LETTER Communicated by Bernardino Romera-Paredes

Theoretical and Experimental Analyses of Tensor-Based Regression and Classification

Kishan Wimalawarne
kishanwn@gmail.com
Department of Computer Science, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8552, Japan

Ryota Tomioka
tomioka@ttic.edu
Toyota Technological Institute at Chicago, Chicago, IL 60637, U.S.A.

Masashi Sugiyama
sugi@k.u-tokyo.ac.jp
Department of Complexity Science and Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan

We theoretically and experimentally investigate tensor-based regression and classification. Our focus is regularization with various tensor norms, including the overlapped trace norm, the latent trace norm, and the scaled latent trace norm. We first give dual optimization methods using the alternating direction method of multipliers, which is computationally efficient when the number of training samples is moderate. We then theoretically derive an excess risk bound for each tensor norm and clarify their behavior. Finally, we perform extensive experiments using simulated and real data and demonstrate the superiority of tensor-based learning methods over vector- and matrix-based learning methods.

1 Introduction

A wide range of real-world data takes the format of matrices and tensors, for example, recommendation (Karatzoglou, Amatriain, Baltrunas, & Oliver, 2010), video sequences (Kim, Wong, & Cipolla, 2007), climates (Bahadori, Yu, & Liu, 2014), genomes (Sankaranarayanan, Schomay, Aiello, & Alter, 2015), and neuroimaging (Zhou, Li, & Zhu, 2013). A naive way to learn from such matrix and tensor data is to vectorize them and apply ordinary regression or classification methods designed for vectorial data. However, such a vectorization approach would lead to loss in structural information of matrices and tensors such as low-rankness.

The objective of this letter is to investigate regression and classification methods that directly handle tensor data without vectorization. Low-rank

Neural Computation **28, 686–715** (2016)
doi:10.1162/NECO_a_00815

structure of data has been successfully utilized in various applications, such as missing data imputation (Cai, Candès, & Shen, 2010), robust principal component analysis (Candès, Li, Ma, & Wright, 2011), and subspace clustering (Liu, Lin, & Yu, 2010). Instead of lowrankness of data itself, in this letter we consider its dual: learning coefficients of a regressor and a classifier. Low-rankness in learning coefficients means that only a subspace of feature space is used for regression and classification.

For matrices, regression and classification have been studied in Tomioka and Aihara (2007) and Zhou and Li (2014) in the context of EEG data analysis. It was experimentally demonstrated that directly learning matrix data by low-rank regularization can significantly improve performance compared to learning after vectorization. Another advantage of using low-rank regularization in the context of EEG data analysis is that analyzing singular value spectra of learning coefficients is useful in understanding activities of brain regions.

More recently, an inductive learning method for tensors has been explored (Signoretto, Dinh, De Lathauwer, & Suykens, 2013). Compared to the matrix case, learning with tensors is inherently more complex. For example, the multilinear ranks of tensors make it more complicated to find a proper low-rankness of a tensor compared to a matrix, which has only one rank. So far, several tensor norms such as the overlapped trace norm or the tensor nuclear norm (Liu, Musialski, Wonka, & Ye, 2009), the latent trace norm (Tomioka & Suzuki, 2013), and the scaled latent trace norm (Wimalawarne, Sugiyama, & Tomioka, 2014) have been proposed and demonstrated to perform well for various tensor structures. However, theoretical analysis of tensor learning in inductive learning settings has not been much investigated yet. Another challenge in inductive tensor learning is efficient optimization strategies, since tensor data often have much higher dimensionalities than matrix and vector data.

We theoretically and experimentally investigate tensor-based regression and classification with regularization by the overlapped trace norm, the latent trace norm, and the scaled latent trace norm. We first provide their dual formulations and propose optimization procedures using the alternating direction method of multipliers (Bertsekas, 1996), which is computationally efficient when the number of data samples is moderate. We then derive an excess risk bound for each tensor regularization, which allows us to theoretically understand the behavior of tensor norm regularization. More specifically, we elucidate that the excess risk of the overlapped trace norm is bounded with the average multilinear ranks of each mode, that of the latent trace norm is bounded with the minimum multilinear rank among all modes, and that of the scaled latent trace norm is bounded with the minimum ratio between multilinear ranks and mode dimensions. Finally, for simulated and real tensor data, we experimentally investigate the behavior of tensor-based regression and classification methods. The experimental results are in concordance with our theoretical findings, and

tensor-based learning methods compare favorably with vector- and matrix-based methods.

The remainder of this letter is organized as follows. In section 2, we formulate the problem of tensor-based supervised learning and review the overlapped trace norm, the latent trace norm, and the scaled latent trace norm. In section 3, we derive dual optimization algorithms based on the alternating direction method of multipliers. In section 4, we theoretically give an excess risk bound for each tensor norm. In section 5, we give experimental results on both artificial and real-world data and illustrate the advantage of tensor-based learning methods. In section 6, we conclude.

Throughout the paper, we use standard tensor notation following Kolda and Bader (2009). We represent a K-way tensor as $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ that consists of $N = \prod_{k=1}^{K} n_k$ elements. A mode-k fiber of $\mathcal{W}$ is an n_k-dimensional vector that can be obtained by fixing all except the kth index. The mode-k unfolding of tensor $\mathcal{W}$ is represented as $W_{(k)} \in \mathbb{R}^{n_k \times N/n_k}$, which is obtained by concatenating all the N/n_k mode-k fibers along its columns. The spectral norm of a matrix X is denoted by $\|X\|_{\mathrm{op}}$, the maximum singular value of X. The operator $\langle \mathcal{W}, \mathcal{X} \rangle$ is the sum of element-wise multiplications of $\mathcal{W}$ and $\mathcal{X}$, that is, $\langle \mathcal{W}, \mathcal{X} \rangle = \mathrm{vec}(\mathcal{W})^\top \mathrm{vec}(\mathcal{X})$. The Frobenius norm of a tensor $\mathcal{X}$ is defined as $\|\mathcal{X}\|_{\mathrm{F}} = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$.

2 Learning with Tensor Regularization

In this section, we put forward inductive tensor learning models with tensor regularization and review different tensor norms used for low-rank regularization.

2.1 Problem Formulation. Our focus in this letter is regression and classification of tensor data. We consider a data set $(\mathcal{X}_i, y_i), i = 1, \ldots, m$, where $\mathcal{X}_i \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is a covariate tensor and $y_i \in \mathbb{R}$ for regression, while $y_i \in \{-1, 1\}$ for classification. We consider the following learning model for a tensor norm $\|\cdot\|_\star$:

$$\min_{\mathcal{W}, b} \sum_{i=1}^{m} l(\mathcal{X}_i, y_i, \mathcal{W}, b) + \lambda \|\mathcal{W}\|_\star, \tag{2.1}$$

where $l(\mathcal{X}_i, y_i, \mathcal{W}, b)$ is the loss function. The squared loss,

$$l(\mathcal{X}_i, y_i, \mathcal{W}, b) = (y_i - (\langle \mathcal{W}, \mathcal{X}_i \rangle + b))^2, \tag{2.2}$$

is used for regression, and the logistic loss,

$$l(\mathcal{X}_i, y_i, \mathcal{W}, b) = \log(1 + \exp(-y_i(\langle \mathcal{W}, \mathcal{X}_i \rangle + b)), \tag{2.3}$$

is used for classification. $b \in \mathbb{R}$ is the bias term, and $\lambda \geq 0$ is the regularization parameter. If $\|\cdot\|_\star = \|\cdot\|_2$ or $\|\cdot\|_1$, then the above problem is equivalent to ordinary vector-based l_2- or l_1-regularization.

To understand the effect of tensor-based regularization, it is important to investigate the low-rankness of tensors. When a matrix $W \in \mathbb{R}^{n_1 \times n_2}$ is being considered, its trace norm is defined as

$$\|W\|_{\text{tr}} = \sum_{j=1}^{J} \sigma_j, \tag{2.4}$$

where σ_j is the j^{th} singular value and J is the number of nonzero singular values ($J \leq \min(n_1, n_2)$). A matrix is called law rank if $J < \min(n_1, n_2)$. The matrix trace norm, equation 2.4 is a convex envelope to the matrix rank and it is commonly used in matrix low-rank approximation (Recht, Fazel, & Parrilo, 2010).

As in matrices, the rank property is also available for tensors, but it is more complicated due to its multidimensional structure. The mode-k rank r_k of a tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is defined as the rank of mode-k unfolding $W_{(k)}$, and the multilinear rank of $\mathcal{W}$ is given as $(r_1, \ldots, r_K)$. The mode-i of a tensor $\mathcal{W}$ is called low rank if $r_i < n_i$.

2.2 Overlapped Trace Norm. One of the earliest definitions of a tensor norm is the tensor nuclear norm (Liu, Musialski, Wonka, & Ye, 2009) or the *overlapped trace norm* (Tomioka & Suzuki, 2013), which can be represented for a tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ as

$$\|\mathcal{W}\|_{\text{overlap}} = \sum_{k=1}^{K} \|W_{(k)}\|_{\text{tr}}. \tag{2.5}$$

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 due to the fact that the same tensor is unfolded over different modes and regularized with the trace norm.

One of the popular applications of the overlapped trace norm is tensor completion (Gandy, Recht, & Yamada, 2011; Liu et al., 2009), where missing entries of a tensor are imputed. Another application is *multilinear multitask learning* (Romera-Paredes, Aung, Bianchi-Berthouze, & Pontil, 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 both tensor completion (Tomioka & Suzuki, 2013) and multilinear multitask learning (Wimalawarne et al., 2014). They have shown that the prediction error of overlapped trace norm regularization is bounded by the average mode-k ranks, 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 poor performance when the multilinear ranks have high variations.

