Joint covariate selection for grouped classification

Technical report 743

Department of Statistics, UC Berkeley

Guillaume Obozinski gobo@stat.berkeley.edu
Ben Taskar taskar@cis.upenn.edu
Michael I. Jordan jordan@stat.berkeley.edu

Abstract

We address the problem of recovering a common set of covariates that are relevant simultaneously to several classification problems. We propose a joint measure of complexity for the group of problems that couples covariate selection. By penalizing the sum of ℓ_2-norms of the blocks of coefficients associated with each covariate across different classification problems, we encourage similar sparsity patterns in all models. To fit parameters under this regularization, we propose a blockwise boosting scheme that follows the regularization path. As the regularization coefficient decreases, the algorithm maintains and updates concurrently a growing set of covariates that are simultaneously active for all problems. We show empirically that this approach outperforms independent ℓ_1-based covariate selection on several data sets, both in accuracy and number of selected covariates.

1 Introduction

The problem of covariate selection for regression and classification has been the focus of a substantial literature [9]. As with many model selection problems, the problem is rendered difficult by the disparity between the large number of models to be considered and the comparatively small amount of data available to evaluate these models. One approach to the problem focuses on procedures that search within the exponentially-large set of all subsets of components of the covariate vector, using various heuristics such as *forward* or *backward selection* to limit the search [6]. Another approach treats the problem as a parameter estimation problem in which the shrinkage induced by a constraint on the ℓ_1 norm of the parameter vector yields estimates in which certain components are equal to zero [17, 8, 5]. A virtue of the former approach is that it focuses on the qualitative decision as to whether a covariate is relevant to the problem at hand, a decision which is conceptually distinct from parameter estimation. A virtue of the latter approach is its computational tractability.

In this paper, we focus on a problem setting in which these virtues appear to be better aligned than they are in general regression and classification problems. In particular, we focus on situations involving multiple, related data sets in which the same set of covariates are present in each data set but where the responses differ. In this multi-response setting it is natural to associate a notion of "relevance" to a covariate that is conceptually distinct from the numerical value of a parameter.

For example, a particular covariate may appear with a positive coefficient in predicting one response variable and with a negative coefficient in predicting a different response. We would clearly want to judge such a covariate as being "relevant" to the overall class of prediction problems without making a commitment to a specific value of a parameter. In general we wish to "borrow strength" across multiple estimation problems in order to support a decision that a covariate is to be selected.

Our focus on this paper is on classification. Consider, in particular the following pattern recognition problem. Assume that we are given a data set consisting of pixel-level or stroke-level representations of handwritten characters and we wish to classify a given character into one of a fixed set of classes. In this *optical character recognition* (OCR) problem, there are several thousands of covariates, most of which are irrelevant to the classification decision of character identity. To support the choice of relevant covariates in this high-dimensional problem, we consider an extended version of the problem in which we assume that multiple data sets are available, one for each individual in a set of writers. We expect that even though the styles of individual writers may vary, there should be a common subset of image features (pixels, strokes) that form a shared set of useful covariates across writers.

As another example of our general setting, consider a DNA microarray analysis problem in which the covariates are levels of gene expression and the responses are phenotypes or cellular processes [12]. Given the high-dimensional nature of microarray data sets, covariate selection is often essential both for scientific understanding and for effective prediction. Our proposal is to approach the covariate selection problem by considering multiple related phenotypes—e.g., related sets of cancers—and seeking to find covariates that are useful in predicting these multiple response variables.

Our approach to the simultaneous covariate selection problem is a straightforward adaptation of ℓ_1 shrinkage methods such as LASSO. Briefly, for each data set $\{(x_i^l, y_i^l) : i = 1, \ldots, N_l\}$, where $l \in \{1, \ldots, L\}$ indexes data sets, we fit a model involving a parameter vector w^l. Define the jth *relevance vector* by taking the jth component of each of the parameter vectors, ranging over l. We now define a regularization term that is an ℓ_1 sum of the ℓ_2 norms of the relevance vectors. Each of these ℓ_2 norms can be viewed as assessing the overall relevance of a particular covariate. The ℓ_1 sum then enforces a selection among covariates based on these norms.

This approach is a particular case of a general methodology in which block norms are used to define groupings of variables in regression and classification problems [22, 15, 13, 21, 23]. However, the focus in this literature differs from ours in that it is concerned with grouping variables within a single regression or classification problem. For example, in a polynomial regression we may wish to group the linear, quadratic and cubic terms corresponding to a specific covariate and select these terms jointly. Similarly, in an ANOVA model we may wish to group the indicator variables corresponding to a specific factor. The block-norm approach to these problems is based on defining block norms involving hybrids of ℓ_1, ℓ_2 and ℓ_∞ norms as regularization terms. Our method is a particular application of this general framework to a multi-response setting in which a group is defined by fixing a specific covariate and ranging over multiple response variables.

Given that our motivation is high-dimensional covariate selection, computational efficiency is an important concern. Regularization-based methods generally requiring determining the value of one or more regularization parameters, and naive approaches that evaluate such parameters on a grid can be infeasible computationally. In the linear regression setting, this problem has been addressed via homotopy-based methods that evaluate entire regularization paths efficiently [7, 14]. An important virtue of the block-norm approach is that it is often possible to extend such homotopy-based methods to other loss functions and regularization terms. Indeed, in the current paper we provide a theoretical result establishing the convergence of a homotopy-based method for our particular ℓ_1/ℓ_2 block norm.

In the machine learning literature, the general problem of estimating models from multiple, related data sets is often referred to as "transfer learning," or "multi-task learning," and there have been a variety of proposals for covariate selection in the transfer learning setting [2, 11, 18]. Our work

is similar in spirit to this line of work, but differs in technical detail in its focus on homotopy-based methods and block-norm regularization.

The paper is organized as follows. In Section 2, we present our proposed regularization scheme and the corresponding optimization problem. In Section 3 we discuss homotopy methods as a preliminary to the algorithm that we present in Section 4. We present experimental results in Section 5 and conclude with a discussion in Section 6.

2 Joint complexity measure

We assume a group of L classification problems or "tasks" and a set of data samples $\{(x_i^l, y_i^l) \in \mathcal{X} \times \mathcal{Y},\ i = 1, \ldots, N_l,\ l = 1, \ldots, L\}$ where the superscript l indexes tasks and the subscript i indexes the i.i.d. observations for each task. We assume that the common covariate space $\mathcal{X}$ is $\Re^K$ and the outcome space $\mathcal{Y}$ is $\{0, 1\}$.

Let $w^l \in \Re^K$ be the parameter vector for a statistical model for task l, and let $J^l(w^l, x^l, y^l)$ be a loss function on example (x^l, y^l) for task l. Typical loss functions for linear classification models include log-likelihood, exponential and hinge losses. A standard approach to obtaining sparse estimates of the parameters w^l is to solve a ℓ_1-regularized empirical risk minimization problem:

$$\min_{w^l} \sum_{i=1}^{N_l} J^l(w^l, x_i^l, y_i^l) + \lambda \|w^l\|_1.$$

Solving each of these problems independently across tasks is equivalent[1] to solving the global problem obtained by summing the objectives:

$$\min_{W} \sum_{l=1}^{L} \sum_{i=1}^{N_l} J^l(w^l, x_i^l, y_i^l) + \lambda \sum_{l=1}^{L} \|w^l\|_1, \tag{1}$$

where $W = (w_k^l)_{l,k}$ is the matrix whose rows are the vectors w^l and whose columns are the vectors w_k of the coefficients associated with covariate k across classification tasks. Solving this optimization problem would lead to individual sparsity patterns for each w^l.

