized MTL (Micchelli and Pontil, 2004; Evgeniou et al., 2005; Maurer, 2006b) can be obtained by taking $\Omega(\boldsymbol{W}) = \frac{1}{2} \|\boldsymbol{D}^{1/2} \boldsymbol{W}\|_F^2$, where $\|.\|_F$ is a Frobenius norm, $\boldsymbol{D} := \boldsymbol{L} + \eta \boldsymbol{I}$, $\boldsymbol{L}$ is the relevant graph Laplacian, and $\eta > 0$ is a regularization constant. On balance, all these MTL models can be considered as norm regularized models. Note that in the sequel, we let q^* be the Hölder conjugate exponent of q, *i.e.* $1/q + 1/q^* = 1$.

4.2 General Bound on the LRC

Now, we can provide the main results on general LRC bounds for any MTL hypothesis space of the form $\Omega(\boldsymbol{W}) = \frac{1}{2} \|\boldsymbol{W}\|^2$ for a norm $\|.\|$. In what follows, the Hilbert-Schmidt operator $\phi(X) \otimes \phi(X) : \mathcal{H} \to \mathcal{H}$ is defined as $\phi(X) \otimes \phi(X)(\boldsymbol{u}) = \langle \phi(X), \boldsymbol{u} \rangle \phi(X)$.

Theorem 11 (LRC bounds for MTL models with norm regularizers) *Let the regularizer* $\Omega(\boldsymbol{W})$ *in* (9) *be given as an appropriate norm* $\|.\|$*, whose dual is denoted by* $\|.\|_*$. *Let the kernels be uniformly bounded, that is,* $\|k\|_\infty \leq \mathcal{K} < \infty$*, 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 eigen-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 in non-increasing order. Then, for any given positive operator $\boldsymbol{D}$ on $\mathbb{R}^T$, any $r > 0$ and any non-negative integers $h_1, \ldots, h_T$:*

$$\mathfrak{R}(\mathcal{F}, r) \leq \min_{0 \leq h_t \leq \infty} \left\{ \sqrt{\frac{r \sum_{t=1}^{T} h_t}{nT}} + \frac{\sqrt{2}R}{T} \mathbb{E}_{X,\sigma} \left\| \boldsymbol{D}^{-1/2} \boldsymbol{V} \right\|_* \right\}, \tag{11}$$

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 Using the LRC's definition, we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}, r) &= \frac{1}{nT} \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{\boldsymbol{f}=(f_1,\ldots,f_T) \in \mathcal{F}, \\ P\boldsymbol{f}^2 \leq r}} \sum_{i=1}^{n} \left\langle (\boldsymbol{w}_t)_{t=1}^{T}, \left(\sigma_t^i \phi(X_t^i) \right)_{t=1}^{T} \right\rangle \right\} \\
&= \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{\boldsymbol{f} \in \mathcal{F}, \\ P\boldsymbol{f}^2 \leq r}} \left\langle (\boldsymbol{w}_t)_{t=1}^{T}, \left(\sum_{j=1}^{\infty} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^{T} \right\rangle \right\} \\
&\leq \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{P\boldsymbol{f}^2 \leq r} \left\langle \left(\sum_{j=1}^{h_t} \sqrt{\lambda_t^j} \left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^{T}, \right.\right. \\
&\qquad \left.\left. \left(\sum_{j=1}^{h_t} \sqrt{\lambda_t^j}^{-1} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^{T} \right\rangle \right\} \qquad (12) \\
&+ \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{\boldsymbol{f} \in \mathcal{F}} \left\langle (\boldsymbol{w}_t)_{t=1}^{T}, \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right)_{t=1}^{T} \right\rangle \right\} \qquad (13) \\
&= A_1 + A_2,
\end{aligned}$$

where A_1 and A_2 stand respectively for the first (12) and second (13) term of the previous bound.

Step 1. Controlling A_1: By applying the Cauchy-Schwartz (C.S.) inequality on A_1, one gets

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

$$\left(\sum_{t=1}^{T}\sum_{j=1}^{h_t}\left(\lambda_t^j\right)^{-1}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle^2\right)^{\frac{1}{2}}\Bigg]\Bigg\}.$$

With the help of Jensen's inequality and taking advantage of 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 $P\boldsymbol{f}^2\leq r$ together imply that $\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\lambda_t^j\left\langle\boldsymbol{w}_t,\boldsymbol{u}_t^j\right\rangle^2\leq r$ (see Lemma C.1 in the Appendix for the proof), we can further bound A_1 as

$$A_1\leq\sqrt{\frac{r\sum_{t=1}^{T}h_t}{nT}}. \tag{14}$$

Step 2. Controlling A_2: We now use Hölder's inequality to bound the second term A_2 as follows:

$$\begin{aligned}A_2&=\frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\langle(\boldsymbol{w}_t)_{t=1}^T,\left(\sum_{j>h_t}\left\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i),\boldsymbol{u}_t^j\right\rangle\boldsymbol{u}_t^j\right)_{t=1}^{T}\right\rangle\right\}\\&=\frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\langle\boldsymbol{D}^{1/2}\boldsymbol{W},\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\rangle\right\}\\&\overset{\text{Hölder}}{\leq}\frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\boldsymbol{f}\in\mathcal{F}}\left\|\boldsymbol{D}^{1/2}\boldsymbol{W}\right\|\cdot\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*\right\}\\&\leq\frac{\sqrt{2}R}{T}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*.\end{aligned} \tag{15}$$

Combining (15) and (14) completes the proof. ■

In what follows, we demonstrate the power of Theorem 11 by applying it to derive the LRC bounds for some popular MTL models, including group norm, Schatten norm and graph regularized models, which have been extensively studied in the MTL literature; for example, see (Maurer, 2006b; Argyriou et al., 2007a,b, 2008a; Li et al., 2015; Argyriou et al., 2014).

4.3 Group Norm Regularized MTL

We first consider the MTL scheme, which captures the inter-task relationships by the group norm regularizer $\frac{1}{2}\|\boldsymbol{W}\|_{2,q}^2:=\frac{1}{2}\left(\sum_{t=1}^{T}\|\boldsymbol{w}_t\|_2^q\right)^{2/q}$ (Argyriou et al., 2007a, 2008a; Lounici et al., 2009; Romera-Paredes et al., 2012). Its associated hypothesis space takes the form

$$\mathcal{F}_q:=\left\{X\mapsto\left[\langle\boldsymbol{w}_1,\phi(X_1)\rangle,\ldots,\langle\boldsymbol{w}_T,\phi(X_T)\rangle\right]^T:\frac{1}{2}\|\boldsymbol{W}\|_{2,q}^2\leq R_{max}^2\right\}. \tag{16}$$

Before presenting the result for this case, we point out that A_1 does not depend on the hypothesis space's $\boldsymbol{W}$-constraint. Therefore, the bound for A_1 is independent of the the choice of reqularizers we consider in this study. However, A_2 can be further bounded in a manner that depends on the regularization function.

We start off with a useful lemma which helps with bounding A_2 for the group norm hypothesis space (16). The proof of this lemma, which is based on the application of the Khintchine (C.1) and Rosenthal (C.2) inequalities, is presented in Appendix C.

Lemma 12 *Assume that the kernels in* (9) *are uniformly bounded, that is,* $\|k\|_\infty \leq \mathcal{K} < \infty$*. Then, for the group norm regularizer* $\frac{1}{2}\|\boldsymbol{W}\|_{2,q}^2$ *in* (16) *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*

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

Corollary 13 *Using Theorem 11, for any* $1 < q \leq 2$*, the LRC of function class* $\mathcal{F}_q$ *in* (16) *can be bounded as*

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{q^*}}, \frac{2eq^{*2}R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}. \quad (17)$$

Proof Sketch: We use Lemma 12 to upper-bound A_2 for the group norm hypothesis space (16) as

$$A_2(\mathcal{F}_q) \leq \sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}. \quad (18)$$

Now, combining (14) and (18) provides the following bound on $\mathfrak{R}(\mathcal{F}_q, r)$

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{r\sum_{t=1}^{T}h_t}{nT}} + \sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}. \quad (19)$$

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

$$(\star)\sqrt{\alpha_1} + \sqrt{\alpha_2} \leq \sqrt{2(\alpha_1 + \alpha_2)} \quad (20)$$

$$(\star\star) \quad l_q - to - l_p: \quad \|\boldsymbol{a}_1\|_q = \langle \boldsymbol{1}, \boldsymbol{a}_1^q\rangle^{\frac{1}{q}} \overset{\text{Hölder's}}{\leq} \left(\|\boldsymbol{1}\|_{(p/q)^*}\|\boldsymbol{a}_1^q\|_{(p/q)}\right)^{\frac{1}{q}} = T^{\frac{1}{q}-\frac{1}{p}}\|\boldsymbol{a}_1\|_p \quad (21)$$

$$(\star\star\star) \quad \|\boldsymbol{a}_1\|_s + \|\boldsymbol{a}_2\|_s \leq 2^{1-\frac{1}{s}}\|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_s \leq 2\|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_s, \quad (22)$$

one obtains the desired result. See Appendix C for the detailed proof.

Remark 14 *Since the LRC bound above is non-monotonic in q, it is more practical to state the above bound in terms of* $\kappa \geq q$*; note that 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 $\kappa \in [q, 2]$,*

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{\kappa^*}}, \frac{2e\kappa^{*2}R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{\kappa^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}\kappa^* T^{\frac{1}{\kappa^*}}}{nT}. \tag{23}$$

Remark 15 (Sparsity-inducing group norm) *The use of the group norm regularizer $\frac{1}{2}\|\boldsymbol{W}\|_{2,1}^2$ encourages a sparse representation that is shared across multiple tasks (Argyriou et al., 2007b, 2008a). 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, the bound gets minimized for $\kappa^* = \log T$. For this particular choice of κ^*, it is easy to show that*

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

Remark 16 ($L_{2,q}$ Group norm regularizer with $q \geq 2$) *For any $q \geq 2$, Theorem 11 provides an LRC bound for the function class $\mathcal{F}_q$ in* (16) *given as*

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT}\left\|\left(\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{q^*}}, \frac{2R_{max}^2}{T}\lambda_t^j\right)\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}, \tag{24}$$

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

Proof Using $\boldsymbol{D} = \boldsymbol{I}$, and $\|.\| = \|.\|_{2,q}$ in (15) gives

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

$$= \frac{\sqrt{2}R_{max}}{T}\left(\sum_{t=1}^{T}\left(\sum_{j>h_t}\frac{\lambda_t^j}{n}\right)^{\frac{q^*}{2}}\right)^{\frac{1}{q^*}} = \sqrt{\frac{2R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^{T}\right\|_{\frac{q^*}{2}}}.$$

By applying (20), (21) and (22), this last result together with the bound for A_1 in (14) yields the result. ■

To investigate the tightness of the bound in (17), we derive the corresponding lower bound, which holds for the LRC of $\mathcal{F}_q$ with $q \geq 1$. The proof of this result can be found in Appendix C.

Theorem 17 (Lower bound) *Consider the hypothesis space shown in* (16)*. The following lower bound holds for the local Rademacher complexity of $\mathcal{F}_q$ for any $q \geq 1$. There is an absolute constant c such that, if $\lambda_t^1 \geq 1/(nR_{max}^2)$ $\forall t$, then, for all $r \geq \frac{1}{n}$ and $q \geq 1$,*

$$\mathfrak{R}(\mathcal{F}_{q,R_{max},T}, r) \geq \sqrt{\frac{c}{nT^{1-\frac{2}{q^*}}}\sum_{j=1}^{\infty}\min\left(rT^{1-\frac{2}{q^*}}, \frac{R_{max}^2}{T}\lambda_1^j\right)}. \tag{25}$$

To make a clear comparison between the lower bound in (25) and the upper bound in (17), we assume identical eigenvalue tail sums $\sum_{j\geq\infty}\lambda_t^j$ for all tasks. In this case, 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 (25), 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 and confirms that the LRC upper bounds that we have obtained are reasonably tight.