To overcome the weakness of the overlapped trace norm, recent research in tensor norms has led to new norms such as the latent trace norm (Tomioka & Suzuki, 2013) and the scaled latent trace norm (Wimalawarne et al., 2014).

2.3 Latent Trace Norm. Tomioka and Suzuki (2013) proposed the latent trace norm as

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

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.

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.

2.4 Scaled Latent Trace Norm. Recently Wimalawarne et al. (2014) proposed the scaled latent trace norm as an extension of the latent trace norm:

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

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.

The scaled latent norm has the ability to overcome this problem by its scaling with the mode dimensions such that it is able to work with the relative ranks of the tensor. In the context of multilinear multitask learning, it has been shown that the scaled latent trace norm works well for tensors with high variations in multilinear ranks and mode dimensions compared to the overlapped trace norm and the latent trace norm (Wimalawarne et al., 2014).

The inductive learning setting mentioned in equation 2.1 with the overlapped trace norm has been studied previously in Signoretto et al. (2013). However, theoretical analysis and performance comparison with other tensor norms have not been conducted yet. Similarly to tensor decomposition (Tomioka & Suzuki, 2013) and multilinear multitask learning (Wimalawarne et al., 2014), tensor-based regression and classification may also be improved by regularization methods that can work with high variations in multilinear ranks and mode dimensions.

In the following sections, to make tensor-based learning more practical and to improve the performance, we consider formulation 2.1 with the overlapped trace norm the latent trace norm and the scaled latent trace norm and give computationally efficient optimization algorithms and excess risk bounds.

3 Optimization

In this section, we consider the dual formulation for equation 2.1 and propose computationally efficient optimization algorithms. Since optimization of equation 2.1 with regularization using the overlapped trace norm has already been studied in Signoretto et al. (2013), we do not discuss it here. Our main focus in this section is optimization of equation 2.1 with regularization using the latent trace norm and the scaled latent trace norm.

Let us consider the formulation equation 2.1 for a data set $(\mathcal{X}_i, y_i) \in \mathbb{R}^{n_1 \times \cdots \times n_K} \times \mathbb{R}, i = 1, \ldots, m$ with latent and scaled latent trace norm regularization as follows:

$$P(\mathcal{W}, b) = \min_{\mathcal{W}^{(1)} + \ldots + \mathcal{W}^{(K)} = \mathcal{W}, b} \sum_{i=1}^{m} l(\mathcal{X}_i, y_i, \mathcal{W}, b) + \sum_{k=1}^{K} \lambda_k \|W_{(k)}^{(k)}\|_{\mathrm{tr}}, \tag{3.1}$$

where, for $k = 1, \ldots, K$ and for any given regularization parameter λ, $\lambda_k = \lambda$ in the case of the latent trace norm and $\lambda_k = \frac{\lambda}{\sqrt{n_k}}$ in the case of the scaled latent trace norm, respectively. $W_{(k)}^{(k)}$ is the unfolding of $\mathcal{W}^{(k)}$ on its kth mode. It is worth noticing that the application of the latent and scaled latent trace norms requires optimizing over K latent tensors, which contain KN variables in total. For large K and N, solving the primal problem, equation 3.1,

can be computationally expensive, especially in nonlinear problems such as logistic regression, since they require computationally expensive optimization methods such as gradient descent or the Newton method. If the number of training samples m is $m \ll KN$, solving the dual problem of equation 3.1 could be computationally more efficient. For this reason, we focus on optimization in the dual below.

The dual formulation of equation 3.1 can be written as follows (its detailed derivation is given in appendix A):

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \mathcal{V}^{(1)}, \cdots, \mathcal{V}^{(K)}} D(-\boldsymbol{\alpha}) + \sum_{k=1}^{K} \delta_{\lambda_k}(V_{(k)}^{(k)}) \\ &\text{subject to} \quad \mathcal{V}^{(k)} = \sum_{i=1}^{m} \alpha_i \mathcal{X}_i \quad (k = 1, \ldots, K,) \\ &\qquad\qquad \sum_{i=1}^{m} \alpha_i = 0, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^m$ are dual variables corresponding to the training data set $(\mathcal{X}_i, y_i), i = 1, \ldots, m$, $D(-\boldsymbol{\alpha})$ is the conjugate loss function defined as

$$D(-\boldsymbol{\alpha}) = \sum_{i=1}^{m} \frac{1}{2}\alpha_i^2 - \alpha_i y_i$$

in the case of regression with the squared loss (Tomioka, Suzuki, & Sugiyama, 2011), and

$$D(-\boldsymbol{\alpha}) = \sum_{i=1}^{m} y_i\alpha_i \log(y_i\alpha_i) + (1 - y_i\alpha_i)\log(1 - y_i\alpha_i)$$

with constraint $0 \leq y_i\alpha_i \leq 1$ in the case of classification with the logistic loss (Tomioka et al., 2011) and δ_{λ_k} is the indicator function defined as $\delta_{\lambda_k}(V) = 0$ if $\|V\|_{\text{op}} \leq \lambda_k$ and $\delta_{\lambda_k}(V) = \infty$ otherwise. The constraint $\sum_{i=1}^{m} \alpha_i = 0$ is due to the bias term b. Here, the auxiliary variables $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(N)}$ are introduced to remove the coupling between the indicator functions in the objective function (see appendix A for details).

The alternating direction method of multipliers (ADMM) (Gabay & Mercier, 1976; Boyd, Parikh, Chu, Peleato, & Eckstein, 2011) has been previously used to solve primal problems of tensor decomposition (Tomioka, Suzuki, Hayashi, & Kashima, 2011) and multilinear multitask learning (Romera-Paredes et al., 2013) with the overlapped trace norm

regularization. Optimization in the dual for tensor decomposition problems with the latent and scaled latent trace norm regularization has been solved using ADMM in Tomioka, Suzuki, Hayashi et al. (2011). Here, we also adopt ADMM to solve equation 3.2 and describe the formulation and the optimization steps in detail.

With the introduction of dual variables $\mathcal{W}^{(k)} \in \mathbb{R}^{n_1 \times \cdots \times n_K},\ k = 1, \ldots, K$ (corresponding to the primal variables of equation 3.1), $b \in \mathbb{R}$, and parameter $\beta > 0$, the augmented Lagrangian function for equation 3.2 is defined as:

$$\begin{aligned} &L(\boldsymbol{\alpha}, \{\mathcal{V}^{(k)}\}_{k=1}^K, \{\mathcal{W}^{(k)}\}_{k=1}^K, b) \\ &\quad = D(-\boldsymbol{\alpha}) + \sum_{k=1}^K \left(\delta_{\lambda_k}(V_{(k)}^{(k)}) + \left\langle W_{(k)}^{(k)}, \sum_{i=1}^m \alpha_i X_{i(k)} - V_{(k)}^{(k)} \right\rangle \right. \\ &\qquad \left. + \frac{\beta}{2} \left\| \sum_{i=1}^m \alpha_i X_{i(k)} - V_{(k)}^{(k)} \right\|_{\mathrm{F}}^2 \right) + b \sum_{i=1}^m \alpha_i + \frac{\beta}{2} \left\| \sum_{i=1}^m \alpha_i \right\|_{\mathrm{F}}^2. \end{aligned}$$

This ADMM formulation is solved for variables $\boldsymbol{\alpha}$, $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(k)}$, $\mathcal{W}^{(1)}, \ldots, \mathcal{W}^{(k)}$, and b by considering subproblems for each variable. Below, we give the solution for each variable at iterative step $t + 1$.

The first subproblem to solve is for $\boldsymbol{\alpha}$ at step $t + 1$:

$$\boldsymbol{\alpha}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \{\mathcal{V}^{(k)t}\}_{k=1}^K, \{\mathcal{W}^{(k)t}\}_{k=1}^K, b^t),$$

where $\{\mathcal{V}^{(k)t}\}_{k=1}^K$, $\{\mathcal{W}^{(k)t}\}_{k=1}^K$, and b^t are the solutions obtained at step t.

Depending on the conjugate loss $D(-\boldsymbol{\alpha})$, the solution for $\boldsymbol{\alpha}$ differs. In the case of regression with the squared loss, equation 2.2, the augmented Lagrangian can be minimized with respect to $\boldsymbol{\alpha}$ by solving the following linear equation:

$$(K\beta \bar{X}\bar{X}^\top + I + \beta \mathbf{1}_m \mathbf{1}_m^\top)\boldsymbol{\alpha}^{t+1} = (\mathbf{y} - \bar{X}\mathrm{vec}(\bar{\mathcal{W}}^t) + \beta \bar{X}\mathrm{vec}(\bar{\mathcal{V}}^t) - \mathbf{1}_m b^t),$$

where $\bar{X} = [\mathrm{vec}(\mathcal{X}_1)^\top; \cdots; \mathrm{vec}(\mathcal{X}_m)^\top] \in \mathbb{R}^{m \times N}$, $\bar{\mathcal{V}}^t = \sum_{k=1}^K \mathcal{V}^{(k)t}$, $\bar{\mathcal{W}}^t = \sum_{k=1}^K \mathcal{W}^{(k)t}$, $\mathbf{y} = (y_1, \ldots, y_m)^\top$, and $\mathbf{1}_m$ is the m-dimensional vector of all ones. Note that in the above system of equations, the coefficient matrix multiplied with $\boldsymbol{\alpha}$ does not change during optimization. Thus, it can be efficiently solved at each iteration by precomputing the Cholesky factorization of the matrix.