Our approach aims instead to select covariates globally. We achieve this by encouraging several w_k to be zero. We thus propose to solve the problem

$$\min_{W} \sum_{l=1}^{L} \sum_{i=1}^{N_l} J^l(w^l, x_i^l, y_i^l) + \lambda \sum_{k=1}^{K} \|w_k\|_2, \tag{2}$$

in which we penalize the ℓ_1-norm of the vector of ℓ_2-norms of the covariate-specific coefficient vectors. In Fig. 1 (Left) we provide a pictorial representation of this regularization. Note that this ℓ_1/ℓ_2 regularization scheme reduces to ℓ_1 regularization if the group is reduced to one task, and can thus be seen an extension of the ℓ_1 regularization where instead of summing the absolute values of coefficients associated with covariates we sum the Euclidean norms of coefficient blocks.

The ℓ_2-norm is used here as a measure of magnitude and one could also generalize to ℓ_1/ℓ_p-norms by considering ℓ_p-norms for $1 \leq p \leq \infty$. The choice of p should depend on how much covariate sharing we wish to impose among classification problems, from none ($p = 1$) to full sharing ($p = \infty$). Indeed, increasing p corresponds to allowing better "group discounts" for sharing the same covariate, from $p = 1$, where the cost grows linearly with the number of classification problems that use a covariate, to $p = \infty$, where only the most demanding classification matters.

[1] Provided the regularization coefficient λ is the same across classification problems.

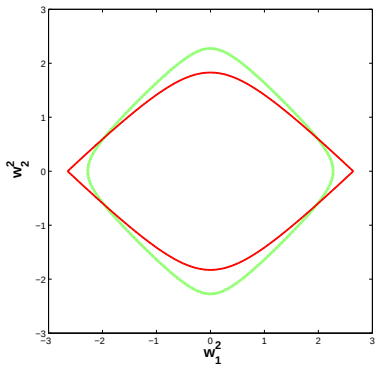

Figure 1: (Left) Norm ball induced on the coefficients (w_1^2, w_2^2) for task 2 as covariate coefficients for task 1 vary: thin red contour for $(w_1^1, w_2^1) = (0, 1)$ and thick green contour for $(w_1^1, w_2^1) = (0.5, 0.5)$.

The shape of the unit "ball" of the ℓ_1/ℓ_2-norm is difficult to visualize. It clearly has corners that, in a manner analogous to the ℓ_1 norm, tend to produce sparse solutions. One way to appreciate the effect of the ℓ_1/ℓ_2 norm is to consider two classification problems based on two covariates and to observe (see Fig. 1) the ball of the norm induced on w^2 when w^1 varies under the constraint that $\|w^1\|_1 = 1$ in an ℓ_1/ℓ_2 ball of size $2\sqrt{2}$ (which is the value of the ℓ_1/ℓ_2 norm if $w_i^1 = w_i^2 = 1$). If a covariate k has a non-zero coefficient in w^1 then the induced norm on w^2 is smooth around $w_k^2 = 0$. Otherwise, it has sharp corners, which encourages w_k^2 to be set to zero.

3 A regularization path

In the optimization problem presented in Eqn. (2) the amount of sparsity is controlled by the regularization coefficient λ. As λ ranges from 0 to ∞, the sparsity of solutions typically progresses through several levels, although this is not guaranteed in general. In most practical cases the appropriate amount of sparsity relevant to a problem is not known in advance and the parameter λ has to be chosen using cross-validation. Re-solving a regularized problem for several values of the parameter is often a waste of computational resources. Interestingly, it has been shown that for a class of regularized problems the regularization path (i.e., the set of all solutions obtained by varying the regularization coefficient) is piecewise linear (cf. [16]). This class includes ℓ_1 and ℓ_2 regularized support vector machines and ℓ_1 regularized least-squares regression. Various algorithms (among them LARS [7]) have been proposed to construct the regularization path efficiently for some of these problems. In general, however, the path is not piecewise linear and can only be approximated. Following the regularization path can be thought of as traversing the parameter space from the origin to the unregularized solution and there is a natural trade-off between speed-of-traversal and closeness to the path. We are interested in simple, fast and scalable algorithms that approximate the regularization path of Eqn. (2). In the case of ℓ_1 regularization it has been noticed that the beginning of the regularization path is very close to the path obtained by coordinate descent (ϵ-boosting or stagewise forward selection [10]). Zhao and Yu [24] have shown that a simple modification of boosting can approximate the LASSO regularization path. In contrast, traditional path-following techniques require the computation of the Hessian of the combined objective so as to move along the path by a combination of *prediction steps* (along the tangent to the path) and *correction steps* (which correct for the first order approximation of the *prediction steps*). Inspired by the work of Zhao and Yu, we capitalize on the sparsity of the solution at the onset of the regularization path and propose an

algorithm in which *correction steps* are extremely cheap computationally.

4 Blockwise boosted LASSO algorithm

An inspection of the Karush-Kuhn-Tucker (KKT) conditions of the optimization problem in Eqn. (2) shows how the sparsity of the solution can lead to an efficient construction of the path. Let us denote $J(W) = \sum_{l=1}^{L} \sum_{i=1}^{N_l} J^l(w^l, x_i^l, y_i^l)$. Then the KKT conditions for Eqn. (2) are as follows:

$$\begin{array}{lll} \text{either} & w_k^* = 0, & \|\nabla_{w_k} J(W^*)\|_2 \leq \lambda \\ \text{or} & w_k^* \propto -\nabla_{w_k} J(W^*), & \|\nabla_{w_k} J(W^*)\|_2 = \lambda, \end{array}$$

where the quantities of interest $\nabla_{w_k} J(W)$ are partial gradients in each of the subspaces of the relevance vectors. In words, only the "active" covariates—those whose gradients are no less than λ—participate in the solution (i.e., have non-zero parameters). In particular if $\lambda \geq \lambda_0 = \max_k \|\nabla_{w_k} J(0)\|_2$ then the zero vector is a solution to our problem.

This analysis suggests an algorithm in which we gradually decrease the regularization coefficient from λ_0 and populate an active set with inactive covariates as they start to violate the approximate KKT conditions. We write the latter generically as follows:

$$\begin{array}{rl} \text{either} & w_k = 0, \quad \|\nabla_{w_k} J(W)\| < \lambda + \xi_0 \\ \text{or} & \left\| \nabla_{w_k} J(W) + (\lambda - \xi) \frac{w_k}{\|w_k\|} \right\| \leq \xi, \end{array} \tag{3}$$

where ξ_0 is a slack parameter. Algorithm 1 uses this idea to maintain the constraints in Eqn. (3) with $\xi_0 = 0$, updating the regularization parameter only if none of the inactive covariates violate the KKT conditions at the end of the previous iteration.

