

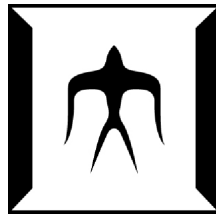
論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Supervised Machine Learning for Tensor Structured Models with Scaled Latent Trace Norm Regularization
著者(和文)	WimalawarneKishan
Author(English)	Kishan Wimalawarne
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10237号, 授与年月日:2016年3月26日, 学位の種別:課程博士, 審査員:杉山 将,徳永 健伸,篠田 浩一,村田 剛志,藤井 敦
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10237号, Conferred date:2016/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

Supervised Machine Learning for Tensor Structured Models with Scaled Latent Trace Norm Regularization

Kishan Wimalawarne

February 2016



Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

Thesis Committee:

Masashi Sugiyama, Chair

Koichi Shinoda

Takenobu Tokunaga

Tsuyoshi Murata

Atsushi Fujii

*Submitted in partial fulfillment of
the requirements for the degree of*

Doctor of Engineering

Keywords: Tensor, multilinear rank, scaled latent trace norm, regularization, regression, classification, multilinear multitask learning, excess risk bounds

To my parents

Abstract

In machine learning the structure of the data and structure of relationships among learning problems can play an important role. As the popularity of machine learning increases more and more challenging complex data structures are becoming available and required to be analysed. In this thesis we study the importance of learning by preserving structure of data and better ways of modelling relationships among related learning tasks.

We focus on higher dimensional arrays or tensors that are frequently found in many application domains. As with matrices, one of the important features of a tensor is the multilinear rank. Estimation and exploitation of the multilinear rank of tensors would allow us to build good learning models for tensors especially if the tensor is low rank. Yet compared to matrices exploiting the low rankness of a tensor is difficult due to the high dimensional structure of tensors. We look into existing low rank inducing tensor norms such as the overlapped trace norm and the latent trace norm that have been previously used to regularize learning models to understand their limitations. We find that both of these norms have a limitation that they do not consider the relative rank compared to mode dimensions. We propose a new norm called the *scaled latent trace norm* which explicitly takes the relative rank compared to mode dimensions when regularized.

The first problem that is investigated in this thesis is the fundamental question of identifying the optimal way to learn with tensor data. We challenge the common approach of converting data into vectors in order to use ordinary vector based learning models. We demonstrate using simple regression and classification models that by learning directly with tensor data without converting them to vectors and by applying low rank regularization methods we can outperform existing vector based learning models. To do this we extend regression and clas-

sification models with different tensor norms such as the overlapped trace norm, the latent trace norm and the scaled latent trace norm. Our theoretical analysis based on the excess risk bounds for each of these tensor norms allows us to infer how the the excess risk for each tensor norm is related to the multilinear rank of the weight tensor. We propose to solve these regularized tensor learning problems using the state of art optimisation method of alternating direction method of multipliers. Through toy experiments and real world experiments we demonstrate that our theoretical results match with our experimental results and that the direct learning with tensors is better than the vectorised learning.

The second topic that is studied in this thesis is multilinear multitask learning. In this topic we investigate how to structure multiple related tasks together in tensor format such that the information sharing among the related tasks leads to better performances among individual tasks. In order to study multilinear multitask learning deeply we extend multilinear multitask learning with the latent trace norm and the scaled latent trace norm regularizations. We derive excess risk bounds to show how the multilinear rank of the task weight tensor is related to the excess risk for each of the tensor norm regularizations. Using the alternating direction method of multipliers we propose to solve multilinear multitask learning problems. Through experiments on toy and real world problems, we show that the scaled latent trace norm is more capable of giving better performances.

We believe that the research described in this thesis can lead to more interesting research directions in the future. After stating our conclusions, we provide many possible future investigations that can be interesting to many researchers.

Acknowledgments

First and foremost I want to deeply thank my PhD supervisor Prof. Masashi Sugiyama for accepting me as his student and for his guidance and support during my study at the Tokyo Institute of Technology. I think I was truly fortunate to join a highly reputed research laboratory as the Sugiyama laboratory which laid the foundation for me to pursue high quality research. It was always a great experience to work with Prof. Masashi Sugiyama due to his immense kindness and support even during his many busy schedules. The guidance that came from his vast experience and knowledge made me learn how to think as a good researcher and learn what and how to research. I also want to thank Prof. Masashi Sugiyama for financial supports provided such as research assistantships and travel supports. I think I will always be in debt to Prof. Masashi Sugiyama for all his support and guidance.

I also want to thank Prof. Ryota Tomioka for offering me a visiting studentship at the Toyota Technological Institute at Chicago during March and April in 2014. While working with Prof. Ryota Tomioka I was able to deepen my knowledge on tensors and many aspects of machine learning. His wonderful teachings, advise and guidance enabled me to understand many of my weaknesses and to improve myself as a better researcher. I also want to thank his financial support during my stay at the Toyota Technological Institute at Chicago.

I want to thank Prof. Tsuyoshi Murata and Prof. Takenobu Tokunaga for offering working space in their laboratories and all the members of their laboratories for their support. I also want to thank my thesis committee members Prof. Koichi Shinoda, Prof. Tsuyoshi Murata, Prof. Takenobu Tokunaga and Prof. Atsushi Fujii for accepting to evaluate my thesis and thesis defence.

I was very fortunate to be selected to receive the MEXT Monbukagakusho

scholarship to support my studies during my first three years. The MEXT Monbukagakusho scholarship helped me to live a comfortable life in Tokyo such that I could concentrate on my research without worrying about finances. I want to thank the Japanese Government and the Tokyo Institute of Technology for the selecting me as a scholarship recipient. I also want to thank Hitokuse Inc. and Freakout Inc. for providing me internship opportunities.

Over the years many members of the Sugiyama lab have supported me immensely in my studies and many aspects of daily life and I want to appreciate their kindness and friendship. I want to make a special mention about the secretaries Ms. Ayako Tamai, Ms. Yasuyo Obana and Ms. Yuko Kawashima for their assistance in many administrative works. I also want to thank the Kendo club at the Tokyo Institute of Technology for letting me join and it was immensely helpful to relax and socialize which made my university life much enjoyable.

Last but not least, I want to thank my parents for their love and great sacrifices all their lives without which I may not have been able to achieve anything.

Contents

Abstract	v
Acknowledgments	vii
List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Machine Learning	1
1.2 Supervised Learning	3
1.2.1 Supervised Learning with Vector Data	5
1.2.2 Supervised Learning with Matrix Data	6
1.2.3 Supervised Learning with Tensor Data	7
1.3 Multitask Learning	8
1.3.1 Matrix based Multitask Learning	8
1.3.2 Tensor based Multitask Learning	9
1.4 Low Rank Tensor Norms	10
1.5 Contribution of This Thesis	12
1.6 Organization of This Thesis	14
2 Scaled Latent Trace Norm	15
2.1 Tensors	15
2.2 Tensor Unfolding	16
2.3 Basic Tensor Algebra	18
2.4 Tensor Decompositions and Tensor Rank	18
2.4.1 CANDECOMP/PARAFAC Decomposition	19
2.4.2 Tucker Decomposition	19
2.5 Low Rank Tensor Norms	20
2.5.1 Overlapped Tensor Trace Norm	21
2.5.2 Latent Trace Norm	22

2.6	New Norm: Scaled Latent Trace Norm	23
2.7	Conclusion	25
3	Tensor based Multitask Learning	27
3.1	Introduction	27
3.2	Multilinear Multitask Learning	28
3.3	Optimisation	29
3.3.1	Optimisation with Latent Trace Norms	29
3.3.2	Optimisation with the Overlapped Trace Norm	31
3.4	Theory	33
3.5	Experimental Results	44
3.5.1	Synthetic data sets	44
3.5.2	Restaurant data set	47
3.5.3	School data set	48
3.6	Conclusion	49
4	Conclusions and Future Work	51
	Bibliography	55

List of Figures

1.1	Supervised learning	4
1.2	(a) Matrix structured multitask learning and (b) multilinear multi-task learning.	10
2.1	An example of a 3-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$	16
2.2	Fibers of a 3-mode tensor. In (a), all fibers denoted as $\mathcal{X}(:, j, k)$, in (b), all fibers denoted as $\mathcal{X}(i, :, k)$ and in (c), all fibers denoted as $\mathcal{X}(i, j, :)$	16
2.3	Slices of a 3-mode tensor. In (a), all slices denoted as $\mathcal{X}(:, :, k)$, in (b), all slices denoted as $\mathcal{X}(i, :, :)$ and in (c), all slices denoted as $\mathcal{X}(:, j, :)$	17
2.4	CP decomposition of a 3-mode tensor.	19
2.5	Tucker decomposition of tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. The tensor \mathcal{C} is the core tensor and matrices $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$ are component matrices.	20
3.1	Results for the synthetic data sets.	46
3.2	Results for the real world data sets.	48

List of Tables

3.1 Sample complexities of matrix trace norm in various settings. The common factor $1/\epsilon^2$ is omitted from the sample complexities. The sample complexities are defined with respect to $|S|$ for matrix completion and m for multitask learning. 42

3.2 Sample complexities of the overlapped trace norm, latent trace norm, and the scaled latent trace norm in various settings. The common factor $1/\epsilon^2$ is omitted from the sample complexities. The sample complexities are defined with respect to $m|S|$ for tensor completion and multilinear multitask learning. In the heterogeneous cases, we assume $P \leq r < r'$. We define $\|\mathbf{r}\|_{1/2} = (\sum_{k=1}^3 \sqrt{r_k}/K)^2$ and $N := n_1 n_2 n_3$ 43

Chapter 1

Introduction

In this thesis we investigate learning with tensor data and tensor structured learning models based on their low rank properties. Our research focuses on tensor norm regularization to exploit low rankness to learn with tensor data and tensor structured problems. We explore weaknesses and limitations of existing tensor norms and propose a new norm called the *scaled latent trace norm* that can overcome limitations of existing tensor norms. As successful applications of this tensor norm we explore tensor regression and classification and multilinear multitask learning. In this chapter we give an overview of our motivations and contributions we make on the above mentioned topics.

1.1 Machine Learning

Machine learning can be put forward as the process that takes data representing some experience as input and output some expertise as a computer program which can improve itself with more experience (Mitchell, 1997; Shalev-Shwartz and Ben-David, 2014). More simply we can think of machine learning as a process of transforming experiences in some domain into an expertise. Machine learning is a complex process involving many branches of mathematics such as statistics, numerical optimisation and analysis and computer science such as software engineering, databases and distributed computing.

There are many ways of learning depending on the characteristic of the data and the acquisitions of data and some of the popular methods of learning are as

follows (Shalev-Shwartz and Ben-David, 2014; Chapelle et al., 2006; Sutton and Barto, 1998; Cesa-Bianchi and Lugosi, 2006).

- Supervised learning
- Semi-supervised learning
- Unsupervised learning
- Reinforcement and bandits learning

One major method of categorising learning is based on the availability of data. If all the data instances have labels and available as a batch, learning from such data can be categorised as supervised learning (Shalev-Shwartz and Ben-David, 2014) and regression and classification are popular examples of supervised learning. If none of the data instances have any labels then the learning from such data is known as unsupervised learning (Shalev-Shwartz and Ben-David, 2014). Clustering, principal component analysis and independent component analysis are examples of unsupervised learning. If data is partially labelled then learning can be performed by benefiting from the data with labels while efficiently using the unlabelled data. Semi-supervised learning (Chapelle et al., 2006) has been modelled on several assumptions on the unlabelled data such as the smoothness assumption, cluster assumption and manifold assumptions to extend supervised learning models. In some applications all the data instances may not be available as a batch for learning and the data instances and their labels may become available when interacting with the environment. Reinforcement learning (Sutton and Barto, 1998) and bandits learning (Cesa-Bianchi and Lugosi, 2006) methods are able to efficiently learn in these situations where other batch based learning methods would not have the capacity to learn efficiently.

In machine learning data plays an important role and capturing the information from data is often crucial in building efficient prediction models. Irrespective of whether the data are labelled or not, the format of the data can also provide additional information. In general data may come in formats of vectors, matrices or tensors and their structural properties can also give additional information in addition to the numerical values of their elements. Low rankness is an important structural property of matrices and tensors which is utilised in many learning problems such inductive learning (Signoretto et al., 2013b), data imputation (Cai

et al., 2010), robust principal component analysis (Candès et al., 2011) and subspace clustering (Liu et al., 2010).

In this thesis we focus on learning utilising the structural properties of data with a focus on tensor data. In order to investigate the importance of the structural properties of tensors in learning we focus on supervised learning. More specifically we look into supervised learning with tensor structured problems using regularization with low rank inducing norms. Next we look more deeply into supervised learning.

1.2 Supervised Learning

Supervised learning is one of the popular methods of learning. The purpose of supervised learning is to infer a function based on completely labelled training data. What we mean by completely labelled training data in the machine learning jargon is that each raw data instance acquired from the environment for training comes with a label or a label is assigned by a domain expert. Regression, classification and ranking are well known examples of supervised learning models.

Figure 1.1 shows the process of supervised learning. In supervised learning, the labelled data goes through the process of training a model. The validation process allows to find the optimal parameters of the model to fit the training data. Once a model has been learned and when new unseen test data are available the prediction process is able to predict the labels of the test data using the trained model.

Many of the supervised learning models employ regularization methods and a heap of norms that can be used for regularization have been invented to acquire the best performance out of learning models and training data. Additionally training data also comes in many formats such as vectors, matrices and tensors. Depending on the format of data such as vector, matrix or tensor and also the properties of data such as sparseness and low rankness we many have to formulate different supervised learning models using appropriate regularization methods. We will distinguish these models in the next subsections.

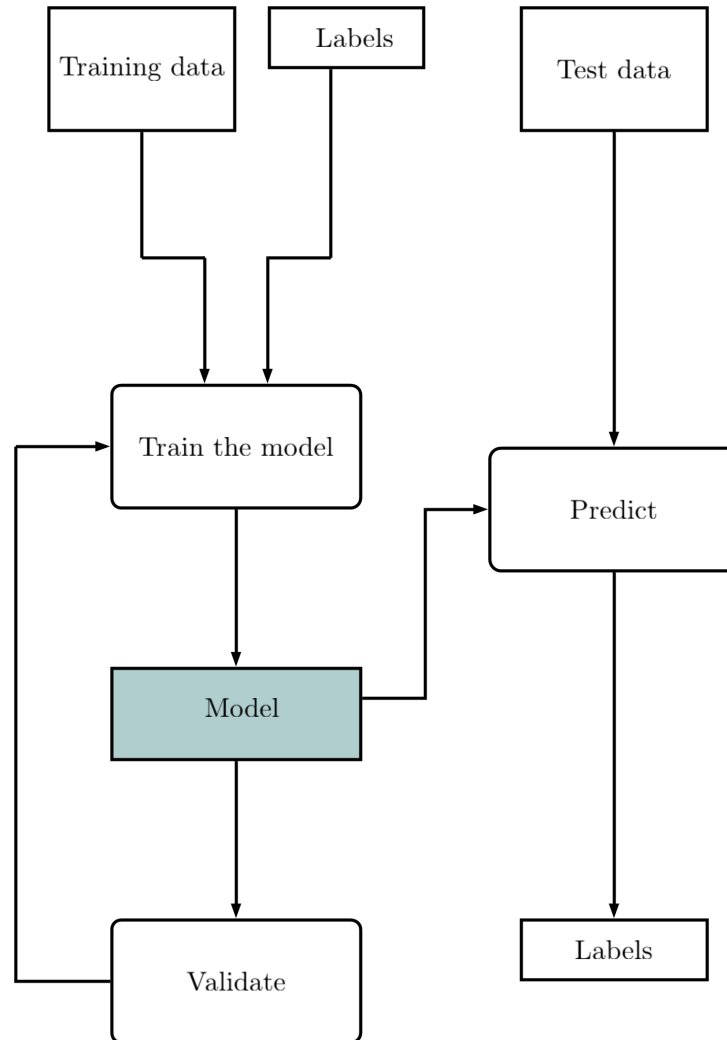


Figure 1.1: Supervised learning

1.2.1 Supervised Learning with Vector Data

The most common approach to model supervised learning problems is with vector formatted data. This could be due to the abundance of vector formatted data and the simplicity of designing learning models. A typical learning model for a data set $(x_i, y_i) \in \mathbb{R}^D \times \mathbb{R}$, $i = 1, \dots, m$ can be modelled as follows:

$$R(w) = \min_w \sum_{i=1}^m l(\langle x_i, w \rangle, y_i) + \lambda \|w\|_{\text{vnorm}},$$

where $l(\cdot, \cdot)$ can be any loss function, λ is a regularization parameter and $\|\cdot\|_{\text{vnorm}}$ can be any vector based norm such as the l_2 , l_1 or any structured norms (Jenatton et al., 2011). Vector norm regularizations have many applications such as to learn from ill-posed problems (Tikhonov and Arsenin, 1977), learn from sparse data (Tibshirani, 1996) and feature selection (Jenatton et al., 2011). Using cross validation (James et al., 2014) an optimal λ can be obtained to best fit the data into the model.