For classification with the logistic loss, equation 2.3, the Newton method is used to find the solution for $\boldsymbol{\alpha}^{t+1}$, which requires the gradient and the

Hessian of $L(\boldsymbol{\alpha}, \{\mathcal{V}^{(k)}\}_{k=1}^K, \{\mathcal{W}^{(k)}\}_{k=1}^K, b)$:

$$\frac{\partial L(\boldsymbol{\alpha}, \{\mathcal{V}^{(k)}\}_{k=1}^K, \{\mathcal{W}^{(k)}\}_{k=1}^K, b)}{\partial \alpha_i}$$

$$= y_i \log\left(\frac{y_i\alpha_i}{1-y_i\alpha_i}\right) + \sum_{k=1}^K \langle \mathcal{W}^{(k)t}, \mathcal{X}_i\rangle$$

$$+ \beta \sum_{k=1}^K \left\langle \mathcal{X}_i, \sum_{i=1}^m \mathcal{X}_i \alpha_i^{t+1} - \mathcal{V}^{(k)t} \right\rangle + b + \beta \sum_{i=1}^m \alpha_i,$$

$$\frac{\partial^2 L(\boldsymbol{\alpha}, \{\mathcal{V}^{(k)}\}_{k=1}^K, \{\mathcal{W}^{(k)}\}_{k=1}^K, b)}{\partial \alpha_i \partial \alpha_j} = \begin{cases} \frac{1}{y_i\alpha_i(1-y_i\alpha_i)} + K\beta\langle \mathcal{X}_i, \mathcal{X}_i\rangle + \beta & (i=j), \\ K\beta\langle \mathcal{X}_i, \mathcal{X}_j\rangle + \beta & (i \neq j). \end{cases}$$

Next, we update $\mathcal{V}^{(k)}$ at step $t+1$ by solving the following subproblem:

$$\mathcal{V}^{(k)^{t+1}} = \underset{\mathcal{V}^{(k)}}{\operatorname{argmax}}\ L(\boldsymbol{\alpha}^{t+1}, \mathcal{V}^{(k)}, \{\mathcal{V}^{(j)t}\}_{j\neq k}^K, \{\mathcal{W}^{(k)t}\}_{k=1}^K, b^t)$$

$$= \operatorname{proj}_{\lambda_k}\left(\frac{W_{(k)}^{(k)t}}{\beta} + \sum_{i=1}^m \alpha_i^{t+1} X_{i(k)}\right), \tag{3.3}$$

where $\operatorname{proj}_\lambda(W) = U\min(S,\lambda)V^\top$ and $W = USV^\top$.

Finally, we update the dual variables $\mathcal{W}^{(k)}$ and b at step $t+1$ as

$$W_{(k)}^{(k)t+1} = W_{(k)}^{(k)t} + \beta\left(\sum_{i=1}^m \alpha_i^{t+1} X_{i(k)} - V_{(k)}^{(k)t+1}\right), \tag{3.4}$$

$$b^{t+1} = b^t + \beta \sum_{i=1}^m \alpha_i^{t+1}. \tag{3.5}$$

Note that step 3.3 and step 3.4 can be combined as

$$W_{(k)}^{(k)t+1} = \operatorname{prox}_{\beta\lambda_k}\left(W_{(k)}^{(k)t} + \beta \sum_{i=1}^m \alpha_i^{t+1} X_{i(k)}\right),$$

where $\operatorname{prox}_\lambda(W) = U\max(S-\lambda, 0)V^T$ and $W = USV^\top$. This allows us to avoid computing singular values and the associated singular vectors that are smaller than the threshold λ_k in equation 3.3.

3.1 Optimality Condition. As a stopping condition, we use the relative duality gap (Tomioka, Hayashi, & Kashima, 2011), which can be expressed as

$$\frac{P(\mathcal{W}^t, b^t) - D(-\hat{\boldsymbol{\alpha}}^t)}{P(\mathcal{W}^t, b^t)} \leq \epsilon,$$

where $P(\mathcal{W}^t, b^t)$ is the primal solution at step t of equation 3.1 and ϵ is a predefined tolerance value. $D(-\hat{\boldsymbol{\alpha}}^t)$ is the dual solution at step t of equation 3.2 with $\hat{\boldsymbol{\alpha}}$ obtained by multiplying $\boldsymbol{\alpha}$ with $\min\left(1, \frac{\|V(\boldsymbol{\alpha})_{(k)}\|_{\mathrm{op}}}{\lambda_1}, \ldots, \frac{\|V(\boldsymbol{\alpha})_{(k)}\|_{\mathrm{op}}}{\lambda_K}\right)$, where $\mathcal{V}(\boldsymbol{\alpha}) = \sum_{i=1}^{m} X_i \alpha_i$ and $\|V\|_{\mathrm{op}}$ is the largest singular value of V.

4 Theoretical Risk Analysis

In this section, we theoretically analyze the excess risk for regularization with the overlapped trace norm, the latent trace norm, and the scaled latent trace norm.

We consider a loss function l, which is Lipshitz continuous with constant Λ. Note that this condition is true for both the squared loss and logistic loss functions. Let the training data set be given as $(\mathcal{X}_i, y_i) \in \mathbb{R}^{n_1 \times \cdots \times n_K} \times \mathrm{Y}$, $i = 1, \ldots, m$, where $\mathrm{Y} \in \mathbb{R}$ for regression and $\mathrm{Y} \in \{-1, 1\}$ for classification. In our theoretical analysis, we assume that elements of $\mathcal{X}_i$ independently follow the standard gaussian distribution.

As the standard formulation (Maurer & Pontil, 2013), the empirical risk without the bias term is defined as

$$\hat{R}(\mathcal{W}) = \frac{1}{m} \sum_{i=1}^{m} l(\langle \mathcal{W}, \mathcal{X}_i \rangle, y_i),$$

and the expected risk is defined as

$$R(\mathcal{W}) = \mathbb{E}_{(\mathcal{X}, y) \sim \mu} l(\langle \mathcal{W}, \mathcal{X} \rangle, y),$$

where μ is the probability distribution from which $(\mathcal{X}_i, y_i)$ are sampled.

The optimal $\mathcal{W}^0$ that minimizes the expected risk is given as

$$\mathcal{W}^0 = \arg\min_{\mathcal{W}} R(\mathcal{W}) \qquad \text{subject to } \|\mathcal{W}\|_\star \leq B_0, \tag{4.1}$$

where $\|\cdot\|_\star$ is either the overlapped trace norm, the latent trace norm, or the scaled latent trace norm. The optimal $\hat{\mathcal{W}}$ that minimizes the empirical risk is denoted as

$$\hat{\mathcal{W}} = \arg\min_{\mathcal{W}} \hat{R}(\mathcal{W}) \qquad \text{subject to } \|\mathcal{W}\|_\star \leq B_0. \tag{4.2}$$

Lemma 1 provides an upper bound of the excess risk for tensor-based learning problems (see appendix B for its proof), where $\|\mathcal{W}\|_{\star^*}$ is the dual norm of $\|\mathcal{W}\|_\star$ for $\star = \{\text{overlap, latent, scaled}\}$:

Lemma 1. *For a given Λ-Lipchitz continuous loss function l and for any $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ such that $\|\mathcal{W}\|_\star \leq B_0$ for problems 4.1 and 4.2, the excess risk for a given training data set $(\mathcal{X}_i, y_i) \in \mathbb{R}^{n_1 \times \cdots \times n_K} \times \mathbb{R}, i = 1, \ldots, m$ is bounded with probability at least $1 - \delta$ as*

$$R(\hat{\mathcal{W}}) - R(\mathcal{W}^0) \leq \frac{2}{m} \Lambda B_0 \mathbb{E}\|\mathcal{M}\|_{\star^*} + \sqrt{\frac{log(\frac{2}{\delta})}{2m}}, \tag{4.3}$$

where $\mathcal{M} = \sum_{i=1}^m \sigma_i \mathcal{X}_i$ and $\sigma_i \in \{-1, 1\}$ are Rademacher random variables.

Theorem 1 gives an excess risk bound for overlapped trace norm regularization (its proof is also included in appendix B), which is based on the inequality $\|\mathcal{W}\|_{\text{overlap}} \leq \sum_{k=1}^K \sqrt{r_k} \|\mathcal{W}\|_{\text{F}}$ given in Tomioka and Suzuki (2013):

Theorem 1. *With probability at least $1 - \delta$, the excess risk of learning using the overlapped trace norm regularization for any $\mathcal{W}^0 \in \mathbb{R}^{n_1 \times \cdots \times n_K}$, $K \geq 3$ with $\|\mathcal{W}^0\|_{\text{F}} \leq B$, multilinear ranks $(r_1, \ldots, r_K)$, and estimator $\hat{\mathcal{W}} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ with $B_0 \leq B \sum_{k=1}^K \sqrt{r_k}$ is bounded as*

$$R(\hat{W}) - R(W^0) \leq 2\Lambda \frac{B}{\sqrt{m}} \left(\sum_{k=1}^K \sqrt{r_k} \right) \min_k (\sqrt{n_k} + \sqrt{n_{\setminus k}}) + \sqrt{\frac{log\,(\frac{2}{\delta})}{2m}}, \tag{4.4}$$

where $n_{\setminus k} = \prod_{j \neq k}^K n_j$.

