Classifying Matrices with a Spectral Regularization

Ryota Tomioka RYOTAT@FIRST.FHG.DE
Dept. of Mathematical Informatics, IST, The University of Tokyo, 113-8656 Tokyo, Japan.
Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany.

Kazuyuki Aihara AIHARA@SAT.T.U-TOKYO.AC.JP
Institute of Industrial Sciences, The University of Tokyo, 153-8505 Tokyo, Japan.

Abstract

We propose a method for the classification of matrices. We use a linear classifier with a novel regularization scheme based on the spectral ℓ_1-norm of its coefficient matrix. The spectral regularization not only provides a principled way of complexity control but also enables automatic determination of the rank of the coefficient matrix. Using the Linear Matrix Inequality technique, we formulate the inference task as a single convex optimization problem. We apply our method to the motor-imagery EEG classification problem. The method not only improves upon conventional methods in the classification performance but also determines a subspace in the signal that concentrates discriminative information without any additional feature extraction step. The method can be easily generalized to regression problems by changing the loss function. Connections to other methods are also discussed.

1. Introduction

In this paper, we consider the following linear regression model over matrices:

$$f(X; W, b) = \mathrm{Tr}\left[W^\top X\right] + b \qquad (1)$$

where $X \in \mathbb{R}^{R\times C}$ is a $R \times C$ matrix input for which we'd like to predict its label y; for example X may be any multi-sensor recording of time series, image, or any other vectorial data. We call $W \in \mathbb{R}^{R\times C}$ a *weight matrix*. The goal is to infer the weight matrix W and the bias $b \in \mathbb{R}$ from the training examples $\{X_i, y_i\}_{i=1}^n$.

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

The simplest approach would be to just ignore the fact that the input X is a matrix and concatenate all e.g., columns into a long vector. Alternatively, one can define a problem specific inner product between matrices and perform the inference in a Hilbert space. We take a yet another approach; we choose a metric (or a norm) for matrices which is used to (a) control the complexity of the weight matrix when viewed from the primal problem, and at the same time to (b) control the deviation of the empirical statistics from their predictions from a dual point of view. We use the norm dual-norm pair, the spectral ℓ_1-norm and ℓ_∞-norm for the weight matrix W and the input X, respectively.

Our framework is a maximum a posteriori (MAP) estimation with a Laplacian prior on the singular values of the weight matrix. The Laplacian prior enforces a low rank solution. That is, if we rewrite Eq. (1) using the singular value decomposition of $W = \sum_{c=1}^r \sigma_c \boldsymbol{u}_c \boldsymbol{v}_c^\top$, the linear combination,

$$f(X; W, b) = \sum_{c=1}^{r} \sigma_c \boldsymbol{u}_c^\top X \boldsymbol{v}_c + b,$$

consists of small number ($r \ll R, C$) of components. Therefore the discriminative information is concentrated in a few projections of the input $\boldsymbol{u}_c^\top X \boldsymbol{v}_c$ ($c = 1, \ldots, r$).

In the next section, we formulate our approach as a convex optimization problem using the Linear Matrix Inequality (LMI) technique, and derive the dual problem for efficient optimization. In Sec. 3 we apply our method to motor-imagery EEG classification problem. The proposed method not only (a) shows improved classification accuracy over conventional methods but also (b) shows that a good classification is possible with a small subspace in the data. In Sec. 4 related methods with other regularization or loss functions are discussed. We summarize the paper in the last section. Some proofs and the details of the implementation are

provided in the appendix.

2. Method

2.1. Logistic Regression Problem

Let us specialize the model (Eq. (1)) for a binary classification problem. We model the logit of the posterior class probability with a linear function:

$$\log \frac{P(y=+1|X)}{P(y=-1|X)} = \mathrm{Tr}\left[W^\top X\right] + b. \tag{2}$$

We minimize the negative log-likelihood of Eq. (2) with a *spectral* ℓ_1*-norm* penalization term, which is written as follows:

$$\begin{aligned}(\mathrm{P}) \quad \min_{W\in\mathbb{R}^{R\times C}, b\in\mathbb{R}, \boldsymbol{z}\in\mathbb{R}^n} & \sum_{i=1}^n \ell_{\mathrm{LR}}(z_i) + \lambda \|W\|_1, \\ \text{s.t.} \quad & y_i\left(\mathrm{Tr}\left[W^\top X_i\right] + b\right) = z_i \quad (i=1,\dots,n), \end{aligned} \tag{3}$$

where $z_i\,(i=1,\dots,n)$ are called the latent variables, λ is the regularization constant, and $\mathrm{Tr}\,[\cdot]$ denotes the trace. Here the *logistic loss* ℓ_{LR} and the *spectral* ℓ_1*-norm* of a matrix $\|W\|_1$ are defined as follows:

$$\begin{aligned}\ell_{\mathrm{LR}}(z) &:= \log\left(1+\exp(-z)\right), \\ \|W\|_1 &:= \sum_{c=1}^r \sigma_c\left[W\right],\end{aligned}$$

where $\sigma_c\left[W\right]$ $(c=1,\dots,r)$ is the c-th singular value of a matrix $W\in\mathbb{R}^{R\times C}$ and r is the rank of W.