Supervised learning with vector based data has been extensively studied theoretically in machine learning (Bunea et al., 2007; Kakade et al., 2009; Maurer and Pontil, 2012). Most fundamentally the excess risk bounds (Bartlett et al., 2006) are useful in understanding how different vector based regularization methods behave based on the different properties of data. Given a supervised learning model, the excess risk bound can be defined as the difference of the empirical training loss for a specific data set of finite samples and the expected training loss provided data samples can be acquired from a data distribution. In recent years analysis using Rademacher complexities (Bartlett and Mendelson, 2002) has become popular with deriving excess risk bounds since it allows us to derive data dependent bounds. Employing the Rademacher complexity, data dependent bounds have been derived for many vector based norms such as l_2 , l_1 and l_p/l_q (Kakade et al., 2009) and structured norms such as Lasso, group Lasso, overlapping group norms (Maurer and Pontil, 2012).

It is a common practice to convert data to vectors irrespective of their original format and apply vector based models for learning. Though this process enables us to reduce the learning with any data format into a vector based learning which provides convenience, it can also lead to loss of critical information about the

data. This is especially relevant with matrix or tensor structured data since considering their structural properties in learning can provide additional information that would lead to better learning and predictions.

1.2.2 Supervised Learning with Matrix Data

Many real world data exists naturally in the format of matrices (e.g., images and recommendation data). When learning with these matrix data it is a common practice to convert these data into vectors and apply vector based learning models as described previously. This process of converting to vectors can often lead to loss of important structural properties such as the low rankness.

Matrix based supervised learning has been proposed by several researchers (e.g., (Tomioka and Aihara, 2007; Zhou and Li, 2014)) to directly learn from matrix data without converting them to vectors. For instance the loss function for matrix based regression for data $(X, y) \in \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}$ can be represented as $l(\langle X, W \rangle, y)$ where $W \in \mathbb{R}^{n_1 \times n_2}$ and $\langle X, W \rangle = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X_{i,j} W_{i,j}$. As a consequence learning coefficient matrix W can be regularized as a matrix using matrix regularization methods such as the matrix *trace norm* (Recht et al., 2010).

The trace norm or the nuclear norm of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ is defined as

$$\|X\|_{\text{tr}} = \sum_{j=1}^J \sigma_j, \quad (1.1)$$

where $\sigma_j \geq 0$ is the j^{th} singular value and J is the number of non-zero singular values ($J \leq \min(n_1, n_2)$). A matrix is called *low rank* if $J < \min(n_1, n_2)$. The matrix trace norm (1.1) is a convex envelope to the matrix rank and it is commonly used in matrix low-rank approximation (Recht et al., 2010).

To formally define matrix based supervised learning let us consider training data $(X_i, y_i) \in \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}, i = 1, \dots, m$ and a learning model can be expressed as follows:

$$R(W) = \min_W \sum_{i=1}^m l(\langle X_i, W \rangle, y_i) + \lambda \|W\|_{\text{tr}},$$

where $l(\cdot, \cdot)$ is any loss function. It has been shown that better prediction accuracies can be achieved with this matrix based learning model for EEG classification (Tomioka and Aihara, 2007). In addition (Tomioka and Aihara, 2007) has also

shown that low rank regularization in the context of EEG data analysis allows us to analyse the *singular value spectra* of learning coefficients that help understand the activities of brain regions.

1.2.3 Supervised Learning with Tensor Data

A much less studied but a commonly found format of real world data is tensor data. Sequence data (Liu et al., 2013), spatio-temporal data (Bahadori et al., 2014) and brain-computer interface (BCI) data (Onishi et al., 2012) are few examples of tensor data. Similarly to matrices tensor data can also be used directly in learning models without converting them to vectors and the learning coefficients can be regularized using suitable low rank tensor norms. Recently tensor based inductive learning has received some attention (Signoretto et al., 2013b) but it has not been studied in depth with latest developments of low rank tensor norms and no theoretical analysis on tensor based supervised learning has been performed.

In order to model tensor based learning let us consider a dataset $(\mathcal{X}_i, y_i) \in \mathbb{R}^{n_1 \times \dots \times n_K} \times \mathbb{R}, i = 1, \dots, m$ and similarly to the matrix based supervised learning we can define the tensor inductive learning models as follows:

$$R(\mathcal{W}) = \min_{\mathcal{W}} \sum_{i=1}^m l(\langle \mathcal{X}_i, \mathcal{W} \rangle, y_i) + \lambda \|\mathcal{W}\|_{\text{tnorm}}, \quad (1.2)$$

where $\langle \mathcal{X}, \mathcal{W} \rangle = \sum_{i_1=1}^{n_1} \dots \sum_{i_K=1}^{n_K} X_{i_1, \dots, i_K} W_{i_1, \dots, i_K}$ and $\|\cdot\|_{\text{tnorm}}$ is any low rank tensor norm. The $\|\cdot\|_{\text{tnorm}}$ can be either the *overlapped trace norm* (Liu et al., 2009; Tomioka and Suzuki, 2013) or the *latent trace norm* (Tomioka and Suzuki, 2013).

In Signoretto et al. (2013b) it has been proposed to use the overlapped trace norm (Liu et al., 2009; Tomioka and Suzuki, 2013) as the low rank tensor regularization method. Theoretical studies by Tomioka and Suzuki (2013) on application of the overlapped trace norm for tensor decomposition has shown that it can result in good performances when the multilinear ranks of the tensor have less variation. Further research by Tomioka and Suzuki (2013) with the latent trace norm has shown that it can produce better performances with tensors that have high variations in multilinear ranks. Depending on the training data and its tensor structure it will result in a learning coefficient tensor with some specific multilinear ranks

which the overlapped trace norm alone may not be sufficiently optimise. This makes it necessary to explore tensor based supervised learning in detail in relation to other tensor norms. Also the non-existence of theoretical study on tensor based learning for different norms is another limitation in acquiring understanding how these different tensor norms would behave.

1.3 Multitask Learning

In this section we overview multitask learning (Caruana, 1997; Baxter, 2000; Ando and Zhang, 2005). Multitask learning is a popular learning approach that learns multiple related learning tasks together by allowing to share information among tasks. We focus our attention on low rank based multitask learning models.

1.3.1 Matrix based Multitask Learning

The matrix based multitask learning based on spectral properties or low rank structure was first developed in Argyriou et al. (2007) and it was further studied in Chen et al. (2012). The basic idea behind these methods was to arrange all the learning coefficients of tasks as a matrix and regularize it using the matrix trace norm. The application of the trace norm allow us to find a low rank subspace in the coefficient matrix which allows information sharing among tasks.

Let us define the trace norm regularized matrix multitask learning more formally. We consider T tasks with each task having training data as $(x_{it}, y_{it}) \in \mathbb{R}^n \times \mathbb{R}$, $t = 1, \dots, T$, $i = 1, \dots, m_t$ and the matrix multitask learning models can be formulated as follows:

$$R(W) = \min_{w_1, \dots, w_T} \sum_{t=1}^T \sum_{i=1}^{m_t} l(\langle x_{it}, w_t \rangle, y_{it}) + \lambda \|W\|_{\text{tr}},$$

where $W = [w_1; \dots; w_T]$.

Advances in theoretical machine learning has been moving along the direction of multitask learning as well (Baxter, 2000). Recently theoretical analysis based on excess risk bounds for trace norm based matrix multitask learning has been conducted using the Rademacher complexity (Maurer and Pontil, 2013) and it has

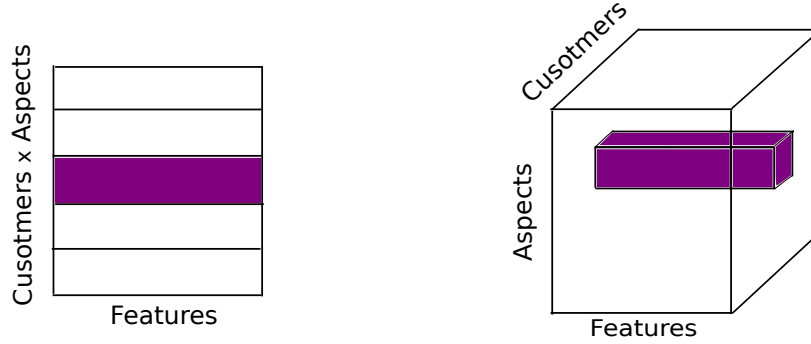
brought forward that the the collective excess risk of all tasks is bounded with the rank of the coefficient matrix. This is a significant result since it has shown that information sharing can occur among tasks provided the coefficient matrix is low rank. Additionally theoretical aspects of matrix online based multitask learning (Kakade et al., 2012; Cavallanti et al., 2010) and multitask dictionary learning (Maurer et al., 2014) have been well studies.

1.3.2 Tensor based Multitask Learning

Recently the multilinear multitask learning (Romera-Paredes et al., 2013) has been proposed which extends the matrix based multitask learning model to tensor structured multitask learning models. Multilinear multitask learning is able to exploit the structural information more than the matrix based multitask learning due to the higher dimensional structure of tensors. An advantage of multilinear multitask learning is that it allows imputation of missing tasks (Romera-Paredes et al., 2013).

As an example of a multilinear multitask learning problem, let us consider a recommendation system where customers give ratings to different aspects (quality of food, service,....etc) of different restaurants (Romera-Paredes et al., 2013). If we consider single task learning problems we can learn a model for each customer for predicting his ratings for each aspect. But among different customers and among different aspects there may be shared information which cannot be utilised when considering tasks in isolation. In order to take advantage of such shared information learning all the tasks together as a multitask learning problem can be used. If we consider the matrix based multitask learning we discussed previously we can arrange all learning coefficients of customers and ratings along a dimension of a matrix as in Figure 1.2 (a) and apply low rank regularization. But this arrangement does not take the full structural information since it does not separate the relationship with customer and aspects. In order to fully utilize these structural information the most optimised approach would be to have customers on one mode and ratings on another mode of a tensor as shown in the Figure 1.2 (b).

The multilinear multitask learning with three dimensional (3-mode) tensor structure can be defined for $n_1 n_2$ tasks with training data $(x_{ipq}, y_{ipq}) \in \mathbb{R}^d \times$



(a) Matrix multitask learning

(b) Multilinear multitask learning

Figure 1.2: (a) Matrix structured multitask learning and (b) multilinear multitask learning.

\mathbb{R} , $p = 1, \dots, n_1, q = 1, \dots, n_2, i = 1, \dots, m_{pq}$ as follows:

$$R(W) = \min_{\mathcal{W}} \sum_{p=1}^{n_1} \sum_{q=1}^{n_2} \sum_{i=1}^{m_{pq}} l(\langle x_{ipq}, w_{pq} \rangle, y_{ipq}) + \lambda \|\mathcal{W}\|_{\text{tnorm}}, \quad (1.3)$$

where $\mathcal{W} = \mathbb{R}^{d \times n_1 \times n_2}$ and $w_{pq} = \mathcal{W}(:, p, q)$. Similarly to the supervised learning model setting different tensor trace norms can be applied to $\|\cdot\|_{\text{tnorm}}$.

The original multilinear multitask learning (Romera-Paredes et al., 2013) has been modelled with regularization using the overlapped trace norm. Multilinear multitask learning problems may often result in irregular "flat" tensors due to smaller number of tasks compared to a larger feature space. Due to this reason original proposal of the application of the overlapped norm regularization may not capture the true low rankness of many multilinear multitask learning problems. Due to its recent development, multilinear multitask learning has not been analysed theoretically prior to our investigations that we elucidate in later chapters.

1.4 Low Rank Tensor Norms

Let us inspect low rank regularization more closely. The rank of a matrix is defined as the maximum number of independent rows or columns (Golub and

Van Loan, 1996). Though there are different ways to define the rank of a tensor (as we will discuss in Chapter 2) such as by CP decomposition (Kiers, 2000b) or Tucker decomposition (Tucker, 1966c), the rank of a tensor can also be viewed as the maximum number of independent components that a tensor can be decomposed. If the rank of a matrix or a tensor is less than the maximum number of possible independent components or in other words less than the mode dimensions then that matrix or the tensor is low rank.

The low rankness of a matrix or a tensor is an important structural property. Similarly to matrices (Recht et al., 2010), constraining the rank of tensor is also a NP hard problem (Hillar and Lim, 2013). It is a standard practice to use the convex envelope of the matrix trace norm (1.1) to approximate the rank of a matrix which can also be used with tensor norms (Liu et al., 2009). Due to the higher dimensions of tensors compared to matrices, estimating the ranks of a tensor can be more difficult computationally and also designing norms to find the low rankness among multiple modes can be challenging. In many applications such as missing data imputation (Cai et al., 2010), robust principal component analysis (Candès et al., 2011), and subspace clustering (Liu et al., 2010) low rank structure of *data* has been successfully utilized. The problems that we discuss in this thesis, the tensor based regression and classification and the multilinear multitask learning consider the low rankness of the dual of data which are the learning coefficients. In these settings the low rankness of the data is not necessary.

As we discussed in the previous sections tensor based learning problems have been modelled in previous researches solely using the overlapped trace norm. This is a major limitation since the learning coefficient tensors can result with high variations in multilinear ranks. The recently proposed latent trace norm (Tomioka and Suzuki, 2013) has shown to be more robust when working with tensors with high variations in multilinear ranks. Application of the latent trace norm regularization to supervised learning with tensor data and multilinear multitask learning may help to overcome limitations of the overlapped trace norm regularization.

A common limitation that both the overlapped trace norm and the latent trace norm have is that both of these norms do not consider the relative rank with respect to mode dimensions. This limitation becomes significant with irregular "flat" tensors with some modes having high dimensions and some modes having signif-

icantly small dimensions compared to others. In these tensors it is possible that along some of modes with small dimensions their ranks can be close to their mode dimension while along high dimensional modes their ranks can be much smaller than their mode dimensions. In some instances it may arise that the ranks of the small dimensional modes are smaller than the ranks of high dimensional modes. For tensors with the above mentioned behaviour the relative rank compared to the mode dimensions can be crucial in understanding their true low rankness. Since both the overlapped trace norm and the latent trace norm do not consider the relative low rankness they may perform inaccurately with tensors with high variations in multilinear ranks and mode dimensions.

1.5 Contribution of This Thesis

The most important contribution of our work is the development of a new tensor norm called the *scaled latent trace norm*. The scaled latent trace norm is an extension of the latent trace norm with scaling of each trace norm by the square root of the mode dimension. The motivation to define this norm is to overcome limitations of existing norms such as the overlapped trace norm and the latent trace norm when regularising tensors with high variations in multilinear ranks and mode dimensions. From a mathematical point of view we identify that in order to specify the low rankness of a tensor based on its multilinear rank it is important to compare the ranks relative to the mode dimensions. More specifically we claim that the true low rankness of a tensor is along the mode with the minimum of relative rank with respect to the mode dimensions and not on the mode with lowest of the multilinear rank. We show that the proposed norm is able to identify the modes with lowest rank *relative* to its mode dimensions which makes it perform better with tensor regularized problems compared to other norms.