In theorem 2, we give an excess risk bound for the latent trace norm (its proof is also included in appendix B), which uses the inequality $\|\mathcal{W}\|_{\text{latent}} \leq \sqrt{\min_k r_k} \|\mathcal{W}\|_{\text{F}}$ given in Tomioka and Suzuki (2013):

Theorem 2. *With probability at least $1 - \delta$, the excess risk of learning using the latent norm regularization for any $\mathcal{W}^0 \in \mathbb{R}^{n_1 \times \cdots \times n_K}$, $K \geq 3$ with $\|\mathcal{W}^0\|_{\text{F}} \leq B$, multilinear ranks $(r_1, \ldots, r_K)$, and estimator $\hat{\mathcal{W}} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ with $B_0 \leq B \sqrt{\min_k r_k}$ is bounded as*

$$R(\hat{W}) - R(W^0) \leq 2\Lambda B \sqrt{\frac{\min_k r_k}{m}} \left(\max_k (\sqrt{n_k} + \sqrt{n_{\setminus k}}) + 1.5\sqrt{2\,log(K)} \right) + \sqrt{\frac{log(\frac{2}{\delta})}{2m}}, \tag{4.5}$$

where $n_{\setminus k} = \prod_{j \neq k}^K n_j$.

Theorem 2 shows that the excess risk for the latent trace norm, equation 4.5 is bounded by the minimum multilinear rank. If $n_1 = \cdots = n_K$, the latent trace norm is always better than the overlapped trace norm in terms of the excess risk bounds because $\sqrt{\min_k r_k} < \sum_{k=1}^{K} \sqrt{r_k}$. If the dimensions $n_1, \ldots, n_K$ are not the same, the overlapped trace norm could be better.

Finally, we bound the excess risk for the scaled latent trace norm based on the inequality $\|\mathcal{W}\|_{\text{scaled}} \leq \sqrt{\min_k \left(\frac{r_k}{n_k}\right)} \|\mathcal{W}\|_{\text{F}}$ given in Wimalawarne et al. (2014):

Theorem 3. *With probability at least $1 - \delta$, the excess risk of learning using the scaled latent trace norm regularization for any $\mathcal{W}^0 \in \mathbb{R}^{n_1 \times \cdots \times n_K}$, $K \geq 3$ with $\|\mathcal{W}^0\|_{\text{F}} \leq B$, multilinear ranks $(r_1, \ldots, r_K)$, and estimator $\hat{\mathcal{W}} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ with $B_0 \leq B\sqrt{min_k\left(\frac{r_k}{n_k}\right)}$ is bounded as*

$$R(\hat{W}) - R(W^0) \leq 2\Lambda B \sqrt{\frac{1}{m} \min_k \left(\frac{r_k}{n_k}\right)} \left(\max_k (n_k + \sqrt{N}) + 1.5\sqrt{2\log(K)} \right) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}. \tag{4.6}$$

Theorem 3 shows that the excess risk for regularization with the scaled latent trace norm is bounded with the minimum of multilinear ranks relative to their mode dimensions. Similar to the latent trace norm, the scaled latent trace norm would also perform better than the overlapped norm when the multilinear ranks have large variations. If we consider a flat tensor, the modes with small dimensions may have ranks comparable to their dimensions. Although these modes have the lowest mode-k rank, they do not impose a low-rank structure. In such cases, our theory predicts that the scaled latent trace norm performs better because it is sensitive to the mode-k rank relative to its dimension.

As a variation, we can also consider a mode-wise scaled version of the overlapped trace norm defined as $\|\mathcal{W}\|_{\text{soverlap}} := \sum_{k=1}^{K} \frac{1}{\sqrt{n_k}} \|W_{(k)}\|_{\text{tr}}$. It can be easily seen that $\|\mathcal{W}\|_{\text{soverlap}} \leq \sum_{k=1}^{K} \sqrt{\frac{r_k}{n_k}} \|\mathcal{W}\|_{\text{F}}$ holds, and with the same conditions as in theorem 1, we can upper-bound the excess risk for the scaled overlapped trace norm regularization as

$$R(\hat{W}) - R(W^0) \leq 2\Lambda \frac{B}{\sqrt{m}} \left(\sum_{k=1}^{K} \sqrt{\frac{r_k}{n_k}} \right) \min_k (n_k + \sqrt{N}) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}. \tag{4.7}$$

Note that when all modes have the same dimensions, equation 4.7 coincides with equation 4.4. Compared with bound 4.6, the scaled latent norm would perform better than the scaled overlapped norm regularization since $\min_k \sqrt{\frac{r_k}{n_k}} < \sum_{k=1}^{K} \sqrt{\frac{r_k}{n_k}}$.

5 Experiments

We conducted several experiments using simulated and real-world data to evaluate the performance of tensor-based regression and classification methods with regularizations using different tensor norms. We discuss simulations for tensor-based regression in section 5.1 and experiments with real-world data for tensor classification in section 5.2. For all experiments, we use a Matlab environment on a 2.10 GHz (2×8 cores) Intel Xeon E5-2450 server machine with 128 GB memory.

5.1 Tensor Regression with Artificial Data. We report the results of artificial data experiments on tensor-based regression.

We generated three different three-mode tensors as weight tensors $\mathcal{W}$ with different multilinear ranks and mode dimensions. We created two homogeneous tensors with equal mode dimensions of $n_1 = n_2 = n_3 = 10$ with different multilinear ranks $(r_1, r_2, r_3) = (3, 3, 3)$ and $(r_1, r_2, r_3) = (3, 5, 8)$. The third weight tensor is an inhomogeneous case with mode dimensions of $n_1 = 4, n_2 = n_3 = 10$ and multilinear ranks $(r_1, r_2, r_3) = (3, 4, 8)$. To generate these weight tensors, we use the Tucker decomposition (Kolda & Bader, 2009) of a tensor as $\mathcal{W} = \mathcal{C} \times_{k=1}^{3} U^{(k)}$, where $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor and $U^{(k)} \in \mathbb{R}^{r_k \times n_k}$ are component matrices. We sample elements of the core tensor $\mathcal{C}$ from a standard gaussian distribution, choose component matrices $U^{(k)} \in \mathbb{R}^{r_k \times n_k}$ to be orthogonal matrices, and generate $\mathcal{W}$ by mode-wise multiplication of the core tensor and component matrices.

To create training samples $\{\mathcal{X}_i, y_i\}_{i=1}^{n}$, we first create the random tensors $\mathcal{X}_i$ generated with each element independently sampled from the standard gaussian distribution and obtain $y_i = \langle \mathcal{W}, \mathcal{X}_i \rangle + \nu_i$, where ν_i is noise drawn from the gaussian distribution with mean zero and variance 0.1. In our experiments, we use cross-validation to select the regularization parameter from the range 0.01 to 100 at intervals of 0.1. For comparison, we have also simulated matrix regularized regressions for each mode unfolding. Also, we experimented with cross-validation among matrix regularization on each unfolded matrix to understand whether it can find the correct mode for regularization. As the baseline vector-based learning method, we use ridge regression (i.e., l_2-regularized least-squares).

Figure 1 shows the performance of homogeneous tensors with equal mode dimensions $n_1 = n_2 = n_3 = 10$ and equal multilinear ranks $(r_1, r_2, r_3) = (3, 3, 3)$. We see that the overlapped trace norm and the scaled overlapped trace norm (due to equal mode dimensions) perform the best

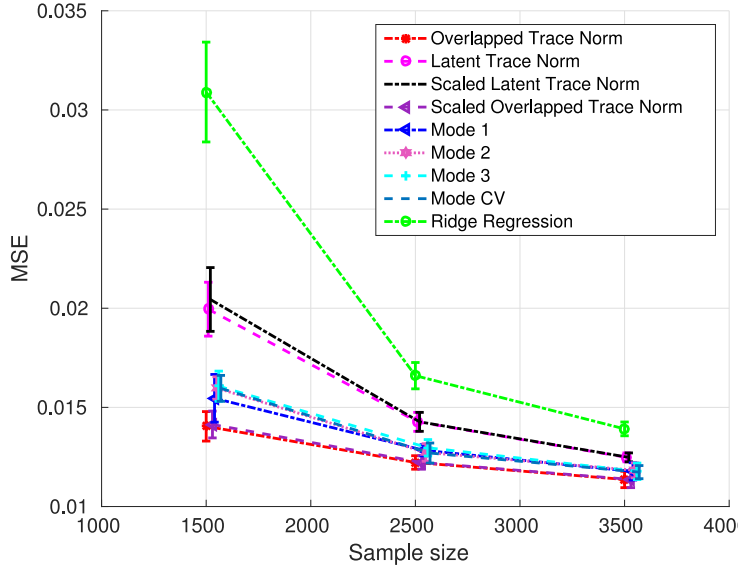


Figure 1: Simulation results of tensor regression based on homogeneous weight tensor of equal mode dimensions $n_1 = n_2 = n_3 = 10$ and equal multilinear ranks $(r_1, r_2, r_3) = (3, 3, 3)$

equally, while both latent norms perform equally (since mode dimensions are equal) but inferior to the overlapped norm. Also, the regression results from all matrix regularizations with individual modes perform better than the latent and the scaled latent norm regularized regression models. Due to the equal multilinear ranks and equal mode dimensions, it results in equal performance with cross-validation among each mode-wise unfolded matrix regularization.

Figure 2 shows the performances of homogeneous tensors with equal mode dimensions $n_1 = n_2 = n_3 = 10$ and unequal multilinear ranks $(r_1, r_2, r_3) = (3, 5, 8)$. In this case, both the latent and the scaled latent norms also perform equally since tensor dimensions are the same. The mode-1 regularized regression models give the best performance since they have the lowest rank; regularization with the latent and scaled latent norms gives the next best performance. The mode-wise cross-validation correctly coincides with the mode-1 regularization. The overlapped trace norm and the scaled overlapped trace (due to equal mode dimensions) perform equally poorly compared to the latent and the scaled latent trace norms.

