Sparse learning with duality gap guarantee

Ryota Tomioka[*]
tomioka@sg.cs.titech.ac.jp

Masashi Sugiyama
sugi@cs.titech.ac.jp

Department of Computer Science,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology,
2-12-1-W8-74, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan.

Abstract

We propose a general regularized empirical risk minimization framework for sparse learning which accommodates popular regularizers such as lasso, group lasso, and the trace norm. Within this framework, we develop two optimization algorithms. The first method is based on squared penalties added to the empirical risk and is solved using a subgradient-based L-BFGS quasi-Newton method. The second method is based on constraints imposed on sparsity-inducing norms and is solved using a gradient projection method. A notable advantage of our approaches is that a simple way to access the dual objective value is available, which is useful in tracking the progress of optimization and deciding when to terminate the optimization procedure.

1 Introduction

Convex regularizers that induce sparsity (e.g., lasso [1], group lasso [2], and trace norm [3, 4, 5]) are useful tools in many applications especially when the interpretability of the learned model is important. They can be used in combination with various convex loss functions. Optimization is convex, which contrasts strikingly to the non-convexity of feature selection in general.

In this paper we focus on the combination of general differentiable convex loss functions and the trace norm regularization. The trace norm[1] is used to reduce the rank of a coefficient matrix in collaborative filtering, multi-input multi-output prediction, and classification over matrices. So far there is no optimization method other than the interior-point method [5, 7], which often scales badly for dense input data, that solves this problem rigorously. Our goal is to generalize recently proposed algorithms for lasso and group lasso to trace norm regularization. In fact considerable effort has been recently made for lasso and group lasso problems. Main difficulty arises from the non-differentiability of these regularizers. Orthant-wise limited-memory quasi-Newton method (OWL-QN) [8] solves the lasso regularization; it uses the steepest descent subgradient (see also [9]) in the L-BFGS quasi-Newton method [10] with a specialized line search that takes the discontinuity of the curvature into account. Roth&Fisher [11] proposed an active set method for group lasso regularization; it uses a gradient projection method suggested in [12] and efficiently identifies a small set of active components and avoids solving a larger problem than actually needed.

Our contributions in this paper are the following three: (i) we propose a general framework that enables us to unify these three sparsity inducing regularizers as special cases of *dual norm regularization*; (ii) we propose two formulations that enable us to access the *duality gap*, which is useful in tracking the progress of the algorithm and deciding when to terminate the algorithm; (iii) we provide practical implementations of the two formulations. The first formulation, which we call the

[*] Part of this work was done while RT was with Fraunhofer FIRST.

[1] It is also known as the dual spectral norm [3], nuclear norm [6], and Ky Fan r-norm [7].

squared penalty formulation is solved with a generalized version of OWL-QN [8] for the trace norm that uses steepest descent subgradients in the L-BFGS quasi Newton method (Sec. 2.1); the second formulation, which we call the *norm constraint formulation* is solved with the gradient projection method [13], which is suggested for group lasso in [12] (Sec. 2.2).

A motivating example. In many previous studies (e.g., [1, 2, 5]), *linear norm penalties* are used in empirical risk minimization framework in order to enforce different types of sparsity. In order to show that there is no straightforward way of assessing the duality gap in this conventional formulation, we present the primal and dual problems of lasso (ℓ_1-) regularization as follows:

$$(\mathrm{P}_0^{\mathrm{lasso}}) \quad \underset{\boldsymbol{w}\in\mathbb{R}^d,\{\boldsymbol{z}_i\}_{i=1}^n}{\text{minimize}} \quad \sum_{i=1}^n \ell_i(z_i) + \lambda_0\|\boldsymbol{w}\|_1, \quad \text{subject to} \quad z_i = \boldsymbol{x}_i{}^\top \boldsymbol{w} \quad (i=1,\dots,n).$$

$$(\mathrm{D}_0^{\mathrm{lasso}}) \quad \underset{\{\boldsymbol{\alpha}_i\}_{i=1}^n}{\text{maximize}} \quad -\sum_{i=1}^n \ell_i^*(\alpha_i), \quad \text{subject to} \quad \|\textstyle\sum_{i=1}^n \alpha_i \boldsymbol{x}_i\|_\infty \le \lambda_0.$$