Remark 18 *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 $\hat{f}$. As shown in Section 4 of Bartlett et al. (2004), a direct analysis of the empirical minimizer can lead to sharper bounds compared to the LRC-based bounds. Consequently, based on Theorem 8 in Bartlett et al. (2004), there might be cases in which the local Rademacher complexity bounds are constants, while $P\hat{f}$ is a decreasing function of the number of samples n. Moreover, it is shown in the same paper that, under some mild conditions on the loss function ℓ, a similar argument also holds for the class of loss functions $\{\ell_f - \ell_{f^*} : f \in \mathcal{F}\}$.*

4.4 Schatten Norm Regularized MTL

Argyriou et al. (2007b) developed a spectral regularization framework for MTL, wherein the L_{S_q}-Schatten norm $\frac{1}{2}\|\boldsymbol{W}\|_{S_q}^2 := \frac{1}{2}\left[\text{tr}\left(\boldsymbol{W}^T\boldsymbol{W}\right)^{\frac{q}{2}}\right]^{\frac{2}{q}}$ is studied as a concrete example that corresponds to performing ERM in the following hypothesis space:

$$\mathcal{F}_{S_q} := \left\{X \mapsto \left[\langle\boldsymbol{w}_1, \phi(X_1)\rangle, \ldots, \langle\boldsymbol{w}_T, \phi(X_T)\rangle\right]^T : \frac{1}{2}\|\boldsymbol{W}\|_{S_q}^2 \leq R'^2_{max}\right\}. \tag{26}$$

Corollary 19 *For any* $1 < q \leq 2$ *in* (26)*, the LRC of function class* $\mathcal{F}_{S_q}$ *is bounded as*

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

The proof is provided in Appendix C.

Remark 20 (Sparsity-inducing Schatten norm (trace norm)) *Trace norm regularized MTL, corresponding to Schatten norm regularization with* $q = 1$ *(Maurer and Pontil, 2013; Pong et al., 2010), imposes a low-rank structure on the spectrum of* $\boldsymbol{W}$*. It can also be interpreted as low dimensional subspace learning (Argyriou et al., 2008b; Kumar and Daume III, 2012; Kang et al., 2011). 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* $q^* = 2$*, we get*

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

Remark 21 (L_{S_q} **Schatten norm regularizer with** $q \geq 2$**)** *For any* $q \geq 2$*, Theorem 11 provides an LRC bound for the function class* $\mathcal{F}_{S_q}$ *in* (26) *as*

$$\mathfrak{R}(\mathcal{F}_{S_q}, r) \leq \sqrt{\frac{4}{nT}\left\| \left(\sum_{j=1}^{\infty} \min\left(rT^{1-\frac{2}{q^*}}, \frac{2R'^2_{max}}{T}\lambda_t^j\right)\right)_{t=1}^T \right\|_{\frac{q^*}{2}}}, \tag{27}$$

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

Proof We first bound the expectation $\mathbb{E}_{X,\sigma}\left\|\boldsymbol{V}\right\|_{S_{q^*}}$. 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 follows

$$\begin{aligned}
\mathbb{E}_{X,\sigma}\left\|\boldsymbol{V}\right\|_{S_{q^*}} &= \mathbb{E}_{X,\sigma}\left\| \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i \boldsymbol{U}_t^i \right\|_{S_{q^*}} \\
&\overset{\text{Jensen}}{\leq} \left[\operatorname{tr}\left(\left(\sum_{t,s=1}^T \sum_{i,j=1}^n \mathbb{E}_{X,\sigma}\left(\sigma_t^i \sigma_s^j \boldsymbol{U}_t^{i^T} \boldsymbol{U}_s^j\right)\right)^{\frac{q^*}{2}} \right) \right]^{\frac{1}{q^*}} \\
&= \left[\operatorname{tr}\left(\left(\sum_{t=1}^T \sum_{i=1}^n \mathbb{E}_X\left(\boldsymbol{U}_t^{i^T} \boldsymbol{U}_t^i\right)\right)^{\frac{q^*}{2}} \right) \right]^{\frac{1}{q^*}}
\end{aligned}$$

$$= \left[\operatorname{tr}\left(\left(\operatorname{diag}\left(\mathbb{E}_X \sum_{i=1}^{n} \sum_{j>h_1} \langle \frac{1}{n}\phi(X_1^i), \boldsymbol{u}_1^j \rangle^2, \ldots, \mathbb{E}_X \sum_{i=1}^{n} \sum_{j>h_T} \langle \frac{1}{n}\phi(X_T^i), \boldsymbol{u}_T^j \rangle^2 \right) \right)^{\frac{q^*}{2}} \right) \right]^{\frac{1}{q^*}}$$

$$= \left[\operatorname{tr}\left(\left(\frac{1}{n} \operatorname{diag}\left(\sum_{j>h_1} \lambda_1^j, \ldots, \sum_{j>h_T} \lambda_T^j \right) \right)^{\frac{q^*}{2}} \right) \right]^{\frac{1}{q^*}}$$

$$= \sqrt{\frac{1}{n}} \left(\sum_{t=1}^{T} \left(\sum_{j>h_t} \lambda_t^j \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} = \sqrt{\frac{1}{n} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}}.$$

One can derive the final result by replacing this last expression into (11) and by utilizing (20), (21) and (22). ■

4.5 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}. Here we consider the following MTL graph regularizer (Maurer, 2006b)

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

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:

$$\mathcal{F}_G := \Big\{ X \mapsto [\Big\langle \boldsymbol{w}_1, \phi(X_1) \Big\rangle, \ldots, \Big\langle \boldsymbol{w}_T, \phi(X_T) \Big\rangle]^T : \frac{1}{2} \|\boldsymbol{D}^{1/2} \boldsymbol{W}\|_F^2 \leq R''^2_{max} \Big\}. \tag{28}$$

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

Corollary 22 *For any given positive operator $\boldsymbol{D}$ in* (28)*, the LRC of $\mathcal{F}_G$ is bounded by*

$$\mathfrak{R}(\mathcal{F}_G, r) \leq \sqrt{\frac{4}{nT} \Big\| \Big(\sum_{j=1}^{\infty} \min\Big(r, \frac{2\boldsymbol{D}_{tt}^{-1} R''^2_{max}}{T} \lambda_t^j \Big) \Big)_{t=1}^{T} \Big\|_1}. \tag{29}$$

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

See Appendix C for the proof.

Remark 23 *Note that if one considers a strongly convex norm of $\boldsymbol{W}$, an alternative proof strategy can be used to bound the A_2 term in* (13)*. This strategy is based on the duality of strong convexity and strong smoothness (Theorem 3 in Kakade et al. (2012)) along with the application of the Fenchel-Young inequality. This approach results in $A_2 \leq \mathcal{A}_{ub} := \sqrt{\frac{2}{\mu}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*^2}$, where μ is the strong convexity parameter. For the strongly convex cases considered in our study (e.g. $\frac{1}{2}\left\|\boldsymbol{W}\right\|_{2,q}^2$ or $\frac{1}{2}\left\|\boldsymbol{W}\right\|_{S_q}^2$ for any $q \in (1, 2]$), it holds that $\mu \leq 1$ (see Theorem 16 and Corollary 19 in Kakade et al. (2009)). Now, comparing $\sqrt{2}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*$ in* (15) *with $\mathcal{A}_{ub}$, one can easily verify that $\sqrt{2}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_* \overset{Jensen's}{\leq} \sqrt{\frac{2}{\mu}\mathbb{E}_{X,\sigma}\left\|\boldsymbol{D}^{-1/2}\boldsymbol{V}\right\|_*^2}$ for any $\mu \leq 1$. Therefore, for the strongly convex norms we considered here, Hölder's inequality yields slightly tighter bounds for the MT-LRC.*

5. Excess Risk Bounds for Norm Regularized MTL Models

In this section we will provide excess risk bounds for the hypothesis spaces considered earlier. Note that, due to space limitations, the proofs are provided only for the hypothesis space $\mathcal{F}_q$ with $q \in (1, 2]$. However, for the cases involving the $L_{2,q}$-group norm with $q = 1$ or $q \geq 2$, as well as the L_{S_q}-Schatten and graph norms, the proofs can be obtained in a very similar fashion. More specifically, by using the LRC bounds of Remark 16, Corollary 19, Remark 21 and Corollary 22, one can follow the same proof steps shown in this section to arrive at the results pertaining to these cases.

Theorem 24 (Excess risk bound for an $L_{2,q}$ group norm regularized MTL) *Assume that $\mathcal{F}_q$ in* (16) *is a convex class of functions with ranges in $[-b, b]$ and let the loss function ℓ of Problem* (9) *satisfy Assumption 8. Let $\hat{\boldsymbol{f}}$ be any element of $\mathcal{F}_q$ for $1 < q \leq 2$ satisfying $P_n\ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f}\in\mathcal{F}_q} P_n\ell_{\boldsymbol{f}}$. Assume, moreover, that k is a positive definite kernel on $\mathcal{X}$ such that $\|k\|_\infty \leq \mathcal{K} < \infty$. Denote by r^* the fixed point of $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$. Then, for any $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$, the excess loss of function class $\mathcal{F}_q$ is bounded as*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} \max\left((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}}\right) + \left(\frac{2^{\beta+3}B^2K^\beta x}{nT}\right)^{\frac{1}{2-\beta}} + \frac{48LBbx}{(2-\beta)nT},$$

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

$$r^* \leq \min_{0\leq h_t\leq\infty} \frac{B^2\sum_{t=1}^T h_t}{nT} + 4BL\sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^T\right\|_{\frac{q^*}{2}}} + \frac{4\sqrt{2\mathcal{K}e}R_{max}BLq^*T^{\frac{1}{q^*}}}{nT}, \tag{30}$$

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

Proof First, notice that $\mathcal{F}_q$ is convex and, therefore, it is star-shaped around any of its elements. Hence, according to Lemma 3.4 in Bartlett et al. (2005)—which indicates that the local Rademacher complexity of the star-hull of any function class $\mathcal{F}$ is a sub-root function—$\mathfrak{R}(\mathcal{F}_q, r)$ is a sub-root function. Moreover, because of the symmetry of the σ_t^i's distribution 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 in (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 (19) 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

$$\alpha := \frac{B^2 \sum_{t=1}^T h_t}{nT}, \text{ and } \gamma := 2BL \sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{2\sqrt{2\mathcal{K}e}R_{max}BLq^*T^{\frac{1}{q^*}}}{nT}. \tag{31}$$

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

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

$$\sum_{j>h_t} \lambda_t^j \leq d_t \sum_{j>h_t} j^{-\alpha_t} \leq d_t \int_{h_t}^{\infty} x^{-\alpha_t} dx = d_t \left[\frac{1}{1-\alpha_t} x^{1-\alpha_t} \right]_{h_t}^{\infty} = -\frac{d_t}{1-\alpha_t} h_t^{1-\alpha_t}. \tag{32}$$

Note that via the $l_q - to - l_p$ conversion inequality (21), for $p = 1$ and $q = \frac{q*}{2}$, we have

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

which with the help of $\sqrt{\alpha_1} + \sqrt{\alpha_2} \leq \sqrt{2(\alpha_1 + \alpha_2)}$ for any $\alpha_1, \alpha_2 > 0$, and $\|\boldsymbol{a}_1\|_s + \|\boldsymbol{a}_2\|_s \leq 2\|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_s$ for any non-negative vectors $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^T$ and $s = \frac{q^*}{2}$ gives

$$r^* \leq \min_{0 \leq h_t \leq \infty} 2B \sqrt{\left\| \left(\frac{B^2 T^{2-\frac{2}{q^*}} h_t^2}{n^2T^2} - \frac{32 d_t e q^{*2} R_{max}^2 L^2}{nT^2(1-\alpha_t)} h_t^{1-\alpha_t} \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{4\sqrt{2\mathcal{K}e}R_{max}BLq^*T^{\frac{1}{q^*}}}{nT}. \tag{33}$$

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

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

Note that substituting the previous expression for $\alpha := \min_{t \in \mathbb{N}_T} \alpha_t$ and $d = \max_{t \in \mathbb{N}_T} d_t$ into (33), we can upper-bound the fixed point of r^* as

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

which implies that

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

It can be seen that the convergence rate can be as slow as $O\left(\frac{q^* T^{1/q^*}\sqrt{d}}{T\sqrt{n}}\right)$ (for small α, where at least one $\alpha_t \approx 1$), and as fast as $O(\frac{1}{n})$ (when $\alpha_t \to \infty$, for all t). The bound obtained for the fixed point together with Theorem 24 provides a bound for the excess risk, which leads to the following remark. Note that in the sequel, we assume that the data distribution of each task is concentrated and uniform on the same M-dimensional unit sphere. This implies that (by symmetry) the eigenvalues must all be equal and they sum up to 1. Thus, for each task t, $\lambda_t^j = \frac{1}{M}$. On the other hand, we assumed earlier that $\lambda_t^j \leq d_t j^{-\alpha}$ for all $1 \leq j \leq M$. Therefore, choosing $j = M$, we are forced to set $d = M^{\alpha-1}$.