Algorithm 1 Maintain approximate KKT conditions

while $\lambda^t > \lambda_{min}$ **do**
  Set $j^* = \text{argmax}_k \|\nabla_{w_k} J(W^t)\|$
  Update $w_{j^*}^{(t+1)} = w_{j^*}^{(t)} - \epsilon u^t$ with $u^t = \frac{\nabla_{w_{j^*}} J}{\|\nabla_{w_{j^*}} J\|}$
  if $\|\nabla_{w_{j^*}} J(W^t)\| > \lambda^t$ **then**
    $\lambda^{t+1} = \lambda^t$
  else
    $\lambda^{t+1} = \min(\lambda^t, \frac{J(W^t) - J(W^{t+1})}{\epsilon})$
  end if
  Add j^* to the active set
  Enforce Eqn. (3) only for covariates of the active set
end while

With an appropriate smoothness assumption on the gradient of J, we can derive a simpler algorithm which is similar to the generalized Boosted LASSO algorithm of Zhao and Yu [24]. Indeed, assuming J to be twice differentiable and assuming that the spectrum of the Hessian of J is bounded above by μ_{max}, i.e., $\forall x,\ x^\top H x \leq \mu_{max} \|x\|^2$, Algorithm 2 maintains Eqn. (3) for any ξ_0 such that $\xi_0 \geq \frac{1}{2} \epsilon \mu_{max}$.

Algorithm 2 Blockwise boosted LASSO

while $\lambda^t > \lambda_{min}$ **do**
  Set $j^* = \text{argmax}_k \|\nabla_{w_k} J(W^t)\|$
  Update $w_{j^*}^{(t+1)} = w_{j^*}^{(t)} - \epsilon u^t$ with $u^t = \frac{\nabla_{w_{j^*}} J}{\|\nabla_{w_{j^*}} J\|}$
  $\lambda^{t+1} = \min(\lambda^t, \frac{J(W^t) - J(W^{t+1})}{\epsilon})$
  Add j^* to the active set
  Enforce (3) only for covariates of the active set
end while

Proposition 1. *Assuming J to be twice differentiable and strictly convex, for all η there exists ϵ such that iterates W^t of Algorithm 2 obey $\|W^t - W(\lambda^t)\| \leq \eta$ for every t such that $\lambda^{t+1} < \lambda^t$, where $W(\lambda^t)$ is the unique solution to Eqn. (2). Moreover, the algorithm converges (provided the active set is not pruned) in a finite number of iterations to a regularization coefficient no greater than any prespecified $\lambda_{min} > 0$.*

A proof of this proposition is presented in the appendix.

Remarks:

- For both algorithms the step in the direction u^t performs descent on J. In practice we set $u^t = \nabla_{w_{j^*}} J / \|\nabla_{w_{j^*}} J\|$. It is also possible to take a coordinate descent step, where, if A is the active set, $u^t = \nabla_{W_{A \cup j^*}} J / \|\nabla_{W_{A \cup j^*}} J\|_{\ell_1/\ell_2}$. Other possibilities include *gradient-related* descent directions.[2] In particular u^t could be along the gradient or be closer to a typical *prediction step* and use the Hessian of the joint objective restricted to the active set.

- For simplicity, we have presented the algorithms using a fixed step size ϵ, but in practice we recommend using an adaptive step size determined by a line search limited to the segment $(0, \epsilon]$. This allows us to explore the end of the path where the regularization coefficient becomes exponentially small. Our proofs extend to this case (cf. Lemma 3).

- If we understand the "active set" as the set of covariates with non-zero coefficients it is possible for a covariate to enter and later exit the set, which would require pruning. The analysis of this case is more delicate and we do not consider it here. In practice, the case of parameters returning to zero is rare, and its consideration would not yield a significant speed-up of the algorithm.

- Each of the two algorithms that we have presented has its own advantages. For Algorithm 1, the step size can be taken relatively large without influencing the precision of the approximation to the path at points where the regularization decreases. On the other hand, for a relatively small step size, Algorithm 2 maintains a more steady progression along the path.

- In the formulation of the algorithms we have assumed that we can solve the problem of enforcing the constraints (3) on the active set. In our implementations, we use blockwise coordinate descent to implement such a solution. We have to be cautious here because the objective we are optimizing is non-differentiable. Pathological examples of continuous non-differentiable functions have been produced for which steepest descent combined with exact line search does not converge to a stationary point ([3], p.633). In our case, however, as we show in Lemma 6, these pathologies cannot arise and the method necessarily converges to a global minimum of the objective function.

[2] I.e., directions such that $\liminf_t -u^t \cdot \frac{\nabla J(W^t)}{\|\nabla J(W^t)\|} > \delta > 0$ (cf. [3], p. 35).

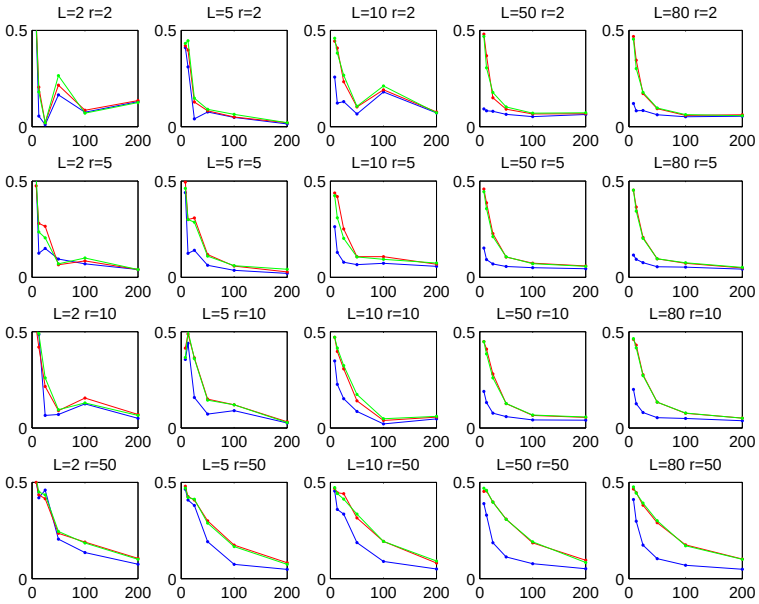

Figure 2: Misclassification error represented as a function of the number n of samples used for training, in plots with increasing number of tasks (from left to right: $L = 2, 5, 10, 50, 80$) and increasing number of discriminative covariates (from top to bottom: $r = 2, 5, 10, 50$ out of 300 covariates total) and for three different algorithms based on either independent ℓ_1 regularization (green), ℓ_1/ℓ_1 regularization (red) or ℓ_1/ℓ_2 regularization (blue).

5 Experiments and applications

We first investigate the behavior of the algorithm with simulated data that satisfies the assumptions underlying our model and analysis. We then turn to experiments with real data, focusing on optical character recognition. We also briefly present extensions to multi-class classification and subspace selection.

5.1 Synthetic data

We consider a number L of binary classifications tasks all based on a common covariate space of dimension K. A common subset of $r \ll K$ covariates defines a common subspace $\mathcal{D}$ within which each pairs of classes can be discriminated. For every classification task, both class-conditional densities are Gaussian in $\mathcal{D}$ and each of the $K - r$ other dimensions consist of noise uniformly distributed on the interval $[0, 1]$. The covariance matrix of each class is drawn from an $r \times r$-dimensional Wishart distribution, $\mathcal{W}(r, r, Id)$, with r degrees of freedom. The two classes are separated by a vector $\delta = \mu_1 - \mu_0$ constructed as follows: a random vector is drawn uniformly in $\{-1, 0, 1\}^r \backslash \{\mathbf{0}\}$ and then normalized so that so that the mean of the Mahalanobis distances for both covariance matrices is a fixed value $c = \frac{1}{2}\sqrt{\delta^\top \Sigma_0 \delta} + \frac{1}{2}\sqrt{\delta^\top \Sigma_1 \delta}$. We picked $c = 3$ in our experiments which corresponds to well-separated classes. Note that by construction, the coordinates of δ are only non-zero on a subset of the r common dimensions, so that the set of covariates that separates the classes is not exactly the same for each classification.