We investigate supervised learning with tensor data extensively by applying our newly proposed scaled latent trace norm along with the latent trace norm and the overlapped trace norm. We derive excess risk bounds for all the tensor norm regularized learning models which is missing in existing research literature. Our theoretical bounds derived using Rademacher complexity analysis are able to show the relationship of multilinear ranks and tensor dimensions with the excess

risk bounds for each tensor norm. We show that for the overlapped trace norm regularization, the excess risk is bounded with the sum of the square root of multilinear ranks of the tensor, for the latent trace norm regularization the excess risk is bounded with the minimum of the multilinear ranks and for the scaled latent trace norm the excess risk is bounded with the ratio of minimum rank to mode dimension. We show that the scaled latent trace norm is superior compared to other norms with tensors having high variation in multilinear ranks and mode dimensions since it considers the relative rank with respect to the mode dimension. We also propose optimisation methods for solve the dual problems of tensor based inductive learning models using the alternating direction method of multipliers (ADMM) (Gabay and Mercier, 1976; Boyd et al., 2011). Through simulation experiments on tensor based regression and tensor based classification on real world data such as image sequences and brain computer interface (BCI) data we validate our theoretical results and show the efficiency of tensor based learning compared to vector and matrix based learning. An important conclusion that we arrive when learning with tensor data is that learning from tensor data without converting to vectors and exploiting low rankness lead to better performances.

We also extend multilinear multitask learning by applying the scaled latent trace norm and the latent trace norm regularizations. We derive excess risk bounds for multilinear multitask learning to understand the theoretical properties of regularization with tensor norms. Based on the bounds we derive we show that for the overlapped trace norm regularization, the excess risk is bounded with average of the square root of multilinear ranks of the tensor, for the latent trace norm regularization the excess risk is bounded with the minimum of multilinear ranks and for the scaled latent trace norm the excess risk is bounded with the ratio of the minimum rank to mode dimension. Similarly to supervised learning, these bounds show that the scaled latent trace norm is superior compared to other norms in multilinear multitask learning with tensors having high variation in multilinear ranks and mode dimensions due to the consideration of the relative rank with respect to the mode dimension. Similarly to the proposed optimisation methods used for tensor based supervised learning we propose to solve the dual problems for multilinear multitask learning using the alternating direction method of multipliers. As in the supervised learning with tensor data we provide excess risk bounds for

all regularizations with tensor norms. Our experiments with real world data show that our proposed scaled latent trace norm performs the best compared to other tensor norms.

1.6 Organization of This Thesis

This thesis is organised in the following structure.

In Chapter 2, we first provide an overview on tensors which is needed to understand our contributions we make in this thesis. We review existing tensor norms, the overlapped trace norm and the latent trace norm. We next move on to define the *scaled latent trace norm* as published in Wimalawarne et al. (2014) and describe its properties.

In Chapter 3, we focus on supervised learning with tensor data using tensor norm regularizations. First we propose regression and classification of tensor data with all tensor norms such as the overlapped trace norm, the latent trace norm and the newly defined scaled latent trace norm. Next we provide optimisation methods to solve tensor based regression and classification for each of the tensor norms using the ADMM. We next analyse the excess risk bounds for supervised learning setting for all tensor norm regularizations. In the experiments section we first provide simulation experiments with tensor based regression to understand how different tensor norms perform under different multilinear ranks and tensor mode dimensions. Next we demonstrate real data experiments with hand gesture recognition data and BCI data. Finally we have the conclusions of the chapter.

In Chapter 4, we describe our proposed extensions of the multilinear multi-task learning with the latent trace norm and the scaled latent trace norm. First we describe the proposed models and provide the details of optimisation procedures. Next we provide excess risk bounds related to all the multilinear multitask learning problems with all the tensor norms. Next we describe our simulation and real world experiments for multilinear multitask learning. Finally we have the conclusions.

In Chapter 5, we describe the conclusion of our research and discusses future research directions.

Chapter 2

Scaled Latent Trace Norm

The main purpose of this chapter is to introduce a new tensor norm called the *scaled latent trace norm*. First we discuss basic concepts of tensors and tensor norms which is essential in understanding research described in this thesis. After the brief introduction to fundamental concepts of tensors we review existing tensor norms such as the overlapped trace norm and the latent trace norm. Next we define our new norm, the *scaled latent trace norm* which is an extension of the latent trace norm. Further we describe basic properties such as the duality of the scaled latent trace norm.

2.1 Tensors

A tensor is a multi-dimensional array (Kolda and Bader, 2009). Let us consider $\mathcal{X} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ which is a K -way or K th order tensor. If K is equal to 1 or 2 it will result in a vector or a matrix respectively and if K is more than or equal to 3 it results in a high dimensional tensor. The total number of elements of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ is $N = \prod_{k=1}^K n_k$. Extending the convention of denoting elements in vectors and matrices an element of the tensor can be represented as $\mathcal{X}_{i_1, \dots, i_K}$ for $(i_1, \dots, i_K) \in [n_1] \times \dots \times [n_K]$. An example of a visualisation possibly 3-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is shown in Figure 2.1.

A fiber of a tensor is the vector obtained after fixing all indexes except one (Figure 2.2). Fibers are analogous to rows and columns of a matrix. For a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ its fibers can be represented as $\mathcal{X}_{:,j,k}$, $\mathcal{X}_{i,: ,k}$ and $\mathcal{X}_{i,j,:}$.

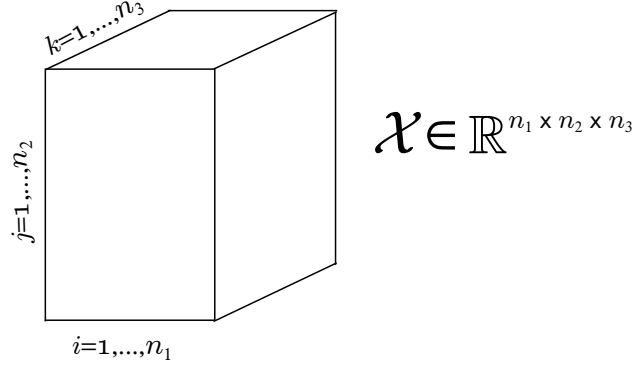
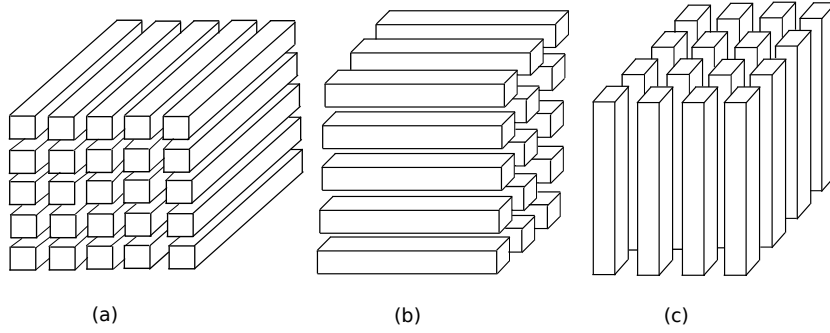
Figure 2.1: An example of a 3-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ 

Figure 2.2: Fibers of a 3-mode tensor. In (a), all fibers denoted as $\mathcal{X}(:, j, k)$, in (b), all fibers denoted as $\mathcal{X}(i, :, k)$ and in (c), all fibers denoted as $\mathcal{X}(i, j, :)$

A slice of a tensor is any matrix that can be obtained by fixing all indices except two (Figure 2.3). For a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ slices can be represented as $\mathcal{X}_{::,k}$, $\mathcal{X}_{i,::}$ and $\mathcal{X}_{:,j,:}$.

2.2 Tensor Unfolding

A very useful operation associated with tensors is the tensor unfolding. The mode- k unfolding (also known as matricization or flattening) of tensor \mathcal{X} is represented as $X_{(k)} \in \mathbb{R}^{n_k \times N/n_k}$ which is obtained by concatenating all the N/n_k mode- k fibers along its columns.

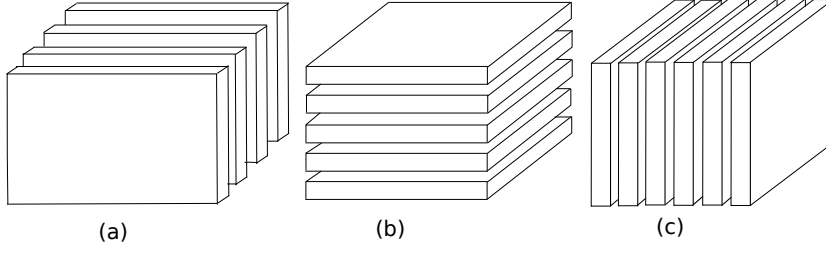


Figure 2.3: Slices of a 3-mode tensor. In (a), all slices denoted as $\mathcal{X}(:, :, k)$, in (b), all slices denoted as $\mathcal{X}(i, :, :)$ and in (c), all slices denoted as $\mathcal{X}(:, j, :)$

In order to understand tensor unfolding easily let us consider an example. We consider a 3-mode tensor $\mathcal{W} \in \mathbb{R}^{4 \times 3 \times 2}$ which is made by slices $\mathcal{X}(:, :, 1) = X_1$ and $\mathcal{X}(:, :, 2) = X_2$ which are

$$X_1 = \begin{bmatrix} u_1 & u_2 & u_3 \\ u_4 & u_5 & u_6 \\ u_7 & u_8 & u_9 \\ u_{10} & u_{11} & u_{12} \end{bmatrix},$$

$$X_2 = \begin{bmatrix} v_1 & v_2 & v_3 \\ v_4 & v_5 & v_6 \\ v_7 & v_8 & v_9 \\ v_{10} & v_{11} & v_{12} \end{bmatrix}.$$

The unfolding of \mathcal{X} on each of the modes can be represented as $X_{(1)}$, $X_{(2)}$ and $X_{(3)}$ as follows:

$$X_{(1)} = \begin{bmatrix} u_1 & u_2 & u_3 & v_1 & v_2 & v_3 \\ u_4 & u_5 & u_6 & v_4 & v_5 & v_6 \\ u_7 & u_8 & u_9 & v_7 & v_8 & v_9 \\ u_{10} & u_{11} & u_{12} & v_{10} & v_{11} & v_{12} \end{bmatrix},$$

$$X_{(2)} = \begin{bmatrix} u_1 & u_4 & u_7 & u_{10} & v_1 & v_4 & v_7 & v_{10} \\ u_2 & u_5 & u_8 & u_{11} & v_2 & v_5 & v_8 & v_{11} \\ u_3 & u_6 & u_9 & u_{12} & v_3 & v_6 & v_9 & v_{12} \end{bmatrix},$$

$$X_{(3)} = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} & u_{11} & u_{12} \\ v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 & v_8 & v_9 & v_{10} & v_{11} & v_{12} \end{bmatrix}.$$

2.3 Basic Tensor Algebra

We review few basic tensor algebraic methods related to tensors in this section.

The k -mode product of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_K}$ with a matrix $U \in \mathbb{R}^{m \times n_k}$ is defined as $\mathcal{X} \times_k U \in \mathbb{R}^{n_1 \times \dots \times n_{k-1} \times m \times n_{k+1} \times \dots \times n_K}$

$$\mathcal{Y} = \mathcal{X} \times_k U \Leftrightarrow Y_{(k)} = UX_{(k)}.$$

We denote the outer product using an operator \circ for $v^{(1)} \in \mathbb{R}^{n_1}, \dots, v^{(K)} \in \mathbb{R}^{n_K}$ which leads to a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ as

$$\mathcal{X} = v^{(1)} \circ v^{(2)} \circ \dots \circ v^{(K)},$$

where each element $x_{i_1, i_2, \dots, i_K}, i_1 \in [n_1], \dots, i_K \in [n_K]$ is

$$x_{i_1, i_2, \dots, i_K} = v_{i_1}^{(1)} v_{i_2}^{(2)} \dots v_{i_K}^{(K)}.$$

The inner product of two tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_K}$ is defined as

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{n_1} \dots \sum_{i_K=1}^{n_K} x_{i_1, \dots, i_K} y_{i_1, \dots, i_K},$$

and the Frobenius norm of a tensor \mathcal{W} can be expressed as

$$\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}.$$

2.4 Tensor Decompositions and Tensor Rank

In this section we discuss two popular tensor decomposition methods namely the CANDECOMP/PARAFAC decomposition and the Tucker decomposition.

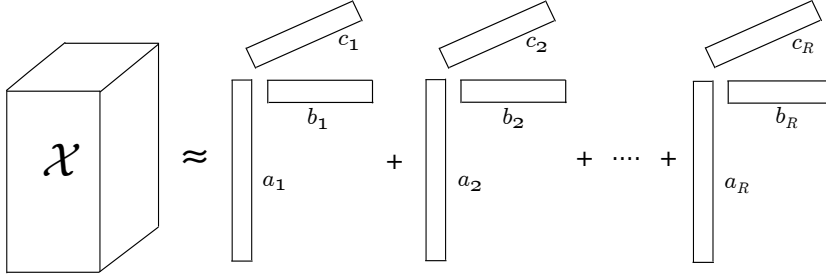


Figure 2.4: CP decomposition of a 3-mode tensor.

2.4.1 CANDECOMP/PARAFAC Decomposition

The CANDECOMP/PARAFAC decomposition (Kiers, 2000b) or commonly known as CP decomposition decomposes a tensor into a sum of finite number of rank one tensors. Given a 3-mode tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ we can write its CP decomposition as

$$\mathcal{X} \approx \sum_{r=1}^R a_r \circ b_r \circ c_r,$$

where $a_r \in \mathbb{R}^{n_1}$, $b_r \in \mathbb{R}^{n_2}$ and $c_r \in \mathbb{R}^{n_3}$. In Figure 2.4, a graphical description of a CP decomposition is given. The CP decomposition can be extended for any K -mode tensor.

The smallest number R of rank one tensors that can generate \mathcal{X} as their sum based on the CP decomposition is defined as the *rank*, $R = \text{Rank}(\mathcal{X})$ of that tensor.

2.4.2 Tucker Decomposition

A more commonly used method of tensor decomposition is the Tucker decomposition (Tucker, 1966c; Kolda and Bader, 2009). For illustrative purposes we consider a 3-way tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and its tucker decomposition can be shown as in the Figure 2.5.

As shown in Figure 2.5, the Tucker decomposition decomposes a tensor into a core tensor $\mathcal{C} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ and component matrices $U^{(1)} \in \mathbb{R}^{n_1 \times m_1}$, $U^{(2)} \in \mathbb{R}^{n_2 \times m_2}$ and $U^{(3)} \in \mathbb{R}^{n_3 \times m_3}$. Formally the Tucker decomposition can be written as

$$\mathcal{X} = \mathcal{C} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}, \quad (2.1)$$

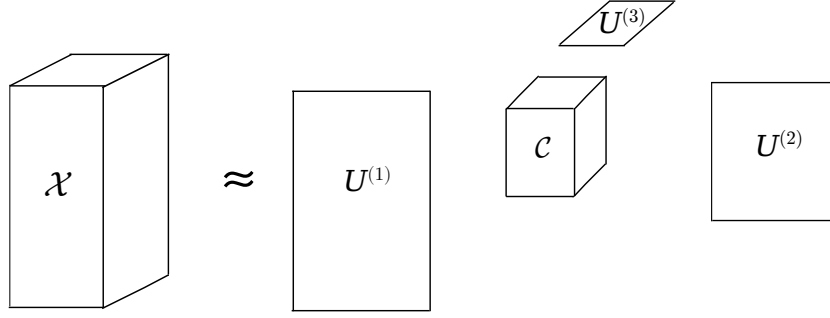


Figure 2.5: Tucker decomposition of tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. The tensor \mathcal{C} is the core tensor and matrices $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$ are component matrices.

or as element-wise computation as

$$x_{i,j,k} = \sum_{i_1=1}^{m_1} \sum_{j_1=1}^{m_2} \sum_{k_1=1}^{m_3} c_{i,j,k} u_{i,i_1}^{(1)} u_{j,j_1}^{(2)} u_{k,k_1}^{(3)}. \quad (2.2)$$

Based on the Tucker decomposition we can define the *multilinear rank* also known as *n-rank* for a tensor. The mode- k rank r_k of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ is defined as the rank of mode- k unfolding that is $X_{(k)}$ and the multilinear rank of \mathcal{X} is given as (r_1, \dots, r_K) . The CP decomposition can be viewed as a special case of the Tucker decomposition when $m_1 = \dots = m_K$ and the core tensor is superdiagonal (Kolda and Bader, 2009). This indicated that the multilinear rank based on the Tucker decomposition is a more general definition to capture the ranks of a tensor.