Remark 25 (Excess risk bounds for selected norm regularized MTL problems) *Assume that $\mathcal{F}$ is a class of functions with ranges in $[-b, b]$. Let the loss function ℓ of Problem* (9) *satisfy Assumption 8. Additionally, assume that k is a positive definite kernel on $\mathcal{X}$, such that $\|k\|_\infty \leq \mathcal{K} < \infty$. Also, denote $\alpha := \min_{t\in\mathbb{N}_T} \alpha_t$ and $d := max_{t\in\mathbb{N}_T} d_t$. Then, for any $\boldsymbol{f} \in \mathcal{F}$, $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$,*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} \max\left((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}} \right) + \left(\frac{2^{\beta+3} B^2 K^\beta x}{nT} \right)^{\frac{1}{2-\beta}} + \frac{48LBbx}{(2-\beta)nT}, \tag{34}$$

where, for $\mathcal{F} \in \left\{\mathcal{F}_q, \mathcal{F}_{S_q}, \mathcal{F}_G\right\}$, $\hat{\boldsymbol{f}}$ is such that $P_n \ell_{\hat{\boldsymbol{f}}} = \inf_{\boldsymbol{f}\in\mathcal{F}} P_n \ell_{\boldsymbol{f}}$ and r^* is the fixed point of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}, \frac{r}{4L^2})$. Furthermore, r^* can be bounded for each of the three hypothesis spaces as follows:

- *Group norm: For any $1 < q \leq 2$,*

  $$r^* \leq \min_{\kappa\in[q,2]} 14\sqrt{\frac{\alpha+1}{\alpha-1}} \left(\kappa^{*2} R_{max}^2 L^2 \right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} \left(T^{\frac{2}{\kappa}} \right)^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} + \frac{10\sqrt{\mathcal{K}} R_{max} BL\kappa^* T^{\frac{1}{\kappa^*}}}{nT}. \tag{35}$$

  Also, for any $q \geq 2$, we have

  $$r^* \leq 8\sqrt{\frac{\alpha+1}{\alpha-1}} \left(R_{max}^2 L^2 \right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} \left(T^{\frac{2}{q}} \right)^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}. \tag{36}$$

- *Schatten norm: For any $1 < q \leq 2$,*

  $$r^* \leq 8\sqrt{\frac{\alpha+1}{\alpha-1}} \left(q^* R'^2_{max} L^2 \right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}. \tag{37}$$

  Note that for the trace norm, we would have $q^* = 2$ in the previous bound (see Remark 20). Additionally, for any $q \geq 2$, it holds

  $$r^* \leq 8\sqrt{\frac{\alpha+1}{\alpha-1}} \left(R'^2_{max} L^2 \right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} \left(T^{\frac{2}{q}} \right)^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}. \tag{38}$$

- *Graph regularizer: For any positive operator* $\boldsymbol{D}$*,*

$$r^* \leq 8\sqrt{\frac{\alpha+1}{\alpha-1}}\left(R''^2_{max}L^2\boldsymbol{D}^{-1}_{max}\right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}. \tag{39}$$

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

6. Discussion

In this section, we investigate the convergence rate of our LRC-based excess risk bounds, which were established in the previous section. We also discuss related works and provide a new excess risk bound by employing a rather different approach, which exhibits the benefit of a MTL regularizer at the expense of a slower convergence rate in terms of the number of examples per task n. Note that, for the purpose of this section, we will assume that $\beta = 1$, which hold for many loss function classes, see Bartlett et al. (2004) for a discussion.

6.1 Convergence Rates

In order to facilitate a more concrete comparison of convergence rates, we will assume the same spherical M-dimensional data distribution for each task t; this assumption leads to $\lambda^j_t = \frac{1}{M}$, or equivalently $d = M^{\alpha-1}$. Furthermore, we will concentrate only on the parameters R, n, T, q^*, M and α and we will assume that all the other parameters are fixed and, hence, hidden in the big-O notation. Thus, for our LRC-based bounds we have

$$\begin{aligned}
&\text{Group norm:} && \text{(a) } \forall\kappa \in [q,2], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R^2_{max}\kappa^{*2})^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} \left(T^{\frac{2}{\kappa}}\right)^{-\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right).\\
& && \text{(b) } \forall q \in [2,\infty], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R^2_{max})^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} \left(T^{\frac{2}{q}}\right)^{-\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right).\\
&\text{Schatten norm:} && \text{(c) } \forall q \in (1,2], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R'^2_{max})^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right).\\
& && \text{(d) } \forall q \in [2,\infty], && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R'^2_{max})^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} \left(T^{\frac{2}{q}}\right)^{-\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right).\\
&\text{Graph:} && \text{(e)} && P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R''^2_{max})^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right).
\end{aligned} \tag{40}$$

A close appraisal of the results in (40) 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, where the rates (exponents) of n and T sum up to -1. Note that the trade-off is determined by the value of α, which can facilitate faster n-rates and, simultaneously, 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 characteristics of the group norm. Now, consider the following two cases:

- M is large (high-dimensional data distribution): Note that in the case of very large M, $\alpha > 1$; Also, large M implies small α, that is, $\alpha \to 1$. In this case we get dimension-independent bounds, which should be considered as an advantage for the case of high-dimensional data distribution.

$$\text{Group norm:} \quad \text{(a) } \forall\kappa \in [q,2], \quad P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R^2_{max}\kappa^{*2})^{\frac{1}{2}} \left(T^{\frac{2}{\kappa}}\right)^{-\frac{1}{2}} n^{-\frac{1}{2}}\right).$$

$$\begin{array}{lll} & \text{(b) } \forall q \in [2,\infty], & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R^2_{max})^{\frac{1}{2}} \left(T^{\frac{2}{q}}\right)^{-\frac{1}{2}} n^{-\frac{1}{2}}\right). \\ \text{Schatten-norm:} & \text{(c) } \forall q \in (1,2], & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R'^2_{max} q^*)^{\frac{1}{2}} T^{\frac{-1}{2}} n^{-\frac{1}{2}}\right). \\ & \text{(d) } \forall q \in [2,\infty], & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R'^2_{max})^{\frac{1}{2}} \left(T^{\frac{2}{q}}\right)^{-\frac{1}{2}} n^{-\frac{1}{2}}\right). \\ \text{Graph:} & \text{(e)} & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left((R''^2_{max})^{\frac{1}{2}} T^{\frac{-1}{2}} n^{-\frac{1}{2}}\right). \end{array}$$

- M is small (low-dimensional data distribution): This case happens when the decay rate α is fast ($\alpha \to \infty$), which gives the following rates

$$\begin{array}{lll} \text{Group norm:} & \text{(a) } \forall \kappa \in [q,2], & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left(Mn^{-1}\right). \\ & \text{(b) } \forall q \in [2,\infty], & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left(Mn^{-1}\right). \\ \text{Schatten-norm:} & \text{(c) } \forall q \in (1,2], & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left(Mn^{-1}\right). \\ & \text{(d) } \forall q \in [2,\infty], & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left(Mn^{-1}\right). \\ \text{Graph:} & \text{(e)} & P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) = O\left(Mn^{-1}\right). \end{array}$$

Note that, most likely, a more realistic case lies somewhere in between these two extreme cases, which can be interpreted as follows: when the data is relatively low-dimensional (small M and fast decay of eigenvalues), we will have bounds with fast rates in n. However, MTL may offer little advantage in this case due to the corresponding slow rates in T. This analysis confirms the general belief that MTL proffers a potential advantage if there are many tasks with little data per task and are sampled from high-dimensional data distributions.

6.2 Comparisons to Related Works

It is interesting to compare our local bound for the trace norm regularized MTL with the GRC-based excess risk bound provided in Maurer and Pontil (2013), 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 $\boldsymbol{W}$ than the one we mentioned earlier; in our notation, this space reads as

$$\mathcal{F}'_{S_1} := \left\{ \boldsymbol{W} : \frac{1}{2} \|\boldsymbol{W}\|^2_{S_1} \leq T R'^2_{max} \right\}. \tag{41}$$

The form of this space is based on the premise that, assuming a common vector $\boldsymbol{w}$ for all tasks, the regularizer should not be a function of the number of tasks (Maurer and Pontil, 2013). Given the task-averaged covariance operator $C := 1/T \sum_{t=1}^{T} J_t = 1/T \sum_{t=1}^{T} \mathbb{E}\left(\phi(X_t) \otimes \phi(X_t)\right)$, the excess risk bound in Maurer and Pontil (2013) reads as

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

Under the aforementioned M-dimensional data distributions and by using the hypothesis space of (41), our local bound for the trace norm for any $\alpha > 1$ is given as

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 6400K\sqrt{\frac{\alpha+1}{\alpha-1}} \left(R'^2_{max} L^2\right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb+16BK)Bx}{nT}. \tag{42}$$

Now, let λ_t^{max} be the maximum eigenvalue of the trace operator J_t. Also, let $\lambda_{max} := \max_{t\in\mathbb{N}_T}\{\lambda_t^{max}\}$. It is easy to verify that $\mathbf{tr}(J_t) \leq M\lambda_t^{max}$ and $\|C\|_\infty \leq \lambda_{max} = 1/M$, which renders the GRC-based bound in Maurer and Pontil (2013) into the form

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 2\sqrt{2}LR'_{max}\left(\sqrt{\frac{\lambda_{max}}{n}} + 5\sqrt{\frac{\ln(nT)+1}{nT}}\right) + \sqrt{\frac{bLx}{nT}}. \tag{43}$$

One observes that, in both cases, the bound vanishes as $n \to \infty$. However, it does so at a rate of $n^{-\alpha/1+\alpha}$ for our local bound in (42) and at a slower rate of $\sqrt{\ln n/n}$ for the one in (43). Also, we remark that, as $T \to \infty$, both bounds converge to a non-zero limit: our local bound in (42) at a fast rate of $1/T$ and at a the slower rate of $\sqrt{\ln T/T}$ for the bound in (43). More specifically, making the benevolent choices $B = 1$ and $R'_{max}L = 1$ and ignoring the factor of $6400K\sqrt{\frac{\alpha+1}{\alpha-1}}$, the limit of our local bound in (42) as $T \to \infty$ becomes $g(\alpha) := M^{\frac{\alpha-1}{1+\alpha}}n^{\frac{-\alpha}{1+\alpha}}$. One can very easily verify that $g(\alpha)$ is increasing in α (*i.e.* $g'(\alpha) > 0$), if and only if $\ln(Mn^{-\frac{1}{2}}) > 0$, or, equivalently, $M > \sqrt{n}$. In this case the optimal choice of $\alpha \in (1,\infty)$ (*i.e.* $\alpha \approx 1$) makes our local bound of the order $O(\frac{1}{\sqrt{n}})$. In other words, when the data distribution is sufficiently high-dimensional relative to n, the LRC bound fails in competing with the $O(\frac{1}{\sqrt{Mn}})$ GRC bound in Maurer and Pontil (2013). On the other hand, for lower dimensional distributions or sufficiently large n, we obtain a rate of $1/n$ for the LRC bound at the expense of explicit dependence on the dimension. In particular, the local bound remains larger than the GRC bound in (43) until $n = M^3$ and improves only for larger sample sizes per task.