Here we consider a linear model $z = \boldsymbol{x}^\top \boldsymbol{w}$, which is parameterized with the weight vector $\boldsymbol{w} \in \mathbb{R}^d$; the empirical risk is measured by smooth convex loss functions ℓ_i (e.g., logistic regression) at n input points $\boldsymbol{x}_i \in \mathbb{R}^d$ $(i = 1,\dots,n)$. Note that the labels in a standard supervised learning problem is absorbed into the definition of ℓ_i. $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are ℓ_1- and ℓ_∞-norms, respectively. α_i $(i = 1,\dots,n)$ are Lagrangian multipliers associated with the equality constraints in $(\mathrm{P}_0^{\mathrm{lasso}})$. ℓ_i^* is the Legendre transformation of ℓ_i; i.e., $\ell_i^*(\alpha_i) := \max_{z'_i}(\langle\alpha_i, z'_i\rangle - \ell_i(z'_i))$. Thus due to the differentiability of the loss function ℓ_i, there is a natural mapping between z_i and α_i as follows:

$$\alpha_i = \frac{\partial \ell_i(z_i)}{\partial z_i} \quad \text{and} \quad z_i = \frac{\partial \ell_i^*(\alpha_i)}{\partial \alpha_i}. \tag{1}$$

This mapping suggests that the quantity in the constraint in the dual problem $(\mathrm{D}_0^{\mathrm{lasso}})$ is the gradient of the primal loss term. However we cannot use this mapping to obtain the dual objective value in the linear penalty formulation $(\mathrm{P}_0^{\mathrm{lasso}})$. In fact, α_i obtained using the mapping (Eq. (1)) does not satisfy the dual constraint in general.

There are two formulations with unconstrained dual problems in which we can use the mapping (Eq. (1)) in order to access the dual objective value. The first is the *squared penalty formulation* as follows:

$$(\mathrm{P}_{\mathrm{sq}}^{\mathrm{lasso}}) \quad \underset{\boldsymbol{w}\in\mathbb{R}^d,\{\boldsymbol{z}_i\}_{i=1}^n}{\text{minimize}} \quad \sum_{i=1}^n \ell_i(z_i) + \frac{\lambda}{2}\|\boldsymbol{w}\|_1^2, \quad \text{subject to} \quad z_i = \boldsymbol{x}_i{}^\top \boldsymbol{w} \quad (i=1,\dots,n),$$

$$(\mathrm{D}_{\mathrm{sq}}^{\mathrm{lasso}}) \quad \underset{\{\boldsymbol{\alpha}_i\}_{i=1}^n}{\text{maximize}} \quad -\sum_{i=1}^n \ell_i^*(\alpha_i) - \frac{1}{2\lambda}\|\textstyle\sum_{i=1}^n \alpha_i \boldsymbol{x}_i\|_\infty^2 ,$$

The other is the *norm constraint formulation* as follows:

$$(\mathrm{P}_{\mathrm{con}}^{\mathrm{lasso}}) \quad \underset{\boldsymbol{w}\in\mathbb{R}^d,\{\boldsymbol{z}_i\}_{i=1}^n}{\text{minimize}} \quad \sum_{i=1}^n \ell_i(z_i), \quad \text{subject to} \quad \|\boldsymbol{w}\|_1 \le C, \quad z_i = \boldsymbol{x}_i{}^\top \boldsymbol{w} \quad (i=1,\dots,n),$$

$$(\mathrm{D}_{\mathrm{con}}^{\mathrm{lasso}}) \quad \underset{\{\boldsymbol{\alpha}_i\}_{i=1}^n}{\text{maximize}} \quad -\sum_{i=1}^n \ell_i^*(\alpha_i) - C\|\textstyle\sum_{i=1}^n \alpha_i \boldsymbol{x}_i\|_\infty .$$

In the next section, we generalize these formulations for general sparsity-inducing norms and solve them using the subgradient L-BFGS method (Sec. 2.1) and the gradient projection method (Sec. 2.2), respectively.

2 General sparse learning framework

In this section, we consider the above mentioned two formulations for general sparsity-inducing norms that include lasso, group lasso, and trace norm regularization as special cases. In the squared penalty formulation, the key idea is to define these norms as *dual norms* of other norms; this enables us to compute the subdifferential of the norms in a systematic manner. The dual norm $\|\cdot\|_*$ of a norm $\|\cdot\|$ is defined as $\|\boldsymbol{w}\|_* = \max_{\boldsymbol{q}\in\mathbb{R}^d,\,\|\boldsymbol{q}\|\le 1}\langle\boldsymbol{q},\boldsymbol{w}\rangle$. In the norm constraint formulation, there is an efficient algorithm that computes the projection with the trace norm constraint as well as lasso and group lasso constraints (see Sec. 3).