Using two auxiliary positive semidefinite matrices $Q_1\in\mathbb{S}_+^R$ and $Q_2\in\mathbb{S}_+^C$,[1] we can rewrite the spectral ℓ_1-norm, which is convex but a non-differentiable function, with LMI, as follows:

$$(\mathrm{P}') \quad \min_{\substack{W\in\mathbb{R}^{R\times C},\\ Q_1\in\mathbb{S}_+^R, Q_2\in\mathbb{S}_+^C,\\ b\in\mathbb{R}, \boldsymbol{z}\in\mathbb{R}^n}} \sum_{i=1}^n \ell_{\mathrm{LR}}(z_i) + \lambda\left(\mathrm{Tr}\left[Q_1\right] + \mathrm{Tr}\left[Q_2\right]\right), \tag{4}$$

$$\text{s.t.} \quad y_i\left(\mathrm{Tr}\left[W^\top X_i\right] + b\right) = z_i \quad (i=1,\dots,n), \tag{5}$$

$$\begin{bmatrix} Q_1 & -\frac{1}{2}W \\ -\frac{1}{2}W^\top & Q_2 \end{bmatrix} \succeq 0. \tag{6}$$

Here the norm of the weight matrix W is replaced by the trace of the matrices Q_1 and Q_2 that satisfy the positive semidefinite constraint (Eq. (6)). One can easily check that the minimum is attained at $Q_1 = \frac{1}{2}USU^\top$ and $Q_2 = \frac{1}{2}VSV^\top$, where $W = USV^\top$ is the singular value decomposition of W. Now we can clearly see that the spectrally ℓ_1-regularized logistic regression is a convex problem. In fact, the objective function (Eq. (4)) is a convex function; the equality constraints (Eq. (5)) are linear; the convexity of the positive semidefinite constraint (Eq. (6)) follows from the convexity of the set of positive semidefinite matrices. Moreover, there exists a strictly feasible point e.g., $W=0, b=0$ with strictly positive definite $Q_1 \succ 0$ and $Q_2 \succ 0$. Therefore the strong duality holds from Slater's theorem (Boyd & Vandenberghe, 2004).

[1] We denote by $\mathbb{S}^C$ the set of $C\times C$ symmetric matrices and by $\mathbb{S}_+^C$ the set of $C\times C$ symmetric positive semidefinite matrices.

2.2. Dual Problem

The Lagrange dual problem of the problem (P) is written as follows (see appendix A for the derivation):

$$\begin{aligned}(\mathrm{D}) \quad \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \sum_{i=1}^n \ell_{\mathrm{LR}}^*\left(\alpha_i\right) \\ \text{s.t.} \quad & 0\le\alpha_i\le 1 \quad (i=1,\dots,n), \\ & \sum_{i=1}^n \alpha_i y_i = 0, \\ & \left\|\sum_{i=1}^n \alpha_i y_i X_i\right\|_\infty \le \lambda, \end{aligned}$$

where $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^n \in \mathbb{R}^n$ are the Lagrangian multipliers associated with the equality constraints (Eq. (3)).

Here the *dual loss* ℓ_{LR}^* [2] and the *spectral* ℓ_∞*-norm* are defined as follows:

$$\begin{aligned}\ell_{\mathrm{LR}}^*(\alpha) &:= \alpha\log\alpha + (1-\alpha)\log(1-\alpha), \\ \|X\|_\infty &:= \max_{1\le c\le r}\sigma_c\left[X\right].\end{aligned}$$

The dual problem (D) can also be rewritten with LMI as follows:

$$\begin{aligned}(\mathrm{D}') \quad \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \sum_{i=1}^n \ell_{\mathrm{LR}}^*\left(\alpha_i\right), \\ \text{s.t.} \quad & 0\le\alpha_i\le 1 \quad (i=1,\dots,n), \\ & \sum_{i=1}^n \alpha_i y_i = 0, \\ & \begin{bmatrix} \lambda I_R & \sum_{i=1}^n \alpha_i y_i X_i \\ \sum_{i=1}^n \alpha_i y_i X_i^\top & \lambda I_C \end{bmatrix} \succeq 0, \end{aligned} \tag{7}$$

where I_R and I_C denote identity matrices of size R and C, respectively. Notice the conjugacy between the matrix inequalities in the primal and the dual problems (Eqs. (6) and (7), respectively).

[2] The dual loss ℓ^*, which is defined as $\ell^*(\alpha) = -\inf_x\left(\ell(x)+\alpha x\right)$ for convenience, is equivalent to the convex conjugate or Legendre transformation ℓ^c except for the negated domain, i.e., $\ell^*(\alpha) = \ell^c(-\alpha)$.