Another interesting comparison can be performed between our bounds and the one introduced in Maurer (2006b) for a graph regularized MTL. Similar to Maurer (2006b), we consider the following hypothesis space

$$\mathcal{F}'_G = \left\{\boldsymbol{W} : \frac{1}{2}\left\|\boldsymbol{D}^{1/2}\boldsymbol{W}\right\|_F^2 \leq TR''^2_{max}\right\}. \tag{44}$$

Maurer (2006b) provides a bound on the empirical GRC of the aforementioned hypothesis space that can be easily converted to a distribution dependent GRC bound of the form

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

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 same M-dimensional distributional assumptions, it can be shown that

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

where $\lambda_{max} = \frac{1}{M}$ as argued earlier. Furthermore, let $\{\delta_2,\ldots,\delta_T\}$ be the nonzero eigenvalues of$\boldsymbol{L}$ with $\delta_{min} := \min\{\delta_2,\ldots,\delta_T\}$. Then, the GRC-based excess risk bound is obtained as

$$\text{Maurer (2006b)}: \qquad P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq \frac{2LR''_{max}}{\sqrt{n}}\sqrt{2M\lambda_{max}\left(\frac{1}{\delta_{min}} + \frac{1}{T\eta}\right)} + \sqrt{\frac{bLx}{nT}}$$

$$\tag{45}$$

Also, based on Remark 25, the LRC-based bound is given as

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq 6400K\sqrt{\frac{\alpha+1}{\alpha-1}}\left(R''^2_{max}L^2\boldsymbol{D}^{-1}_{max}\right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(48Lb+16BK)Bx}{nT}. \tag{46}$$

The above results show that, when $n \rightarrow \infty$, both GRC and LRC bounds approach zero, albeit at different rates: the global bound at a rate of $\sqrt{1/n}$ and the local one at a faster rate of $n^{-\alpha/\alpha+1}$, since $\alpha > 1$. Additionally, both bounds approach non-zero limits as $T \rightarrow \infty$. Nevertheless, the global bound does so at a rate of $\sqrt{1/T}$ and the local one at a faster rate of $1/T$. Furthermore, similar to the previous case, it can be shown that at the limit $T \rightarrow \infty$, for high-dimensional data distribution (large M, small $\alpha \approx 1$), both local and global bounds yield the same convergence rate of $O(\frac{1}{\sqrt{n}})$. However, for low number of dimensions relative to n (in specific, for $M < n^{\frac{1}{3}}$), our bound improves over the GRC bound.

6.3 A Different Technique for The Trace Norm Regularized Space $\mathcal{F}'_{S_1}$

In what follows, we show that, by applying a rather different proof technique (departing from Theorem 11), we can obtain an excess risk bound for the MTL space $\mathcal{F}'_{S_1}$ in (41), which aims at slower rates in n and T, but exhibits the benefits of a multi-task regularizer. Recall that $\mathcal{F}'_{S_1}$ is given as

$$\mathcal{F}'_{S_1} := \left\{ X \mapsto [\langle \boldsymbol{w}_1, \phi(X_1)\rangle, \ldots, \langle \boldsymbol{w}_T, \phi(X_T)\rangle]^T : \frac{1}{2}\|\boldsymbol{W}\|^2_{S_1} \leq TR'^2_{max} \right\}. \tag{47}$$

Also recall that, from Theorem 11, it can be shown that the LRC of $\mathcal{F}'_{S_1}$ can be bounded as

$$\mathfrak{R}(\mathcal{F}'_{S_1}, r) \leq \min_{0 \leq h_t \leq \infty} \left\{ \sqrt{\frac{r\sum_{t=1}^T h_t}{nT}} + \sqrt{\frac{2R'^2_{max}}{n^2T}} \mathbb{E}_{X,\sigma} \left\| \boldsymbol{V}' \right\|_{S_\infty} \right\}, \tag{48}$$

where

$$\boldsymbol{V}' := \left(\sum_{j > h_t} \left\langle \sum_{i=1}^n \sigma^i_t \phi(X^i_t), \boldsymbol{u}^j_t \right\rangle \boldsymbol{u}^j_t \right)^T_{t=1}. \tag{49}$$

Now, following an approach similar to the one applied in Maurer and Pontil (2013), we will bound $\mathbb{E}_{X,\sigma} \|\boldsymbol{V}'\|_{S_\infty}$ to yield the next theorem. Note that $\|.\|_{S_\infty}$ stands for the operator norm on the separable Hilbert space $\mathcal{H}$.

Theorem 26 *Assume that the conditions of Theorem 24 hold for the hypothesis space $\mathcal{F}'_{S_1}$ in* (47). *Also, denote by r^* the fixed point of $2BL\mathfrak{R}(\mathcal{F}'_{S_1}, \frac{r}{4L^2})$. Then, for any $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$, the excess loss of function class $\mathcal{F}'_{S_1}$ is bounded as*

$$P(\ell_{\hat{\boldsymbol{f}}} - \ell_{\boldsymbol{f}^*}) \leq (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} \max\left((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}} \right) + \left(\frac{2^{\beta+3}B^2K^\beta x}{nT} \right)^{\frac{1}{2-\beta}} + \frac{48LBbx}{(2-\beta)nT},$$

for

$$r^* \leq \min_{0 \leq h_t \leq \infty} \left\{ \frac{B^2 \sum_{t=1}^{T} h_t}{nT} + 4BL\sqrt{\frac{2R'^2_{max}\lambda_h}{n}} + 24BL\sqrt{\frac{2R'^2_{max}\mathcal{K}\left(\ln(nT)+1\right)}{nT}} \right\}, \tag{50}$$

where $\lambda_h := \max_{t\in\mathbb{N}_T}\{\lambda_t^{h_t}\}$ *and where* $h_1, \ldots, h_T$ *are arbitrary non-negative integers.*

The proof of the results is provided in Appendix D.

By considering the same M-dimensional data distribution, the bound in (50) becomes

$$r^* \leq 6BLR'_{max}\left(\sqrt{\frac{1}{Mn}} + 6\sqrt{\frac{\mathcal{K}\left(\ln(nT)+1\right)}{nT}}\right). \tag{51}$$

It can be seen that, when the number of tasks T approaches ∞, the above bound simplifies to

$$r^* \leq \frac{6BLR'_{max}}{\sqrt{Mn}}.$$

In the sequel, we compare the two bounds (37) and (51) for the trace norm regularized MTL models in terms of their convergence rates.

Remark 27 *Using two different techniques, we proved the two following bounds on the fixed point* r^* *of the local Rademacher complexity* $2BL\mathfrak{R}(\mathcal{F}'_{S_1}, \frac{r}{4L^2})$*:*

- *Our approach*

$$r^* \leq 12\sqrt{\frac{\alpha+1}{\alpha-1}}\left(R'^2_{max}L^2\right)^{\frac{1}{1+\alpha}} M^{\frac{\alpha-1}{1+\alpha}} B^{\frac{2\alpha}{\alpha+1}} n^{\frac{-\alpha}{1+\alpha}}. \tag{52}$$

- *MP approach (Maurer and Pontil, 2013)*

$$r^* \leq 6BLR'_{max}\left(\sqrt{\frac{1}{Mn}} + 6\sqrt{\frac{\mathcal{K}\left(\ln(nT)+1\right)}{nT}}\right). \tag{53}$$

As a reminder, the proof of Theorem 11 refers to two terms: A_1*, which embodies a variance constraint, and* A_2*, which constitutes a MTL regularization constraint. The aforementioned bounds were derived by using two different approaches to bound the* A_2 *term, namely the LRC-based approach for* (52) *and the MP technique for* (53)*. In the case of* (52)*, due to the LRC-based approach, the variance constraint (*A_1 *term) plays a dominant role in the overall bound and, thus, yields faster rates in* n *for any* $\alpha > 1$*. However, it offers no improvements in the limit* $T \to \infty$*, since this bound does not decrease with increasing* T*. In contrast, using the MP technique, the MTL regularization constraint (*A_2 *term) is dominant in* (53)*. While this prevents obtaining faster rates in terms of the number of samples, it potentially offers the advantages of MTL for large* T *and high-dimensional data distributions.*

Acknowledgments

NY acknowledges financial support from National Science Foundation (NSF) grant No. 1161228, and No. 1200566. MK acknowledges support from the German Research Foundation (DFG) award KL 2698/2-1 and from the Federal Ministry of Science and Education (BMBF) award 031B0187B. YL acknowledges support from the Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. ZDSYS201703031748284). Finally, GCA acknowledges partial support from the US National Science Foundation (NSF) under Grant No. 1560345. 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.

Appendices

Appendix A. Proof of Theorem 1

This section presents the proof of Theorem 1. We first provide some useful foundations used in the derivation of our result in Theorem 1.

Theorem A.1 (Theorem 2 in Boucheron et al. (2003)) *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 $Z := g(X_1, \ldots, X_n)$. Let $X'_1, \ldots, X'_n$ denote an independent copy of $X_1, \ldots, X_n$, and $Z'_i := g(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)$, which is obtained by replacing the variable X_i with X'_i. Define the random variable $V^+ := \sum_{i=1}^n \mathbb{E}'\big[\big(Z - Z'_i\big)_+^2\big]$, where $(u)_+ := \max\{u, 0\}$, and $\mathbb{E}'[\cdot] := \mathbb{E}[\cdot|X]$ denotes the expectation only w.r.t. the variables $X'_1, \ldots, X'_n$. Let $\theta > 0$ and $\lambda \in (0, 1/\theta)$. Then,*

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

Definition A.2 (Section 3.3 in Boucheron et al. (2013)) *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 $i \in \mathbb{N}_n$,*

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

and

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

Theorem A.3 (Theorem 6.12 in Boucheron et al. (2013)) *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*

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

Lemma A.4 (Lemma 2.11 in Bousquet (2002)) *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 $x \geq 0$,

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

Lemma A.5 (Contraction property in Bartlett et al. (2005)) *Let ϕ be a Lipschitz function with Lipschitz constant $L \geq 0$, that is, $|\phi(a) - \phi(b)| \leq L|a - b|$, $\forall a, b \in \mathbb{R}$. Let $X_1, \ldots, X_n$ be n independent random variables. Then, for every real-valued function class $\mathcal{F}$, it holds*

$$\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i \phi(f(X_i)) \leq L \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(X_i). \tag{A.1}$$

Note that, in Theorem 17 of Maurer (2006a), it has been shown that the result of this lemma also holds for classes of vector-valued functions.