Figure 3 shows the performance of inhomogeneous tensors with mode dimensions $n_1 = 4, n_2 = n_3 = 10$ and multilinear ranks $(r_1, r_2, r_3) = (3, 4, 8)$. In this case, we can see that the scaled latent trace norm outperforms all

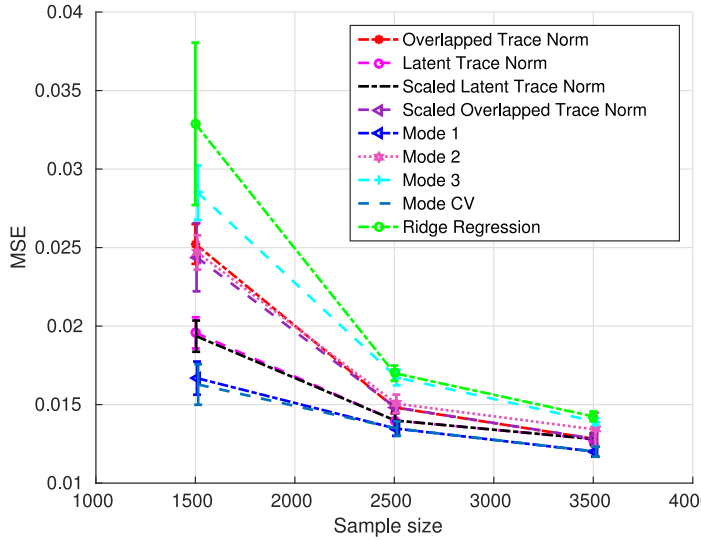


Figure 2: Simulation results of tensor regression based on homogeneous weight tensor of equal mode sizes $n_1 = n_2 = n_3 = 10$ and unequal multilinear rank $(r_1, r_2, r_3) = (3, 5, 8)$

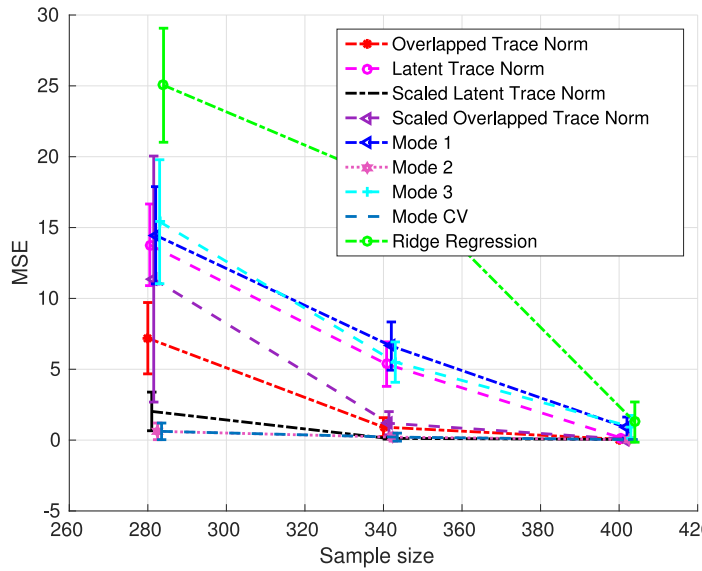


Figure 3: Simulation results of tensor regression based on inhomogeneous weight tensor of equal mode sizes $n_1 = 4$, $n_2 = n_3 = 10$ and multilinear rank $(r_1, r_2, r_3) = (3, 4, 8)$

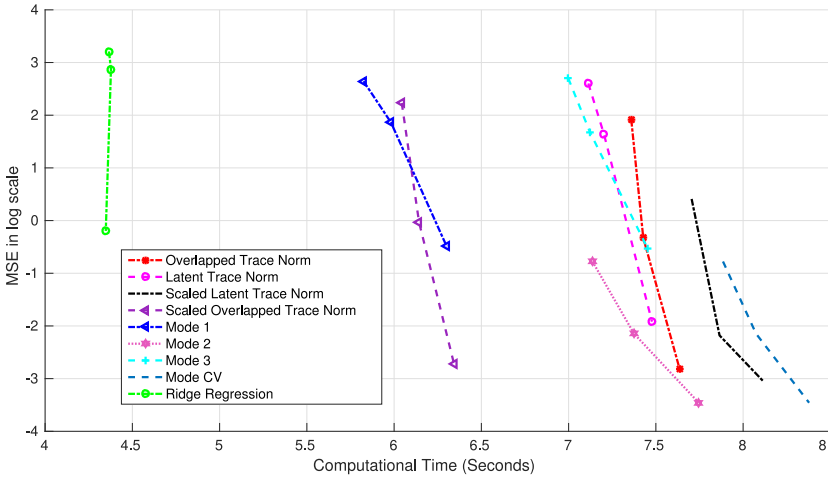


Figure 4: Computation times in seconds for toy experiment with inhomogeneous tensors with mode dimensions $n_1 = 4, n_2 = n_3 = 10$ and multilinear rank $(r_1, r_2, r_3) = (3, 4, 8)$

other tensor norms. The latent trace norm performs poorly since it fails to find the mode with the lowest rank. This agrees well with our theoretical analysis. As shown in equation 4.5, the excess risk of the latent trace norm is bounded with the minimum of multilinear ranks, which is on the first mode in the current setup and is high ranked. The scaled latent trace norm is able to find the mode with the lowest rank since it takes the relative rank with respect to the mode dimension as in equation 4.6. If we look at the individual mode regularizations, we see that the best performance is given with the second mode, which has the lowest rank with respect to the mode dimension, and the worst performance is given with the first mode, which is high ranked compared to other modes. Here, the mode-wise cross-validation is again as good as mode-2 regularization. The overlapped trace norm performs poorly compared to the scaled latent trace norm, and the scaled overlapped trace norm performs worse than the overlapped trace norm.

It is also worth noticing in these experiments that ridge regression performed worse than all the tensor regularized learning models. This highlights the need to employ low-rank-inducing norms for learning with tensor data without vectorization to get the best performance.

Figure 4 shows the computation time for the toy regression experiment with inhomogeneous tensors with mode dimensions $n_1 = 4, n_2 = n_3 = 10$ and multilinear ranks $(r_1, r_2, r_3) = (3, 4, 8)$ (computation time for other setups showed similar tendency and thus we omit the results). For each data

Figure 5: Samples of hand motion sequences of left/flat and left/spread.

set, we measured the computation time of training regression models, cross-validation for model selection, and predicting output values for test data. We can see that methods based on tensor norms and matrix norms are computationally much more expensive compared to ridge regression. However, as we saw, they achieve higher accuracy than ridge regression. It is worth noticing that mode-wise cross-validation is computationally more expensive compared to the scaled latent trace norm and other tensor norms. This computational advantage and comparable performance with respect to the best mode-wise regularization make the scaled latent trace norm a useful regularization method for tensor-based regression, especially for tensors with high variations in its multilinear ranks.

5.2 Tensor Classification for Hand Gesture Recognition. Next, we report the results of experiments on tensor classification with the *Cambridge hand gesture data set* (Kim et al., 2007).

The Cambridge hand gesture data set contains image sequences from nine gesture classes. These gesture classes include three primitive hand shapes of flats, spread, and V-shape, and three different hand motions of rightward, leftward, and contrast. Each class has 100 image sequences with different illumination conditions and arbitrary motions of two people. Previously, the tensor canonical correlation (Kim et al., 2007) was used to classify these hand gestures.

To apply tensor classification, we first build action sequences as tensor data by sampling S images with equal time intervals from each sequence. This makes each sequence a tensor of $20 \times 20 \times S$, where the first two modes are downsampled images as in (Kim et al., 2007) and S is the number of sampled images. In our experiments, we set S at 5 or 10. We consider binary classification and choose visually similar sequences of left/flat and left/spread (see Figure 5), which we found to be difficult to classify. We apply standardization of data by mean removal and variance normalization to all the data. We randomly sample data into a training set of 120 data elements, use a validation set of 40 data elements to select the optimal regularization parameter, and finally use a test set of 40 elements to evaluate the learned classifier. In addition to the tensor regularized learning models, we also trained classifiers with matrix regularization with unfolding on each mode separately. As a baseline vector-based learning method, we have

Table 1: Classification Error of Experiments with the Hand Gesture Data Set.

	Tensor Dimensions	
Norm	(20,20,5)	(20,20,10)
Overlapped trace norm	0.1375 (0.0530)	**0.0775 (0.0343)**
Latent trace norm	**0.1275 (0.0416)**	**0.0875 (0.0429)**
Scaled latent trace norm	**0.1075 (0.0409)**	**0.1000 (0.0500)**
Scaled overlapped trace norm	**0.1275 (0.0416)**	**0.0850 (0.0444)**
Mode-1	**0.1050 (0.0438)**	**0.0975 (0.0463)**
Mode-2	**0.1275 (0.0777)**	**0.0850 (0.0489)**
Mode-3	**0.1175 (0.0409)**	0.1075 (0.0602)
Mode-wise CV	0.1475 (0.0671)	0.1025 (0.0381)
Logistic regression (l_2)	0.1500 (0.0565)	0.1425 (0.0457)

Note: The bold figures indicate comparable accuracies among classifiers after a t-test with a significance of 0.05.

used the l_2-regularized logistic regression. We also trained mode-wise cross-validation (CV) with individual mode regularization (mode-wise CV). We selected regularization parameters as 50 splits in logarithmic scale from 0.01 to 500. We repeated the learning procedure for 10 sample sets for each classifier, the results are shown in Table 1.

In both experiments for $S = 5$ and 10, we see that tensor norm regularized classification performs better than the vectorized learning method. With a tensor structure of (20, 20, 5), we can see that the mode-1 gives the best performance; the scaled latent trace norm, latent trace norm, scaled overlapped trace norm, mode-2, and mode-3 are comparable. We observed that with the tensor structure of (20, 20, 5), the resulting weight tensor after learning its third mode becomes full rank. The scaled latent trace norm performed as well as mode-1 since it could identify the mode with the minimum rank relative to its mode dimension, the first mode in the current setup. The overlapped trace norm performs poorly due to large variations in the multilinear ranks and tensor dimensions.