2.3. Interior Point Method

Both the primal and dual problems ((P′) and (D′), respectively) can be solved by interior point methods (Boyd & Vandenberghe, 2004). The basic idea of the interior point methods is to keep the solution strictly *inside* the feasible region; the inequality constraints are replaced with a smooth function ϕ that grows infinitely high at the boundary of the feasible region, called a *barrier function*; the strength of the barrier function is controlled so that the optimum is attained as a limit. For example, the barrier function for the dual problem (D′) is written as follows:

$$\phi(\boldsymbol{\alpha}) := -\left(\log\det\begin{bmatrix}\lambda I_R & \sum_{i=1}^n \alpha_i y_i X_i \\ \sum_{i=1}^n \alpha_i y_i X_i^\top & \lambda I_C\end{bmatrix} + \sum_{i=1}^n \log\alpha_i + \sum_{i=1}^n \log(1-\alpha_i)\right).$$

Note that the negative log of the determinant of a matrix becomes infinitely large as the matrix becomes close to singularity. The dual problem (D′) is particularly attractive to solve when the input matrix is large because the number of variables in the dual problem is only the number of samples. Now we have the following equality constrained minimization problem,

$$(\mathrm{D_B}) \qquad \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^n \ell^*_{\mathrm{LR}}(\alpha_i) + \frac{1}{t}\phi(\boldsymbol{\alpha}),$$
$$\text{s.t.} \qquad \sum_{i=1}^n \alpha_i y_i = 0,$$

where the parameter t controls the strength of the barrier. For any finite t, the solution of the problem ($\mathrm{D_B}$) is kept strictly *inside* the feasible region. The optimal of the original dual problem (D′) is attained as the limit $t \to \infty$. Each step in an interior point method is an equality constrained minimization ($\mathrm{D_B}$) which we can solve efficiently using Newton's method. More detail on the implementation are found in appendix B.

3. Application

3.1. Motor-Imagery EEG Classification Problem

As an application of the proposed matrix coefficient logistic regression, we consider the motor-imagery EEG classification problem in the context of Brain-Computer Interface (BCI). The task is to classify *imagined* movement from a single-trial measurement of EEG signals. The underlying physiology is a spatially localized band-power modulation known as Event Related Desynchronization (ERD, see (Pfurtscheller & da Silva, 1999)). Therefore we take the short-time variance-covariance matrix for each trial (3 seconds) as the input matrix. From now on the input X and the weight matrix W are assumed to be symmetric, i.e., $X \in \mathbb{S}^C$ and $W \in \mathbb{S}^C$, where C is the number of electrodes.

3.2. Experimental Settings

We use 60 BCI experiments (Blankertz et al., 2006) from 29 subjects where the subjects performed three imaginary movements, namely "right hand" (R), "left hand" (L) and "foot" (F) according to the visual cue presented on the screen, except 9 experiments where only two classes were performed. Since we focus on binary classification, all the pairwise combination of the performed classes produced $162\,(= 51\cdot 3 + 9)$ datasets. Each dataset contains 70 to 600 trials (at median 280) of imaginary movements. All the recordings come from the calibration measurements, i.e., no feedback was presented to the subjects. The signal was recorded from the scalp with multi-channel EEG amplifiers using 32, 64 or 128 channels. The signal was sampled at 1000Hz and down-sampled to 100Hz before the processing. We reduce the number of channels to at most 49 channels[3] because we were not able to handle datasets with larger number of channels with the initial implementation using `CVX` (see Sec. B.2); the specialized implementation we present in Sec. B.3 is able to handle these datasets but all the results presented in Sec 3.4 is produced using the `CVX` implementation.

The signal is band-pass filtered at 7-30Hz and the interval 500-3500ms after the appearance of visual cue is cut out from the continuous EEG signal as a trial $S \in \mathbb{R}^{C\times T}$, where C is the number of electrodes and T is the number of sampled time points. The input matrix X is defined as $X = \tilde{S}\tilde{S}^\top \in \mathbb{S}^C$, where $\tilde{S} = \frac{1}{\sqrt{T-1}} S\left(I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^\top\right)$ is the input signal after centering and scaling. The training data is whitened before applying all the methods. For the prediction of the test data, coefficients including the whitening operation $W = \Sigma_P^{-1/2}\tilde{W}\Sigma_P^{-1/2}$ for ℓ_1- and ℓ_2-regularized LR and $\boldsymbol{w}_c = \Sigma_P^{-1/2}\tilde{\boldsymbol{w}}_c$ $(c = 1, 2)$ for the rank=2 approximated LR are used, where $\Sigma_P^{-1/2}$ is the whitening and $\tilde{W}$ and $\tilde{\boldsymbol{w}}_c$ denote the solution on the whitened data (the methods are explained in the next section). Note that we did *not* whitened the training and test data *jointly*, which could have improved the performance.

The performance is measured by a chronological validation; that is, all methods are trained on the first

[3]The following 49 channels are used: F7, F5, F3, F1, Fz, F2, F4, F6, F8, FC5, FC3, FC1, FCz, FC2, FC4, FC6, T7, C5, C3, C1, Cz, C2, C4, C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO3, POz, PO4, O1, Oz, and O2.