2.5 Low Rank Tensor Norms

One of the main concepts that we discuss in this thesis is the low-rankness of tensors. Before we consider tensors we first discuss the low rankness of matrices. When considering a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ its *trace norm* (Recht et al., 2010) is defined as

$$\|X\|_{\text{tr}} = \sum_{j=1}^J \sigma_j, \quad (2.3)$$

where σ_j is the j^{th} singular value and J is the number of non-zero singular values ($J \leq \min(n_1, n_2)$). A matrix is called a *low rank* matrix if $J < \min(n_1, n_2)$. The matrix trace norm (2.3) is a convex envelope of the rank of a matrix and it is commonly used in matrix low rank approximation (Recht et al., 2010).

As in matrices, the rank property is also available for tensors but it is more complicated due to its multidimensional structure. In order to define the low rankness of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ we can consider its multilinear rank (r_1, \dots, r_K) and if $r_k < n_k$ for any mode- k then the tensor is a low rank tensor. In recent years much effort has been put to develop convex envelopes of the tensor ranks that can be used in tensor regularizations. Some of the previously developed tensor norm are the overlapped trace norm and the latent trace norm. We will discuss them in the next two subsections.

2.5.1 Overlapped Tensor Trace Norm

One of the earliest definitions of a tensor norm is the *tensor nuclear norm* (Liu et al., 2009) or the *overlapped trace norm* (Tomioka and Suzuki, 2013), which can be represented for a tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ as

$$\|\mathcal{W}\|_{\text{overlap}} = \sum_{k=1}^K \|W_{(k)}\|_{\text{tr}}. \quad (2.4)$$

The overlapped trace norm can be viewed as a direct extension of the matrix trace norm since it unfolds a tensor on each of its modes and computes the sum of trace norms of the unfolded matrices. Regularization with the overlapped trace norm can also be seen as an overlapped group regularization method due to the fact that the same tensor is unfolded over different modes and regularized with the trace norm.

The following inequality is important for theoretical analysis in next chapters and it captures how the overlapped norm of a tensor can be related to its Forbenous norm,

$$\|\mathcal{W}\|_{\text{overlap}} \leq \left(\sum_{k=1}^K \sqrt{r_k} \right) \|\mathcal{W}\|_F. \quad (2.5)$$

One of the popular applications of the overlapped trace norm is *tensor completion* (Gandy et al.; Liu et al., 2009), where missing entries of a tensor are

imputed. Another application is *multilinear multitask learning* (Romera-Paredes et al., 2013), where multiple vector-based linear learning tasks with a common feature space are arranged as a tensor feature structure and the multiple tasks are solved together with constraints to minimize the multilinear ranks of the tensor feature.

Theoretical analyses on the overlapped norm have been carried out for tensor decomposition (Tomioka et al., 2011b). Based on their research when recovering a tensor \mathcal{W} from a noisy tensor $\hat{\mathcal{W}}$ of mode dimensions $n_1 \times \cdots \times n_K$ with multilinear rank (r_1, \dots, r_K) , the recovery error ($\|\hat{\mathcal{W}} - \mathcal{W}\|_F$) scales as $\mathcal{O}((\frac{1}{K} \sum_{k=1}^K \sqrt{r_k})^2 (\frac{1}{K} \sum_{k=1}^K 1/\sqrt{n_k})^2)$. We can see that the recovery error of overlapped trace norm regularization is bounded by the average mode- k ranks (Tomioka et al., 2011b) which can be large if some modes are close to full rank even if there are low-rank modes. Thus, these studies imply that the overlapped trace norm performs well when the multilinear ranks have small variations, and it may result in a poor performance when the multilinear ranks have high variations.

To overcome the weakness of the overlapped trace norm, recent research in tensor norms has lead to new norms such as the *latent trace norm* (Tomioka and Suzuki, 2013).

2.5.2 Latent Trace Norm

Recently the latent trace norm (Tomioka and Suzuki, 2013) has been proposed which is defined as

$$\|\mathcal{W}\|_{\text{latent}} = \inf_{\mathcal{W}^{(1)} + \mathcal{W}^{(2)} + \dots + \mathcal{W}^{(K)} = \mathcal{W}} \sum_{k=1}^K \|\mathcal{W}_{(k)}^{(k)}\|_{\text{tr}}.$$

The latent trace norm takes a mixture of K latent tensors which is equal to the number of modes, and regularizes each of them separately. In contrast to the overlapped trace norm, the latent tensor trace norm regularizes different latent tensors for each unfolded mode and this gives the tendency that the latent tensor trace norm picks the latent tensor with the lowest rank.

Following inequality is also useful in theoretical analysis and it shows how the latent trace norm can be bounded with the Frobenius norm,

$$\|\mathcal{W}\|_{\text{latent}} \leq \min_k \sqrt{r_k} \|\mathcal{W}\|_F. \quad (2.6)$$

In general, the latent trace norm results in a mixture of latent tensors and the content of each latent tensor would depend on the rank of its unfolding. In an extreme case, for a tensor with all its modes full except one mode, regularization with the latent tensor trace norm would result in making the latent tensor with the lowest mode become prominent while others become zero. When recovering a tensor \mathcal{W} from a noisy tensor $\hat{\mathcal{W}}$ of mode dimensions $n_1 \times \dots \times n_K$ with multilinear rank (r_1, \dots, r_K) with the latent trace norm regularization, the recovery error ($\|\hat{\mathcal{W}} - \mathcal{W}\|_F$) (Tomioka and Suzuki, 2013) scales as $\mathcal{O}(\frac{\min_k r_k}{\min_k n_k})$. This also shows that latent trace norm can select the mode with lowest multilinear rank.

2.6 New Norm: Scaled Latent Trace Norm

As we discussed in the introduction a major limitation of the overlapped trace norm and the latent trace norm is that both of these norms do not consider the relative low rankness with respect to mode dimensions. We define a new norm called the *scaled latent trace norm* to overcome these limitations as published in Wimalawarne et al. (2014). Our proposal is to extend the latent trace norm by scaling each mode wise trace norm of matrix unfolding with inverse of its mode dimension as follows

$$\|\mathcal{W}\|_{\text{scaled}} = \inf_{\mathcal{W}^{(1)} + \mathcal{W}^{(2)} + \dots + \mathcal{W}^{(K)} = \mathcal{W}} \sum_{k=1}^K \frac{1}{\sqrt{n_k}} \|W_{(k)}^{(k)}\|_{\text{tr}}.$$

Compared to the latent trace norm, the scaled latent trace norm takes the rank relative to the mode dimension. A major drawback of the latent trace norm is its inability to identify the rank of a mode relative to its dimension. If a tensor has a mode where its dimension is smaller than other modes yet its relative rank with respect to its mode dimension is high compared to other modes, the latent trace norm could incorrectly pick the smallest mode.

For a given tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ with multilinear rank (r_1, \dots, r_K) , the relative rank with respect to its dimensions can be defined as $(r_1/n_1, \dots, r_K/n_K)$. Due to the scaling of the trace norm of each unfolding with the inverse of its mode dimension, the scaled latent trace norm is able to find the relative rank on each mode. This fact will be demonstrated in excess risk bounds that we derive in the

next two chapters. The following theorem gives an useful inequality between the scaled latent trace norm and the Frobenius norm of a tensor and it shows that the scaled latent trace norm is bounded with the minimum of the relative rank with respect to its Frobenius norm.

Theorem 2.1. *For a tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ following inequality holds*

$$\|\mathcal{W}\|_{\text{scaled}} \leq \left(\min_k \sqrt{\frac{r_k}{n_k}} \right) \|\mathcal{W}\|_F, \quad (2.7)$$

where (r_1, \dots, r_K) is its multilinear rank.

Proof. We again use the singleton decomposition argument used in Tomioka and Suzuki (2013), where we assume that the scaled latent trace norm find the correct low rank mode k such that $\mathcal{W} = \mathcal{W}^{(k)}$ and $\mathcal{W}^{(i)} = 0$ for all $i \neq k$.

$$\begin{aligned} \|\mathcal{W}\|_{\text{scaled}} &= \inf_{\mathcal{W}^{(1)} + \mathcal{W}^{(2)} + \dots + \mathcal{W}^{(K)} = \mathcal{W}} \sum_{k=1}^K \frac{1}{\sqrt{n_k}} \|\mathcal{W}^{(k)}\|_{\text{tr}} \\ &= \min_k \frac{1}{\sqrt{n_k}} \|\mathcal{W}^{(k)}\|_{\text{tr}} \quad (\text{Singleton decomposition}) \\ &= \min_k \frac{1}{\sqrt{n_k}} \sum_{i=1}^J \sigma_i \quad (\sigma_i - \text{non zero singular values } J \leq \min(n_k, n_{\setminus k})) \\ &\leq \min_k \frac{1}{\sqrt{n_k}} \sqrt{\sum_{i=1}^J 1^2} \sqrt{\sum_{i=1}^J \sigma_i^2} \quad (\text{Cauchy - Schwarz inequality}) \\ &= \left(\min_k \sqrt{\frac{r_k}{n_k}} \right) \|\mathcal{W}\|_F. \end{aligned}$$

□

Next theorem put forward the dual of the scaled latent trace norm.

Theorem 2.2. *The dual of the scaled latent trace norm of a tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \dots \times n_K}$ is*

$$\|\mathcal{W}\|_{\text{scaled}^*} = \max_k \sqrt{n_k} \|\mathcal{W}^{(k)}\|_{\text{op}}. \quad (2.8)$$

Proof. Let

$$\|\mathcal{Y}\|_{\text{scaled}} = \inf_{\mathcal{Y}^{(1)} + \mathcal{Y}^{(2)} + \dots + \mathcal{Y}^{(K)} = \mathcal{Y}} \sum_{k=1}^K \frac{1}{\sqrt{n_k}} \|Y^{(k)}\|_{\text{tr}}.$$

In order to derive the duality we use the singleton decomposition argument used in Tomioka and Suzuki (2013), where we assume that the scaled latent trace norm find the correct low rank mode k such that $\mathcal{W} = \mathcal{W}^{(k)}$ and $\mathcal{W}^{(i)} = 0$ for all $i \neq k$.

Let the dual of $\|\mathcal{W}\|_{\text{scaled}}$ be $\|\mathcal{W}\|_{\text{scaled}^*}$ and then,

$$\begin{aligned} \|\mathcal{W}\|_{\text{scaled}^*} &= \sup_{\mathcal{W}} \langle \mathcal{W}, \mathcal{Y} \rangle \quad \text{s.t.} \quad \|\mathcal{Y}\|_{\text{scaled}} \leq 1 \\ &= \sup_{\mathcal{W}} \sum_{k=1}^K \langle W^{(k)}, Y^{(k)} \rangle \quad \text{s.t.} \quad \inf_{\mathcal{Y}^{(1)} + \dots + \mathcal{Y}^{(K)} = \mathcal{Y}} \sum_{k=1}^K \frac{1}{\sqrt{n_k}} \|Y^{(k)}\|_{\text{tr}} \leq 1 \\ &= \sup_{W^{(k)}} \langle W^{(k)}, Y^{(k)} \rangle \quad \text{s.t.} \quad \frac{1}{\sqrt{n_k}} \|Y^{(k)}\|_{\text{tr}} \leq 1 \quad (\text{Singleton decomposition}) \\ &= \sup_{W^{(k)}} \left\langle \sqrt{n_k} W^{(k)}, \frac{Y^{(k)}}{\sqrt{n_k}} \right\rangle \quad \text{s.t.} \quad \left\| \frac{Y^{(k)}}{\sqrt{n_k}} \right\|_{\text{tr}} \leq 1 \quad (\text{Rescaling}) \\ &= \max_k \sqrt{n_k} \|W^{(k)}\|_{\text{op}}. \end{aligned}$$

□

2.7 Conclusion

In this chapter we reviewed basic concepts in tensors and existing tensor norms such as the overlapped trace norm and latent trace norm. We described how each of those norms is defined and explained their usage in regularization in machine learning problems and also their limitations. To overcome the limitations with the overlapped norm and the latent trace norm we proposed our new norm the scaled latent trace norm and described its properties such as duality and an important inequality with respect to the Frobenius norm. In the next two chapters we apply these norms to regression and classification of tensor data and multilinear multitask learning.

Chapter 3

Tensor based Multitask Learning

In this chapter we discuss multilinear multitask learning. We extend multilinear multitask learning with tensor regularizations with the latent trace norm and the scaled latent trace norm and derive excess risk bounds for all tensor norm regularizations. With toy experiments and real world data experiments we show that the scaled latent trace norm performs well in many situations especially when the multilinear rank and mode dimensions of coefficient tensors have high variations.

3.1 Introduction

Multilinear multitask learning (Romera-Paredes et al., 2013) is a recently developed multitask learning method that has started gaining attention by researchers. The advantage of multilinear multitask learning is that it arranges tasks in a tensor structure allowing to model relationships along tasks better than its counterpart matrix based multitask learning. As we discussed in Chapter 1, most of the multilinear multitask learning models may have a "flat" structure due to large feature spaces and smaller number of tasks which make it important to consider the relative rank with respect to the mode dimensions when applying low rank tensor regularizations. The originally proposed multilinear multitask learning model by Romera-Paredes et al. (2013) was only based on the overlapped norm regularization which may not be appropriate for "flat" structured tensor models. The lack of theoretical analysis of multilinear multitask learning with different tensor norms is another limitation in understanding how the low rankness of tensors are related

to the information sharing and performance improvement among tasks.

In this chapter, we propose to extend multilinear multitask learning with the latent trace norm and the scaled latent trace norm. We study the excess risk of the three norms through their Rademacher complexities in various settings including matrix completion, matrix multitask learning, and multilinear multitask learning. Our analysis allows us to also study the tensor completion setting, which was only empirically studied in Tomioka et al. (2011a,b). Our analysis consistently shows the advantage of the scaled latent trace norm in various settings in which the dimensions or ranks are heterogeneous. Experiments on both synthetic and real data sets are also consistent with our theoretical findings.

This chapter is organised as follows. First in Section 4.2 we propose our extensions to the multilinear multitask learning. In Section 4.3 we discuss optimisation methods for multilinear multitask learning with the overlapped trace norm and the scaled latent trace norm. Next in Section 4.4 we derive and discuss excess risk bounds for multilinear multitask learning and in Section 4.5 we discuss experimental results. Finally in Section 4.6 we have our conclusions.

3.2 Multilinear Multitask Learning

In multilinear multitask learning, M -dimensional task space is considered, i.e., the parameters of $R = T_2 \times \cdots \times T_{M+1}$ tasks form a $(M + 1)$ -mode tensor $\mathcal{W} \in \mathbb{R}^{T_1 \times T_2 \times \cdots \times T_{M+1}}$, and the parameter vector of task $r = (t_2, \dots, t_{M+1})$ is given by $w_r = \mathcal{W}_{:,t_2,\dots,t_{M+1}}$. For each task r we may have training data $(X_r \in \mathbb{R}^{m_r \times T_1}, y_r \in \mathbb{R}^{m_r})$ where m_r is the number of training samples and some tasks may not have any training data.