Proof of Theorem 1

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

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

we first apply Theorem A.1 to control the log-moment generating function $\log\mathbb{E}\big(e^{\lambda(Z-\mathbb{E}Z)}\big)$. From Theorem A.1, 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 A.1 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 approach to tackle W is to show that it is a self-bounding function and then apply Theorem A.3 to control $\log\mathbb{E}[e^{\frac{\lambda W}{b'}}]$. The Υ term 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 A.4 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. For clarify, 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 $X' := (X'^i_t)_{(t,i)=(1,1)}^{(T,N_t)}$ be an independent copy of $X := (X_t^i)_{(t,i)=(1,1)}^{(T,N_t)}$. Define the quantity $Z'_{s,j}$ by replacing the variable X_s^j in Z with X'^j_s. Then,

$$\begin{aligned} Z'_{s,j} := \sup_{\boldsymbol{f}\in\mathcal{F}} \Big[&\frac{1}{TN_s}\big[\mathbb{E}'f_s(X'^j_s) - f_s(X'^j_s)\big] - \frac{1}{TN_s}\big[\mathbb{E}f_s(X_s^j) - f_s(X_s^j)\big] \\ &+ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}[\mathbb{E}f_t(X_t^i) - f_t(X_t^i)]\Big]. \end{aligned} \tag{A.3}$$

Let $\hat{\boldsymbol{f}} := (\hat{f}_1, \dots \hat{f}_T)$ be such that $Z = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\big[\mathbb{E}\hat{f}_t(X_t^i) - \hat{f}_t(X_t^i)\big]$ and introduce

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

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

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

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

and, therefore,

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

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

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

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

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

$$
\log\mathbb{E}\big[e^{\lambda(Z-\mathbb{E}Z)}\big]\le\frac{\lambda b'}{1-\lambda b'}\log\mathbb{E}\big[e^{\frac{\lambda}{b'}(W+\Upsilon)}\big],\quad\forall\lambda\in(0,1/b'). \tag{A.4}
$$

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 $\{X_t^i, t\in\mathbb{N}_T, i\in\mathbb{N}_{N_t}\}\backslash\{X_s^j\}$. Letting $\tilde{\boldsymbol{f}}:=(\tilde{f}_1,\ldots,\tilde{f}_T)$ be the function achieving the supremum in the definition of W, one can verify that (note that $b'=\frac{2b}{nT}$)

$$
T^2[W-W_{s,j}]\le\frac{1}{N_s^2}[\mathbb{E}\tilde{f}_s(X_s^j)-\tilde{f}_s(X_s^j)]^2\le\frac{4b^2}{n^2}=T^2b'^2. \tag{A.5}
$$

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

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

Also, it can be shown that

$$
\begin{aligned}
\sum_{s=1}^{T}\sum_{i=1}^{N_s}[W-W_{s,j}]&\le\frac{1}{T^2}\sum_{s=1}^{T}\frac{1}{N_s^2}\sum_{i=1}^{N_s}[\mathbb{E}\tilde{f}_s(X_s^j)-\tilde{f}_s(X_s^j)]^2\\
&=\sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{N_t^2}\sum_{i=1}^{N_t}[\mathbb{E}f_t(X_t^i)-f_t(X_t^i)]^2\Big]=W.
\end{aligned} \tag{A.6}
$$

Therefore, according to Definition A.2, W/b' is a b'-self bounding function. Applying Theorem A.3 then gives the following inequality for any $\lambda\in(0,1/b')$:

$$
\log\mathbb{E}e^{\lambda(W/b')}\le\frac{(e^{\lambda b'}-1)}{b'^2}\mathbb{E}W=\frac{(e^{\lambda b'}-1)}{b'^2}\Sigma^2\le\frac{\lambda\Sigma^2}{b'(1-\lambda b')}, \tag{A.7}
$$

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

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

where the first inequality follows from the definition of W and Υ and the second inequality follows from the standard symmetrization technique used to relate the Rademacher complexity to the uniform deviation of empirical averages from their expectation; see Bartlett et al. (2005). The third inequality comes from a direct application of Lemma A.5 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 $\frac{1}{N_t^2} \leq \frac{1}{nN_t}$. Substituting the previous inequality on Σ^2 back into (A.7) gives

$$\log \mathbb{E}e^{\lambda(W/b')} \leq \frac{\lambda}{b'(1-\lambda b')}\Big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT}+\Upsilon\Big], \quad \forall\lambda\in(0,1/b'). \tag{A.8}$$

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

$$\begin{aligned}
\Upsilon :&= \sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\frac{1}{T^2}\sum_{s=1}^T\frac{1}{N_s^2}\sum_{j=1}^{N_s}\mathbb{E}[\mathbb{E}f_s(X_s^j)-f_s(X_s^j)]^2\Big] \\
&\leq \frac{1}{nT^2}\sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\sum_{s=1}^T\mathbb{E}[\mathbb{E}f_s(X_s^1)-f_s(X_s^1)]^2\Big] \\
&\leq \frac{1}{nT^2}\sup_{\boldsymbol{f}\in\mathcal{F}}\Big[\sum_{s=1}^T\mathbb{E}[f_s(X_s^1)]^2\Big] \\
&\leq \frac{r}{nT},
\end{aligned}\tag{A.9}$$

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 the bound on log-moment generating function of Z into tail probabilities. Substituting the bound on $\log\mathbb{E}e^{\lambda W/b'}$ in (A.8) and the bound on Υ in (A.9) back into (A.4) immediately yields the following inequality on the log-moment generating function of Z for any

$\lambda \in (0, 1/2b')$

$$\begin{aligned}\log \mathbb{E}[e^{\lambda(Z-\mathbb{E}Z)}] &\leq \frac{\lambda b'}{1-\lambda b'}\Big[\frac{\lambda}{b'(1-\lambda b')}\big[16(nT)^{-1}b\mathfrak{R}(\mathcal{F}) + \Upsilon\big] + \frac{\lambda\Upsilon}{b'}\Big] \\ &\leq \frac{\lambda b'}{1-\lambda b'}\frac{\lambda}{b'(1-\lambda b')}\Big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT} + 2\Upsilon\Big] \\ &\leq \frac{2\lambda^2}{2(1-2\lambda b')}\Big[\frac{16b\mathfrak{R}(\mathcal{F})}{nT} + \frac{2r}{nT}\Big],\end{aligned} \tag{A.10}$$

where the last inequality uses $(1-\lambda b')^2 \geq 1 - 2\lambda b' > 0$ since $\lambda \in (0, 1/2b')$. That is, the conditions of Lemma A.4 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)

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

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 in Sect. 3

Theorem B.3 is at the core of proving Theorem 9 in Sect. 3. We first present some useful lemmata.

Lemma B.1 *Let $c_1, c_2 > 0$ and $s > q > 0$. Then the equation $x^s - c_1x^q - c_2 = 0$ has a unique positive solution x_0 satisfying*

$$x_0 \leq \Big[c_1^{\frac{s}{s-q}} + \frac{sc_2}{s-q}\Big]^{\frac{1}{s}}.$$

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 Cucker and Zhou (2007). Let x_0 be this unique positive solution. Then, it follows from Young's inequality

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

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 $x_0^s \leq c_1^{\frac{s}{s-q}} + \frac{sc_2}{s-q}$. The inequality $p(x) \geq 0$ for any $x \geq x_0$ then follows immediately from the facts that $p(x_0) = 0$, $\lim_{x \to \infty} p(x) = \infty$ and the uniqueness of roots for the equation $p(x) = 0$. ■

Also, we will need the following lemma for the second step of the proof of Theorem B.3.

Lemma B.2 *Let $K > 1, r > 0, 0 < \beta \leq 1$ and $B \geq 1$. Assume that $\mathcal{F} = \{\boldsymbol{f} := (f_1, \ldots, f_T)\}$ is a vector-valued (β, B)-Bernstein class of functions. Define the re-scaled version of $\mathcal{F}$ as*

$$\mathcal{F}_r := \left\{ \boldsymbol{f}' = (f'_1, \ldots, f'_T) : f'_t := \frac{r f_t}{\max(r, V(\boldsymbol{f}))}, \boldsymbol{f} = (f_t, \ldots, f_T) \in \mathcal{F} \right\}. \tag{B.2}$$

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

$$\forall \boldsymbol{f} \in \mathcal{F} \qquad P\boldsymbol{f} \leq \frac{K}{K - \beta} P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}}}{K}. \tag{B.3}$$

Proof We prove (B.3) 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 $V_r^+ \leq \frac{r^{\frac{1}{\beta}}}{BK}$ leads to

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

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

$$\begin{aligned}
P\boldsymbol{f} &\leq P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}-1} V(\boldsymbol{f})}{BK} \leq P_n \boldsymbol{f} + \frac{r^{\frac{1}{\beta}-1} (P\boldsymbol{f})^{\beta}}{K} \\
&\overset{(B.1)}{\leq} P_n \boldsymbol{f} + \frac{1}{K} \frac{[(P\boldsymbol{f})^{\beta}]^{\frac{1}{\beta}}}{\frac{1}{\beta}} + \frac{1}{K} \frac{(r^{\frac{1}{\beta}-1})^{\frac{1}{1-\beta}}}{\frac{1}{1-\beta}} \\
&= P_n \boldsymbol{f} + \frac{\beta}{K} P\boldsymbol{f} + \frac{(1-\beta) r^{\frac{1}{\beta}}}{K},
\end{aligned}$$

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

$$P\boldsymbol{f} \leq \frac{K}{K-\beta}P_n\boldsymbol{f} + \frac{1-\beta}{K-\beta}r^{\frac{1}{\beta}} \leq \frac{K}{K-\beta}P_n\boldsymbol{f} + \frac{r^{\frac{1}{\beta}}}{K}. \tag{B.5}$$

Eq. (B.3) follows by combining (B.4) and (B.5). ■

Theorem B.3 (LRC-based bounds 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* $\max_{t\in\mathbb{N}_T} \sup_{x\in\mathcal{X}} |f_t(x)| \leq b$. *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 with* $0 < \beta \leq 1$ *and* $B \geq 1$. *Let* ψ *be a sub-root function with fixed point* r^*. *If* $B\mathfrak{R}(\mathcal{F}, r) \leq \psi(r), \forall r \geq r^*$, *then, for any* $K > 1$, *and* $x > 0$, *with probability at least* $1 - e^{-x}$, *every* $\boldsymbol{f} \in \mathcal{F}$ *satisfies*

$$P\boldsymbol{f} \leq \frac{K}{K-\beta}P_n\boldsymbol{f} + (2K)^{\frac{\beta}{2-\beta}} 20^{\frac{2}{2-\beta}} \max\Big((r^*)^{\frac{1}{2-\beta}}, (r^*)^{\frac{1}{\beta}}\Big) + \Big(\frac{2^{\beta+3}B^2K^{\beta}x}{nT}\Big)^{\frac{1}{2-\beta}} + \frac{24Bbx}{(2-\beta)nT}. \tag{B.6}$$

Proof Let $r \geq r^*$ be a fixed real number. Here, we use the vector-valued function class $\mathcal{F}_r$ as defined in (B.2). The proof is broken down into two major steps. The first step applies Theorem 1 and the "peeling" technique (Van De Geer, 1987; Van Der Vaart and Wellner, 1996) to establish an inequality on the uniform deviation over the function class $\mathcal{F}_r$. The second step then uses the Bernstein assumption $V(\boldsymbol{f}) \leq B(P\boldsymbol{f})^{\beta}$ 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 1 to $\mathcal{F}_r$, we need to control the variances and uniform bounds for elements in $\mathcal{F}_r$. We first show that $P\boldsymbol{f}'^2 \leq r, \forall \boldsymbol{f}' \in \mathcal{F}_r$. 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 $f'_t = rf_t/V(\boldsymbol{f})$ and we get

$$P\boldsymbol{f}'^2 = \frac{1}{T}\sum_{t=1}^{T} Pf_t'^2 = \frac{r^2}{\big[V(\boldsymbol{f})\big]^2}\Big(\frac{1}{T}\sum_{t=1}^{T} Pf_t^2\Big) \leq \frac{r^2}{\big[V(\boldsymbol{f})\big]^2}V(\boldsymbol{f}) \leq r.$$

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 $0 \leq r/\max(r, V(\boldsymbol{f})) \leq 1$, it holds that $\max_{t\in\mathbb{N}_T}\sup_{x\in\mathcal{X}}|f'_t(x)| \leq b$ for any $\boldsymbol{f}' \in \mathcal{F}_r$. Applying Theorem 1 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 $\mathcal{F}(u,v) := \big\{\boldsymbol{f} \in \mathcal{F} : u \leq V(\boldsymbol{f}) \leq v\big\}, \forall 0 \leq u \leq v$, and introduce

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

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

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

we obtain

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

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

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

Choosing $\lambda = 4$ 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, 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 $\mathcal{F}$. letting $A := 20\sqrt{r^*}/B + \sqrt{8x/nT}$ and $C := 12bx/nT$, the upper bound of (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 B.2, 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 $\boldsymbol{f} \in \mathcal{F}$,

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

To apply Lemma B.2, 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 B.1 then implies

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

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

$$Pf\le\frac{K}{K-\beta}P_nf+(2K)^{\frac{\beta}{2-\beta}}20^{\frac{2}{2-\beta}}(r^*)^{\frac{1}{2-\beta}}+\Big(\frac{2^{\beta+3}B^2K^\beta x}{nT}\Big)^{\frac{1}{2-\beta}}+\frac{24Bbx}{(2-\beta)nT}. \tag{B.11}$$