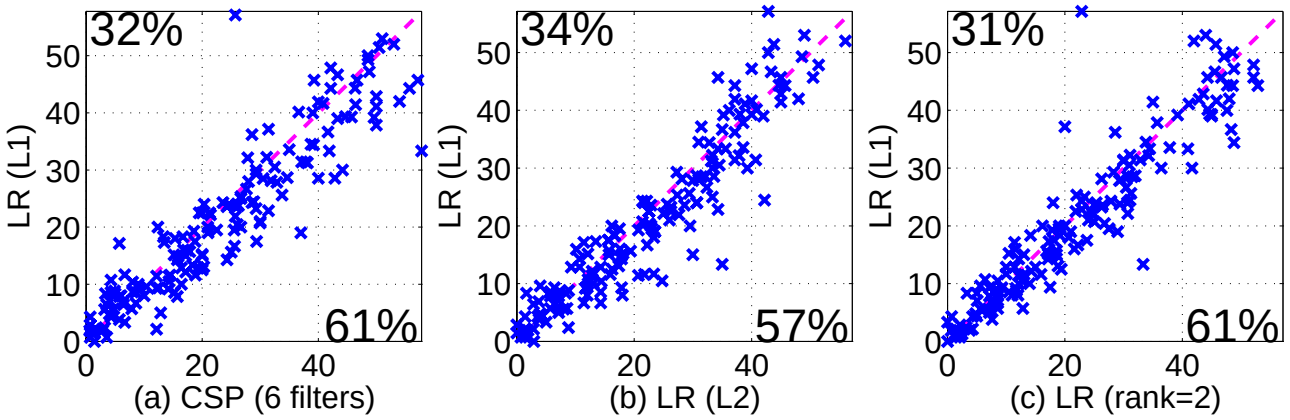

Figure 1. Comparison of the proposed ℓ_1 regularized logistic regression against (a) CSP based classifier, (b) full-rank parameterized LR with Frobenius- (ℓ_2-) norm regularization, and (c) rank=2 parameterized LR. The classification error (in percent unit) is plotted for 162 motor-imagery BCI datasets.

half of the examples and applied on the second half. The regularization constant λ for the logistic regression models is chosen out of 20 candidates log-linearly spaced between 10^{-2} and 10^{2} by 2×10 cross-validation on the training set.

3.3. Conventional Approaches

We compare the spectrally ℓ_1-regularized logistic regression (LR) classifier (see Eq. (2) and (P)) against previously proposed methods, namely, Common Spatial Pattern (CSP) based classifiers (Koles, 1991; Ramoser et al., 2000), ℓ_2-regularized LR and the rank=2 approximated LR (Tomioka et al., 2007).

CSP based classifier, which is a popular technique in BCI, takes a two stage approach. In the first stage, a dimensionality reduction technique called CSP (Koles, 1991) is used; in fact, it is based on simultaneous diagonalization of the covariance Σ_+ and Σ_-, where $\Sigma_\pm$ are the covariance matrices estimated within each class; then the EEG signal is projected into $n_{\rm of}$ components using the top $n_{\rm of}$ eigenvectors selected according to the eigenvalues. In the second stage, a linear discriminant classifier is trained on the log-power of the $n_{\rm of}$ projected signals (see (Ramoser et al., 2000)). We choose $n_{\rm of} = 6$ according to a common practice.

Two previously proposed LRs differ from the proposed method in the following way: ℓ_2-regularized LR uses the Frobenius norm for regularization; in the rank=2 approximated LR, coefficient matrix $W \in \mathbb{S}^C$ is explicitly parameterized with two vectors $\boldsymbol{w}_c \in \mathbb{R}^C$ $(c = 1, 2)$ as follows:

$$W_{\rm r2} =: \frac{1}{2}\left(-\boldsymbol{w}_1\boldsymbol{w}_1^\top + \boldsymbol{w}_2\boldsymbol{w}_2^\top\right). \tag{8}$$

3.4. Classification Performance

In Fig. 1 the proposed ℓ_1-logistic regression (LR) is compared against three methods, namely (a) CSP based classifier (b) ℓ_2-regularized LR and (c) rank=2 constrained LR. For each dataset the chronological test error of the proposed method is plotted against that of the conventional method as a cross. Crosses lying below the diagonal corresponds to datasets that the proposed method outperforms conventional methods. The proportions of datasets lying above/below the diagonal are shown at the top-left/bottom-right corners. All three comparison show that the proposed method significantly outperforms the conventional methods, the p-values are 6.6×10^{-7}, 3.5×10^{-4}, and 2.9×10^{-5}, respectively, based on the paired-sample Wilcoxon signed rank test.

3.5. Extracted Features

Figure 2 (a) illustrates the model selection process for a single dataset. Figure 2 (b) shows the eigenvalue spectrum for various values of λ including the one selected by cross validation. We see that as the regularization become stronger, the number of eigencomponents with eigenvalues significantly larger than zero become smaller. At the selected $\lambda = 8.86$, only two components are used. The average number of the selected components over all datasets were 4.4, which roughly agrees with the common practice of using 4-6 components of CSP decomposition. The two eigenvectors used in the dataset in Fig. 2 are shown in Fig. 3. The coefficients are color-coded and topographically mapped on a scalp viewed from above (nose pointing upwards). We observe a strong focus on the central area, which is known as the motor cortex and the lat-

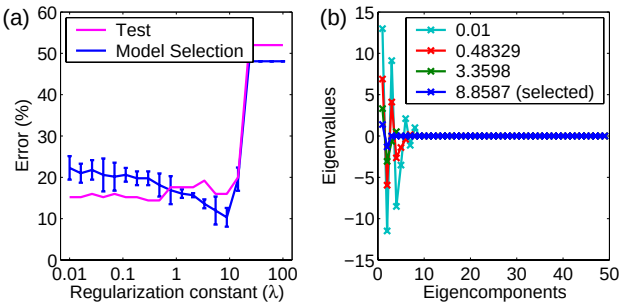

Figure 2. (a) A model selection example. The blue line with error-bars show the cross-validation error on the training set for one dataset; the model $\lambda = 8.86$ is chosen, which gives the minimum error. The magenta line shows the chronological test error. (b) Eigenvalue spectrums for various values of λ.