The results are shown in Fig. 2, where we compare independent ℓ_1, ℓ_1/ℓ_1 and ℓ_1/ℓ_2 regular-

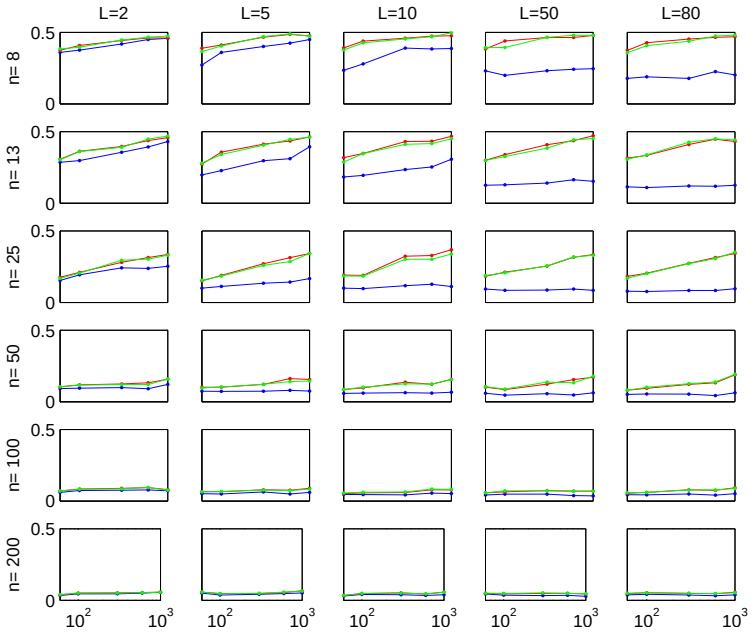

Figure 3: Average misclassification error represented as a function on the log scale of the total number K of covariates, for a fixed number $r = 10$ of discriminative covariates, in plots with increasing number of tasks (from left to right: $L = 2, 5, 10, 50, 80$) and increasing number of datapoints (from top to bottom: $n = 8, 13, 25, 50, 100, 200$) and for three different algorithms based on either independent ℓ_1 regularization (green), ℓ_1/ℓ_1 regularization (red) or ℓ_1/ℓ_2 regularization (blue).

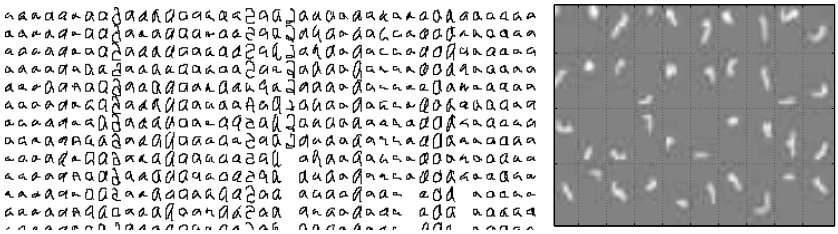

Figure 4: (Left) The letter a written by 40 different people. (Right) Strokes extracted from the data.

izations. The results indicate that the ℓ_1/ℓ_1 and independent ℓ_1 regularizations perform almost identically. This is not surprising because the essential difference between the behavior of these two regularizations is due to the coupling of the regularization coefficients λ_k in the ℓ_1/ℓ_1 case. But the classification problems we generated are of equal difficulty, which means that the amount of regularization that is needed for each problem is presumably the same. On the other hand we see from Fig. 2 that the ℓ_1/ℓ_2 regularization achieves systematically better results, with dramatic improvements for small training set sizes. The relative improvement is generally larger for small training sets, but as the number of informative dimensions increases the best training set size increases as well.

Fig. 3 illustrates that ℓ_1/ℓ_2 is more robust to the number of noisy dimensions than the other regularizations, and suggests that the growth of the error is roughly linear with $\log K$ but that the slope decreases significantly with the number of tasks.

5.2 Writer-specific character recognition

In this section, we consider an application to the problem of the optical character recognition (OCR) of handwritten characters. Consider the problem of discriminating between pairs of letters for different writers. The simplest approach is to pool all the letters from all writers and build a global classifier for each pair; this may be justifiable if we obtain only a few examples of each letter per writer, but enough different writers. Another naive method is to learn a classifier for each writer independently. We compare these naive methods to our ℓ_1/ℓ_2 regularization method.

Data

We used letters from a handwritten words data set by Rob Kassel at the MIT Spoken Language Systems Group.[3] This data set contains samples from more than 180 different writers (see Fig. 4 (Left) for examples). For each writer, however, the number of examples of each letter is rather small: between 4 and 30 depending on the letter. As shown in Fig. 5, the letters are originally represented as 8×16 binary pixel images.

Experimental setup

We built binary classifiers that discriminate between pairs of letters. Specifically we concentrated on the pairs of letters that are difficult to distinguish when written by hand. We compared four discriminative methods, all based on minimizing a loss function, with different regularization schemes as follows:

[3] Available at www.seas.upenn.edu/∼taskar/ocr/.

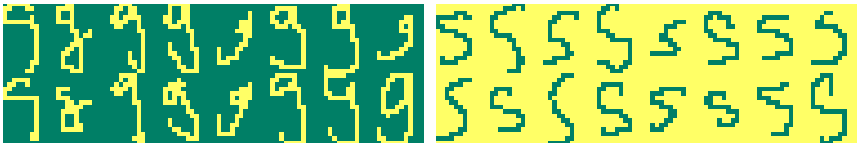

Figure 5: Samples of the letters s and g for one writer.

- **Pooled ℓ_1:** The writers are ignored and all the letters of both classes to be discriminated are pooled. We use Algorithm 2 to follow the regularization path of the ℓ_1-regularized logistic regression.
- **Independent ℓ_1 regularization:** For each writer an independent ℓ_1-regularized logistic regression is learned using Algorithm 2 with blocks of size 1.
- **ℓ_1/ℓ_1-regularization:** The objective function is Eqn. (1) with the log-loss and so the tasks are tied only by the regularization coefficient. The regularization path is obtained by Algorithm 2 with blocks of size 1, where the updates of covariates used in different tasks are interwoven.
- **ℓ_1/ℓ_2-regularization:** The objective is Eqn. (2) with the log-loss; here the covariate selection processes are coupled by the regularization. The regularization path is obtained by Algorithm 2.

Covariates: pixels and strokes

The basic covariates we use are the 8×16 binary pixels. Since individual pixels are often uninformative, we also used a simple, ad hoc procedure to generate combinations of contiguous pixels (strokes) that appeared in the images. To produce a stroke, we selected a random image and a random filled pixel and we followed a biased random walk on the filled pixels of the image. We used an second-order Gaussian Markov model of strokes in which the velocity varies slowly to bias for low-curvature lines. We followed walks of lengths 2, 4 and 6. To produce realistically thick strokes we also included the pixels of the letters that are neighbors of the stroke. The obtained stroke was finally smoothed by convolution with a simple kernel combining only neighboring pixels. To construct a set of strokes for the task of discriminating between two letters we extracted 500 strokes in the training set from letters of each of these two types and 100 strokes from other letter types as well. The total number of strokes we generated in each of our experiments was on the order of a thousand. The "strokes" selected by our algorithm for the g vs s classification are shown in Fig. 4 (Right).