If $r^*>r_0$, Lemma B.1 implies that $A\sqrt{r^*}+C\le(r^*)^{\frac{1}{\beta}}/(BK)$. We now take $r=r^*$ in (B.9) to get $V_{r^*}^+\le A\sqrt{r^*}+C\le(r^*)^{\frac{1}{\beta}}/(BK)$, from which—via Lemma B.2—we obtain that

$$Pf\le\frac{K}{K-\beta}P_nf+\frac{r_*^{\frac{1}{\beta}}}{K}. \tag{B.12}$$

Note that inequality (B.6) follows immediately by combining (B.11) and (B.12). ■

Proof of Theorem 9

Note that the proof of this theorem relies on the results of Theorem B.3. Introduce the following class of excess loss functions

$$\mathcal{H}_{\mathcal{F}}^*:=\{h_f=(h_{f_1},\dots,h_{f_T}),h_{f_t}:(X_t,Y_t)\mapsto\ell(f_t(X_t),Y_t)-\ell(f_t^*(X_t),Y_t),f\in\mathcal{F}\}. \tag{B.13}$$

It can be shown that

$$\max_{t\in\mathbb{N}_T}\sup_{x\in\mathcal{X}}|h_{f_t}(x,y)|=\max_{t\in\mathbb{N}_T}\sup_{x\in\mathcal{X}}|\ell(f_t(x),y)-\ell(f_t^*(x),y)|\le L\max_{t\in\mathbb{N}_T}\sup_{x\in\mathcal{X}}|f_t(x)-f_t^*(x)|\le 2Lb.$$

Also, Assumption 8 implies that

$$P(\ell_f-\ell_{f^*})^2\le L^2P(f-f^*)^2\le B'L^2\big(P(\ell_f-\ell_{f^*})\big)^\beta,\quad\forall h_f\in\mathcal{H}_{\mathcal{F}}^*,$$

By letting $B:=\max(B'L^2,1)$, we have for all $h_f\in\mathcal{H}_{\mathcal{F}}^*$,

$$Ph_f^2\le V(h_f):=L^2P(f-f^*)^2\le B\big(P(\ell_f-\ell_{f^*})\big)^\beta=B(Ph_f)^\beta.$$

which implies that $\mathcal{H}_{\mathcal{F}}^*$ is a (β, B)-Bernstein class of vector-valued functions. In addition, for any $r \geq r^*$, one can verify that

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

where the second to last inequality is due to Lemma A.5. Applying Theorem B.3 to the function class $\mathcal{H}_{\mathcal{F}}^*$ completes the proof.

Appendix C. Proofs of the results in Sect. 4: "Local Rademacher Complexity Bounds for Norm Regularized MTL Models"

Lemma C.1 *Assume that the conditions of Theorem 11 hold. Then, for ever $\boldsymbol{f} \in \mathcal{F}$,*

(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 We first prove part (**a**). Given the eigen-decomposition $\mathbb{E}(\phi(X_t)\otimes\phi(X_t)) = \sum_{j=1}^{\infty}\lambda_t^j\boldsymbol{u}_t\otimes\boldsymbol{u}_t^j$ for each task $t \in \mathbb{N}_T$, we obtain

$$\begin{aligned}
P\boldsymbol{f}^2 =& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\langle \boldsymbol{w}_t, \phi(X_t)\rangle\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\langle \boldsymbol{w}_t\otimes\boldsymbol{w}_t, \phi(X_t)\otimes\phi(X_t)\rangle\right) \\
=& \frac{1}{T}\sum_{t=1}^{T}\langle \boldsymbol{w}_t\otimes\boldsymbol{w}_t, \mathbb{E}_X\left(\phi(X_t)\otimes\phi(X_t)\right)\rangle = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\lambda_t^j\left\langle \boldsymbol{w}_t\otimes\boldsymbol{w}_t, \boldsymbol{u}_t^j\otimes\boldsymbol{u}_t^j\right\rangle \\
=& \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\lambda_t^j\left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j\right\rangle\left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j\right\rangle = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}\lambda_t^j\left\langle \boldsymbol{w}_t, \boldsymbol{u}_t^j\right\rangle^2 \leq r.
\end{aligned}$$

Now, we turn to part (**b**). From the independence among the elements of the sequence $\{\sigma_t^i\}_{\substack{i\in\mathbb{N}_n\\ t\in\mathbb{N}_T}}$, it follows that

$$\begin{aligned}
&\mathbb{E}_{X,\sigma}\left\langle \frac{1}{n}\sum_{i=1}^{n}\sigma_t^i\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle^2 = \frac{1}{n^2}\mathbb{E}_{X,\sigma}\sum_{i,k=1}^{n}\sigma_t^i\sigma_t^k\left\langle \phi(X_t^i), \boldsymbol{u}_t^j\right\rangle\left\langle \phi(X_t^k), \boldsymbol{u}_t^j\right\rangle \\
&\overset{\boldsymbol{\sigma}_t \text{i.i.d.}}{=} \frac{1}{n^2}\mathbb{E}_X\left(\sum_{i=1}^{n}\left\langle \phi(X_t^i), \boldsymbol{u}_t^j\right\rangle^2\right) = \frac{1}{n}\left\langle \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_X\left(\phi(X_t^i)\otimes\phi(X_t^i)\right), \boldsymbol{u}_t^j\otimes\boldsymbol{u}_t^j\right\rangle \\
&= \frac{1}{n}\sum_{l=1}^{\infty}\lambda_t^l\left\langle \boldsymbol{u}_t^l\otimes\boldsymbol{u}_t^l, \boldsymbol{u}_t^j\otimes\boldsymbol{u}_t^j\right\rangle = \frac{\lambda_t^j}{n}.
\end{aligned}$$

■

The next lemmata are used in the proof of the LRC bound for the $L_{2,q}$-group norm regularized MTL in Corollary 13.

Lemma C.2 (Khintchine-Kahane Inequality in (Peshkir and Shiryaev, 1995)) *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*

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^{n} \sigma_i v_i \right\|_{\mathcal{H}}^{p} \leq \left(c \sum_{i=1}^{n} \|v_i\|_{\mathcal{H}}^2 \right)^{\frac{p}{2}}. \tag{C.1}$$

where $c := \max\{1, p-1\}$. The inequality also holds for p in place of c.

Lemma C.3 (Rosenthal-Young Inequality; Lemma 3 of (Kloft and Blanchard, 2012)) *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*

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^{q} \leq c_q \left[\left(\frac{B}{n} \right)^{q} + \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} X_i \right)^{q} \right]. \tag{C.2}$$

Proof of Lemma 12

For the group norm regularizer $\|\boldsymbol{W}\|_{2,q}$, we can further bound the expectation term in (15) for $\boldsymbol{D} = \boldsymbol{I}$ as follows

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

$$\overset{\text{Jensen}}{\leq} \sqrt{\frac{q^*}{n}} \left(\sum_{t=1}^{T} \mathbb{E}_X \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^{n} \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}}. \tag{C.3}$$

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 C.3), which yields

$$\begin{aligned}
\mathbb{E} &\overset{\text{R+Y}}{\leq} \sqrt{\frac{q^*}{n}} \left(\sum_{t=1}^{T} (eq^*)^{\frac{q^*}{2}} \left(\left(\frac{\mathcal{K}}{n} \right)^{\frac{q^*}{2}} + \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_X \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right) \right)^{\frac{1}{q^*}} \\
&= \sqrt{\frac{q^*}{n}} \left(\sum_{t=1}^{T} (eq^*)^{\frac{q^*}{2}} \left(\left(\frac{\mathcal{K}}{n} \right)^{\frac{q^*}{2}} + \left(\sum_{j>h_t} \lambda_t^j \right)^{\frac{q^*}{2}} \right) \right)^{\frac{1}{q^*}}.
\end{aligned} \tag{C.4}$$

The last quantity can be further bounded using the sub-additivity of $\sqrt[q^*]{\cdot}$ as shown next

$$\begin{aligned}
\mathbb{E} &\leq q^* \sqrt{\frac{e}{n}} \left[\left(T \left(\frac{\mathcal{K}}{n} \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} + \left(\sum_{t=1}^{T} \left(\sum_{j>h_t} \lambda_t^j \right)^{\frac{q^*}{2}} \right)^{\frac{1}{q^*}} \right] \\
&= q^* \sqrt{\frac{e}{n}} \left[T^{\frac{1}{q^*}} \sqrt{\frac{\mathcal{K}}{n}} + \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}^{\frac{1}{2}} \right] \\
&= \frac{\sqrt{\mathcal{K}}eq^*T^{\frac{1}{q^*}}}{n} + \sqrt{\frac{eq^{*2}}{n} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}}.
\end{aligned} \tag{C.5}$$

Proof of Corollary 13

Combining (14), (15) and Lemma 12 provides the next bound on $\mathfrak{R}(\mathcal{F}_q, r)$

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_q, r) &\leq \sqrt{\frac{r \sum_{t=1}^{T} h_t}{nT}} + \sqrt{\frac{2eq^{*2}R_{max}^2}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT} \qquad \text{(C.6)} \\
&\overset{(\star)}{\leq} \sqrt{\frac{2}{nT} \left(r \sum_{t=1}^{T} h_t + \frac{2eq^{*2}R_{max}^2}{T} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_{\frac{q^*}{2}} \right)} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}
\end{aligned}$$

$$\overset{(\star\star)}{\leq} \sqrt{\frac{2}{nT}\left(rT^{1-\frac{2}{q^*}}\left\|(h_t)_{t=1}^T\right\|_{\frac{q^*}{2}} + \frac{2eq^{*2}R_{max}^2}{T}\left\|\left(\sum_{j>h_t}\lambda_t^j\right)_{t=1}^T\right\|_{\frac{q^*}{2}}\right)} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}$$

$$\overset{(\star\star\star)}{\leq} \sqrt{\frac{4}{nT}\left\|\left(rT^{1-\frac{2}{q^*}}h_t + \frac{2eq^{*2}R_{max}^2}{T}\sum_{j>h_t}\lambda_t^j\right)_{t=1}^T\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT},$$

where in steps $(\star)$, $(\star\star)$ and $(\star\star\star)$ we applied the corresponding inequalities shown next, which hold for all non-negative numbers α_1 and α_2, any non-negative vectors $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^T$, any p, q such that $0 \leq q \leq p \leq \infty$ and any $s \geq 1$.