2.1 Squared penalty formulation

We consider the following regularized empirical risk minimization problem:

$$(\mathrm{P}_{\mathrm{sq}}) \qquad \underset{\boldsymbol{w}\in\mathbb{R}^d,\{\boldsymbol{z}_i\}_{i=1}^n}{\text{minimize}} \qquad \sum_{i=1}^n \ell_i(\boldsymbol{z}_i) + \frac{\lambda}{2}\|\boldsymbol{w}\|_*^2,$$
$$\text{subject to} \qquad \boldsymbol{z}_i = \boldsymbol{x}_i{}^\top \boldsymbol{w} \qquad (i = 1, \ldots, n),$$

We can consider $\boldsymbol{w}$ to be a vector, a vectorized matrix, or a matrix, which correspond, e.g., to standard classification, classification over matrices, and multiclass classification, respectively. Similarly the input $\boldsymbol{x}_i$ can be a vector, a vectorized matrix, or a matrix. Note that we denote by $\|\cdot\|_*$ the norm that measures the complexity (ℓ_1-norm in lasso) as the dual norm of $\|\cdot\|$ (ℓ_∞-norm in lasso). The dual problem of the above optimization problem $(\mathrm{P}_{\mathrm{sq}})$ can be written as follows:

$$(\mathrm{D}_{\mathrm{sq}}) \qquad \underset{\{\boldsymbol{\alpha}_i\}_{i=1}^n}{\text{maximize}} \qquad -\sum_{i=1}^n \ell_i^*(\boldsymbol{\alpha}_i) - \frac{1}{2\lambda}\left\|\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\alpha}_i\right\|^2,$$

where $\boldsymbol{\alpha}_i$ is the Lagrangian multiplier associated with the equality constraints in $(\mathrm{P}_{\mathrm{sq}})$ and has the same dimensionality as $\boldsymbol{z}_i$. Now using the mapping in Eq. (1), the dual objective value can be readily calculated. We define the primal objective function $f^{\mathrm{sq}}(\boldsymbol{w})$ by substituting the "subject to" line into the objective functions in $(\mathrm{P}_{\mathrm{sq}})$ and the dual objective function $g^{\mathrm{sq}}(\{\boldsymbol{\alpha}_i\}_{i=1}^n)$ as in $(\mathrm{D}_{\mathrm{sq}})$.

In general, when a convex function f is non-differentiable at a point $\boldsymbol{w}_0$, the gradient at $\boldsymbol{w}_0$ is not uniquely defined. A subgradient $\boldsymbol{g}$ is the normal vector of a tangent plane of f at $\boldsymbol{w}_0$ as follows:

$$f(\boldsymbol{w}) \geq f(\boldsymbol{w}_0) + \langle \boldsymbol{g}, \boldsymbol{w} - \boldsymbol{w}_0\rangle \qquad \forall \boldsymbol{w} \in \mathbb{R}^d.$$

The *subdifferential* $\partial f(\boldsymbol{w}_0)$ is the set of all the subgradients at $\boldsymbol{w}_0$. The directional derivative $\nabla_{\boldsymbol{d}} f(\boldsymbol{w}_0)$ in the direction $\boldsymbol{d}$ can be computed as follows:

$$\nabla_{\boldsymbol{d}} f(\boldsymbol{w}_0) = \max_{\boldsymbol{g}\in\partial f(\boldsymbol{w}_0)} \langle \boldsymbol{g}, \boldsymbol{d}\rangle\,.$$