With the tensor structure (20, 20, 10), the overlapped trace norm gives the best performance. In this case, we found that the multilinear ranks are close to each other, which made the overlapped trace norm give better performance. The scaled latent trace norm, latent trace norm, scaled overlapped trace norm, mode-1, and mode-2 gave a performance comparable to that with the overlapped trace norm.

5.3 Tensor Classification for Brain Computer Interface. As our second tensor classification, we experimented with a motor-imagery EEG classification problem in the context of brain-computer interface (BCI). The objective of the experiments was to classify movements imagined by person using the EEG signals captured in that instance. For our experiments, we

used the data from the BCI competition IVa (Dornhege, Blankertz, Curio, & Müller, 2004). Previous research by Tomioka and Aihara (2007) has considered channel × channel as a matrix of the EEG signal and classified it using logistic regression with low-rank matrix regularization. Our objective is to model EEG data as tensors to incorporate more information and learn to classify using tensor regularization methods.

The BCI competition IVa data set consists of BCI experiments of five people. Though BCI experiments have used 256 channels, we use signals from only 49 channels following Tomioka and Aihara (2007) and preprocess each signal from each channel with Z different band-pass filters (Butterworth filters). Let $S_i \in \mathbb{R}^{C\times T}$, where C denotes the number of channels and T denotes the time, be the matrix obtained by processing with the ith filter. As in Tomioka and Aihara (2007), each S_i is further processed to make centering and scaling as $\hat{S}_i = \frac{1}{\sqrt{T-1}} S_i(I_T - 11^\top)$. Then we obtain $X_i = \hat{S}_i \hat{S}_i^\top$, a channel × channel matrix (in our setting, it is 49 × 49). We arrange all X_i, $i = 1, \ldots, Z$ to form a tensor of dimensions $Z \times 49 \times 49$.

For our experiments, we used $Z = 5$ different bandpass Butterworth filters with cutoff frequencies of (7, 10), (9 12), (11 14), (13 16), and (15 18) with scaling by 50, which resulted in a signal converted into a tensor of dimensions $5 \times 49 \times 49$. We split the data used in the competition into training and validation sets with a proportion of 80:20; the rest of the data we used for testing. As in the previous experiment, we used logistic regression with all the tensor norms, individual mode unfolded matrix regularizations, and cross-validation with unfolded matrix regularization. We also used vector-based logistic regression with l_2-regularization for comparison. To compare tensor-based methods with the previously proposed matrix approach (Tomioka & Aihara, 2007), we averaged tensor data over the frequency mode and applied classification with matrix trace norm regularization. For all experiments, we selected all regularization parameters in 50 splits in logarithmic scale from 0.01 to 500. We show the validation and test errors for the tensor norms in appendix C in Figure 6.

The results of the experiment are given in Table 2, which strongly indicate that vector-based logistic regression is clearly outperformed by the overlapped and scaled latent trace norms. Also, in most cases, the averaged matrix method performs poorly compared to the optimal tensor structured regularization methods. Mode-1 regularization performs poorly since mode-1 was high ranked compared to the other modes. Similarly, the latent trace norm gives poor performance since it cannot properly regularize since it does not consider the rank relative to the mode dimension. For all subjects, mode-2 and mode-3 unfolded regularizations result in the same performance due to the symmetry of each X_i resulting in same rank along mode-2 and mode-3 unfoldings. For subject *aa*, the scaled latent norm, mode-1, mode-2, and mode-wise cross-validation give the best or comparable performance. In subject *al*, the scaled overlapped trace norm gives the best performance, and in subject *av*, both the overlapped trace norm and the

Table 2: Classification Error of Experiments with the BCI Competition IVa Data Set.

Norm	Subject *aa*	Subject *al*	Subject *av*	Subject *aw*	Subject *ay*	Average Time (seconds)
Overlapped trace norm	0.2205 (0.0139)	0.0178 (0.0)	**0.3244 (0.0132)**	**0.0603 (0.0071)**	**0.1254 (0.0190)**	17,986 (1489)
Scaled overlapped trace norm	0.2295 (0.0270)	**0.0018 (0.0056)**	**0.3235 (0.0160)**	0.1022 (0.0192)	0.2532 (0.0312)	18,118 (1608)
Latent trace norm	0.3107 (0.0210)	0.0339 (0.0056)	0.3735 (0.0218)	0.1549 (0.0381)	0.4008 (0.0)	20,021 (14024)
Scaled latent trace norm	**0.2080 (0.0043)**	0.0179 (0.0)	0.3694 (0.0182)	0.0804 (0.0)	0.1980 (0.0476)	77,123 (149024)
Mode-1	0.3205 (0.0174)	0.0339 (0.0056)	0.3739 (0.0211)	0.1450 (0.0070)	0.4020 (0.0038)	5,737 (3238)
Mode-2	**0.2035 (0.0124)**	0.0285 (0.0225)	0.3653 (0.0186)	0.0790 (0.0042)	0.1794 (0.0025)	5,195 (1446)
Mode-3	**0.2035 (0.0124)**	0.0285 (0.0225)	0.3653 (0.0186)	0.0790 (0.0042)	0.1794 (0.0025)	5,223 (1452)
Mode-wise CV	**0.2080 (0.0369)**	0.0428 (0.0305)	0.3545 (0.0125)	0.1008 (0.0227)	**0.1452 (0.0224)**	14,473 (4142)
Averaged matrix	0.2732 (0.0286)	0.0178 (0.0)	0.4030 (0.2487)	0.1366 (0.0056)	0.1825 (0.0)	1,936 (472)
Logistic regression (l_2)	0.3161 (0.0075)	0.0179 (0.0)	0.3684 (0.0537)	0.2241 (0.0432)	0.4040 (0.0640)	72 (62)

Note: The bold numbers in columns *aa*, *al*, *av*, *aw*, and *ay* indicate comparable accuracies among classifiers after a *t*-test with a significance of 0.05.

scaled overlapped trace norm give comparable performances. In subjects *aw* and *ay*, the overlapped trace norm gives the best performance.

In contrast to the computation time for regression experiments, in this experiment, we see that the computation time for tensor trace norm regularizations is more expensive compared to the mode-wise regularization. Also, the mode-wise cross-validation is computationally less expensive than the scaled latent trace norm and other tensor trace norms. This is a slight drawback with the tensor norms, though they tend to have higher classification accuracy.

6 Conclusion and Future Work

In this letter, we have studied tensor-based regression and classification with regularization using the overlapped trace norm, the latent trace norm, and the scaled latent trace norm. We have provided dual optimization methods, theoretical analysis, and experimental evaluations to understand tensor-based inductive learning. Our theoretical analysis on excess risk bounds showed the relationship of excess risks with the multilinear ranks and dimensions of the weight tensor. Our experimental results on both simulated and real data sets further confirmed the validity of our theoretical analyses. From the theoretical and empirical results, we can conclude that the performance of regularization with tensor norms depends on the multilinear ranks and mode dimensions, where the latent and scaled latent norms are more robust in tensors with large variations of multilinear ranks.

Our research opens up many future research directions. For example, an important direction is improvement of optimization methods. Optimization over the latent tensors that results in the use of the latent trace norm and the scaled latent trace norm increases the computational cost compared to the vectorized methods. Also, computing multiple singular value decompositions and solving Newton optimization subproblems (for logistic regression) at each iterative step are computationally expensive. This is evident from our experimental results on computation time for regression and classification. It would be an important direction to develop computationally more efficient methods for learning with tensor data to make it more practical.

Regularization with a mixture of norms is common in both vector-based (e.g., the elastic net; Zou & Hastie, 2003) and matrix-based regularizations (Savalle, Richard, & Vayatis, 2012). It would be an interesting research direction to combine sparse regularization (the l_1-norm) to existing tensor norms. There is also a recent research direction to develop new composite norms such the (k, q)-trace norm (Richard, Obozinski, & Vert, 2014). Development of composite tensor norms can be useful for inductive tensor learning to obtain sparse and low-rank solutions.

Appendix A: Dual Formulations

In this appendix, we derive the dual formulation of the latent trace norms. We consider a training data set $(\mathcal{X}_i, y_i)$, $i = 1, \ldots, m$, where $\mathcal{X}_i \in \mathbb{R}^{n_1 \times \cdots \times n_K}$. To derive the dual for the latent trace norms, we rewrite the primal for the regression of equation 3.1 as

$$\min_{\mathcal{W}^{(1)}+\cdots+\mathcal{W}^{(K)}=\mathcal{W},b} \sum_{i=1}^{m} \frac{1}{2}(y_i - z_i)^2 + \sum_{k=1}^{K} \lambda_k \|W_{(k)}^{(k)}\|_{\mathrm{tr}}$$

$$\text{subject to} \quad z_i = \left\langle \sum_{k=1}^{K} \mathcal{W}^{(k)}, \mathcal{X}_i \right\rangle + b, \quad i = 1, \ldots, m.$$