Results

We fitted classification models for discriminating 9 pairs of letters for 40 different writers according to the four schemes presented in Section 5.2. We conducted experiments with the two types of covariate sets proposed (pixels and strokes). The error rates of the classifiers obtained are reported in Table 1.

For the pixel covariates, the ℓ_1/ℓ_2 regularization method improves significantly on pooling and on the other regularization methods. Indeed, it improves in all cases except one, with an improvement over ℓ_1 regularization that is greater than 50% in many cases.

For the stroke covariates the improvement due to the ℓ_1/ℓ_2 regularization is less pronounced. There is a clear improvement over pooling and over ℓ_1/ℓ_1; on the other hand, ℓ_1 and ℓ_1/ℓ_2 regularizations perform comparably.

Table 1: Average 0-1 loss on the test set. The bold font indicates the best performing method in each row among the different methods: ℓ_1/ℓ_2, ℓ_1/ℓ_2, ℓ_1/ℓ_1, independent (id.) ℓ_1 or pooled ℓ_1.

	strokes : error(%)				pixels: error (%)			
Task	ℓ_1/ℓ_2	ℓ_1/ℓ_1	id.ℓ_1	pool	ℓ_1/ℓ_2	ℓ_1/ℓ_1	id.ℓ_1	pool
c/e	**2.5**	3.0	3.3	3.0	**4.0**	8.5	9.0	4.5
g/y	8.4	11.3	**8.1**	17.8	**11.4**	16.1	17.2	18.6
g/s	3.3	3.8	**3.0**	10.7	**4.4**	10.0	10.3	6.9
m/n	4.4	4.4	**3.6**	4.7	**2.5**	6.3	6.9	4.1
a/g	**1.4**	2.8	2.2	2.8	**1.3**	3.6	4.1	3.6
i/j	**8.9**	9.5	9.5	11.5	12.0	14.0	14.0	**11.3**
a/o	**2.0**	2.9	2.3	3.8	**2.8**	4.8	5.2	4.2
f/t	**4.0**	5.0	6.0	8.1	**5.0**	6.7	6.1	8.2
h/n	**0.9**	1.6	1.9	3.4	**3.2**	14.3	18.6	5.0

Our interpretation of these results is that classifiers based on the weaker features (pixels) benefit more from the sharing among tasks than those based on the stronger features (strokes). As support for this interpretation, consider Fig. 6, where we present the parameters w^l learned for each of the 40 writers. The top two rectangles contain the parameters for the pixel covariates, with the results from ℓ_1/ℓ_2 regularization on the left and the results from independent ℓ_1 regularization on the right. It is clear that the sharing induced by the ℓ_1/ℓ_2 regularization has yielded parameters that are more discriminative in this case. On the other hand, in the case of stroke covariates (the lower two rectangles), we see that the parameters induced by independent ℓ_1 are already quite discriminative; thus, there appears to be less to gain from shrinkage among tasks in this case. Note also (from Table 1) that the overall error rate from the classifiers based on pixels is significantly higher than that of the classifiers based on strokes. Finally, for this problem pooling does not perform well presumably because the inter-writer variance of the letters is large compared to the inter-class variance.

Another advantage of the ℓ_1/ℓ_2-regularization is that it yields a more compact representation than the other methods (with the exception of pooling). This is particularly noticeable for the stroke representation where fewer than 50 features are typically retained for the ℓ_1/ℓ_2-regularization versus versus three to five times as many for the other regularization schemes.

5.3 Multi-class classification

Multi-class classification can be viewed as a multiple response problem in which a set of responses share a set of covariates. This is certainly an appropriate perspective if the multi-class classification problem is approached (as is often done) by fitting a set of binary classifiers, but it is also appropriate if a single multi-class classifier is fit by a single "polychotomous" logistic regression. In either case, it may be useful to find covariates that are useful across the set of discriminations. Our ℓ_1/ℓ_2 regularization applies directly to this setting; indeed, the methodology that we have presented thus far makes no reference to the fact that the loss function is a sum of losses across tasks. We can thus replace this loss function with any joint loss function (e.g., the polychotomous logit). In the remainder of this section we investigate the use of ℓ_1/ℓ_2 regularization in two multi-class classification domains.

5.3.1 Digit classification

We conducted a multi-class classification experiment using the "multi-feature digit" data set from the University of California Irvine repository [19]. This data set of 2000 entries contains 200 examples of each of the 10 digits. The data are represented by 649 covariates of different types (76 Fourier coefficients, 216 profile correlations, 64 Karhunen-Love coefficients, 240 pixel averages in 2×3 windows, 47 Zernike moments and 6 morphological features). We compared models based on

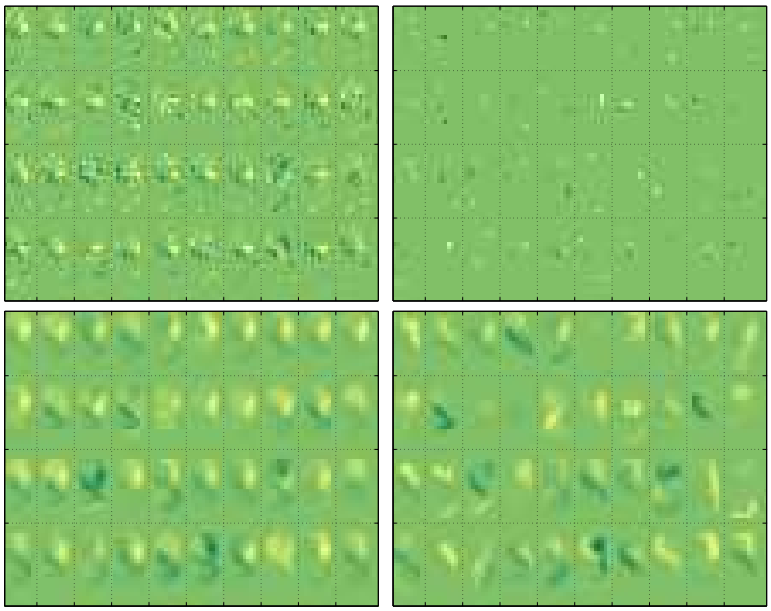

Figure 6: Plots of the discriminative masks learned for the classification of g vs s under ℓ_1/ℓ_2 regularization (Left) and independent ℓ_1 regularization (Right), based on either pixel covariates (Top) or stroke covariates (Bottom). The better masks capture the closure of the circle in g and the diagonal stroke of s as discriminative features of these letters.

polychotomous logistic regression fitted with ℓ_1/ℓ_2 and ℓ_1/ℓ_1 regularizations and the classification obtained by combining individually regularized logistic regressions (using the ℓ_1 norm). To focus on the data-poor regime in which regularization methods would appear to be of most value, we used only 1/10 of the data to fit the model and retained the rest for testing. We replicated the experiment ten times.

Our results indicate that ℓ_1/ℓ_2 regularization is clearly superior for this problem compared to the other regularization methods. The average error rate obtained was 2.9% ($\hat{\sigma} = 0.24\%$) for ℓ_1/ℓ_2, versus 4.2% ($\hat{\sigma} = 0.65\%$) for ℓ_1/ℓ_1 and 4.1% ($\hat{\sigma} = 0.65\%$) for separate binary classifications.