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

$(\star\star)\quad l_p - to - l_q : \quad \|\boldsymbol{a}_1\|_q = \langle \mathbf{1}, \boldsymbol{a}_1^q \rangle^{\frac{1}{q}} \overset{\text{Hölder}}{\leq} \left(\|\mathbf{1}\|_{(p/q)^*}\|\boldsymbol{a}_1^q\|_{(p/q)}\right)^{\frac{1}{q}} = T^{\frac{1}{q}-\frac{1}{p}}\|\boldsymbol{a}_1\|_p$

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

Since inequality $(\star\star\star)$ holds for any non-negative h_t, it follows that

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT}\left\|\left(\min_{h_t\geq 0} rT^{1-\frac{2}{q^*}}h_t + \frac{2eq^{*2}R_{max}^2}{T}\sum_{j>h_t}\lambda_t^j\right)_{t=1}^T\right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\mathcal{K}e}R_{max}q^*T^{\frac{1}{q^*}}}{nT}$$

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

Proof of Theorem 17

By considering the hypothesis space in (16) and the MT-LRC's definition, we have

$$\mathfrak{R}(\mathcal{F}_{q,R_{max},T}, r) = \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\substack{P\boldsymbol{f}^2\leq r,\\ \|\boldsymbol{W}\|_{2,q}^2\leq 2R_{max}^2}}\sum_{t=1}^T\left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^n\sigma_t^i\phi(X_t^i)\right\rangle\right\}$$

$$= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\substack{1/T\sum_{t=1}^T\mathbb{E}\langle\boldsymbol{w}_t,\phi(X_t)\rangle^2\leq r,\\ \|\boldsymbol{W}\|_{2,q}^2\leq 2R_{max}^2}}\sum_{t=1}^T\left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^n\sigma_t^i\phi(X_t^i)\right\rangle\right\}$$

$$\geq \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{\sup_{\substack{\forall t\ \mathbb{E}_X\langle\boldsymbol{w}_t,\phi(X_t)\rangle^2\leq r,\\ \|\boldsymbol{W}\|_{2,q}^2\leq 2R_{max}^2,\\ \|\boldsymbol{w}_1\|_2=\ldots=\|\boldsymbol{w}_t\|_2}}\sum_{t=1}^T\left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^n\sigma_t^i\phi(X_t^i)\right\rangle\right\}$$

$$
\begin{aligned}
&= \frac{1}{T}\mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{\forall t\ \mathbb{E}_X\langle \boldsymbol{w}_t,\phi(X_t)\rangle^2\le r,\\ \forall t\ \|\boldsymbol{w}_t\|_2^2\le 2R_{max}^2T^{-\frac{2}{q}}}} \sum_{t=1}^{T} \left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^{n} \sigma_t^i\phi(X_t^i)\right\rangle \right\}\\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{\forall t\ \mathbb{E}_X\langle \boldsymbol{w}_t,\phi(X_t)\rangle^2\le r,\\ \forall t\ \|\boldsymbol{w}_t\|_2^2\le 2R_{max}^2T^{-\frac{2}{q}}}} \left\langle \boldsymbol{w}_t, \frac{1}{n}\sum_{i=1}^{n} \sigma_t^i\phi(X_t^i)\right\rangle \right\}\\
&= \mathbb{E}_{X,\sigma}\left\{ \sup_{\substack{\mathbb{E}_X\langle \boldsymbol{w}_1,\phi(X_1)\rangle^2\le r,\\ \|\boldsymbol{w}_1\|_2^2\le 2R_{max}^2T^{-\frac{2}{q}}}} \left\langle \boldsymbol{w}_1, \frac{1}{n}\sum_{i=1}^{n} \sigma_1^i\phi(X_1^i)\right\rangle \right\}\\
&= \mathfrak{R}(\mathcal{F}_{1,R_{max}T^{-\frac{1}{q}},1}, r).
\end{aligned}
$$

According to Mendelson (2003), it can be shown that there is a constant c such that if $\lambda_t^1 \geq \frac{1}{nR_{max}^2}$, then, for all $r \geq \frac{1}{n}$, it holds that $\mathfrak{R}(\mathcal{F}_{1,R_{max}T^{-\frac{1}{q}},1}, r) \geq \sqrt{\frac{c}{n}\sum_{j=1}^{\infty}\min\left(r, R_{max}^2T^{-\frac{2}{q}}\lambda_1^j\right)}$, which provides the desired result after some algebraic manipulations. The following lemma is used in the proof of the LRC bounds for the L_{S_q}-Schatten norm regularized MTL in Corollary 19.

Lemma C.4 (Khintchine's inequality for arbitrary matrices in Tomczak-Jaegermann (1974)) *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 $p \geq 2$,*

$$
\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{Q}_i\right\|_{S_p}^{p}\right] \leq p^{p/2}\left(\sum_{i=1}^{n}\|\boldsymbol{Q}_i\|_{S_p}^2\right)^{p/2}. \tag{C.7}
$$

Proof of Corollary 19

In order to find an LRC bound for an L_{S_q}-Schatten norm regularized hypothesis space of (26), one just needs to bound the expectation term in (11). Define $\boldsymbol{U}_t^i$ as the matrix with T columns, whose only non-zero t^{th} column equals $\sum_{j>h_t}\left\langle \frac{1}{n}\phi(X_t^i), \boldsymbol{u}_t^j\right\rangle \boldsymbol{u}_t^j$. Recall that, for the Schatten norm regularized hypothesis space of (26), it holds that $\boldsymbol{D} = \boldsymbol{I}$. Therefore, we will have that

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

$$\begin{aligned}
&\overset{\text{(C.7)}}{\leq} \mathbb{E}_X \left\{ (q^*)^{q^*/2} \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \left\| \boldsymbol{U}_t^i \right\|_{S_{q^*}}^2 \right)^{q^*/2} \right\}^{\frac{1}{q^*}} \\
&= \sqrt{q^*}\,\mathbb{E}_X \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \left\| \sum_{j>h_t} \left\langle \frac{1}{n}\phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j \right\|^2 \right)^{1/2} \\
&= \sqrt{q^*}\,\mathbb{E}_X \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{j>h_t} \frac{1}{n^2} \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{1/2} \\
&\overset{\text{Jensen}}{\leq} \frac{\sqrt{q^*}}{n} \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{j>h_t} \lambda_t^j \right)^{\frac{1}{2}} = \sqrt{\frac{q^*}{n} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_1 }\,. \qquad \text{(C.8)}
\end{aligned}$$

Proof of Corollary 22

Similar to the proof of Corollary 19, for the graph regularized hypothesis space depicted in (28), one can bound the expectation term in (11) in this manner

$$\begin{aligned}
&\mathbb{E}_{X,\sigma} \left\| \boldsymbol{D}^{-1/2}\boldsymbol{V} \right\|_* = \mathbb{E}_{X,\sigma} \left[\mathrm{tr} \left(\boldsymbol{V}^T \boldsymbol{D}^{-1} \boldsymbol{V} \right) \right]^{\frac{1}{2}} \\
&\overset{\text{Jensen}}{\leq} \mathbb{E}_X \left(\frac{1}{n^2} \sum_{t,s=1}^{T,T} \sum_{i,l=1}^{n,n} \sum_{j>h_t} \sum_{k>h_s} \boldsymbol{D}_{st}^{-1} \mathbb{E}_\sigma \left(\sigma_t^i \sigma_s^l \right) \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \left\langle \phi(X_s^l), \boldsymbol{u}_s^k \right\rangle \left\langle \boldsymbol{u}_t^j, \boldsymbol{u}_s^k \right\rangle \right)^{\frac{1}{2}} \\
&= \mathbb{E}_X \left(\frac{1}{n} \sum_{t=1}^{T} \boldsymbol{D}_{tt}^{-1} \sum_{j>h_t} \frac{1}{n} \sum_{i=1}^{n} \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{1}{2}} \\
&\overset{\text{Jensen}}{\leq} \left(\frac{1}{n} \sum_{t=1}^{T} \boldsymbol{D}_{tt}^{-1} \sum_{j>h_t} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_X \left\langle \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle^2 \right)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{n}} \left(\sum_{t=1}^{T} \sum_{j>h_t} \boldsymbol{D}_{tt}^{-1} \lambda_t^j \right)^{\frac{1}{2}} = \sqrt{ \frac{1}{n} \left\| \left(\boldsymbol{D}_{tt}^{-1} \sum_{j>h_t} \lambda_t^j \right)_{t=1}^{T} \right\|_1 }\,. \qquad \text{(C.9)}
\end{aligned}$$

The remainder of the derivation is similar to that of Corollary 13 and is omitted for brevity.

Appendix D. Proof of the results in Sect. 6: "Discussion"

In what follows, we provide some general results that imply Theorem 26. More specifically, we restate two concentration results for sums of non-negative operators with finite-dimensional ranges. Towards this end, we will say that two operators A and B are related as $A \preceq B$, if $B - A$ is a positive semi-definite operator.

Theorem D.1 (Theorem A.3 in Maurer and Pontil (2016)) *Consider the separable Hilbert space $\mathcal{H}$. Let $\mathcal{M} \subseteq \mathcal{H}$ be a subspace of finite dimension d. Also, consider the finite sequence A_k of random, independent, self-adjoint operators on $\mathcal{H}$. Assume that, for all $m \in \mathbb{N}$, $k \in \mathbb{N}_N$ and some $R \geq 0$, it holds that $A_k \succeq 0$, Ran$(A_k) \subseteq M$ almost surely and*

$$\mathbb{E}A_k^m \preceq m! R^{m-1} \mathbb{E}A_k.$$

Then,

$$\sqrt{\mathbb{E}\| \sum_k A_k \|_{S_\infty}} \leq \sqrt{\|\mathbb{E} \sum_k A_k \|_{S_\infty}} + \sqrt{R\left(\ln \dim(M) + 1\right)}. \tag{D.1}$$

Lemma D.2 (Lemma A.4 in Maurer and Pontil (2016)) *Let $a_1, \ldots, a_n \in \mathbb{R}^d$. Let*

$$\alpha := \sum_{i=1}^n \|a_i\|^2,$$

and define a rank-one operator Q_x on H, such that $Q_x v = \langle v, x \rangle x$. Also, let $D := \sum_{i=1}^n \sigma_i a_i$. Then, for any $p \geq 1$, it holds that

$$\mathbb{E}[(Q_D)^p] \preceq (2p-1)!! \alpha^{p-1} \mathbb{E}[Q_D],$$

where $(2p-1)!! := \prod_{i=1}^p (2i-1) = (2p-1)(2(p-1)-1) \times \ldots \times 5 \times 3 \times 1$.

Theorem D.3 (Theorem 7 in Maurer and Pontil (2013)) *Consider the independent random operators $A_1, \ldots, A_N$, which satisfy $0 \preceq A_k \preceq I$, $\forall k$. Also, assume that, for some $d \in \mathbb{N}$, it holds that*

$$\text{dimSpan}(\text{Ran}(A_1), \ldots, \text{Ran}(A_N)) \leq d,$$

almost surely. Then

$$\sqrt{\mathbb{E}\| \sum_{k=1}^N A_k \|_{S_\infty}} \leq \sqrt{\|\mathbb{E} \sum_{k=1}^N A_k \|_{S_\infty}} + \sqrt{6\left(\ln\left(4d^2\right) + 1\right)}. \tag{D.2}$$

Proof of Theorem 26

We first proceed to bound $\mathbb{E}_\sigma \|\boldsymbol{V}'\|_{S_\infty}$. Let $\boldsymbol{D}_t$ be the random vector $\sum_{j>h_t} \left\langle \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \boldsymbol{u}_t^j \right\rangle \boldsymbol{u}_t^j$, and recall that the rank-one operator $Q_{\boldsymbol{D}_t}$ is such that $Q_{\boldsymbol{D}_t} v := \langle v, \boldsymbol{D}_t \rangle \boldsymbol{D}_t$. Then, it is clear that $\boldsymbol{V}'^* \boldsymbol{V}' = \sum_{t=1}^T Q_{\boldsymbol{D}_t}$ and, by using Jensen's inequality, we have

$$\mathbb{E}_\sigma \|\boldsymbol{V}'\|_{S_\infty} \leq \sqrt{\mathbb{E}_\sigma \| \sum_{t=1}^T Q_{\boldsymbol{D}_t} \|_\infty}.$$

Note that $\boldsymbol{D}_t$ is the projection of $\sum_{i=1}^n \sigma_t^i \phi(X_t^i)$ onto the space spanned by $\left(\boldsymbol{u}_t^j\right)_{j>h_t}$. Since $\sum_{i=1}^n \sigma_t^i \phi(X_t^i)$ belongs to the space spanned by $\left(\phi(X_t^i)\right)_{i=1}^n$, we know that $\boldsymbol{D}_t$ belongs to a subspace of dimension at most n. It then follows that Ran$(Q_{\boldsymbol{D}_1}), \ldots,$ Ran$(Q_{\boldsymbol{D}_T})$ lie in a subspace of dimension at most nT. Thus, Lemma D.2 for $\alpha_t := \sum_{i=1}^n \|\phi(X_t^i)\|^2$ yields

$$\mathbb{E}_\sigma[(Q_{\boldsymbol{D}_t})^m] \preceq (2m-1)!! \alpha_t^{m-1} \mathbb{E}_\sigma[Q_{\boldsymbol{D}_t}] \preceq m! \left(2 \max_t \alpha_t\right)^{m-1} \mathbb{E}_\sigma[Q_{\boldsymbol{D}_t}],$$