The first proposal of the multilinear multitask learning (Romera-Paredes et al., 2013) has been developed using the overlapped trace norm regularization to regression tasks as follows:

$$\min_{\mathcal{W}} \sum_{r=1}^R \|X_r w_r - y_r\|^2 + \lambda \sum_{k=1}^{M+1} \|W_{(k)}\|_{\text{tr}}. \quad (3.1)$$

In this chapter we propose to extend (3.1) by applying the latent trace norm

and the scaled latent trace norm as follows:

$$\min_{\mathcal{W}^{(1)+\dots+\mathcal{W}^{(M+1)}=\mathcal{W}}} \sum_{r=1}^R \left\| X_r \left(\sum_{k=1}^{M+1} (w_r^{(k)}) \right) - y_r \right\|^2 + \sum_{k=1}^{M+1} \lambda_k \|W_{(k)}^{(k)}\|_{\text{tr}}, \quad (3.2)$$

where for the latent trace norm, $\lambda_k = \lambda$ and for the scaled latent trace norm, $\lambda_k = \frac{\lambda}{\sqrt{T_k}}$ for $k = 1, \dots, M + 1$ for any given regularization parameter λ .

In the next section we discuss optimisation strategies for solving above two multilinear multitask learning problems.

3.3 Optimisation

In this section we discuss optimisation of (3.1) and (3.2).

3.3.1 Optimisation with Latent Trace Norms

We first discuss optimisation procedure for our proposed approach (3.2). Similarly to the optimisation methods that we used in the previous chapter we solve the dual formulation of (3.2) using the alternating direction method of multipliers (ADMM).

First we express the (3.2) with the introduction of auxiliary variables $\mathcal{Z}^{(k)}$, $k = 1, \dots, M + 1$ as follows:

$$\begin{aligned} \min_{\mathcal{W}^{(1)+\dots+\mathcal{W}^{(M+1)}=\mathcal{W}}} \sum_{r=1}^R \left\| X_r \left(\sum_{k=1}^{M+1} (w_r^{(k)}) \right) - y_r \right\|^2 + \sum_{k=1}^{M+1} \lambda_k \|Z_{(k)}^{(k)}\|_{\text{tr}}, \\ \text{s.t. } \mathcal{W}^{(k)} = \mathcal{Z}^{(k)}, \quad k = 1, \dots, M + 1. \end{aligned} \quad (3.3)$$

The dual problem for the above problem (3.3) can be written as follows:

$$\begin{aligned} \min_{\alpha} \sum_{r=1}^R \left(\frac{\lambda}{2} \|\alpha_r\|^2 - \alpha_r^\top y_r \right) + \sum_{k=1}^{M+1} \delta_{\gamma_k}(Z_{(k)}^{(k)}), \\ \mathfrak{X}_{(1)}^\top(\alpha) = [X_1^\top \alpha_1, \dots, X_R^\top \alpha_R] \\ \mathcal{Z}^{(k)} = \mathfrak{X}^\top(\alpha), \quad k = 1, \dots, M + 1, \end{aligned} \quad (3.4)$$

where $\alpha = [\alpha_1, \dots, \alpha_R]$, the $\delta_{\gamma_k}(V)$ function is the indicator function such that $\delta_{\gamma_k}(V) = 0$ if $\|V\|_{\text{op}} \leq \gamma_k$ and $\delta_{\gamma_k}(V) = +\infty$ otherwise ($\|V\|_{\text{op}}$ is the maximum

singular value of V). To obtain the solution for the latent trace norm, we have to make $\gamma_k = 1$ and to obtain the solution for the scaled latent trace norm we have to set $\gamma_k = \sqrt{T_k}$ for $k = 1, \dots, M + 1$.

By introducing $W^{(k)} \in \mathbb{R}^{T_1 \times T_2 \times \dots \times T_{M+1}}$, $k = 1, \dots, M + 1$ as dual variables the augmented Lagrangian for dual problem (3.4) can be written as follows:

$$\begin{aligned} \mathcal{L} = & \sum_{r=1}^R \left(\frac{\lambda}{2} \|\alpha_r\|^2 - \alpha_r^\top y_r \right) + \left(\sum_{k=1}^{M+1} \delta_{\gamma_k}(Z_{(k)}^{(k)}) \right. \\ & \left. + \sum_{k=1}^{M+1} \langle W_{(k)}^{(k)}, Z_{(k)}^{(k)} - \mathfrak{X}_{(k)}^\top(\boldsymbol{\alpha}) \rangle + \frac{\beta}{2} \sum_{k=1}^{M+1} \sum_{r=1}^R \|z_r^{(k)} - X_r^\top \alpha_r\|^2 \right), \end{aligned}$$

where $z_r^{(k)}$ is the fiber corresponding to the task r in $\mathcal{Z}^{(k)}$. Note that the dual variables in the above dual problem lead to the solutions for $\mathcal{W}^{(k)}$ $k = 1, \dots, M + 1$ in primal problem (3.2).

The solution for α_r for iteration $t + 1$ results in the following,

$$\alpha_r^{t+1} = (\beta X_r X_r^\top + \lambda I)^{-1} \left(y_r + \sum_{k=1}^{M+1} X_r w_r^{(k)t} + \sum_{k=1}^{M+1} \beta X_r z_r^{(k)t} \right),$$

where $w_r^{(k)}$ is the task r of $W^{(k)}$. For tasks with no training instances, $\alpha_r = 0$.

Solutions for each $\mathcal{Z}^{(k)}$ at iteration $t + 1$ can be obtained by solving

$$\delta_{\gamma_k}(Z_{(k)}^{(k)t+1}) - \frac{\beta}{2} \left\| Z_{(k)}^{(k)t} + \left(\mathfrak{X}_{(k)}^\top(\boldsymbol{\alpha}^t) - \frac{W_{(k)}^{(k)t}}{\beta} \right) \right\|^2,$$

which results in

$$Z_{(k)}^{(k)t+1} = \text{proj}_{\gamma_k} \left(\mathfrak{X}_{(k)}^\top(\boldsymbol{\alpha}^t) - \frac{W_{(k)}^{(k)t}}{\beta} \right),$$

where $\text{proj}_\mu(W) = U \min(S, \mu) V^\top$ for a given matrix $W = USV^\top$.

Finally the update for each $\mathcal{W}^{(k)}$ at iteration $t + 1$ is as follows:

$$W_{(k)}^{(k)t+1} = W_{(k)}^{(k)t} + \beta (Z_{(k)}^{(k)t+1} - \mathfrak{X}_{(k)}^\top(\boldsymbol{\alpha}^t)).$$

Stopping Condition

The relative duality gap (Tomioaka et al., 2011a) can be used as the stopping condition for above dual problem. Let $p(\mathcal{W}^t)$ be the primal objective (3.2) and

$D(-\alpha^t) = \sum_{r=1}^R \left(\frac{\lambda}{2} \|\alpha_r\|^2 - \alpha_r^\top y_r \right)$ be the computable dual objective (3.4) at the iteration step t . The computation of the $D(-\alpha^t)$ dual objective requires α^t to satisfy the constraints $\|Z_{(k)}^{(k)}\|_{\text{op}} \leq \gamma_k$ for each $k = 1, \dots, M+1$ as given in (3.4). Let $\sigma_{k,1}$ be the maximum eigenvalue of $Z_{(k)}^{(k)}$ and let a shrinkage factor c be defined as $c = \min \left(1, \frac{\gamma_1}{\sigma_{1,1}}, \frac{\gamma_2}{\sigma_{1,2}}, \dots, \frac{\gamma_{M+1}}{\sigma_{1,M+1}} \right)$. Then the scaling $\alpha^* = c\alpha$ makes sure that each $\|Z_{(k)}^{(k)}\|_{\text{op}} \leq \gamma_k$ for each $k = 1, \dots, M+1$. The relative duality gap stopping condition at step t is given as follows,

$$\frac{p(\mathcal{W}^t) - D(-\alpha^{*t})}{p(\mathcal{W}^t)} \leq \epsilon,$$

where ϵ is a tolerance.

3.3.2 Optimisation with the Overlapped Trace Norm

We now look at the optimisation of the primal problem of multilinear multi-task learning with the overlapped trace norm regularization (3.1). The primal problem for the multilinear multitask learning with the overlapped trace norm regularization (3.1) can be reexpressed with introduction of auxiliary variables $\mathcal{Z}^{(k)}, k = 1, \dots, M+1$ as follows,

$$\begin{aligned} \min_{\mathcal{W}} \sum_{r=1}^R \frac{1}{2} \|X_r w_r - y_r\|^2 + \lambda \sum_{k=1}^{M+1} \|Z_{(k)}^{(k)}\|_{\text{tr}} \\ \text{s.t. } \mathcal{W} = \mathcal{Z}^{(k)}, \quad k = 1, \dots, M+1. \end{aligned} \quad (3.5)$$

Formulating the above problem as an ADMM problem with introduction of Lagrangian multipliers $C^{(k)} \in \mathbb{R}^{T_1 \times T_2 \times \dots \times T_{M+1}}$ and parameter $\beta > 0$, the augmented Lagrangian function is defined as follows:

$$\mathcal{L} = \sum_{r=1}^R \frac{1}{2\lambda} \|X_r w_r - y_r\|^2 + \sum_{k=1}^{M+1} \left(\|Z_{(k)}^{(k)}\|_{\text{tr}} + \langle C_{(k)}^{(k)}, W_{(k)} - Z_{(k)}^{(k)} \rangle + \frac{\beta}{2} \|W_{(k)} - Z_{(k)}^{(k)}\|_F^2 \right).$$

Differentiating with respect to w_r and solving for each update at iteration $t+1$, we obtain the following:

$$w_r^{t+1} = \left(\frac{X_r^T X_r}{\lambda} + (M+1)\beta I \right)^{-1} \left(\frac{X_r^T y_r}{\lambda} - \sum_{k=1}^{M+1} c_r^{(k)t} + \beta \sum_{k=1}^{M+1} z_r^{(k)t} \right),$$

where $c_r^{(k)}$ is the fiber with respect to the task r in $\mathcal{C}^{(k)}$ and $z_r^{(k)}$ is the fiber with respect to the task r in $\mathcal{Z}^{(k)}$.

For tasks without any training instances, the w_r update at iteration $t + 1$ is given as follows:

$$w_r^{t+1} = \frac{1}{(M+1)\beta} \left(- \sum_{k=1}^{M+1} c_r^{(k)t} + \beta \sum_{k=1}^{M+1} z_r^{(k)t} \right).$$

The update for each $Z^{(k)}$ at iteration $t + 1$ can be obtained by solving

$$Z_{(k)}^{(k)t+1} = \min_{Z_{(k)}^{(k)}} \left\| Z_{(k)}^{(k)t} \right\|_{\text{tr}} + \frac{\beta}{2} \left\| Z_{(k)}^{(k)t} - \left(\frac{C_{(k)}^{(k)t+1}}{\beta} + W_{(k)}^{t+1} \right) \right\|_F^2,$$

which results in

$$Z_{(k)}^{(k)t+1} = \mathbf{S}_{1/\beta} \left(W_{(k)}^{t+1} + \frac{C_{(k)}^{(k)t+1}}{\beta} \right),$$

where $\mathbf{S}_\alpha(X) = U \text{diag}((\mu - \alpha)_+) V^\top$ for a matrix X with singular value decomposition of $X = U \text{diag}(\mu) V^\top$.

The updates for each $C^{(k)}$ at iteration $t + 1$ are as follows:

$$C_{(k)}^{(k)t+1} = C_{(k)}^{(k)t} + \beta(W_{(k)}^{t+1} - Z_{(k)}^{(k)t+1}).$$

Optimality Condition

We propose to use the subgradient (Boyd and Vandenberghe, 2004) as the stopping criterion for the primal solution. Let us write (3.5) by introducing Lagrangian multipliers $\mathcal{C}^{(k)}$ without the proximity terms as follows:

$$\mathcal{L} = \min_W \sum_{r=1}^R \frac{1}{2\lambda} \|X_r w_r - y_r\|^2 + \sum_{k=1}^{M+1} (\|Z_{(k)}^{(k)}\|_{\text{tr}} + \langle C_{(k)}^{(k)}, W_{(k)} - Z_{(k)}^{(k)} \rangle). \quad (3.6)$$

By differentiating the above \mathcal{L} with respect to each $Z^{(k)}$ we get,

$$\partial \mathcal{L} = \partial \|Z_{(k)}^{(k)}\|_{\text{tr}} - C_{(k)}^{(k)} \ni 0.$$

Taking the differentiation of the objective function (3.5) with respect to $W_{(k)}$ and substituting the subgradient of $C_{(k)}^{(k)} \in \partial \|Z_{(k)}^{(k)}\|_{\text{tr}}$ in place of $\partial \|W_{(k)}\|_{\text{tr}}$, we can

have the following stopping condition:

$$\left\| \sum_{r=1}^R \left(X_r^T (X_r w_r - y_r) + \sum_{k=1}^{M+1} C^{(k)r} \right) \right\|^2 \leq \epsilon_1,$$

where ϵ_1 is some tolerance value. In addition we also have to check

$$\|W_{(k)} - Z_{(k)}^{(k)}\|_F \leq \epsilon_2 \quad k = 1, \dots, M + 1,$$

where ϵ_2 is a tolerance value.

3.4 Theory

In order to develop theoretical analysis we consider only 3-mode tensors but our theoretical results can be extended to any tensor. Let us consider $T = PQ$ supervised learning tasks. Training samples $(\mathbf{x}_{ipq}, y_{ipq})_{i=1}^{m_{pq}}$ ($(p, q) \in S$) are provided for a relatively small fraction of the task index pairs $S \subset [P] \times [Q]$. Each task is parametrized by a weight vector $\mathbf{w}_{pq} \in \mathbb{R}^d$, which can be collected into a 3-way tensor $\mathcal{W} = (\mathbf{w}_{pq}) \in \mathbb{R}^{d \times P \times Q}$ whose (p, q) fiber is \mathbf{w}_{pq} . We define the learning problem as follows:

$$\hat{\mathcal{W}} = \underset{\mathcal{W} \in \mathbb{R}^{d \times P \times Q}}{\operatorname{argmin}} \hat{L}(\mathcal{W}), \quad \text{subject to} \quad \|\mathcal{W}\|_{\star} \leq B_0, \quad (3.7)$$

where the norm $\|\cdot\|_{\star}$ is either the overlapped trace norm, latent trace norm, or the scaled latent trace norm, and the empirical risk \hat{L} is defined as follows:

$$\hat{L}(\mathcal{W}) = \frac{1}{|S|} \sum_{(p,q) \in S} \frac{1}{m_{pq}} \sum_{i=1}^{m_{pq}} \ell(\langle \mathbf{x}_{ipq}, \mathbf{w}_{pq} \rangle - y_{ipq}).$$

The true risk we are interested in minimizing is defined as follows:

$$L(\mathcal{W}) = \frac{1}{PQ} \sum_{p,q} \mathbb{E}_{(\mathbf{x}, y) \sim P_{pq}} \ell(\langle \mathbf{x}, \mathbf{w}_{pq} \rangle - y),$$

where P_{pq} is the distribution from which the samples $(\mathbf{x}_{ipq}, y_{ipq})_{i=1}^{m_{pq}}$ are drawn from.

The next lemma relates the excess risk $L(\hat{\mathcal{W}}) - L(\mathcal{W}^*)$ with the expected dual norm $\mathbb{E} \|\mathcal{D}\|_{\star}$ through Rademacher complexity.