5.3.2 Classification of cancers

The diagnosis of complex diseases such as cancer can be assisted by genomic information provided by expression microarrays; specifically, microarrays allow us to identify genes that are differentially expressed in different cell lineages or at different stages of a cancer. This is interesting because the relationship between gene expression patterns and the illness is more direct than that of somatic symptoms, but it is also difficult because of the large number of genes and the high levels of noise present in the data. We used the ℓ_1/ℓ_2, ℓ_1/ℓ_1 and independent ℓ_1 regularizations to differentiate four types of skin cancers (studied by [12]) based on gene expression data.

In terms of predictive performance, all three of these regularization schemes performed as well as the best-performing methods studied by [12] and [20]. However, ℓ_1/ℓ_2 regularization achieved this result with a smaller set of non-zero parameters than the other methods: there were 57, 81 and 85 contributing genes to the classifier based on ℓ_1/ℓ_2, ℓ_1/ℓ_1 and independent ℓ_1, respectively. This small gene signature is obviously of importance in the biological setting, where simpler/cheaper tests are desirable and where predictively-important genes may be prioritized for further study. Note also that the parameter values obtained from ℓ_1/ℓ_2 regularization were different qualitatively from those obtained via the other regularizations (see Fig. 7). We found that a striking feature of the sparsity pattern obtained from ℓ_1/ℓ_2 was that several genes used by the other regularizations were eliminated because if the expression of a gene is indicative of a cancer type, then that covariate is encouraged to be also more discriminative for the other cancers. This might be an efficient way to eliminate competing correlated predictors.

5.4 Subspace selection

While we have focused on the problem of selecting covariates, it is also of interest to consider possible extensions of our approach to the problem of selecting general subspaces (i.e., linear combinations of covariates). That is, we may wish to consider situations in which a subspace that is useful across multiple tasks is not aligned with the original covariate coordinate system, such that the models are sparse in a rotated coordinate system. This situation is considered in [1, 2].

We can treat the subspace selection problem within our formalism by making use of random projections. Specifically, we propose the following approach: map all data sets to $\Re^d$ using a common set of d projections on one-dimensional random subspaces. In this new representation of the data, use ℓ_1/ℓ_2 regularization to perform joint covariate selection. The covariates selected in $\Re^d$ correspond to a common relevant subset of directions in the original space. Intuitively, we would expect for this procedure to find projections that are useful across tasks, thus uncovering a common subspace linking the tasks.[4]

[4] In a further theoretical development that is beyond the scope of the present manuscript, we have been able to show that the random projection method approaches a trace-norm regularization as the number of projections grow. This connects the random projection approach with the work of Argyriou et al. [2], who propose to directly optimize the trace norm as the basis of a subspace selection method. Numerically, the approach based on random projections would appear to have some advantages, because the trace-norm regularization is a non-differentiable problem and

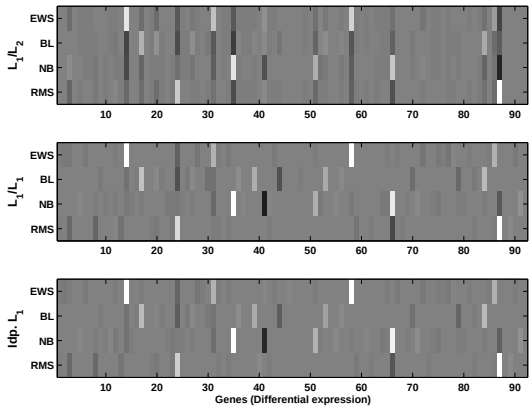

Figure 7: Matrix of parameters obtained from three regularization methods. The ℓ_1/ℓ_2, ℓ_1/ℓ_1 and independent ℓ_1 regularizations use 57, 81 and 85 (respectively) contributing genes to classify four cancer types: EWS, BL, NB, RMS. Note that the ℓ_1/ℓ_2 regularization has an interesting "mikado" pattern (i.e., with alternating, contrasted coefficients columnwise) indicating that a given feature has important opposite effects in the classification of two classes that it discriminates well.

Table 2: Average 0-1 loss on the test set with random projections as a preprocessing step. The bold font indicates the best performing method in each row among the different methods: ℓ_1/ℓ_2, ℓ_1/ℓ_1, independent (id.) ℓ_1 or pooled ℓ_1. The boxed entry indicates conditions in which there was an improvement over the feature selection results in Table 1.

	strokes : error(%)				pixels: error (%)			
Task	ℓ_1/ℓ_2	ℓ_1/ℓ_1	id.ℓ_1	pool	ℓ_1/ℓ_2	ℓ_1/ℓ_1	id.ℓ_1	pool
c/e	**2.0**	3.5	3.3	2.5	**3.5**	7.8	10.3	4.5
g/y	10.3	10.3	**9.3**	16.9	11.6	**9.7**	10.9	21.4
g/s	3.8	4.0	**2.5**	12.0	4.7	6.7	5.0	6.4
m/n	4.1	5.8	**3.6**	5.3	**1.9**	2.8	4.1	
a/g	**0.8**	1.6	1.3	2.5	**0.8**	1.7	1.4	3.9
i/j	**9.2**	9.8	11.1	11.3	**10.3**	12.7	13.5	11.5
a/o	2.7	2.7	**1.9**	4.3	**2.1**	3.1	3.5	4.2
f/t	5.8	**4.1**	5.5	7.5	**6.4**	11.1	9.6	7.1
h/n	0.9	0.6	**0.3**	3.7	**1.8**	3.6	5.0	5.0

Returning to the OCR problem, we conducted an experiment that was identical to the previous experiment, but in which 500 random projections were used to transform the pixel covariates into a new covariate space. Similarly, in the case of the strokes covariates we used 3000 projections. In both cases this yielded roughly four times as many projections as there were dimensions of the original covariate space. The results of this experiment are shown in Table 2. We see that the subspace selection yields an improvement over the earlier feature selection results in the case of the pixel covariates.

6 Discussion

We have presented a regularization scheme for joint covariate selection in grouped classification, where several classification models are fitted simultaneously and make simultaneous choices of rele-

involves the implicit solution of an eigenvalue problem.

vant covariates. We have also proposed a scalable algorithm that follows the regularization path for this problem.

We should emphasize that although classification has been the focus of our presentation, the approach is generic and applies immediately to problems based on other smooth loss functions, including least squares regression and more broadly generalized linear models. More generally, any norm inducing sparse solutions can benefit from a similar approach.

We should also point out that there is a natural extension of our setting to a sequential (online) version of grouped classification. In this case, tasks are presented one after another and each new model is encouraged to share the same sparsity pattern as previous classifiers. This can be achieved by minimizing the sum of the loss on the new task plus a joint ℓ_1/ℓ_2 norm combined with previous classification parameters.

We showed empirically in several examples that the proposed regularization methodology can improve overall classification performance and can also aid in interpretability. We found that the method is particularly useful in cases in which covariates are noisy and in which the amount of data for each classification task is limited.

References

[1] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Proceedings of the 20th Neural Information Processing Systems Conference*. MIT Press, Cambridge, MA, 2006.

[3] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, Nashua, NH, 1999.

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

[5] D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Technical Report 2004–10, Statistics Department, Stanford University, 2004.

[6] N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley-Interscience, New-York, 1998.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[8] W. Fu and K. Knight. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.

[9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[10] R. Hastie, T. Tibshirani and J. Friedman. *Elements of Statistical Learning*. Springer, Berlin, 2001.