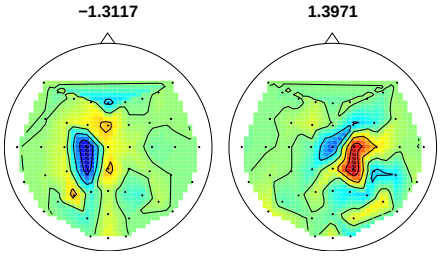

Figure 3. Spatial filter coefficients topographically mapped on a scalp. The two eigenvectors in Fig. 2 (b) at $\lambda = 8.86$ are multiplied by the whitening transformation and plotted. Coefficients are appropriately scaled and color coded as red-green-blue from positive to negative. Eigenvalues are also plotted above the patterns.

eralization corresponding to left- or right-hand motor imagination.

3.6. Rank=2 Approximated Logistic Regression Uses ℓ_1-regularization

Here we show that the sum of the squared norms of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, which is the regularizer in the rank=2 constrained logistic regression (Tomioka et al., 2007), is actually the spectral ℓ_1-norm of W_{r2} (Eq. (8)).

First, we assume that the vectors $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ are orthogonal to each other because the non-orthogonality only increases the regularization term without contributing to the loss term. In fact it was shown that one can always "orthogonalize" a pair $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ without increasing the objective function (Tomioka et al., 2006). Second, we rewrite Eq. (8) as follows:

$$W_{\mathrm{r2}} = \frac{1}{2}\begin{pmatrix} \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|} & \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|} \end{pmatrix} \begin{pmatrix} \|\boldsymbol{w}_1\|^2 & 0 \\ 0 & \|\boldsymbol{w}_2\|^2 \end{pmatrix} \begin{pmatrix} \frac{-\boldsymbol{w}_1^\top}{\|\boldsymbol{w}_1\|} \\ \frac{\boldsymbol{w}_2^\top}{\|\boldsymbol{w}_2\|} \end{pmatrix},$$

where $\|\boldsymbol{w}\| = \left(\sum_{c=1}^{C} w_c^2\right)^{-1/2}$ is the Euclidian norm. We conclude the following from the orthogonality and the definition of the singular values:

$$\|W_{\mathrm{r2}}\|_1 = \frac{1}{2}\left(\|\boldsymbol{w}_1\|^2 + \|\boldsymbol{w}_2\|^2\right).$$

The superiority of rank=2 parameterization to ℓ_2-regularized full rank parameterization reported by (Tomioka et al., 2007) along with our results strongly suggests the appropriateness of the ℓ_1-regularization in this context.

4. Discussion

4.1. Probabilistic Interpretation of the Dual Variables

Since our dual problem is a regularized maximum entropy (MaxEnt) problem, the dual variable $\boldsymbol{\alpha}$ has a natural probabilistic interpretation. First, we reparameterize the dual variable as follows:

$$p_i = \begin{cases} 1 - \alpha_i & (y_i = +1) \\ \alpha_i & (y_i = -1) \end{cases} \qquad (i = 1, \ldots, n).$$

Indeed from the definition of the regression function (Eq. (2)) and the KKT condition (Eq. (11)), we can see that p_i is the probability the classifier assigns to the i-th sample to be in the positive class. We can rewrite the dual problem (D) using $\boldsymbol{p} = \{p_i\}_{i=1}^n$ instead of $\boldsymbol{\alpha}$ as follows,

$$\begin{aligned} &\max_{\boldsymbol{p}\in\mathbb{R}^n} \sum_{i=1}^{n} H_2(p_i) \\ &\text{s.t.} \quad 0 \le p_i \le 1 \quad (i = 1, \ldots, n), \\ &\sum_{i=1}^{n} (y_i - \mathbb{E}[y_i|p_i]) = 0, \qquad (9) \\ &\left\| \sum_{i=1}^{n} (y_i - \mathbb{E}[y_i|p_i]) X_i \right\|_\infty \le \lambda, \qquad (10) \end{aligned}$$

where $H_2(p) = -\ell^*_{\mathrm{LR}}(p)$ is the binary entropy function and $\mathbb{E}[y_i|p_i] := 2p_i - 1$ is the expectation of the label given p_i. It is now clear that the last two constraints (Eqs. (9) and (10)) are bounding the deviation of the empirical statistics from the prediction of the classifier. The deviation of the matrix-valued statistics (Eq. (10)) is measured by the spectral ℓ_∞-norm. The regularization promotes simpler model by allowing a loose fit; the tightness of the fit is controlled by the regularization constant λ. Moreover, from the complementary slackness we can show that the weight matrix has a projection only for subspace that the inequality constraint (Eq. (10)) is satisfied with equality, i.e., subspace corresponding to singular values equal λ.