Lemma 3.1. *We assume that the output y_{ipq} is bounded as $|y_{ipq}| \leq b$, and the number of samples $m_{pq} \geq m > 0$ for the observed tasks. We also assume that the loss function ℓ is Lipschitz continuous with the constant Λ , bounded in $[0, c]$ and $\ell(0) = 0$. Let \mathcal{W}^* be any tensor such that $\|\mathcal{W}^*\|_* \leq B_0$. Then with probability at least $1 - \delta$, any minimizer of (3.7) satisfies the following bound:*

$$L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) \leq 2\Lambda \left(\frac{2B_0}{|S|} \mathbb{E} \|\mathcal{D}\|_{**} + \frac{b\sqrt{\rho}}{\sqrt{|S|m}} \right) + c' \sqrt{\frac{\log(2/\delta)}{2|S|m}},$$

where $c' = c + 1$, $\|\cdot\|_{**}$ is the dual norm of $\|\cdot\|_*$, $\rho := \frac{1}{|S|} \sum_{(p,q) \in S} \frac{m_{pq}}{m}$. The tensor $\mathcal{D} \in \mathbb{R}^{d \times P \times Q}$ is defined as the sum $\mathcal{D} = \sum_{(p,q) \in S} \sum_{i=1}^{m_{pq}} \mathcal{Z}^{ipq}$, where $\mathcal{Z}^{ipq} \in \mathbb{R}^{d \times P \times Q}$ is defined as

$$(p', q') \text{th fiber of } \mathcal{Z}^{ipq} = \begin{cases} \frac{1}{m_{pq}} \sigma_{ipq} \mathbf{x}_{ipq}, & \text{if } p = p' \text{ and } q = q', \\ 0, & \text{otherwise.} \end{cases}$$

Here $\sigma_{ipq} \in \{-1, +1\}$ are Rademacher random variables and the expectation in the above inequality is with respect to σ_{ipq} , the random draw of tasks S , and the training samples $(\mathbf{x}_{ipq}, y_{ipq})_{i=1}^{m_{pq}}$.

Proof. The proof follows a standard argument, which can be found in Bartlett and Mendelson (Bartlett and Mendelson, 2002, Theorem 8).

$$\begin{aligned} L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) &\leq \left(L(\hat{\mathcal{W}}) - \hat{L}(\hat{\mathcal{W}}) \right) + \left(\hat{L}(\hat{\mathcal{W}}) - \hat{L}(\mathcal{W}^*) \right) + \left(\hat{L}(\mathcal{W}^*) - L(\mathcal{W}^*) \right) \\ &\leq \sup_{\|\mathcal{W}\|_* \leq B_0} \left(L(\mathcal{W}) - \hat{L}(\mathcal{W}) \right) + \sqrt{\frac{\log(2/\delta)}{2\rho|S|m}} \quad (\text{w/ probability} \\ &\hspace{15em} \text{at least } 1 - \delta/2) \\ &\leq R(\ell \circ \mathcal{L}_{B_0}) + \left(c + \frac{1}{\sqrt{\rho}} \right) \sqrt{\frac{\log(2/\delta)}{2|S|m}} \quad (\text{w/ probability} \\ &\hspace{15em} \text{at least } 1 - \delta), \end{aligned}$$

where

$$R(\ell \circ \mathcal{L}_{B_0}) := \mathbb{E} \sup_{\|\mathcal{W}\|_* \leq B_0} \frac{2}{|S|} \sum_{(p,q) \in S} \frac{1}{m_{pq}} \sum_{i=1}^{m_{pq}} \sigma_{ipq} \ell(\langle \mathbf{x}_{ipq}, \mathbf{w}_{pq} \rangle - y_{ipq}).$$

In the third line, we used McDiarmid's inequality and introduced Rademacher random variables $\sigma_{ipq} \in \{-1, +1\}$; the expectation is over both the Rademacher random variables and the training samples $(\mathbf{x}_{ipq}, y_{ipq})$. Using the fact that $c + 1/\sqrt{\rho} \leq c + 1 =: c'$, the last term can be upper bounded by the last term in the statement.

We further analyze the first term. Using the Lipschitz continuity of ℓ and the bound on $|y_{ipq}|$, we have

$$R(\ell \circ \mathcal{L}_{B_0}) \leq 2\Lambda \left(R(\mathcal{L}_{B_0}) + \frac{b\sqrt{\sum_{(p,q) \in S} m_{p,q}}}{|S|m} \right),$$

where

$$R(\mathcal{L}_B) = \frac{2}{|S|} \mathbb{E} \sup_{\|\mathcal{W}\|_* \leq B_0} \sum_{(p,q) \in S} \frac{1}{m_{pq}} \sum_{i=1}^{m_{pq}} \sigma_{ipq} \langle \mathbf{x}_{ipq}, \mathbf{w}_{pq} \rangle.$$

Finally, using the definition of \mathcal{D} and Hölder's inequality, we have

$$R(\mathcal{L}_{B_0}) \leq \frac{2B_0}{|S|} \mathbb{E} \|\mathcal{D}\|_{*^*},$$

which concludes the proof. \square

The next theorem computes the expected dual norm $\mathbb{E} \|\mathcal{D}\|_{*^*}$ for the three norms for tensors.

Theorem 3.2. *We assume that $\mathbf{C}_{pq} := \mathbb{E}[\mathbf{x}_{ipq}\mathbf{x}_{ipq}^\top] \preceq \frac{\kappa}{d}\mathbf{I}_d$ and there is a constant $R > 0$ such that $\|\mathbf{x}_{ipq}\| \leq R$ almost surely. Let us define*

$$D_1 := d + PQ, \quad D_2 := P + dQ, \quad D_3 := Q + dP.$$

In order to simplify the presentation, we assume that $\max_k D_k \geq 3$ and $dPQ \geq \max(d^2, P^2, Q^2)$. For the overlapped trace norm, the latent trace norm, and the

scaled latent trace norm, the expectation $\mathbb{E} \|\mathcal{D}\|_{*^*}$ can be bounded as follows:

$$\frac{1}{|S|} \mathbb{E} \|\mathcal{D}\|_{\text{overlap}^*} \leq C \min_k \left(\sqrt{\frac{\kappa}{m|S|dPQ}} D_k \log D_k + \frac{R}{m|S|} \log D_k \right), \quad (3.8)$$

$$\frac{1}{|S|} \mathbb{E} \|\mathcal{D}\|_{\text{latent}^*} \leq C' \left(\sqrt{\frac{\kappa}{m|S|dPQ}} \max_k (D_k \log D_k) + \frac{R}{m|S|} \log(\max_k D_k) \right), \quad (3.9)$$

$$\frac{1}{|S|} \mathbb{E} \|\mathcal{D}\|_{\text{scaled}^*} \leq C'' \left(\sqrt{\frac{\kappa}{m|S|}} \log(\max_k D_k) + \frac{R \sqrt{\max_k n_k}}{m|S|} \log(\max_k D_k) \right), \quad (3.10)$$

where C, C', C'' are constants, $n_1 = d, n_2 = P$, and $n_3 = Q$. Furthermore, if $m|S| \geq R^2(\max_k n_k) \log(\max_k D_k)/\kappa$, the $O(1/m|S|)$ terms in the above inequalities can be dropped.

Proof of inequality (3.8): From Tomioka et al. (Tomioka et al., 2011b, Lemma 1), we have

$$\|\mathcal{D}\|_{\text{overlap}^*} = \inf_{\mathcal{D}^{(1)} + \mathcal{D}^{(2)} + \mathcal{D}^{(3)} = \mathcal{D}} \max_k \|\mathbf{D}_{(k)}^{(k)}\|_{\text{op}},$$

where the infimum is over three tensors $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$, and $\mathcal{D}^{(3)}$ that sum to the original tensor \mathcal{D} , and $\|\cdot\|_{\text{op}}$ is the operator norm (maximal singular value). Since we can take any $\mathcal{D}^{(k)}$ to equal \mathcal{D} , the norm can be upper bounded as follows:

$$\|\mathcal{D}\|_{\text{overlap}^*} \leq \min_k \|\mathbf{D}_{(k)}\|_{\text{op}}.$$

Since the expectation of minimum over k can be upper bounded by the minimum of expectations, we have

$$\mathbb{E} \|\mathcal{D}\|_{\text{overlap}^*} \leq \mathbb{E} \min_k \|\mathbf{D}_{(k)}\|_{\text{op}} \leq \min_k \mathbb{E} \|\mathbf{D}_{(k)}\|_{\text{op}}.$$

Now we upper bound each expectation using Theorem 6.1 in Tropp (Tropp, 2012, see also Remarks 6.3 and 6.5), which states that

$$\Pr \left\{ \|\mathbf{D}_{(k)}\|_{\text{op}} \geq t \right\} \leq \begin{cases} D_k \exp(-3t^2/8\sigma_k^2), & \text{for } t \leq \sigma_k^2/R_k, \\ D_k \exp(-3t/8R_k), & \text{for } t \geq \sigma_k^2/R_k, \end{cases} \quad (3.11)$$

and

$$\mathbb{E}\|\mathbf{D}_{(k)}\|_{\text{op}} \leq C(\sigma_k \sqrt{\log D_k} + R_k \log D_k), \quad (3.12)$$

where C is an absolute constant, and

$$\sigma_k^2 := \max \left(\left\| \sum_{(p,q) \in S} \sum_{i=1}^{m_{pq}} \mathbb{E} \left[\mathbf{Z}_{(k)}^{ipq} \left(\mathbf{Z}_{(k)}^{ipq} \right)^\top \right] \right\|_{\text{op}}, \left\| \sum_{(p,q) \in S} \sum_{i=1}^{m_{pq}} \mathbb{E} \left[\left(\mathbf{Z}_{(k)}^{ipq} \right)^\top \mathbf{Z}_{(k)}^{ipq} \right] \right\|_{\text{op}} \right),$$

$$R_k \geq \left\| \mathbf{Z}_{(k)}^{ipq} \right\|_{\text{op}} \quad (\text{almost surely}).$$

Due to our assumption $\|\mathbf{x}_{ipq}\| \leq R$, we can take $R_k = R/m$. Thus the remaining task is to compute σ_k^2 for $k = 1, 2, 3$.

First for $k = 1$, the unfolding $\mathbf{Z}_{(1)}^{ipq}$ is a $d \times PQ$ matrix that contains $\sigma_{ipq} \mathbf{x}_{ipq}/m_{pq}$ in the column specified by (p, q) . Therefore, using $m_{pq} \geq m$ and $\|\mathbf{C}_{pq}\| \leq \kappa/d$, we obtain

$$\sum_{i=1}^{m_{pq}} \mathbb{E} \left[\mathbf{Z}_{(1)}^{ipq} \left(\mathbf{Z}_{(1)}^{ipq} \right)^\top \right] = \frac{1}{m_{pq}} \mathbf{C}_{pq} \preceq \frac{\kappa}{md} \mathbf{I}_d,$$

from which we have

$$\left\| \sum_{(p,q) \in S} \sum_{i=1}^{m_{pq}} \mathbb{E} \left[\mathbf{Z}_{(1)}^{ipq} \left(\mathbf{Z}_{(1)}^{ipq} \right)^\top \right] \right\|_{\text{op}} \leq \frac{\kappa|S|}{md}. \quad (3.13)$$

Similarly, since the choice of (p, q) is uniform over $[P] \times [Q]$, we have

$$\sum_{i=1}^{m_{pq}} \mathbb{E} \left[\left(\mathbf{Z}_{(1)}^{ipq} \right)^\top \mathbf{Z}_{(1)}^{ipq} \right] = \frac{1}{PQ} \text{diag} \left(\frac{\text{Tr} \mathbf{C}_{pq}}{m_{pq}} \right) \preceq \frac{\kappa}{mPQ} \mathbf{I}_{PQ},$$

from which we have

$$\left\| \sum_{(p,q) \in S} \sum_{i=1}^{m_{pq}} \mathbb{E} \left[\left(\mathbf{Z}_{(1)}^{ipq} \right)^\top \mathbf{Z}_{(1)}^{ipq} \right] \right\|_{\text{op}} \leq \frac{\kappa|S|}{mPQ}. \quad (3.14)$$

Substituting inequalities (3.13) and (3.14) into (3.12), we have

$$\mathbb{E}\|\mathbf{D}_{(1)}\|_{\text{op}} \leq C \left(\sqrt{\frac{\kappa|S|}{mdPQ}} D_1 \log D_1 + \frac{R}{m} \log D_1 \right).$$

Following a similar line of argument, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{D}_{(2)}\|_{\text{op}} &\leq C \left(\sqrt{\frac{\kappa|S|}{mdPQ} D_2 \log D_2} + \frac{R}{m} \log D_2, \right), \\ \mathbb{E}\|\mathbf{D}_{(3)}\|_{\text{op}} &\leq C \left(\sqrt{\frac{\kappa|S|}{mdPQ} D_3 \log D_3} + \frac{R}{m} \log D_3, \right).\end{aligned}$$

Taking the minimum over k and dividing by $|S|$, we obtain inequality (3.8). \square

Note that the assumption that the norm of \mathbf{x}_{ipq} is bounded is natural because the target y_{ipq} is also bounded. The parameter κ in the assumption $\mathbf{C}_{pq} \preceq \kappa/d\mathbf{I}_d$ controls the amount of correlation in the data. Since $\text{Tr}(\mathbf{C}) = \mathbb{E}\|\mathbf{x}_{ipq}\|^2 \leq R^2$, we have $\kappa = O(1)$ when the features are uncorrelated; on the other hand, we have $\kappa = O(d)$, if they lie in a one dimensional subspace. The number of samples $m|S| = \tilde{O}(\max_k n_k)$ is enough to drop the $O(1/m|S|)$ term even if $\kappa = O(1)$.

Now we state the consequences of Theorem 3.2 for the three norms for tensors. The common assumptions are the same as in Lemma 3.1 and Theorem 3.2. We also assume $m|S| \geq R^2(\max_k n_k) \log(\max_k D_k)/\kappa$ to drop the $O(1/m|S|)$ terms. Let \mathcal{W}^* be any $d \times P \times Q$ tensor with multilinear-rank (r_1, r_2, r_3) and bounded element-wise as $\|\mathcal{W}^*\|_{\ell_\infty} \leq B$.

Corollary 3.3 (Overlapped trace norm). *With probability at least $1 - \delta$, any minimizer of (3.7) with $\|\mathcal{W}\|_{\text{overlap}} \leq B\sqrt{\|\mathbf{r}\|_{1/2}dPQ}$ satisfies the following inequality:*

$$\begin{aligned}L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) &\leq c_1\Lambda B \sqrt{\frac{\kappa}{m|S|} \|\mathbf{r}\|_{1/2} \min_k (D_k \log D_k)} + c_2\Lambda b \sqrt{\frac{\rho}{m|S|}} \\ &\quad + c_3 \sqrt{\frac{\log(2/\delta)}{m|S|}},\end{aligned}$$

where $\|\mathbf{r}\|_{1/2} = (\sum_{k=1}^3 \sqrt{r_k}/3)^2$ and c_1, c_2, c_3 are constants.

Note that Tomioka et al. (Tomioka et al., 2011b) obtained a bound that depends on $(\sum_{k=1}^3 \sqrt{D_k}/3)^2$ instead of $\min(D_k \log D_k)$. Although the minimum may look better than the average, our bound has the worse constant $K = 3$ hidden in c_1 . The $\log D_k$ factor allows us to apply the above result to the setting of tensor completion as we show below.

Corollary 3.4 (Latent trace norm). *With probability at least $1 - \delta$, any minimizer of (3.7) with $\|\mathcal{W}\|_{\text{latent}} \leq B\sqrt{\min_k r_k d P Q}$ satisfies the following inequality:*

$$L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) \leq c'_1 \Lambda B \sqrt{\frac{\kappa}{m|S|} \min_k r_k \max_k (D_k \log D_k)} + c_2 \Lambda b \sqrt{\frac{\rho}{m|S|}} \\ + c_3 \sqrt{\frac{\log(2/\delta)}{m|S|}},$$

where c'_1, c_2, c_3 are constants.

Proof of inequality (3.9): From Tomioka et al. (Tomioka and Suzuki, 2013, Lemma 1), we know that

$$\|\mathcal{D}\|_{\text{latent}^*} = \max_k \|\mathbf{D}_{(k)}\|_{\text{op}}.$$

Combining inequality (3.11) with a union bound, we have

$$\Pr \{ \|\mathcal{D}\|_{\text{latent}^*} \geq t \} \leq 3(\max_k D_k) \max \left(\exp \left(-\frac{3t^2}{8 \max_k \sigma_k^2} \right), \exp \left(-\frac{3t}{8 \max_k R_k} \right) \right),$$

from which we have

$$\mathbb{E} \|\mathcal{D}\|_{\text{latent}^*} \leq C \left(\max_k \sigma_k \sqrt{\log(\max_k D_k) + \log 3} + \max_k R_k (\log(\max_k D_k) + \log 3) \right) \\ (3.15) \\ \leq C' \left(\max_k \sigma_k \sqrt{\log(\max_k D_k)} + \frac{R}{m} \log(\max_k D_k) \right).$$

Here we used $R_k = R/m$ and the simplifying assumption that $\max_k D_k \geq 3$ in the second inequality. Finally, using $\sigma_k \leq \sqrt{\kappa|S|D_k/(mdPQ)}$ as in the proof of inequality (3.8), we obtain inequality (3.9).