[11] T. Jebara. Multi-task feature and kernel selection for SVMs. In *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA, 2004.

[12] J. Khan, J. Wei, M. Ringnér, and al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.

[13] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. Technical Report 131, Mathematics Department, Swiss Federal Institute of Technology Zürich, 2007.

[14] M.R. Osborne, B. Presnell, and B.A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–403, 2000.

[15] M.Y. Park and T. Hastie. Regularization path algorithms for detecting gene interactions. Technical Report 2006-13, Department of Statistics, Stanford University, 2006.

[16] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35(3):1012–1030, 2007.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

[18] A. Torralba, K.P. Murphy, and W.T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762–769. IEEE Computer Society, Washington, DC, 2004.

[19] M. van Breukelen, R.P.W. Duin, D.M.J. Tax, and J.E. den Hartog. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4):381–386, 1998.

[20] B. Wu. Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics*, 22(5):472–476, 2005.

[21] Kim Y., Kim J., and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375–390, 2006.

[22] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49–67, 2006.

[23] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. Technical Report 703, Statistics Department, UC Berkeley, 2007.

[24] P. Zhao and B. Yu. Boosted lasso. Technical Report 702, Statistics Department, UC Berkeley, 2006.

A Appendix

In this appendix we prove Proposition 1, showing that the path-following algorithms that we have presented guarantee steady progress along the path and guaranteeing that the latter is well approximated. While we focus on Algorithm 2, it is also worth noting that with minor modifications the proof also goes through for Algorithm 1.

The proof proceeds via a sequence of lemmas. Lemma 2 justifies the update rule $\lambda^{t+1} = \min(\lambda^t, \epsilon^{-1}[J(W^t) - J(W^{t+1})])$ by showing that it ensures that each time the regularization coefficient λ^t is updated, the solution satisfies approximate KKT conditions and is thus, by Lemma 1, reasonably close to the path. The algorithm is designed to move along the path smoothly in parameter space, by taking a bounded step. Lemmas 3 and 4 establish that the progression is steady in terms of λ^t and that the algorithm terminates after a finite number of steps. More precisely, Lemma 3 shows that the regularization decreases by at least a constant amount $\epsilon\mu_{min}$ at almost each iteration and therefore becomes smaller than $\epsilon\mu_{min}$ after a finite number of steps. Lemma 4 establishes additionally that even the part of the path corresponding to small values of the regularization can be reached efficiently after a finite number of steps if a bounded line search method is used to determine the step size of the descent steps on J. Lemmas 5 and 6 show that steepest descent converges for our regularized objective, in spite of its non-differentiability.

All lemmas assume that J is convex, continuously twice differentiable ($\mathcal{C}^2$) and that, as a consequence, the spectrum of its Hessian is uniformly bounded above and below respectively by $\mu_{\max}$ and $\mu_{\min}$ on some fixed compact set. Lemmas 3 and 4 assume that Algorithm 2 is used without pruning the active set A (i.e., once a point is inserted in A it stays in A). For a function F, we denote by $\partial F(x)$ the set of subgradients of the function at x and $\partial_k F(x)$ the set of subgradients in the k^{th} subspace.

Lemma 1. *Let T be any convex function, and $G(x) = \lambda T(x) + J(x)$. Then let $g \in \partial G(x)$ be a subgradient of G at x and x^* the unique minimum of G , then*

$$\|x^* - x\| \leq 2\frac{\|g\|}{\mu_{min}}$$

Proof. This is a simple extension of a standard result in optimization (cf. [4], pp. 459-460). Combining a Taylor expansion of J with a convexity inequality for the norm we get:

$$\begin{aligned} \exists \xi, \quad J(x^*) &\geq J(x) + \nabla J(x)(x^* - x) + \frac{1}{2}(x^* - x)^\top H(\xi)(x^* - x) \\ T(x^*) &\geq T(x) + t^\top (x^* - x) \qquad \text{with } t \in \partial T(x) \end{aligned}$$

So that with $g = t + \nabla J(x)$

$$\begin{aligned} \exists \xi, \quad G(x^*) &\geq G(x) + g^\top (x^* - x) + \frac{1}{2}(x^* - x)^\top H(\xi)(x^* - x) \\ 0 \geq G(x^*) - G(x) &\geq g^\top (x^* - x) + \frac{1}{2}\mu_{min}\|x^* - x\|^2 \\ \frac{1}{2}\mu_{min}\|x^* - x\|^2 &\leq \|g\|\|x^* - x\| \end{aligned}$$

which yields the desired result.

Lemma 2. *Let ξ_0 in Eqn. (3) satisfy $\xi_0 \geq \max(2\xi, \frac{1}{2}\epsilon\mu_{\max})$. Then for all t such that $\lambda^{t+1} < \lambda^t$ the approximate KKT conditions hold just before the gradient step at iteration t; as a consequence $\|W^t - W(\lambda^t)\| \leq \sqrt{K}\frac{2\xi_0}{\mu_{\min}}$ where $W(\lambda^t)$ is the optimal solution of Eqn. (2) for the regularization coefficient λ^t.*

Proof. The approximate KKT conditions are explicitly enforced by the algorithm in the active set. Moreover, $\forall k \notin A$, $w_k = 0$ and

$$\|\nabla_{w_k} J(W^t)\| \leq \|\nabla_{w_{j^*}} J(W^t)\| \leq \|\nabla J(W^t)\| \leq \frac{1}{\epsilon}(J(W^t) - J(W^{t+1})) + \frac{1}{2}\epsilon\mu_{\max} \leq \lambda^t + \xi_0$$

using a second order Taylor expansion of J. This shows the first part of the lemma. As we argue now, these approximate KKT conditions imply that there exists a subgradient of our regularized objective of size at most $\sqrt{K}\xi_0$, which by Lemma 1 implies the result. Indeed for every covariate k such that $w_k \neq 0$, $\|\nabla_{w_k} J(W)\| \leq 2\xi + \lambda$; then for every covariate such that $w_k = 0$, since the subgradient set of $\lambda\|\cdot\|_2$ at 0 is the Euclidean ball of radius λ, given that $\|\nabla_{w_k} J(W)\| \leq \lambda + \xi_0$, one can choose a subgradient of the ℓ_2-norm such that the corresponding partial subgradient of the regularized objective with respect to w_k is of norm less than ξ_0. Since the subgradient of the norms can be chosen independently in each subspace, we have a subgradient $g = (g_1, \cdots, g_K)$ such that $\max_k \|g_k\| \leq \xi_0$ and therefore $\|g\| \leq \sqrt{K}\xi_0$.

Lemma 3. *If we use steps of fixed size ϵ, after a finite number of steps λ^t becomes smaller than $\frac{1}{2}\epsilon\mu_{min}$.*

Proof. Except for a number of iterations bounded by K, at the beginning of each iteration of the algorithm, we have $\|\nabla_{w_{j_t^*}} J(W^t)\| \leq \lambda^t$. Indeed, any active covariate k satisfies $\|\nabla_{w_k} J(W^t)\| \leq \lambda^t$ after the approximate KKT conditions are enforced at the end of the previous iteration, and if some inactive covariate has a gradient larger than λ^t then the largest gets incorporated in the active set, which can only happen once for every covariate if there is no pruning. For all steps t such that $\|\nabla_{w_{j_t^*}} J(W^t)\| \leq \lambda^t$, if the step taken is $\epsilon\, u^t$ with u^t a unit vector in subspace k, then, using a Taylor expansion of J, the update of the regularization satisfies

$$\lambda^{t+1} = \frac{J(W^t) - J(W^{t+1})}{\epsilon} \leq \|\nabla_{w_{j_t^*}} J(W^t)\| - \frac{\epsilon}{2}\mu_{min} \leq \lambda^t - \frac{\epsilon}{2}\mu_{min}$$

So if steps of fixed size ϵ are used, then, after a finite number of steps λ^t becomes smaller than $\frac{1}{2}\epsilon\mu_{min}$.