Thus the steepest descent direction $\boldsymbol{d}^s$ is obtained as follows:

$$\boldsymbol{d}^s = \underset{\|\boldsymbol{d}\|_2\leq 1}{\mathrm{argmin}} \max_{\boldsymbol{g}\in\partial f(\boldsymbol{w}_0)} \langle \boldsymbol{g}, \boldsymbol{d}\rangle = -\underset{\boldsymbol{g}\in\partial f(\boldsymbol{w}_0)}{\mathrm{argmin}} \|\boldsymbol{g}\|_2, \quad (\|\cdot\|_2 \text{ is the Euclidian norm})$$

because $\min_{\|\boldsymbol{d}\|_2\leq 1}\max_{\boldsymbol{g}\in\partial f(\boldsymbol{w}_0)}\langle \boldsymbol{g}, \boldsymbol{d}\rangle = \max_{\boldsymbol{g}\in\partial f(\boldsymbol{w}_0)}\min_{\|\boldsymbol{d}\|_2\leq 1}\langle \boldsymbol{g}, \boldsymbol{d}\rangle = \max_{\boldsymbol{g}\in\partial f(\boldsymbol{w}_0)}(-\|\boldsymbol{g}\|^2)$. We call $\boldsymbol{g}^s = \mathrm{argmin}_{\boldsymbol{g}\in\partial f(\boldsymbol{w}_0)}\|\boldsymbol{g}\|_2$ the steepest descent subgradient.

We solve the primal problem $(\mathrm{P}_{\mathrm{sq}})$ using the steepest descent subgradient in L-BFGS quasi Newton method as in [8] (see [9] for the general case in which the steepest descent subgradient is not necessarily available). The dual norm formulation is particularly useful in explicitly writing the subdifferential of $f(\boldsymbol{w})$ and finding the steepest descent subgradient. In fact, the subdifferential of a function defined as the point-wise maximum of linear functions is the set of maximizers at each point [13]. The subdifferential $\partial f^{\mathrm{sq}}(\boldsymbol{w})$ of the primal objective $f^{\mathrm{sq}}(\boldsymbol{w})$ is an affine transformation of $\partial\|\boldsymbol{w}\|_*$ as follows:

$$\partial f^{\mathrm{sq}}(\boldsymbol{w}) = \boldsymbol{g}^\ell + \lambda\|\boldsymbol{w}\|_*\partial\|\boldsymbol{w}\|_*, \tag{2}$$

with $\boldsymbol{g}^\ell = \sum_{i=1}^n \partial\ell_i(\boldsymbol{x}_i{}^\top\boldsymbol{w})/\partial\boldsymbol{w}$. Note that even when $-\boldsymbol{g}^s$ is a descent direction, $\boldsymbol{d} = -\boldsymbol{B}\boldsymbol{g}^s$ may not be a descent direction ($\boldsymbol{B}$ is an approximate inverse Hessian), because the directional derivative $\nabla_{\boldsymbol{d}} f(\boldsymbol{w})(\geq \langle \boldsymbol{g}^s, \boldsymbol{d}\rangle)$ can be positive in general; we may switch to gradient descent in this situation because we have the steepest descent direction; the direction-finding algorithm proposed in [9] can also be used.

Remark The proposed formulation is equivalent to the conventional formulation that uses $\lambda'\|\boldsymbol{w}\|_*$ as the penalty term instead of $\lambda\|\boldsymbol{w}\|_*^2$. In fact the subdifferential of the primal objective function in the conventional formulation can be written as $\boldsymbol{g}^\ell + \lambda'\partial\|\boldsymbol{w}\|_*$ with $\boldsymbol{g}^\ell = \sum_{i=1}^n \partial\ell_i(\boldsymbol{x}_i{}^\top\boldsymbol{w})/\partial\boldsymbol{w}$; comparing this expression with Eq. (2), we see that two regularization constants are mapped as $\lambda' = \lambda\|\boldsymbol{w}\|_*$. However our formulation enables us to quickly access the duality gap.

2.2 Norm constraint formulation

The norm constraint formulation for general sparsity-inducing norm $\|\cdot\|_*$ is written as follows:

$$
(\mathrm{P}_{\mathrm{con}}) \qquad \underset{\boldsymbol{w}\in\mathbb{R}^d, \{\boldsymbol{z}_i\}_{i=1}^n}{\text{minimize}} \quad \sum_{i=1}^n \ell_i(\boldsymbol{z}_i),
$$
$$
\text{subject to} \quad \|\boldsymbol{w}\|_* \leq C, \quad \boldsymbol{z}_i = {\boldsymbol{x}_i}^\top \boldsymbol{w} \qquad (i = 1, \ldots, n),
$$

and its dual problem can be written as follows:

$$
(\mathrm{D}_{\mathrm{con}}) \qquad \underset{\{\boldsymbol{\alpha}_i\}_{i=1}^n}{\text{maximize}} \qquad -\sum_{i=1}^n \ell_i^*(\boldsymbol{\alpha}_i) - C \left\| \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{\alpha}_i \right\|.
$$

We define the primal objective function $f^{\mathrm{con}}(\boldsymbol{w})$ by substituting the "subject to" line into the objective functions in $(\mathrm{P}_{\mathrm{con}})$ and the dual objective function $g^{\mathrm{con}}(\{\boldsymbol{\alpha}_i\}_{i=1}^n)$ as in $(\mathrm{D}_{\mathrm{con}})$.