More general discussion on regularized MaxEnt can be found in (Dudík et al., 2004; Altun & Smola, 2006)

4.2. Variations on Regularization

We obtain many previously proposed methods by choosing a different regularization. First, if we choose the Frobenius norm (ℓ_2-norm) instead of the spectral ℓ_1-norm, we obtain a MAP estimation problem with a Gaussian prior on the regression coefficients, which is often called a kernel logistic regression (see (Jaakkola & Haussler, 1999) for a more general discussion). In the context of EEG classification (see Sec. 3) the kernel can be identified as $K(\tilde{S}_i, \tilde{S}_j) := \mathrm{Tr}\left[\tilde{S}_i \tilde{S}_i^\top \tilde{S}_j \tilde{S}_j^\top\right]$. Second, if we choose the element-wise ℓ_1-norm, i.e., $\|W\|_{\mathrm{elm}} = \sum_{i \geq j} |w_{ij}|$, we obtain the popular ℓ_1-regularization known as LASSO (Tibshirani, 1996). Note that both regularizations ignore the fact that the input is a matrix. Indeed, they are invariant to any permutation of the *elements* of the matrix. On the other hand our spectral ℓ_1-regularization is specifically defined for matrices; it is only invariant to orthogonal transformation from both sides, e.g., *row* or *column* permutations.

4.3. Alternative loss function

The general framework we have presented in Sec. 2 allow us to use any loss functions for classification. For example, if we take the hinge loss, we obtain a Semidefinite Program (SDP), because the hinge loss $\ell_{\mathrm{H}}(z) = \max\{0, 1 - z\}$ is a piecewise linear function. The advantages of the hinge loss are the following two; first there are plenty of optimization algorithm in the market to efficiently solve SDP; second, because the hinge loss enforces sparsity also in the dual variables, it might be useful when we face a flood of examples, e.g., in an online training of a classifier, where one needs a criterion to discard some of the examples.

In Fig. 4 the hinge loss $\ell_{\mathrm{H}}(z)$ and the dual hinge loss $\ell_{\mathrm{H}}^*(\alpha) := -\inf_z(\alpha z + \ell_{\mathrm{H}}(z))$ are compared to the counterparts in logistic regression. The difference appears strikingly in the dual loss, although it might seem to be negligible in the primal (Fig. 4 (a)). It is also interesting to compare what the KKT conditions of the two approaches tell. The dual variables in the logistic regression *never* reach $\alpha = 0$ or 1 because they have a probabilistic interpretation as the certainty of the classifier (see Sec. 4.1). On the other hand, in the case of hinge loss, all samples other than support vectors become either totally inactive $\alpha = 0$ or margin errors $\alpha = 1$.

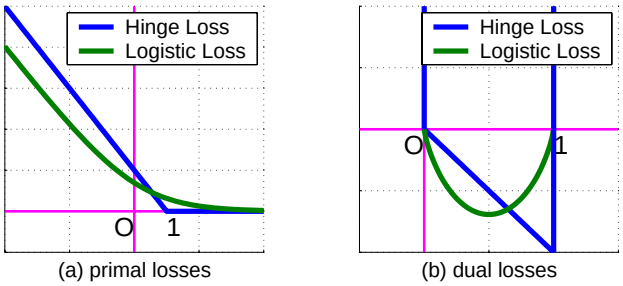

Figure 4. Comparison of the hinge loss to the logistic loss.

5. Conclusion

In this paper, we have proposed a new framework for the classification of matrices. We have introduced the spectral ℓ_1-norm and the ℓ_∞-norm in the space of the weight matrix and the input matrices, respectively. The spectral ℓ_1-regularization in the weight matrix space provides not only a principled way of complexity control but also an elegant low rank solution that concentrates all the discriminative information in a few components. Using the LMI technique we have formulated the inference task as a single convex optimization problem. The optimization can be efficiently done using an interior point method in the dual formulation.

We have applied our method to the single trial EEG classification problem in the context of BCI. The proposed method significantly outperforms conventional methods that use (a) pre-specified number of hidden components or (b) non-sparse ℓ_2-regularization with the same logistic regression model. Moreover, the method automatically produces a decomposition of the signal into small number of components; the number of components is automatically selected. The proposed method sidesteps the explicit rank constraints, which usually result in non-convex optimization, by using the ℓ_1-regularization on the singular values of the weight matrix.

Our current direction is to apply the method to other multiple-sensor recordings e.g., fMRI signals or computer vision problems. The extension of the method to regression problems is also an interesting task.

Acknowledgments

We thank Klaus-Robert Müller, Motoaki Kawanabe, and Benjamin Blankertz for helpful discussions. This research was partially supported by MEXT, Grant-in-Aid for JSPS fellows, 17-11866 and Grant-in-Aid for Scientific Research on Priority Areas, 17022012, by BMBF-grant FKZ 01IBE01A, and by the IST Programme of the European Community, under the PASCAL Network of Excel-

lence, IST-2002-506778. This publication only reflects the authors' views.