Corollary 3.5 (Scaled latent trace norm). *With probability at least $1 - \delta$, any minimizer of (3.7) with $\|\mathcal{W}\|_{\text{scaled}} \leq B\sqrt{\min_k (r_k/n_k) d P Q}$ satisfies the following inequality:*

$$L(\hat{\mathcal{W}}) - L(\mathcal{W}^*) \leq c''_1 \Lambda B \sqrt{\frac{\kappa}{m|S|} \min_k \left(\frac{r_k}{n_k} \right) d P Q \log(\max_k D_k)} + c_2 \Lambda b \sqrt{\frac{\rho}{m|S|}} \\ + c_3 \sqrt{\frac{\log(2/\delta)}{m|S|}},$$

where $n_1 = d, n_2 = P, n_3 = Q$, and c''_1, c_2, c_3 are constants.

Proof of inequality (3.10): Following the proof of (Tomioka and Suzuki, 2013, Lemma 1), we have

$$\|\mathcal{D}\|_{\text{scaled}^*} = \max_k \sqrt{n_k} \|\mathbf{D}_{(k)}\|_{\text{op}},$$

where $n_1 = d$, $n_2 = P$, and $n_3 = Q$. Thus, replacing σ_k and R_k with $\sqrt{n_k}\sigma_k$ and $\sqrt{n_k}R/m$ in inequality (3.15), respectively, we have

$$\mathbb{E} \|\mathcal{D}\|_{\text{scaled}^*} \leq C' \left(\max_k (\sqrt{n_k}\sigma_k) \sqrt{\log(\max_k D_k)} + \frac{R\sqrt{\max_k n_k}}{m} \log(\max_k D_k) \right).$$

Finally, since $n_k D_k = n_k^2 + dPQ \leq 2dPQ$, we have

$$\sqrt{n_k}\sigma_k \leq \sqrt{\frac{\kappa|S|n_k D_k}{mdPQ}} \leq \sqrt{\frac{2\kappa|S|}{m}},$$

which gives inequality (3.10).

The last claim of the theorem is true, because $m|S| \geq R^2(\max_k n_k)(\log_k D_k)/\kappa$ implies

$$m|S| \geq \frac{R^2}{\kappa} \frac{dPQ}{n_k^2 + dPQ} n_k \log D_k = \frac{R^2}{\kappa} \frac{dPQ}{D_k} \log D_k,$$

which gives

$$\sqrt{\frac{\kappa}{m|S|dPQ} D_k \log D_k} \geq \frac{R}{m|S|} \log D_k.$$

We summarize the implications of the above corollaries for different settings in Table 3.1 and Table 3.2. We almost recover the settings for matrix completion (Foygel and Srebro, 2011) and multitask learning (MTL) (Maurer and Pontil, 2013). Note that these simpler problems sometimes disguise themselves as the more general tensor completion or multilinear multitask learning problems. Therefore it is important that the new tensor based norms adapts to the simplicity of the problems in these cases.

Matrix completion is when $d = \kappa = m = r_1 = 1$, and we assume that $r_2 = r_3 = r < P, Q$. The sample complexities are the number of samples $|S|$ that we need to make the leading term in Corollaries 3.3, 3.4, and 3.5 equal ϵ . We can see that the overlapped trace norm and the scaled latent trace norm recover the

known result for matrix completion (Foygel and Srebro, 2011). The plain latent trace norm requires $O(PQ)$ samples because it recognizes the first mode as the mode with the lowest rank 1. Although the rank r of the last two modes are low *relative to their dimensions*, the latent trace norm fails to recognize this. Note that $\|\mathbf{r}\|_{1/2} \leq r$. This is not a contradiction, because in Cor. 3.3, we assume that the overlapped trace norm is bounded, which may or may not be true for matrix completion. In fact, in this case, the overlapped trace norm is an Elastic-net-type regularizer (trace norm + Frobenius norm).

In multitask learning (MTL), we set $P = T$ (the number of tasks) and $Q = 1$. The first and the second modes have a low rank r . We also assume that all the pairs (p, q) are observed ($|S| = T$) as in (Maurer and Pontil, 2013). The sample complexities are defined the same way as above with respect to the number of samples m because $|S|$ is fixed. Our bound for the overlapped trace norm is almost as good as the one in (Maurer and Pontil, 2013) but has an multiplicative $\log(d + T)$ factor (as oppose to their additive $\log(mT)$ term). Also note that the results in (Maurer and Pontil, 2013) can be applied when d is much larger than T . Turning back to our bounds, the scaled latent trace norm *performs as well as knowing the mode with the lowest rank* (the first and the second modes; see also (Tomioka and Suzuki, 2013)). However, similarly to the matrix completion case above, the plain latent trace norm fails to recognize the low-rank-ness of the first two modes, and requires $O(d)$ samples, because the third mode has the lowest rank.

In multilinear multitask learning (MLMTL) (Romera-Paredes et al., 2013), any mode could possibly be low rank but it is a priori unknown. The sample complexities are defined the same way as above with respect to $m|S|$. The homogeneous case is when $d = P = Q$. The heterogeneous case is when the first mode or the third mode is low rank but $P \leq r < d$. Similarly to the above two settings, the overlapped trace norm has a mild dependence on the dimensions but a higher dependence on the rank $\|\mathbf{r}\|_{1/2} \geq \min_k r_k$. The latent trace norm performs as well as knowing the mode that has the lowest rank in the homogeneous case. However, it fails to recognize the mode with the lowest rank relative to its dimension. The scaled latent trace norm does this and although it has a higher logarithmic dependence, it is competitive in both cases.

Table 3.1: Sample complexities of matrix trace norm in various settings. The common factor $1/\epsilon^2$ is omitted from the sample complexities. The sample complexities are defined with respect to $|S|$ for matrix completion and m for multitask learning.







	(n_1, n_2, n_3)	(r_1, r_2, r_3)	(κ, B, S)	Sample complexities (per $1/\epsilon^2$)		
				Overlap	Latent	Scaled
Matrix completion Foygel and Srebro (2011)		$(1, r, r)$	$(1, 1, S)$	$\ \mathbf{r}\ _{1/2}(P+Q)$ $\log(P+Q)$	$PQ \log(PQ)$	$r(P+Q) \log(PQ)$
MTL Maurer and Pontil (2013) (homogeneous case)		(r, d, d)	$(d, 1/\sqrt{d}, d^2)$	$\ \mathbf{r}\ _{1/2} \log(d^2)$	$r \log(d^2)$	$r \log(d^2)$
MTL (heterogeneous case)		(r, P, r')	$(d, 1/\sqrt{d}, PQ)$	$\ \mathbf{r}\ _{1/2} \log(PQ)$	$d \log(dQ)$	$r \log(dQ)$

Table 3.2: Sample complexities of the overlapped trace norm, latent trace norm, and the scaled latent trace norm in various settings. The common factor $1/\epsilon^2$ is omitted from the sample complexities. The sample complexities are defined with respect to $m|S|$ for tensor completion and multilinear multitask learning. In the heterogeneous cases, we assume $P \leq r < r'$. We define $\|\mathbf{r}\|_{1/2} = (\sum_{k=1}^3 \sqrt{r_k}/K)^2$ and $N := n_1 n_2 n_3$.

	(n_1, n_2, n_3)	(r_1, r_2, r_3)	(κ, B, S)	Overlap	Latent	Scaled
MLMTL Romera-Paredes et al. (2013) (homogeneous case)		(r_1, r_2, r_3)	$(\kappa, 1, S)$	$\kappa \ \mathbf{r}\ _{1/2} d^2 \log(d^2)$	$\kappa (\min_k r_k) d^2 \log(d^2)$	$\kappa (\min_k r_k) d^2 \log(d^2)$
MLMTL Romera-Paredes et al. (2013) (heterogeneous case)		(r, P, r')	$(\kappa, 1, S)$	$\kappa \ \mathbf{r}\ _{1/2} P Q \log(PQ)$	$\kappa d P Q \log(dQ)$	$\kappa \min(r P Q, d P r') \log(dQ)$
Tensor completion		(r_1, r_2, r_3)	$(1, 1, S)$	$\ \mathbf{r}\ _{1/2} \min_k (D_k \log D_k)$	$\min_k r_k \max_k (D_k \log D_k)$	$\min_k (\frac{r_k}{n_k}) N \log(\max_k D_k)$

Finally, our bounds also hold for tensor completion. Although Tomioka et al. (Tomioka et al., 2011a,b) studied tensor completion algorithms, their analysis assumed that the inputs x_{ipq} are drawn from a Gaussian distribution, which does not hold for tensor completion. Note that in our setting x_{ipq} can be an indicator vector that has one in the j th position uniformly over $1, \dots, d$. In this case, $\kappa = 1$. The sample complexities of different norms with respect to $m|S|$ is shown in the last row of Table 3.2. The sample complexity for the overlapped trace norm is the same as the one in (Tomioka et al., 2011b) with a logarithmic factor. The sample complexities for the latent and scaled latent trace norms are new. Again we can see that while the latent trace norm recognize the mode with the lowest rank, the scaled latent trace norm is able to recognize the mode with the lowest rank relative to its dimension.

3.5 Experimental Results

We conducted several experiments to evaluate performances of tensor based multitask learning setting we have discussed in Section 3.4. In Section 3.5.1, we discuss simulation we conducted using synthetic data sets. In Sections 3.5.2 and 3.5.3, we discuss experiments on two real world data sets, namely the Restaurant data set (Vargas-Govea et al., 2011) and School Effectiveness data set (Bakker and Heskes, 2003; Argyriou et al., 2008). Both of our real world data sets have heterogeneous dimensions (see Figure 3.2) and it is a priori unclear across which mode has the most amount of information sharing.

3.5.1 Synthetic data sets

For simplicity we consider 3-mode tensor based multi-task learning problems for simulations with synthetic data. The true $d \times P \times Q$ tensor W^* was generated by first sampling a $r_1 \times r_2 \times r_3$ core tensor and then multiplying random orthonormal matrix to each of its modes. For each task $(p, q) \in [P] \times [Q]$, we generated training set of m a zero-mean normal distribution with variance 0.1. We used the penalty formulation of (3.7) with the squared loss and selected the regularization parameter λ using two-fold cross validation on the training set from the range 0.01

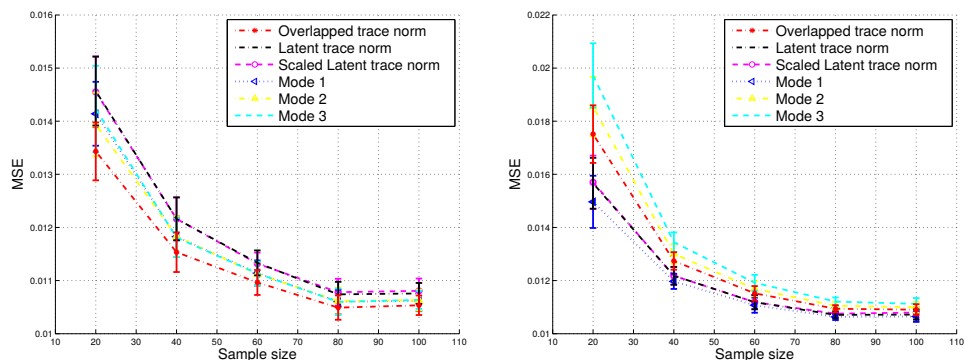
to 10 with the interval 0.1.

In addition to the three norms for tensors we discussed in the previous section, we evaluated the matrix-based multitask learning approaches that penalizes the trace norm of the unfolding of W at specific modes. The conventional convex multitask learning (Argyriou et al., 2006, 2008; Maurer and Pontil, 2013) corresponds to one of these approaches that penalizes the trace norm of the first unfolding $\|W_{(1)}\|_{\text{tr}}$. The convex MLMTL in (Romera-Paredes et al., 2013) corresponds to the overlapped trace norm.

In the first experiment, we chose $d = P = Q = 10$ and $r_1 = r_2 = r_3 = 3$. Therefore, both the dimensions and the multilinear rank are homogeneous. The result is shown in Figure 3.1(a). The overlapped trace norm performed the best, the matrix-based approaches performed next, and the latent trace norm and the scaled latent trace norm were the worst. The scaling of the latent trace norm had no effect because the dimensions were homogeneous. Since the sample complexities for all the methods were the same in this setting (see Table 3.2), the difference in the performances could be explained by a constant factor $K(= 3)$ that is not shown in the sample complexities.

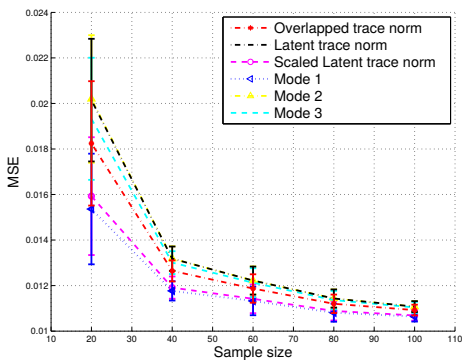
In the second experiment, we chose the dimensions to be homogeneous as $d = P = Q = 10$, but $(r_1, r_2, r_3) = (3, 6, 8)$. The result is shown in Figure 3.1(b). In this setting, the (scaled) latent trace norm and the mode-1 regularization performed the best. The lower the rank of the corresponding mode, the lower were the error of the matrix-based MTL approaches. The overlapped trace norm was somewhat in the middle of the three matrix-based approaches.

In the last experiment, we chose both the dimensions and the multilinear rank to be inhomogeneous as $(d, P, Q) = (10, 3, 10)$ and $(r_1, r_2, r_3) = (3, 3, 8)$. The result is shown in Figure 3.1(c). Clearly the first mode had the lowest rank relative to its dimension. However, the latent trace norm recognizes the second mode as the mode with the lowest rank and performed similarly to the mode-2 regularization. The overlapped trace norm performed better but it was worse than the mode-1 regularization. The scaled latent trace norm performed comparably to the mode-1 regularization.



(a) Synthetic experiment for the case when both the dimensions and the ranks are homogeneous. The true tensor is $10 \times 10 \times 10$ with multilinear rank $(3, 3, 3)$.

(b) Synthetic experiment for the case when the dimensions are homogeneous but the ranks are heterogeneous. The true tensor is $10 \times 10 \times 10$ with multilinear rank $(3, 6, 8)$.



(c) Synthetic experiment for the case when both the dimensions and the ranks are heterogeneous. The true tensor is $10 \times 3 \times 10$ with multilinear rank $(3, 3, 8)$.

Figure 3.1: Results for the synthetic data sets.