In this formulation, there is no need to consider the subgradient in order to solve the primal problem $(\mathrm{P}_{\mathrm{con}})$ because the primal objective function is differentiable. However, we need to make sure that the iterates stay feasible. To this end, we use the gradient projection method [13] which is suggested for group lasso [12]. That is we first take a negative gradient step with a step-size s_g: $\hat{\boldsymbol{w}} := \boldsymbol{w}_t - s_g \boldsymbol{g}$ and project $\hat{\boldsymbol{w}}$ to the constraint set by solving the following minimization problem:

$$
\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2, \quad \text{subject to} \quad \|\boldsymbol{w}\|_* \leq C. \tag{3}
$$

This minimization problem can be efficiently solved for the trace norm as well as lasso and group lasso (see Sec. 3).

3 Trace norm regularization

In this section, we specialize the above formulations to the trace norm regularization; the weight vector $\boldsymbol{w}$ is seen as a matrix of size $d_r \times d_c$ ($d = d_r d_c$). We define the trace norm as the dual norm of the spectral norm [6] as follows:

$$
\|\boldsymbol{w}\|_* := \max_{\boldsymbol{q}\in\mathbb{R}^{d_r\times d_c}} \langle \boldsymbol{q}, \boldsymbol{w} \rangle \quad \text{s.t.} \quad \|\boldsymbol{q}\| := \max_j \sigma_j(\boldsymbol{q}) \leq 1,
$$

where $\boldsymbol{q} \in \mathbb{R}^{d_r\times d_c}$ and $\sigma_j(\boldsymbol{q})$ is the j-th singular value of $\boldsymbol{q}$; thus $\|\boldsymbol{q}\|$ is the spectral norm of $\boldsymbol{q}$. The above dual norm view is particularly useful in computing the steepest descent subgradient as we see below.

3.1 Subdifferential and the steepest descent subgradient

Subdifferential of the trace norm can now be computed as the set of maximizers $\{\boldsymbol{g}\}$ that achieve $\langle \boldsymbol{g}, \boldsymbol{w} \rangle = \|\boldsymbol{w}\|_*$. Noting that the maximizer is not unique in the null space of $\boldsymbol{w}$, we can write the subdifferential as follows:

$$
\partial\|\boldsymbol{w}\|_* = \Big\{ \boldsymbol{g} = \boldsymbol{u}_1 {\boldsymbol{v}_1}^\top + \boldsymbol{u}_0 \boldsymbol{c} {\boldsymbol{v}_0}^\top \in \mathbb{R}^{d_r\times d_c} : \boldsymbol{w} = \boldsymbol{u}_1 \boldsymbol{\sigma} {\boldsymbol{v}_1}^\top,\ \boldsymbol{u}_0 \perp \boldsymbol{u}_1,\ \boldsymbol{v}_0 \perp \boldsymbol{v}_1,
$$
$$
\boldsymbol{c} = \mathrm{diag}(c_1, \ldots c_{r_0}),\ c_j \in [0,1]\ (j = 1, \ldots, r_0) \Big\},
$$

where $\boldsymbol{w} = \boldsymbol{u}_1 \boldsymbol{\sigma} {\boldsymbol{v}_1}^\top$ is the singular value decomposition of $\boldsymbol{w} \in \mathbb{R}^{d_r\times d_c}$ with $\boldsymbol{u}_1 \in \mathbb{R}^{d_r\times r}$, $\boldsymbol{v}_1 \in \mathbb{R}^{d_c\times r}$, and $\boldsymbol{\sigma} \in \mathbb{R}^{r\times r}$, and r is the rank of $\boldsymbol{w}$ (i.e., $\mathrm{diag}(\boldsymbol{\sigma}) > 0$); the singular vectors $\boldsymbol{u}_0 \in \mathbb{R}^{d_r\times r_0}$ and $\boldsymbol{v}_0 \in \mathbb{R}^{d_c\times r_0}$, where $r_0 = \min(d_r, d_c) - r$, corresponding to the zero singular values can be any orthonormal set of vectors that are orthogonal to $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, respectively. $\boldsymbol{c} \in \mathbb{R}^{r_0\times r_0}$ is a diagonal matrix with diagonal elements in the interval $[0,1]$. The steepest descent subgradient of $f^{\mathrm{sq}}(\boldsymbol{w})$ can be computed as follows. First let $\boldsymbol{g}^\ell := \sum_{i=1}^n \partial\ell_i({\boldsymbol{x}_i}^\top \boldsymbol{w})/\partial\boldsymbol{w}$ be the gradient of the loss term and