References

Altun, Y., & Smola, A. (2006). Unifying Divergence Minimization and Statistical Inference via Convex Duality. In *Proc. COLT2006*.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Kunzmann, V., Losch, F., & Curio, G. (2006). The Berlin Brain-Computer Interface: EEG-based communication without subject training. *IEEE Trans. Neural Sys. Rehab. Eng.*, *14*, 147–152.

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Dudík, M., Phillips, S. J., & Schapire, R. E. (2004). Performance Guarantees for Regularized Maximum Entropy Density Estimation. In *Lect. Notes Comput. Sci.*, vol. 3120, 472–486. Springer.

Grant, M., Boyd, S., & Ye, Y. (2006). CVX: Matlab Software for Disciplined Convex Programming. `http://www.stanford.edu/~boyd/cvx/`, Version 1.0RC.

Jaakkola, T. S., & Haussler, D. (1999). Probabilistic kernel regression models. *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann.

Koles, Z. J. (1991). The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalogr. Clin. Neurophysiol.*, *79*, 440–447.

Pfurtscheller, G., & da Silva, F. H. L. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.*, *110*, 1842–1857.

Ramoser, H., Müller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehab. Eng.*, *8*, 441–446.

Sturm, J. F. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, *11*, 625–653. Software available at http://sedumi.mcmaster.ca/.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, *58*, 267–288.

Tomioka, R., Aihara, K., & Müller, K.-R. (2007). Logistic Regression for Single Trial EEG Classification. *Advances in Neural Inf. Proc. Systems (NIPS 06)*. MIT press.

Tomioka, R., Hill, J., Blankertz, B., & Aihara, K. (2006). Adapting Spatial Filtering Methods for Nonstationary BCIs. *Proceedings of 2006 Workshop on Information-Based Induction S ciences (IBIS2006)* (pp. 65–70).

A. Derivation of the Dual Problem

The Lagrangian of the primal problem (P) is written as follows:

$$
\begin{aligned}
g(\boldsymbol{\alpha}) = \min_{\substack{W\in\mathbb{R}^{R\times C},\\ b\in\mathbb{R},\\ \boldsymbol{z}\in\mathbb{R}^n}} \Bigg[& \sum_{i=1}^n \ell_{\mathrm{LR}}(z_i) + \lambda \|W\|_1 \\
& + \sum_{i=1}^n \alpha_i \left(z_i - y_i \left(\mathrm{Tr}\left[W^\top X_i\right] + b\right)\right) \Bigg] \\
= & \sum_{i=1}^n \min_{z_i} \left(\ell_{\mathrm{LR}}(z_i) + \alpha_i z_i\right) \\
& + \lambda \min_{W\in\mathbb{R}^{R\times C}} \left(\|W\|_1 - \mathrm{Tr}\left[W^\top \left(\lambda^{-1}\textstyle\sum_{i=1}^n \alpha_i y_i X_i\right)\right]\right) \\
& + \min_{b\in\mathbb{R}} b\left(-\sum_{i=1}^n \alpha_i y_i\right) \\
= & \sum_{i=1}^n \begin{cases} -\ell^*_{\mathrm{LR}}(\alpha_i) & (0\le\alpha_i\le 1) \\ -\infty & (\text{otherwise}) \end{cases} \\
& + \begin{cases} 0 & (\|\sum_{i=1}^n \alpha_i y_i X_i\|_\infty \le \lambda) \\ -\infty & (\text{otherwise}) \end{cases} \\
& + \begin{cases} 0 & (\sum_{i=1}^n \alpha_i y_i = 0) \\ -\infty & (\text{otherwise}), \end{cases}
\end{aligned}
$$

where the dual loss ℓ^*_{LR} is obtained as follows:

$$
\begin{aligned}
\ell^*_{\mathrm{LR}} &:= -\min_z \left(\log(1+e^{-z}) + \alpha z\right) \\
&= \alpha\log\alpha + (1-\alpha)\log(1-\alpha),
\end{aligned}
$$

where the minimum is attained at

$$
z = \log\frac{1-\alpha}{\alpha}. \tag{11}
$$

We obtain the dual problem (D) by negating the objective function to have a minimization and by making the implicit constraints in the above expression explicit.

B. Implementation

B.1. Specializing for the Symmetric Case

We note that if the input matrix is symmetric $X = X^\top \in \mathbb{S}^C$, the weight matrix becomes symmetric as well $W = W^\top \in \mathbb{S}^C$. Thus the LMIs (Eq. (6) and (7)) can be simplified into two smaller LMIs as follows:

from Eq. (6),

$$
Q_1 = Q_2 = \frac{1}{2}U, \quad W \preceq U, \quad -W \preceq U, \tag{12}
$$

from Eq. (7),

$$
\sum_{i=1}^n \alpha_i y_i X_i \preceq \lambda I_C, \quad -\sum_{i=1}^n \alpha_i y_i X_i \preceq \lambda I_C. \tag{13}
$$