Its Lagrangian can be written by introducing variables $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$, $\alpha_i \in \mathbb{R}$ as

$$\begin{aligned}
G(\boldsymbol{\alpha}) &= \min_{\mathcal{W}, z_1, \ldots, z_m, b} \sum_{i=1}^{m} \frac{1}{2}(y_i - z_i)^2 + \sum_{k=1}^{K} \lambda_k \|W_{(k)}^{(k)}\|_{\mathrm{tr}} \\
&\quad + \sum_{i=1}^{m} \alpha_i \left(z_i - \left\langle \sum_{k=1}^{K} \mathcal{W}^{(k)}, \mathcal{X}_i \right\rangle - b \right) \\
&= \min_{z_1, \ldots, z_m} \sum_{i=1}^{m} \left(\frac{1}{2}(y_i - z_i)^2 + \alpha_i z_i \right) - \min_{b} b \sum_{i}^{m} \alpha_i \\
&\quad + \min_{\mathcal{W}} \sum_{k=1}^{K} \left(\lambda_k \|W_{(k)}^{(k)}\|_{\mathrm{tr}} - \left\langle W_{(k)}^{(k)}, \sum_{i=1}^{m} \alpha_i X_{i(k)} \right\rangle \right) \\
&= \sum_{i=1}^{m} \left(-\frac{1}{2}\alpha_i^2 + \alpha_i y_i \right) + \sum_{k=1}^{K} \begin{cases} 0 & \left\| \sum_{i=1}^{m} \alpha_i X_{i(k)} \right\|_{\mathrm{op}} \leq \lambda_k \\ -\infty & \text{otherwise} \end{cases} \\
&\quad + \begin{cases} 0 & \sum_{i=1}^{m} \alpha_i = 0 \\ -\infty & \text{otherwise} \end{cases} \\
&= \sum_{i=1}^{m} \left(-\frac{1}{2}\alpha_i^2 + \alpha_i y_i \right) + \sum_{k=1}^{K} \delta_{\lambda_k} \left(\sum_{i=1}^{m} \alpha_i X_{i(k)} \right) + \delta \left(\sum_{i=1}^{m} \alpha_i \right).
\end{aligned}$$

We introduce auxiliary variables $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(K)}$ to remove the coupling between the indicator functions. Then the above dual solutions can be

restated as

$$\min_{\boldsymbol{\alpha}, \mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(K)}} \sum_{i=1}^{m} \left(-\frac{1}{2}\alpha_i^2 + \alpha_i y_i \right) + \sum_{k=1}^{K} \delta_{\lambda_k}(V_{(k)}^{(k)})$$

$$\text{subject to} \quad \mathcal{V}^{(k)} = \sum_{i=1}^{m} \alpha_i \mathcal{X}_i, \quad k = 1, \ldots, K,$$

$$\sum_{i=1}^{m} \alpha_i = 0. \tag{A.1}$$

Similarly, we can derive the dual formulation for logistic regression.

Appendix B: Proofs of Theorems in Section 4

Proof of Lemma 1. By using the same approach as the one given in Wimalawarne et al. (2014) and Maurer and Pontil (2013), we rewrite

$$R(\hat{\mathcal{W}}) - R(\mathcal{W}^0) = [R(\hat{\mathcal{W}}) - \hat{R}(\hat{\mathcal{W}})] + [\hat{R}(\hat{\mathcal{W}}) - \hat{R}(\mathcal{W}^0)] + [\hat{R}(\mathcal{W}^0) - R(\mathcal{W}^0)].$$

The second term is always negative, and based on Hoeffding's inequality, with probability at least $1 - \delta/2$, the third term can be bounded as $\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$:

$$R(\hat{\mathcal{W}}) - R(\mathcal{W}^0) \leq R(\hat{\mathcal{W}}) - \hat{R}(\hat{\mathcal{W}}) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}},$$

$$\leq \sup_{\|\mathcal{W}\|_\star \leq B_0} \left(R(\mathcal{W}) - \hat{R}(\mathcal{W}) \right) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$

Further applying McDiarmid's inequality, with probability at least $1 - \delta$, we get the following Rademacher complexity:

$$\mathfrak{R} = \frac{2}{m} \mathbb{E} \sup_{\|\mathcal{W}\|_\star \leq B_0} \sum_{i=1}^{m} \sigma_i l(\langle \mathcal{W}, \mathcal{X}_i \rangle, y_i),$$

where $\sigma_i \in \{-1, 1\}$ are Rademacher variables, which leads to

$$R(\hat{\mathcal{W}}) - R(\mathcal{W}^0)$$

$$\leq \frac{2}{m} \mathbb{E} \sup_{\|\mathcal{W}\|_\star \leq B_0} \sum_{i=1}^{m} \sigma_i l(\langle \mathcal{W}, \mathcal{X}_i \rangle, y_i) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$

$$\leq \frac{2\Lambda}{m}\mathbb{E}\sup_{\|\mathcal{W}\|_\star \leq B_0}\sum_{i=1}^{m}\sigma_i\langle\mathcal{W},\mathcal{X}_i\rangle + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$

$$= \frac{2\Lambda}{m}\mathbb{E}\sup_{\|\mathcal{W}\|_\star \leq B_0}\left\langle\mathcal{W},\sum_{i=1}^{m}\sigma_i\mathcal{X}_i\right\rangle + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$

$$\leq \frac{2\Lambda}{m}\mathbb{E}\sup_{\|\mathcal{W}\|_\star \leq B_0}\|\mathcal{W}\|_\star\|\mathcal{M}\|_{\star^*} + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}} \quad \text{(Hölder's inequality)}$$

$$\leq \frac{2\Lambda B_0}{m}\mathbb{E}\|\mathcal{M}\|_{\star^*} + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}},$$

where $\mathcal{M} = \sum_{i=1}^{m}\sigma_i\mathcal{X}_i$.

Proof of Theorem 1. First, we bound the data-dependent component of $\mathbb{E}\|\mathcal{M}\|_{\text{overlap}^*}$. For this, we use the following duality relationship borrowed from Tomioka and Suzuki (2013):

$$\|\mathcal{M}\|_{\text{overlap}^*} = \inf_{\mathcal{M}^{(1)}+\cdots+\mathcal{M}^{(K)}=\mathcal{M}}\max_k \|M^{(k)}_{(k)}\|_{\text{op}}.$$

Since we can take any $\mathcal{M}^{(k)}$ to equal $\mathcal{M}$, the above norm can be upper-bounded as

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

Furthermore, the expectation of the minimum of k can be upper-bounded by the minimum of the expectation:

$$\mathbb{E}\|\mathcal{M}\|_{\text{overlap}^*} \leq \mathbb{E}\min_k \|M_{(k)}\|_{\text{op}} \leq \min_k \mathbb{E}\|M_{(k)}\|_{\text{op}}. \tag{B.1}$$

Let $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_m\}$ be fixed Rademacher variables. Since each $\mathcal{X}_i$ contains elements following the standard gaussian distribution, it makes each element in $\mathcal{M}$ a sample from $\mathcal{N}(0, \|\boldsymbol{\sigma}\|_2^2)$. Based on the standard methods used in Tomioka, Suzuki, Hayashi et al. (2011), we can express $\|M_{(k)}\|_{\text{op}}$ as

$$\|M_{(k)}\|_{\text{op}} = \sup_{u\in S^{n_k-1}, v\in S^{\prod_{i\neq k} n_i-1}} u^\top M_{(k)} v.$$

Using Gordan's theorem as in Tomioka, Suzuki, Hayashi et al. (2011b), we have

$$\mathbb{E}\|M_{(k)}\|_{\mathrm{op}} \leq \|\boldsymbol{\sigma}\|_2 \min_k(\sqrt{n_k} + \sqrt{n_{\backslash k}}). \tag{B.2}$$

Next, taking the expectation over $\boldsymbol{\sigma}$, we have

$$\mathbb{E}\|\boldsymbol{\sigma}\|_2 \leq \sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\|\boldsymbol{\sigma}\|_2^2} = \sqrt{m}. \tag{B.3}$$

Combining equations B.2 and B.3 with B.1 results in

$$\mathbb{E}\|\mathcal{M}\|_{\mathrm{overlap}^*} \leq \min_k \sqrt{m}(\sqrt{n_k} + \sqrt{n_{\backslash k}}).$$

Finally, the excess loss can be written as

$$R(\hat{W}) - R(W^0) \leq 2\Lambda \frac{B}{\sqrt{m}}\left(\sum_{k=1}^K \sqrt{r_k}\right) \min_k(\sqrt{n_k} + \sqrt{n_{\backslash k}}) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$

We prove the following useful lemma.

Lemma 2. *Let*

$$P(X \geq \mu + t) \leq K\, exp(-t^2/(2\sigma^2)).$$

Then

$$\mathbb{E}X \leq \mu + \frac{3}{2}\sqrt{2\sigma^2\, logK},$$

given logK > 1.

Proof.

$$\begin{aligned}
\mathbb{E}X &\leq \int_0^{\mu+\sqrt{2\sigma^2 \log K}} 1dt + \int_{\mu+\sqrt{2\sigma^2 \log K}}^{\infty} P\big(X \geq t\big)dt \\
&= \mu + \sqrt{2\sigma^2 \log K} + \int_{\sqrt{2\sigma^2 \log K}}^{\infty} P\big(X \geq \mu + t\big)dt \\
&\leq \mu + \sqrt{2\sigma^2 \log K} + \int_{\sqrt{2\sigma^2 \log K}}^{\infty} K \exp(-t^2/(2\sigma^2))dt \\
&= \mu + \sqrt{2\sigma^2 \log K} + \int_{\sqrt{2\sigma^2 \log K}}^{\infty} \exp(-t^2/(2\sigma^2) + \log K)dt
\end{aligned}$$

$$\leq \mu + \sqrt{2\sigma^2 \log K} + \int_{\sqrt{2\sigma^2 \log K}}^{\infty} \exp(-t\sqrt{2\log K/\sigma^2} + 2\log K)dt$$

$$\leq \mu + \sqrt{2\sigma^2 \log K} + \sqrt{\frac{\sigma^2}{2\log K}}$$

$$\leq \mu + \sqrt{2\sigma^2 \log K} + \sqrt{\sigma^2 \frac{\log K}{2}}, \log K > 1$$

$$\leq \mu + \frac{3}{2}\sqrt{2\sigma^2 \log K}.$$

Proof of Theorem 2. To bound the data-dependent component, we use the duality result given in Tomioka and Suzuki (2013):