$\lambda' := \lambda\|\boldsymbol{w}\|_*$. In addition, we define $\boldsymbol{g}_0^\ell = (\boldsymbol{I}_{d_r} - \boldsymbol{u}_1\boldsymbol{u}_1{}^\top)\boldsymbol{g}^\ell(\boldsymbol{I}_{d_c} - \boldsymbol{v}_1\boldsymbol{v}_1{}^\top)$. Then,

$$\begin{aligned}
\min_{\boldsymbol{g}\in\partial f^{\mathrm{sq}}(\boldsymbol{w})} \|\boldsymbol{g}\|_2^2 &= \min_{\boldsymbol{g}\in\partial\|\boldsymbol{w}\|_*} \left\|\boldsymbol{g}^\ell + \lambda'\boldsymbol{g}\right\|_2^2 \\
&= \min_{\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{c}} \|\boldsymbol{g}^\ell + \lambda'\left(\boldsymbol{u}_1\boldsymbol{v}_1{}^\top + \boldsymbol{u}_0\boldsymbol{c}\boldsymbol{v}_0{}^\top\right)\|_2^2 \\
&= \min_{\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{c}} \|\boldsymbol{g}_0^\ell + \lambda'\boldsymbol{u}_0\boldsymbol{c}\boldsymbol{v}_0{}^\top\|_2^2 + c(\boldsymbol{g}^\ell,\boldsymbol{u}_1,\boldsymbol{v}_1) \\
&= \min_{\boldsymbol{c}\in[0,1]^{r_0}} \textstyle\sum_{j=1}^{r_0}\left(\lambda' c_j - \sigma_j(\boldsymbol{g}_0^\ell)\right)^2 + c(\boldsymbol{g}^\ell,\boldsymbol{u}_1,\boldsymbol{v}_1) \\
&= \sum_{j:\sigma_j(\boldsymbol{g}_0^\ell)>\lambda'} (\lambda' - \sigma_j(\boldsymbol{g}_0^\ell))^2 + c(\boldsymbol{g}^\ell,\boldsymbol{u}_1,\boldsymbol{v}_1),
\end{aligned}$$

where the third equality follows from the orthogonality $\boldsymbol{u}_1\perp\boldsymbol{u}_0$ and $\boldsymbol{v}_1\perp\boldsymbol{v}_0$ and $c(\boldsymbol{g}^\ell,\boldsymbol{u}_1,\boldsymbol{v}_1)$ is a constant with respect to $\boldsymbol{u}_0$, $\boldsymbol{c}$, and $\boldsymbol{v}_0$, which we cannot control. The fourth line follows from von Neumann's inequality, which states for $\boldsymbol{w},\boldsymbol{q}\in\mathbb{R}^{d_r\times d_c}$

$$\langle\boldsymbol{w},\boldsymbol{q}\rangle \leq \textstyle\sum_{j=1}^{\min(d_r,d_c)}\sigma_j(\boldsymbol{w})\sigma_j(\boldsymbol{q}),$$

where $\sigma_j(\cdot)$ is the j-th singular value of a matrix; the equality is obtained when the singular vectors of $\boldsymbol{w}$ and $\boldsymbol{q}$ coincide. In the above minimization, we choose $\boldsymbol{u}_0$ and $\boldsymbol{v}_0$ as the right and left singular vectors of $-\boldsymbol{g}_0^\ell$. Finally the last line is obtained by choosing c_j as $c_j = \min(\sigma_j(\boldsymbol{g}_0^\ell)/\lambda', 1)$.