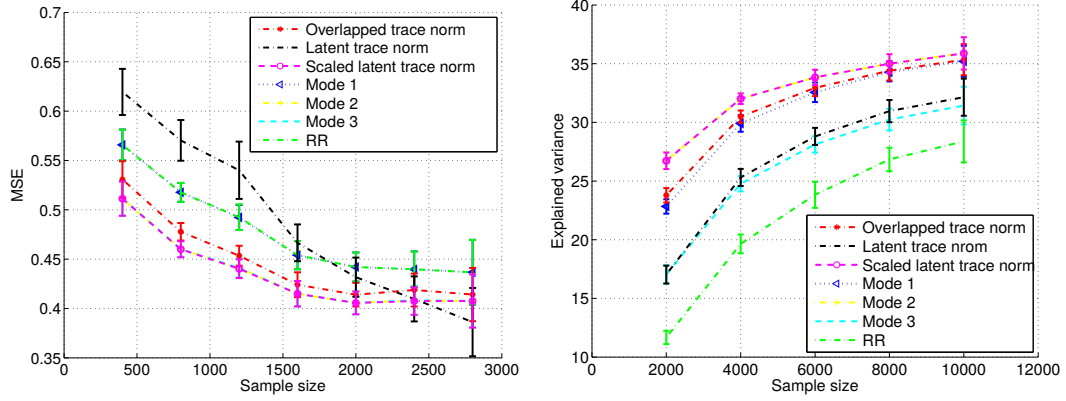
3.5.2 Restaurant data set

The Restaurant data set contains data for a recommendation system for restaurants where different customers have given ratings to different aspects of each restaurant. Three different ratings are given as general rating, rating of food and rating of service with ratings numerically varying from 0,1 and 2. Features that are considered for learning are 45 features of restaurants such as its geographical location, cuisine type and price bands. Following the same approach as in (Romera-Paredes et al., 2013) we modelled the problem as a MLMTL problem with $d = 45$ features, $P = 3$ aspects, and $Q = 138$ customers.

The total number of instances for all the tasks were 3483 and we randomly selected training set of sizes 400, 800, 1200, 1600, 2000, 2400, and 2800. When the size was small many tasks contained no training example. We also selected 250 instances as the validation set and the rest was used as the test set. The regularization parameter for each norm was selected by minimizing the mean squared error on the validation set from the candidate values in the interval [50, 1000] for the overlapped, [0.5, 40] for the latent, [6000, 20000] for the scaled latent norms, respectively.

We also evaluated matrix-based MTL approaches on different modes and ridge regression (Frobenius norm regularization; abbreviated as RR) as baselines. The convex MLMTL in (Romera-Paredes et al., 2013) corresponds to the overlapped trace norm.

The result is shown in Figure 3.2(a). We found the multilinear rank of the solution obtained by the overlapped trace norm to be typically (1, 3, 3). This was consistent with the fact that the performances of the mode-1 regularization and the ridge regression were equal. In other words, the effective dimension of the first mode (features) was one instead of 45. The latent trace norm recognized the first mode as the mode with the lowest rank and it failed to take advantage of the low-rank-ness of the second and the third modes. The scaled latent trace norm was able to perform the best matching with the performances of mode-2 and mode-3 regularization. When the number of samples was above 2400, the latent trace norm caught up with other methods, probably because the effective dimension became higher in this regime.



(a) Result for the $45 \times 3 \times 138$ Restaurant data set. (b) Result for the $24 \times 139 \times 3$ School data set.

Figure 3.2: Results for the real world data sets.

3.5.3 School data set

The data set comes from the inner London Education Authority (ILEA) consisting of examination records from 15362 students at 139 schools in years 1985, 1986, and 1987. We followed (Bakker and Heskes, 2003) for the preprocessing of categorical attributes and obtained 24 features. Previously Argyriou et al. (Argyriou et al., 2008) modeled this data set as a 27×139 matrix-based MTL problem in which the year was modeled as a trinomial attributes. Instead here we model this data set as a $24 \times 139 \times 3$ MLMTL problem in which the third mode corresponds to the year. Following earlier papers, (Bakker and Heskes, 2003; Argyriou et al., 2008), we used the percentage of explained variance, defined as $100 \cdot (1 - (\text{test MSE})/(\text{variance of } y))$ as the evaluation metric.

The results are shown in Figure 3.2(b). First, ridge regression performed the worst because it was not able to take advantage of the low-rank-ness of any mode. Second, the plain latent trace norm performed similarly to the mode-3 regularization probably because the dimension 3 was lower than the rank of the other two modes. Clearly the scaled latent trace norm performed the best matching with the performance of the mode-2 regularization; probably the second mode had the most redundancy. The performance of the overlapped trace norm was comparable or slightly better than the mode-1 regularization. The percentage of the explained variance of the latent trace norm exceeds 30 % around sample size 4000 (around

30 samples per school), which is higher than the Hierarchical Bayes (Bakker and Heskes, 2003) (around 29.5 %) and matrix-based MTL (Argyriou et al., 2008) (around 26.7 %) that used around 80 samples per school.

3.6 Conclusion

In this chapter we extended multilinear multitask learning with the latent trace norm and the scaled latent trace norm. We derived optimisation methods and excess risk bounds for multilinear multitask learning. Through experiments we show that our theory agrees with our experimental results. Most importantly we show that the scaled latent trace norm is best suited for multilinear multitask learning due to their "flat" structure of tensors. In addition to multilinear multitask learning we have derived bounds for matrix multitask learning and tensor completion.

Chapter 4

Conclusions and Future Work

In this thesis we studied tensor norms and applications of tensor norm regularization in machine learning problems. Our most important contribution is the proposal of the scaled latent trace norm which is capable of obtaining better performances when regularizing tensors with high variations in both multilinear ranks and mode dimensions. In this thesis we extensively studied tensor based regression and classification covering optimisation methods, theoretical analysis and experiments. Similarly we have explored tensor based multilinear multitask learning extending previous research (Romera-Paredes et al., 2013) by applying latent trace norm and scaled latent trace norm regularizations with discussion on theoretical analysis and experiments. Learning with tensors is still in its early stages and we believe that research in this thesis opens many interesting future research directions.

An important future direction is to find domains where tensor based learning can be applied. In Chapter 3, we saw that image sequences classification and BCI data classification are well suited for tensor based supervised learning. One of the interesting domains is spatio-temporal data (Benetos and Kotropoulos, 2010) which naturally are tensor formatted data. Similarly to BCI data, data that are in frequency domains such as video and audio data (Bahadori et al., 2014) could also be good applications. When considering multilinear multitask learning popular recommendation systems (Karatzoglou et al., 2010) and web advertising (Ahmed et al., 2012) can be considered since relationships among multiple users, multiple products and multiple actions (click, buy, recommend) would require higher order

tensor structures.

One major observation that we can derive from our work is that no single tensor norm is superior in performance and our theoretical analysis and experimental results show that different tensor norms give better performances depending on the multilinear rank and the structure of the tensor that is regularized. We also observed that unfolded matrices of tensors can lead to better performances comparable to tensor norms provided that the unfolded matrix has the lowest rank. Though we experimented unfolded matrix regularization with cross validation in Chapter 3 it was not reliable with classification since cross validation could fail when using a binary loss function. Recently it has been established that better sample complexities in tensor completion of higher order tensors with number of modes above three can be achieved by square reshaping of tensors to create a balanced matrix (Mu et al., 2013). These factors can be considered when designing improved tensor norms in the future.

There are several further theoretical analysis possible by extending our research. One of the open theoretical questions that have not been addressed in this thesis is why the low rank regularization can be better than the vector based norms. As we have seen in the simulation results with tensor regression it is evident that when the number of training samples is small the low rank tensor regularizations lead to higher accuracies compared to vector based regularizations. However, this is a difficult problem to analyse theoretically using the excess risk bounds used in this thesis. One approach to solve this could be to formulate oracle inequalities (Gaiffas and Lecue, 2011) for tensor and vector norm regularizations. We may prove better bounds on regularizations using combinations of tensor low rank regularizations and vector based regularizations (l_2, l_1) and show that usage of both these norms together could lead to better accuracies. Such analysis would imply that under different conditions such as low rankness and sparseness of data suitable norms must be applied to obtain the best result.

Another major future improvement can be on optimization specially extension of stochastic optimization to tensor regularized learning problems. Many of the existing stochastic optimisation methods such as the regularized dual averaging (RDA) (Xiao, 2010) and the stochastic ADMM (Ouyang et al., 2013) can be directly applied to tensor based regression and classification but they may be

inadequate for speeding up learning since they are only stochastic in terms of selecting data samples. A major drawback in speed could occur in the update steps involving singular values since our implementation requires full singular value decomposition for each mode unfolding. For tensors with large mode dimensions this could be highly computationally expensive and better alternative approaches could be useful when applying to real world implementations. One strategy could be to apply sketching methods and random projections directly to tensors or unfolded matrices to transform them into lower dimensions preserving original ranks (Nguyen et al., 2009). These approaches would lead to designing of new algorithms with two-fold stochastic optimisations which are stochastic on both data and singular value computations capable of giving faster training.

Developing online learning algorithms with tensor norms is another relatively interesting question at least from a theoretical view point. Slightly different from stochastic optimisation, online algorithms such as Perceptron algorithms for vectors (Gentile, 2003) and matrices (Cavallanti et al., 2010) have been designed such that a single training data instance is used only once. As a future work a tensor Perceptron could be developed by exploiting the duality relationship between the overlapped trace norm and the latent trace norm (Tomioka and Suzuki, 2013).

Tensor regularization in the reproducible kernel Hilbert spaces has also been proposed recently (Signoretto et al., 2013a) with applications to multilinear multitask learning and inductive learning settings. In their study the tensor norm that is used for regularization in reproducible kernel Hilbert spaces is the overlapped trace norm and none of the latent trace norms has been considered. Also no theoretical analysis is available for any of the tensor norms in reproducible kernel Hilbert spaces. It is an open research question to understand how the multilinear rank would behave in reproducible kernel Hilbert space relative to its original data domain. Intuitively if the multilinear rank of the tensor based kernel Hilbert space has higher variations, similar approaches such as the latent trace norm and the scaled latent trace norm may be applicable. It would be an interesting future research direction to explore tensor regularization in reproducible kernel Hilbert space both theoretically and experimentally.

Bibliography

Amr Ahmed, Mohamed Aly, Abhimanyu Das, Alexander J. Smola, and Tasos Anastasakos. Web-scale multi-task feature selection for behavioral targeting. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 1737–1741, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398508. URL <http://doi.acm.org/10.1145/2396761.2398508>.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, 2005.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NIPS*, pages 41–48, 2006.

Andreas Argyriou, Charles A. Micchelli, Massimiliano Pontil, and Yiming Ying. A spectral regularization framework for multi-task structure learning. In *NIPS*, 2007.

Andreas Argyriou, Massimiliano Pontil, Yiming Ying, and Charles A. Micchelli. A spectral regularization framework for multi-task structure learning. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Adv. Neural. Inf. Process. Syst. 20*, pages 25–32. Curran Associates, Inc., 2008.

Mohammad Taha Bahadori, Qi (Rose) Yu, and Yan Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3491–3499. Curran Associates, Inc., 2014.

Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *J. Mach. Learn. Res.*, 4:83–99, 2003.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

- Peter L. Bartlett, Michael . Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101: 138–156, 2006.
- Jonathan Baxter. A model of inductive bias learning. *J. Artif. Intell. Res.*, 12: 149–198, 2000.
- E. Benetos and C. Kotropoulos. Non-negative tensor factorization applied to music genre classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):1955–1967, Nov 2010. ISSN 1558-7916. doi: 10.1109/TASL.2010.2040784.
- Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.
- Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011. ISSN 1935-8237. doi: 10.1561/22000000016.
- Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electron. J. Statist.*, 1:169–194, 2007. doi: 10.1214/07-EJS008. URL <http://dx.doi.org/10.1214/07-EJS008>.
- Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20(4): 1956–1982, March 2010. ISSN 1052-6234. doi: 10.1137/080738970. URL <http://dx.doi.org/10.1137/080738970>.
- Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395.
- Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- Giovanni Cavallanti, Nicol Cesa-bianchi, and Manfred Warmuth. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 2010.
- Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.
- Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. 2006.

- Jianhui Chen, Ji Liu, and Jieping Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Trans. Knowl. Discov. Data*, 5(4):22:1–22:31, February 2012. ISSN 1556-4681. doi: 10.1145/2086737.2086742.
- Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011, Budapest, Hungary*, pages 315–340, 2011.
- Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of non-linear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17 – 40, 1976. ISSN 0898-1221. doi: [http://dx.doi.org/10.1016/0898-1221\(76\)90003-1](http://dx.doi.org/10.1016/0898-1221(76)90003-1).
- S. Gaiffas and G. Lecue. Sharp oracle inequalities for high-dimensional matrix prediction. *Information Theory, IEEE Transactions on*, 57(10):6942–6957, Oct 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2136318.
- Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010.
- Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003. ISSN 0885-6125. doi: 10.1023/A:1026319107706.
- Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8.
- Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *J. ACM*, 60(6):45:1–45:39, November 2013. ISSN 0004-5411. doi: 10.1145/2512329. URL <http://doi.acm.org/10.1145/2512329>.
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.
- Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.*, 12:2777–2824, November 2011. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1953048.2078194>.
- Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 793–800. Curran Associates, Inc., 2009.

- Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *J. Mach. Learn. Res.*, 13:1865–1890, June 2012. ISSN 1532-4435.
- Alexandros Karatzoglou, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10*, pages 79–86, 2010. ISBN 978-1-60558-906-0.
- H. A. L. Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14:105–122, 2000b.
- Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 2009.
- Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 663–670, Haifa, Israel, June 2010.
- Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. In *ICCV*, pages 2114–2121, 2009.
- Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor Completion for Estimating Missing Values in Visual Data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):208–220, 2013.
- Andreas Maurer and Massimiliano Pontil. Structured sparsity and generalization. *J. Mach. Learn. Res.*, 13(1):671–690, March 2012. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2503308.2188408>.
- Andreas Maurer and Massimiliano Pontil. Excess risk bounds for multitask learning with trace norm regularization. In *COLT 2013*, 2013.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. An inequality with applications to structured sparsity and multitask dictionary learning. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 440–460, 2014. URL <http://jmlr.org/proceedings/papers/v35/maurer14.html>.
- Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. ISBN 0070428077, 9780070428072.

- Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *CoRR*, abs/1307.5870, 2013.
- Nam H. Nguyen, Thong T. Do, and Trac D. Tran. A fast and efficient algorithm for low-rank approximation of a matrix. In *STOC '09: Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 215–224, New York, NY, USA, 2009. ACM.
- A. Onishi, Anh-Huy Phan, K. Matsuoka, and A. Cichocki. Tensor classification for p300-based brain computer interface. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 581–584, March 2012. doi: 10.1109/ICASSP.2012.6287946.
- Hua Ouyang, Niao He, Long Tran, and Alexander G. Gray. Stochastic alternating direction method of multipliers. In *ICML (1)*, volume 28 of *JMLR Proceedings*, pages 80–88. JMLR.org, 2013.
- Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. doi: 10.1137/070697835.
- Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1444–1452, 2013.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135.
- M. Signoretto, L. De Lathauwer, and J. A. K. Suykens. Learning tensors in reproducing kernel hilbert spaces with multilinear spectral penalties. Technical report, arXiv:1310.4977, 2013a.
- M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. K. Suykens. Learning with tensors: a framework based on convex optimization and spectral regularization. *Mach. Learn.*, 94(3):303–351, 2013b.
- Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.
- R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

- A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed problems*. W.H. Winston, 1977.
- Ryota Tomioka and Kazuyuki Aihara. Classifying matrices with a spectral regularization. In *Proceedings of the 24th International Conference on Machine Learning, ICML '07*, pages 895–902, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273609.
- Ryota Tomioka and Taiji Suzuki. Convex Tensor Decomposition via Structured Schatten Norm Regularization. 2013.
- Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. Estimation of low-rank tensors via convex optimization. Technical report, 2011a.
- Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. Statistical Performance of Convex Tensor Decomposition. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 972–980, 2011b.
- Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, August 2012. ISSN 1615-3375. doi: 10.1007/s10208-011-9099-z. URL <http://dx.doi.org/10.1007/s10208-011-9099-z>.
- L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966c.
- Blanca Vargas-Govea, Gabriel González-Serna, and Rafael Ponce-Medellin. Effects of relevant contextual features in the performance of a restaurant recommender system. In *Proceedings of 3rd Workshop on Context-Aware Recommender Systems*. 2011.
- Kishan Wimalawarne, Masashi Sugiyama, and Ryota Tomioka. Multitask learning meets tensor factorization: task imputation via convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2825–2833. Curran Associates, Inc., 2014.
- Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, December 2010. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1756006.1953017>.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2): 463–483, 2014. ISSN 1467-9868. doi: 10.1111/rssb.12031. URL <http://dx.doi.org/10.1111/rssb.12031>.