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

Since $\mathcal{M}$ consists of elements following the standard gaussian distribution, for each mode k unfolding, we can write a tail bound (Tomioka & Suzuki, 2013) as

$$P\big(\|M_{(k)}\|_{\text{op}} \geq \|\boldsymbol{\sigma}\|_2(\sqrt{n_k} + \sqrt{n_{\setminus k}}) + t\big) \leq \exp(-t^2/(2\|\boldsymbol{\sigma}\|_2^2)).$$

Using a union bound, we have

$$P\big(\max_k \|M_{(k)}\|_{\text{op}} \geq \|\boldsymbol{\sigma}\|_2 \max_k(\sqrt{n_k} + \sqrt{n_{\setminus k}}) + t\big) \leq K\exp(-t^2/(2\|\boldsymbol{\sigma}\|_2^2)),$$

and using lemma 2 when $\log K \geq 1$, we have

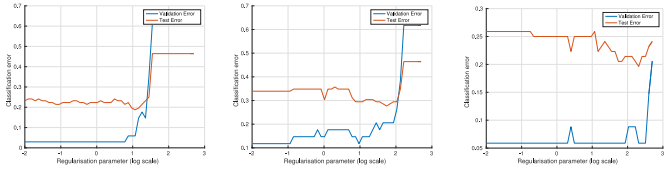
$$\mathbb{E}\max_k \|M_{(k)}\|_{\text{op}} \leq \|\boldsymbol{\sigma}\|_2 \max_k(\sqrt{n_k} + \sqrt{n_{\setminus k}}) + 1.5\sqrt{2\|\boldsymbol{\sigma}\|_2^2 \log(K)}.$$

Similar to equation B.3, taking the expectation over $\|\boldsymbol{\sigma}\|_2$, we arrive at

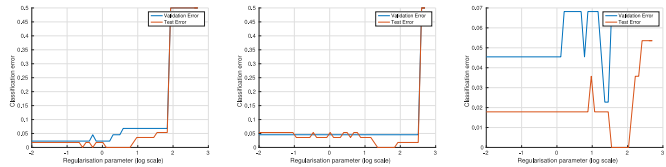
$$\mathbb{E}\max_k \|M_{(k)}\|_{\text{op}} \leq \sqrt{m}\max_k(\sqrt{n_k} + \sqrt{n_{\setminus k}}) + \sqrt{m}1.5\sqrt{2\log(K)},$$

where C is constant. Finally, the excess risk is given as

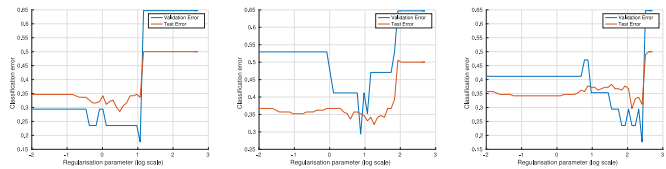
$$R(\hat{W}) - R(W^0) \leq 2\Lambda B\sqrt{\frac{\min_k r_k}{m}}\left(\max_k(\sqrt{n_k} + \sqrt{n_{\setminus k}}) + 1.5\sqrt{2\log(K)}\right) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$



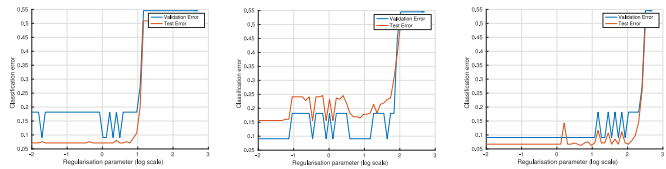
(a) Overlapped trace norm - subject *aa* (b) Latent trace norm - subject *aa* (c) Scaled latent trace norm - subject *aa*



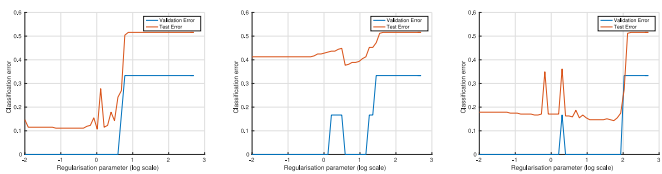
(d) Overlapped trace norm - subject *al* (e) Latent trace norm - subject *al* (f) Scaled latent trace norm - subject *al*



(g) Overlapped trace norm - subject *av* (h) Latent trace norm - subject *av* (i) Scaled latent trace norm - subject *av*



(j) Overlapped trace norm - subject *aw* (k) Latent trace norm - subject *aw* (l) Scaled latent trace norm - subject *aw*



(m) Overlapped trace norm - subject *ay* (n) Latent trace norm - subject *ay* (o) Scaled latent trace norm - subject *ay*

Figure 6: Plots of validation error and test error for BCI data subjects.

Proof of Theorem 3. From Tomioka and Suzuki (2013), we have

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

Using a similar approach to the latent trace norm using lemma 2 and with the additional scaling of $\sqrt{n_k}$, we arrive at the following excess bound for the scaled latent trace norm:

$$R(\hat{W}) - R(W^0) \leq 2\Lambda B\sqrt{\frac{1}{m}\min_k\left(\frac{r_k}{n_k}\right)}\left(\max_k(n_k+\sqrt{N}) + 1.5\sqrt{2\log(K)}\right) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$

Appendix C: Test and Validation Curves for BCI data

We show in Figure 6 the validation errors and test errors for BCI data sets.

Acknowledgments

K.W. acknowledges the Monbukagakusho MEXT Scholarship and KAKENHI 23120004, and M.S. acknowledges the JST CREST program.

References

Bahadori, M. T., Yu, Q. R., & Liu, Y. (2014). Fast multivariate spatio-temporal analysis via low rank tensor learning. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, 27 (pp. 3491–3499). Red Hook, NY: Curran.

Bertsekas, D. P. (1996). *Constrained optimization and Lagrange multiplier methods.* Belmont, MA: Athena Scientific.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Cai, J., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, *20*, 1956–1982.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*(3), 1–37.

Dornhege, G., Blankertz, B., Curio, G., & Müller, K.-R. (2004). Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Transactions on Biomedical Engineering*, *51*(6), 993–1002.

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, *2*(1), 17–40.

Gandy, S., Recht, B., & Yamada, I. (2011). Tensor completion and low-*n*-rank tensor recovery via convex optimization. *Inverse Problems*, *27*(2), 025010.

Karatzoglou, A., Amatriain, X., Baltrunas, L., & Oliver, N. (2010). Multiverse recommendation: *N*-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 79–86). New York: ACM.

Kim, T.-K., Wong, S.-F., & Cipolla, R. (2007). Tensor canonical correlation analysis for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). Piscataway, NJ: IEEE.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500.

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 663–670). Madison, WI: Omnipress.

Liu, J., Musialski, P., Wonka, P., & Ye, J. (2009). Tensor completion for estimating missing values in visual data. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2114–2121). Piscataway, NJ: IEEE.

Maurer, A., & Pontil, M. (2013). Excess risk bounds for multitask learning with trace norm regularization. In *Proceedings of the Annual Conference on Learning Theory 2013* (pp. 55–76). JMLR.org.

Recht, B., Fazel, M., & Parrilo, P. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, *52*(3), 471–501.

Richard, E., Obozinski, G. R., & Vert, J.-P. (2014). Tight convex relaxations for sparse matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 27* (pp. 3284–3292). Red Hook, NY: Curran.

Romera-Paredes, B., Aung, H., Bianchi-Berthouze, N., & Pontil, M. (2013). Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 1444–1452). JMLR.org.

Sankaranarayanan, P., Schomay, T. E., Aiello, K. A., & Alter, O. (2015). Tensor GSVD of patient- and platform-matched tumor and normal DNA copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival. *PLoS ONE*, *10*(4), e0121396.

Savalle, P., Richard, E., & Vayatis, N. (2012). Estimation of simultaneously sparse and low rank matrices. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 1351–1358). Madison, WI: Omnipress.

Signoretto, M., Dinh, Q. T., De Lathauwer, L., & Suykens, J.A.K. (2013). Learning with tensors: A framework based on convex optimization and spectral regularization. *Machine Learning*, *94*(3), 303–351.

Tomioka, R., & Aihara, K. (2007). Classifying matrices with a spectral regularization. In *Proceedings of International Conference on Machine Learning* (pp. 895–902). New York: ACM.

Tomioka, R., Hayashi, K., & Kashima, H. (2011). Estimation of low-rank tensors via convex optimization (Technical report). arXiv 1010.0789.

Tomioka, R., & Suzuki, T. (2013). Convex tensor decomposition via structured Schatten norm regularization. In *Advances in neural information processing systems*, *26* (pp. 1331–1339). Red Hook, NY: Curran.

Tomioka, R., Suzuki, T., Hayashi, K., & Kashima, H. (2011). Statistical performance of convex tensor decomposition. In C.J.C Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *24* (pp. 972–980). Red Hook, NY: Curran.

Tomioka, R., Suzuki, T., & Sugiyama, M. (2011). Super-linear convergence of dual augmented-Lagrangian algorithm for sparsity regularized estimation. *Journal of Machine Learning Research*, *12*, 1537–1586.

Wimalawarne, K., Sugiyama, M., & Tomioka, R. (2014). Multitask learning meets tensor factorization: Task imputation via convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27* (pp. 2825–2833. Red Hook, NY: Curran.

Zhou, H., & Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 463–483.

Zhou, H., Li, L., & Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, *108*(502), 540–552.

Zou, H., & Hastie, T. (2003). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Received February 11, 2015; accepted November 13, 2015.