3.2 Computation of the projection

The projection (Eq. (3)) subject to the bound on the trace norm can be computed as follows. First, similarly to above, using von Neumann's inequality, we obtain

$$\min_{\boldsymbol{w}} \|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2^2 = \min_{\{\sigma_j(\boldsymbol{w})\}_{j=1}^{\bar{r}}} \textstyle\sum_{j=1}^{\bar{r}}\left(\sigma_j(\boldsymbol{w}) - \sigma_j(\hat{\boldsymbol{w}})\right)^2,$$

where $\bar{r} = \min(d_r, d_c)$. The equality is obtained when the singular vectors of $\boldsymbol{w}$ coincide with those of $\hat{\boldsymbol{w}}$. Now all we need to solve is the following minimization problem over the singular values $\{\sigma_j(\boldsymbol{w})\}_{j=1}^{\bar{r}}$:

$$\underset{\{\sigma_j(\boldsymbol{w})\}_{j=1}^{\bar{r}}}{\text{minimize}} \quad \frac{1}{2}\sum_{j=1}^{\bar{r}}\left(\sigma_j(\boldsymbol{w}) - \sigma_j(\hat{\boldsymbol{w}})\right)^2, \quad \text{subject to} \quad \sigma_j(\boldsymbol{w}) \geq 0,\ \sum_{j=1}^{\bar{r}}\sigma_j(\boldsymbol{w}) \leq C,$$

whose dual problem can be written as follows:

$$\underset{\nu\geq 0}{\text{maximize}} \quad -\frac{1}{2}\sum_{j\in J_+}(\nu - \sigma_j(\hat{\boldsymbol{w}}))^2 - C\nu + \sum_{j=1}^{\bar{r}}\frac{1}{2}\sigma_j^2(\hat{\boldsymbol{w}}),$$

where $J_+ := \{j : \sigma_j(\hat{\boldsymbol{w}}) - \nu > 0\}$. For solving the dual problem, we can use the algorithm described in [12, 11] which is a Newton method on the dual variable ν. The algorithm updates the dual variable as $\nu := \frac{\sum_{j\in J_+}\sigma_j(\hat{\boldsymbol{w}}) - C}{|J_+|}$ until convergence. Finally the primal variable can be obtained as $\sigma_j(\boldsymbol{w}) = (\sigma_j(\hat{\boldsymbol{w}}) - \nu)_+$.

4 Experimental results

We apply the methods described above to the six class classification problem in the context of brain-computer interfacing (see [14] for details). The loss function ℓ_i is the negative logarithm of multinomial likelihood, i.e., $\boldsymbol{z}_i = \{z_{i,l}\}_{l=1}^6 \in \mathbb{R}^6$ and $\ell_i(\boldsymbol{z}_i) = -z_{i,y_i} + \log\left(\sum_{l=1}^6 \exp(z_{i,l})\right)$, where $y_i \in \{1,\ldots,6\}$ is the correct class label. Moreover, $\boldsymbol{x}_i \in \mathbb{R}^{6\times(37\times 64)}$ and $\boldsymbol{w}\in\mathbb{R}^{37\times 64}$. The number of samples $n = 2550$. Figure 1 shows the results of the two methods on a simulated data. Figure 2 shows the results of the two methods on a real data provided by J. R. Wolpaw, G. Schalk, and D. Krusienski in the BCI competition III [15]. The L-BFGS method seems to have difficulty in reducing the relative duality gap lower than 10^{-2}; in fact, in both experiments the quasi-Newton direction tends to give insufficient decrease after 500-1000 iterations. However at this point it is not clear whether it is due to the optimization algorithm or the looseness in the evaluation of the lower bound. The gradient projection method performs well for the simulated data but tends to require a large number of iterations for the real data, probably because of the poor scaling of the real problem.

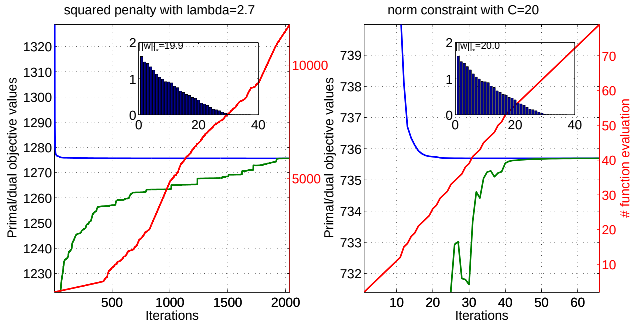

Figure 1: Simulated data. The subgradient L-BFGS method (left) spent 11784 function evaluations and 2240 seconds to achieve RDG=9.7×10^{-6}. The gradient projection method (right) spent 79 function evaluations and 12 seconds to achieve RDG=9.3×10^{-7}. RDG denotes the relative duality gap.