B.2. Primal Optimization

First we show a simple implementation of the primal problem (P′) with the symmetry assumption (Eq. (12)) using `CVX` (Grant et al., 2006). `CVX` is a MATLAB toolbox for convex optimization, which is amazingly simple and intuitive to use. Table 1 shows the complete MATLAB code. Currently `CVX` uses SeDuMi (Sturm, 1999) as the core solver. Unfortunately, the nonlinearity in the dual problem (entropy) is currently not handled in `CVX`; this limitation motivated us to develop a specialized implementation that mainly works on the dual variables (see Secs. 2.3 and B.3).

```
function [W, bias, z]=lrl1(X, Y, lmd)
C = size(X,1);  n = length(Y);
cvx_begin sdp
  variable W(C,C) symmetric;
  variable U(C,C) symmetric;
  variable bias;
  variable z(n);
  minimize sum(log(1+exp(-z)))+lmd*trace(U);
  subject to
  for i=1:n
     Y(i)*(trace(W*X(:,:,i))+bias)==z(i);
  end
  U >= W;    U >= -W;
cvx_end
```

Table 1. An executable MATLAB code for primal (P′) based optimization using `CVX` (Grant et al., 2006).

B.3. Dual Optimization

Here, we present some details on the implementation of the dual optimization explained in Sec. 2.3.

The optimality condition for the equality constrained problem ($\mathrm{D_B}$) is written as follows:

$$\frac{\partial \ell^*_{\mathrm{LR}}(\alpha_i)}{\partial \alpha_i} + \mathrm{Tr}\left[F_1 y_i X_i\right] + \mathrm{Tr}\left[F_2\left(-y_i X_i\right)\right] - \beta_i + \beta_i^* + \nu y_i = 0 \quad (i = 1, \ldots, n), \tag{14}$$

$$F_1\left(\lambda I_C - \sum_{i=1}^n \alpha_i y_i X_i\right) := \frac{1}{t} I_C, \tag{15}$$

$$F_2\left(\lambda I_C + \sum_{i=1}^n \alpha_i y_i X_i\right) := \frac{1}{t} I_C, \tag{16}$$

$$\beta_i \alpha_i := \frac{1}{t} \quad (i = 1, \ldots, n), \tag{17}$$

$$\beta_i^* \left(1 - \alpha_i\right) := \frac{1}{t} \quad (i = 1, \ldots, n), \tag{18}$$

$$\sum_{i=1}^n \alpha_i y_i = 0, \tag{19}$$

where $F_1, F_2 \in \mathbb{S}_+^C$ and $\beta_i, \beta_i^* (\geq 0) \in \mathbb{R}$ $(i = 1, \ldots, n)$ are defined as Eqs. (15)-(18) and $\nu \in \mathbb{R}$ is the Lagrangian multiplier for the equality constraint. Equations (14)-(19) are called the *modified KKT conditions* because if we take $t \to \infty$ we obtain the KKT conditions of the original dual problem (D′) with the symmetry assumption (Eq. (13)). It is known that for any $t > 0$, the solution of the modified KKT conditions gives a primal feasible point $(W^*, b^*) := (F_1 - F_2, \nu)$. Moreover, the duality gap associated with the pair of $\boldsymbol{\alpha}$ and (W^*, b^*) obtained by solving Eqs. (14)-(19) for a given t can be shown to be $2(C+n)/t$ (see (Boyd & Vandenberghe, 2004)).

Algorithm 1 Interior Point Method in the Dual

Input: data $\{X_i, y_i\}_{i=1}^n$ Initialize $t := 2(C + n)/(\mu n \log 2)$
repeat
  $t := \mu t$
  Solve the modified KKT conditions (Eqs. (14)-(19)) by Newton's method for the current t with the tolerance ϵ_1.
until $2(C + n)/t < \epsilon_2$

The overall algorithm is shown in Algorithm 1. We start from a small $t = 2(C + n)/(n \log 2)$, which corresponds to the primal-dual gap associated with the trivial pair $(W, b) = (0, 0)$ and $\boldsymbol{\alpha} = \mathbf{0}$. The algorithm solves the modified KKT conditions (14)-(19) for every t and each time t is multiplied by a constant μ; we empirically choose $\mu = 20$. Additionally, we use the tolerances $\epsilon_1 = \epsilon_2 = 10^{-6}$.

B.4. Computational Costs

The efficiency of the specialized dual optimization (Sec. B.3) against the simple `CVX` implementation based on the primal (Sec. B.2) is shown in Fig. 5. Both implementations are run on a single dataset with $n = 125$ samples. The same tolerance $\epsilon = 1.35 \times 10^{-6}$ ("medium") is used. The specialized implementation is 10 times faster when all the 103 electrodes are used. Note that the improvement is not only due to the dual formulation but also the specialization to our problem.

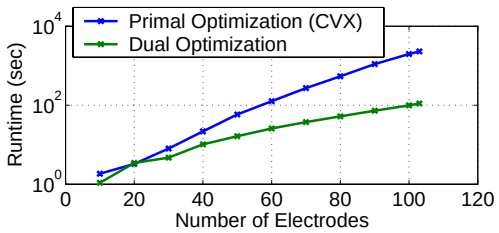

Figure 5. Comparison of the runtimes of the implementations. Both implementations are run on MATLAB 6.5 on a computer with 2.4GHz AMD Opteron and 4GB memory.