Therefore, applying Theorem D.1 with $R = 2\max_t \alpha_t$ and dimension less than nT gives

$$\sqrt{\mathbb{E}_\sigma \| \sum_{t=1}^{T} Q_{\boldsymbol{D}_t} \|_{S_\infty}} \leq \sqrt{\| \mathbb{E}_\sigma \sum_{t=1}^{T} Q_{\boldsymbol{D}_t} \|_{S_\infty}} + \sqrt{2 \max_t \alpha_t \left(\ln(nT) + 1\right)}.$$

Since $\alpha_t = \sum_{i=1}^{n} \|\phi(X_t^i)\|^2 \leq n\mathcal{K}$, we get

$$\sqrt{\mathbb{E}_\sigma \| \sum_{t=1}^{T} Q_{\boldsymbol{D}_t} \|_{S_\infty}} \leq \sqrt{\| \mathbb{E}_\sigma \sum_{t=1}^{T} Q_{\boldsymbol{D}_t} \|_{S_\infty}} + \sqrt{2n\mathcal{K} \left(\ln(nT) + 1\right)}.$$

Now, we define

$$\begin{aligned} B_t := \mathbb{E}_\sigma Q_{\boldsymbol{D}_t} &= \sum_{i=1}^{n} \sum_{j,j' > h_t} \Big\langle \phi(X_t^i), \boldsymbol{u}_t^j \Big\rangle \Big\langle \phi(X_t^i), \boldsymbol{u}_t^{j'} \Big\rangle \boldsymbol{u}_t^i \otimes \boldsymbol{u}_t^{j'} \\ &= \sum_{i=1}^{n} \Big\langle \sum_{j > h_t} \Big\langle \phi(X_t^i), \boldsymbol{u}_t^j \Big\rangle \boldsymbol{u}_t^j \Big\rangle \otimes \Big\langle \sum_{j' > h_t} \Big\langle \phi(X_t^i), \boldsymbol{u}_t^{j'} \Big\rangle \boldsymbol{u}_t^{j'} \Big\rangle = \sum_{i=1}^{n} Q_{\boldsymbol{D}_{t,i}}, \end{aligned}$$

where we introduce $\boldsymbol{D}_{t,i} := \sum_{j > h_t} \langle \phi(X_t^i), \boldsymbol{u}_t^j \rangle \boldsymbol{u}_t^j$. Note that, in taking the expectation with respect to the random variables X_t^i's, Theorem D.1 cannot be utilized, since the covariance may have infinite rank, that is, we might not be able to find a finite-dimensional subspace, which contains the range of all the $Q_{\boldsymbol{D}_{t,i}}$'s. However, since for all $t \in \mathbb{N}_T$ and $i \in \mathbb{N}_n$, it holds that $\|\boldsymbol{D}_{t,i}\| \leq \sqrt{\mathcal{K}}$, all the $Q_{\boldsymbol{D}_{t,i}}$'s satisfy $0 \preceq Q_{\boldsymbol{D}_{t,i}} \preceq \mathcal{K}I$ and they all are rank-one operators. It then follows that

$$\text{dimSpan}\Big((\text{Ran}(\boldsymbol{B}_1), \ldots, \text{Ran}(\boldsymbol{B}_T)) \Big) \leq nT.$$

Therefore, we can apply Theorem D.3 with $d = nT$, which, in conjunction with the Jensen's inequality, yields

$$\begin{aligned} \mathbb{E}_X \mathbb{E}_\sigma \|\boldsymbol{V}'\|_{S_\infty} &\leq \mathbb{E}_X \sqrt{\mathbb{E}_\sigma \| \sum_{t=1}^{T} Q_{\boldsymbol{D}_t} \|_{S_\infty}} \\ &\leq \mathbb{E}_X \sqrt{\| \mathbb{E}_\sigma \sum_{t=1}^{T} Q_{\boldsymbol{D}_t} \|_{S_\infty}} + \sqrt{2n\mathcal{K} \left(\ln(nT) + 1\right)} \\ &\leq \sqrt{\mathbb{E}_X \| \sum_{t=1}^{T} \boldsymbol{B}_t \|_{S_\infty}} + \sqrt{2n\mathcal{K} \left(\ln(nT) + 1\right)} \\ &\leq \sqrt{\| \mathbb{E}_X \sum_{t=1}^{T} \boldsymbol{B}_t \|_{S_\infty}} + \sqrt{6\mathcal{K} \left(\ln(4n^2T^2) + 1\right)} + \sqrt{2n\mathcal{K} \left(\ln(nT) + 1\right)}. \end{aligned} \tag{D.3}$$

After some simplifications, we arrive at

$$\mathbb{E}_{X,\sigma} \|\boldsymbol{V}'\|_{S_\infty} \leq \sqrt{\| \mathbb{E}_X \sum_{t=1}^{T} \boldsymbol{B}_t \|_{S_\infty}} + 6\sqrt{n\mathcal{K} \left(\ln(nT) + 1\right)}. \tag{D.4}$$

Furthermore, it can be shown that

$$\mathbb{E}_X \boldsymbol{B}_t = n \sum_{j>h_t} \lambda_t^j \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j.$$

By considering the task-averaged operator $C = 1/T \sum_{t=1}^{T} J_t = 1/T \sum_{t=1}^{T} \sum_{j=1}^{\infty} \lambda_t^j \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j$ and choosing $\lambda_h := \max_{t \in \mathbb{N}_T} \{\lambda_t^{h_t}\}$, we get

$$\begin{aligned}\mathbb{E}_{X,\sigma} \|\boldsymbol{V}'\|_{S_\infty} &\leq \sqrt{n \| \sum_{t=1}^{T} \sum_{j>h_t} \lambda_t^j \boldsymbol{u}_t^j \otimes \boldsymbol{u}_t^j \|_{S_\infty}} + 6\sqrt{n\mathcal{K}\left(\ln(nT)+1\right)} \\ &= \sqrt{nT\lambda_h} + 6\sqrt{n\mathcal{K}\left(\ln(nT)+1\right)}. \end{aligned} \tag{D.5}$$

This last result, combined with (48), provides the LRC bound for the trace norm regularized class $\mathcal{F}'_{S_1}$

$$\mathfrak{R}(\mathcal{F}'_{S_1}, r) \leq \sqrt{\frac{r \sum_{t=1}^{T} h_t}{nT}} + \sqrt{\frac{2R'^2_{max}\lambda_h}{n}} + 6\sqrt{\frac{2R'^2_{max}\mathcal{K}\left(\ln(nT)+1\right)}{nT}}. \tag{D.6}$$

Finally, using a similar argument as the one in Theorem 24, we get

$$r^* \leq \min_{0 \leq h_t \leq \infty} \left\{ \frac{B^2 \sum_{t=1}^{T} h_t}{nT} + 4BL\sqrt{\frac{2R'^2_{max}\lambda_h}{n}} + 24BL\sqrt{\frac{2R'^2_{max}\mathcal{K}\left(\ln(nT)+1\right)}{nT}} \right\}. \tag{D.7}$$

References

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 *International Conference on Machine learning*, pages 17–24. ACM, 2008.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, pages 41–48, 2007a.

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, 2007b.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008a.

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, 2008b.

Andreas Argyriou, Stéphan Clémençon, and Ruocong Zhang. Learning the graph of relations among multiple tasks. *ICML workshop on New Learning Frameworks and Models for Big Data*, 2014.

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

Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

Peter L Bartlett, Shahar Mendelson, and Petra Philips. Local complexities for empirical risk minimization. In *International Conference on Computational Learning Theory*, pages 270–284. Springer, 2004.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12 (149-198):3, 2000.

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.

Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for hiv therapy screening. In *International Conference on Machine learning*, pages 56–63. ACM, 2008.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *Annals of Probability*, pages 1583–1614, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, Ecole Polytechnique, Paris, 2002.

Bin Cao, Nathan N Liu, and Qiang Yang. Transfer learning for collective link prediction in multiple heterogenous domains. In *International Conference on Machine Learning (ICML-10)*, pages 159–166, 2010.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, pages 308–323. Springer, 2011.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems*, pages 2760–2768, 2013.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*. Cambridge University Press, 2007.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004.

Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637, 2005.

Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. Unpublished Manuscript, 2009.

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.

Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *International Conference on Machine Learning*, pages 521–528, 2011.

Marius Kloft and Gilles Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In *Advances in Neural Information Processing Systems*, pages 2438–2446, 2011.

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.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 02 2002.

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

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 12 2006. doi: 10.1214/009053606000001019.

Vladimir Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, December 2010. ISSN 1532-4435.

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

Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

Yunwen Lei, Lixin Ding, and Yingzhou Bi. Local rademacher complexity bounds based on covering numbers. *arXiv:1510.01463 [cs.AI]*, 2015.

Cong Li, Michael Georgiopoulos, and Georgios C Anagnostopoulos. Multitask classification hypothesis space with improved generalization bounds. *IEEE Transactions on Neural Networks and Learning Systems*, 26(7):1468–1479, 2015.

K Lounici, M Pontil, AB Tsybakov, and SA Van De Geer. Taking advantage of sparsity in multi-task learning. In *Conference on Learning Theory*, 2009.

Yishay Mansour and Mariano Schain. Robust domain adaptation. *Annals of Mathematics and Artificial Intelligence*, 71(4):365–380, 2013.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, June 2009a.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Conference on Uncertainty in Artificial Intelligence*, pages 367–374, 2009b.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, pages 1041–1048. 2009c.

Andreas Maurer. Bounds for linear multi-task learning. *The Journal of Machine Learning Research*, 7:117–139, January 2006a.

Andreas Maurer. The rademacher complexity of linear transformation classes. In *International Conference on Computational Learning Theory*, pages 65–78. Springer, 2006b.

Andreas Maurer. A chain rule for the expected suprema of gaussian processes. *Theoretical Computer Science*, 650:109–122, 2016.

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.

Andreas Maurer and Massimiliano Pontil. Bounds for vector-valued function estimation. *arXiv preprint arXiv:1606.01487*, 2016.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Shahar Mendelson. On the performance of kernel classes. *The Journal of Machine Learning Research*, 4:759–771, 2003.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

Charles A Micchelli and Massimiliano Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing Systems*, pages 921–928, 2004.

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.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Anastasia Pentina and Shai Ben-David. Multi-task and lifelong learning of kernels. In *Algorithmic Learning Theory*, pages 194–208. Springer, 2015.

Anastasia Pentina and Christoph H Lampert. Lifelong learning with non-iid tasks. In *Advances in Neural Information Processing Systems*, pages 1540–1548, 2015.

G Peshkir and Albert Nikolaevich Shiryaev. The khintchine inequalities and martingale expanding sphere of their action. *Russian Mathematical Surveys*, 50(5):849–904, 1995.

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.

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.

Sebastian Thrun and Lorien Pratt. *Learning To Learn*. Springer Science & Business Media, 2012.

I. Tolstikhin, G. Blanchard, and M. Kloft. Localized complexities for transductive learning. In *Conference on Learning Theory*, volume 35, pages 857–884, 2014.

Nicole Tomczak-Jaegermann. The moduli of smoothness and convexity and the rademacher averages of the trace classes s_p $(1 \leq p < \infty)$. *Studia Mathematica*, 2(50):163–182, 1974.

Sara Van De Geer. A new approach to least-squares estimation, with applications. *The Annals of Statistics*, pages 587–602, 1987.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

Christian Widmer, Marius Kloft, and Gunnar Rätsch. Multi-task learning for computational biology: Overview and outlook. In *Empirical Inference*, pages 117–127. Springer, 2013.

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.

Niloofar Yousefi, Michael Georgiopoulos, and Georgios C Anagnostopoulos. Multi-task learning with group-specific feature space sharing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 120–136. Springer, 2015.

Chao Zhang, Lei Zhang, and Jieping Ye. Generalization bounds for domain adaptation. In *Advances in Neural Information Processing Systems 25*, pages 3320–3328. 2012.

Yu Zhang and Dit-Yan Yeung. Multi-task warped gaussian process for personalized age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2622–2629, 2010.