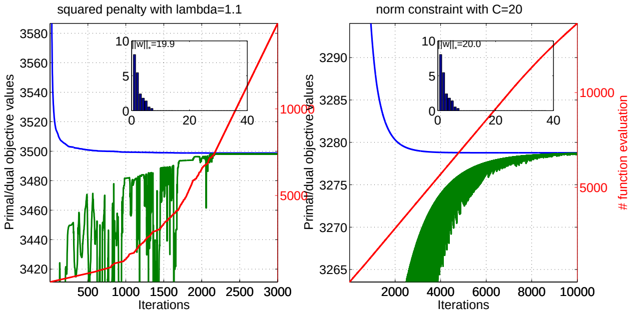

Figure 2: Real data. The subgradient L-BFGS method (left) spent 14935 function evaluations and 2476 seconds to achieve RDG=1.2×10^{-4}. The gradient projection method (right) spent 13650 function evaluations and 2619 seconds to achieve RDG=2.4×10^{-5}. RDG denotes the relative duality gap.

5 Discussion

We have proposed two formulations for sparsity-inducing regularization that includes lasso, group lasso, and trace norm regularization with general convex differentiable loss functions; the proposed formulations enable us to use recently developed techniques for lasso and group lasso, namely subgradient-based L-BFGS method [8, 9] and the gradient projection method [13, 12, 11], also for the trace norm regularization. The differentiability of the loss function provides a simple one-to-one correspondence between primal variable $\boldsymbol{z}$ and dual variable $\boldsymbol{\alpha}$. We have chosen two formulations whose duals are unconstrained, which enables us to access the dual objective value at any primal iterate. This technique may be used in any (including nonconvex) optimization problem that can be separated into several terms, some of which are differentiable.

References

[1] R. Tibshirani, “Regression shrinkage and selection via the lasso”, *J. Roy. Stat. Soc. B*, 58(1): 267–288, 1996.

[2] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables”, *J. Roy. Stat. Soc. B*, 68(1): 49–67, 2006.

[3] M. Fazel, H. Hindi, and S. P. Boyd, “A Rank Minimization Heuristic with Application to Minimum Order System Approximation”, in: *Proc. of the American Control Conference*, 2001.

[4] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola, “Maximum-Margin Matrix Factorization”, in: *Advances in Neural Information Processing Systems 17*, 1329–1336, MIT Press, Cambridge, MA, 2005.

[5] R. Tomioka and K. Aihara, “Classifying Matrices with a Spectral Regularization”, in: *Proc. of the 24th international conference on Machine learning*, 895–902, ACM Press, 2007.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[7] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, “Dimension reduction and coefficient estimation in multivariate linear regression”, *J. Roy. Stat. Soc. B*, 69(3): 329–346, 2007.

[8] G. Andrew and J. Gao, “Scalable training of L1-regularized log-linear models”, in: *Proc. of the 24th international conference on Machine learning*, 33–40, ACM, New York, NY, USA, 2007.

[9] J. Yu, S. V. N. Vishwanathan, S. Günter, and N. N. Schraudolph, “A Quasi-Newton Approach to Nonsmooth Convex Optimization”, 2008, arXiv:0804.3835v3 [stat.ML].

[10] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, 1999.

[11] V. Roth and B. Fischer, “The Group LASSO for Generalized Linear Models: Uniqueness of Solutions and Efficient Algorithms”, in: *Proc. of the 25th international conference on Machine learning*, 2008.

[12] Y. Kim, J. Kim, and Y. Kim, “Blockwise Sparse Regression”, *Stat. Sinica*, 16: 375–390, 2006.

[13] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999, 2nd edition.

[14] R. Tomioka and S. Haufe, “Combined classification and channel/basis selection with L1-L2 regularization with application to P300 speller system”, in: *Proc. of the 4th International BCI Workshop and Training Course*, 2008.

[15] B. Blankertz *et al.*, “The BCI Competition III: Validating Alternative Approaches to Actual BCI Problems”, *IEEE Trans. Neural Sys. Rehab. Eng.*, 14(2): 153–159, 2006, see also the webpage: `http://ida.first.fhg.de/projects/bci/competition_iii/`.