Lemma 4. *If, given the direction* $u^t = \frac{\nabla_{w_{j_t^*}} J(W^t)}{\|\nabla_{w_{j_t^*}} J(W^t)\|}$, *we choose a step size* $\epsilon_t \leq \epsilon$ *which maximizes the decrease* $J(W^t) - J(W^{t+1})$, *then* $\lim_t \lambda^t \leq 2\xi$.

Proof. The beginning of the previous argument is still valid and so there exists t_0 such that $\forall t > t_0$, $\lambda^{t+1} \leq \lambda^t - \frac{1}{2}\epsilon_t \mu_{min}$. So ϵ_t converges to 0. In particular, there exists t_1 such that $\forall t > t_1$, $\epsilon_t < \epsilon$. But if $\epsilon_t < \epsilon$:

$$J(W^t) - J(W^{t+1}) \leq \epsilon_t \nabla_{w_{j_t^*}} J(W^{t+1}) \cdot u^t + \frac{1}{2}\epsilon_t^2 \mu_{max} = \frac{1}{2}\epsilon_t^2 \mu_{max}$$

the last equality being due to the fact that the minimizer is in the interior of $[0, \epsilon]$. But then using $J(W^t) - J(W^{t+1}) \geq \epsilon_t \delta \|\nabla_{w_{j_t^*}} J(W^t)\| - \frac{\epsilon_t^2}{2}\mu_{max}$ we finally get $\lambda^t - 2\xi \leq \|\nabla_{w_{j_t^*}} J(W^t)\| \leq \frac{\epsilon_t}{\delta}\mu_{max} \underset{t}{\rightarrow} 0$.

Lemma 5. *For any non-differentiable point* W^0 *of* $G = J + \lambda\|\cdot\|_{\ell_1/\ell_2}$ *with* $w_k = 0$ *and a descent direction in the* k^{th} *subspace, there exists a neighborhood* $\mathcal{B}_\epsilon$ *of* W^0 *such that, for all* $W \in \mathcal{B}_\epsilon$, *a steepest descent step with line search executed in the the* k^{th} *subspace leads to a point* V^* *such that* $G(W^0) > G(V^*)$.

Proof. Let $\mathcal{B}_\epsilon = \{W \mid \max_j \|w_j - w_j^0\| \leq \epsilon\}$. For clarity denote $g^0 = \nabla_k J(W^0)$, $g = \nabla_k J(W)$, $g_v = \nabla_k J(V)$. Fix η_0 and let $\epsilon_1 = \frac{\|g^0\| + \eta_0}{\lambda}\epsilon$. Since J is assumed $\mathcal{C}_2$, we can choose ϵ such that ϵ_1 is small enough that for all V in $\mathcal{B}_{\epsilon_1}$, $\|g_v - g^0\| < \eta$. Note that the existence of a descent direction in the k^{th} subspace implies $\lambda < \|g^0\|$ and so, provided $\eta_0 < \lambda - \|g^0\|$, then we have $\epsilon_1 \leq \epsilon$ and $\mathcal{B}_\epsilon \subset \mathcal{B}_{\epsilon_1}$.

Since we perform a line search from W, note that the new component in the k^{th} subspace, v_k^*, is of the form $v_k(\gamma) = w_k - \gamma\left(\lambda\frac{w_k}{\|w_k\|} + g\right)$. Consider the suboptimal point in the steepest direction of subspace k given by $v_k = v_k\left(\frac{\|w_k\|}{\lambda}\right) = -\frac{\|w_k\|}{\lambda}g$. Note that for V (resp. V^*) obtained from W by updating coordinate k to $v_k = -\frac{\|w_k\|}{\lambda}g$ (resp. v_k^*), we have $G(V^*) \leq G(V)$.

Moreover $V \in \mathcal{B}_{\epsilon_1}$ since

$$\max_j \|v_j - w_j^0\| \leq \max(\max_j \|w_j - w_j^0\|, \frac{\|g\|\|w_k\|}{\lambda}) < \epsilon^1.$$

But $\forall\, V \in \mathcal{B}_{\epsilon_1}$ by convexity of G,

$$G(W^0) - G(V) \geq -v_k^\top\left(\lambda\frac{v_k}{\|v_k\|} + g_v\right) \geq -\lambda\|v_k\| - v_k^\top g^0 - \eta\|v_k\|$$

and since $v_k = -\gamma g$ with $\gamma = \frac{\|w_k\|}{\lambda}$,

$$\frac{1}{\gamma}\left[(\lambda + \eta)\|v_k\| + v_k^\top g^0\right] \leq (\lambda + \eta)\|g\| - g^\top g^0 \leq \left[(\lambda + 3\eta) - \|g^0\|\right]\|g\|,$$

which is negative as soon as $\eta < \eta_0 \leq \frac{1}{3}(\|g^0\| - \lambda)$.

So for $\eta < \eta_0 \leq \frac{1}{3}(\|g^0\| - \lambda)$, $G(V^*) - G(W^0) \leq G(V^*) - G(W^0) < 0$.

Lemma 6. *For any strictly convex function* $J \in \mathcal{C}_2$ *and* $G = J + \lambda\|\cdot\|_{\ell_1/\ell_2}$, *then blockwise coordinate descent on* G *based on partial directional derivatives combined with exact line searches converges to the global minimum of* G.

Proof. Let $W^{t,k} = (w_1^t, \cdots, w_k^t, w_{k+1}^{t-1}, \cdots, w_K^{t-1})$ and let $g^{t,k} \in \partial_k G(W^{t,k-1})$ be the steepest descent direction.[5] For any cycle t and update of the component in subspace k, the following inequality can be obtained from a Taylor expansion and convexity inequalities:

$$\exists \tilde{W}^{t,k}, \qquad G(W^{t,k-1}) \geq G(W^{t,k}) - \alpha_{t,k} {s^{t,k+1}}^\top g^{t,k} + \frac{1}{2}\alpha_{t,k}^2 {g^{t,k}}^\top H_J^{kk}(\tilde{W}^{t,k}) g^{t,k}$$

where $s^{t,k+1} \in \partial_k G(W^{t,k})$ can be chosen such that ${s^{t,k+1}}^\top g^{t,k} = 0$: indeed, if there was no such $s^{t,k+1}$, given the convexity of the subgradient set, this would contradict that $g^{t,k}$ is a minimum of G along the descent direction. We used the notation H_J^{kk} for the restriction of the Hessian of J to the k^{th} subspace and denoted $\tilde{W}^{t,k-1}$ a point on the segment $[W^{t,k-1}, W^{t,k}]$. Given that $(W^{t,k-1})_t$ is in a compact set (because G is coercive) we can extract a converging subsequence. On a compact set we have ${g^{t,k}}^\top H_J^{kk}(\tilde{W}^{t,k}) g^{t,k} \geq \mu_{\min} \|g^{t,k}\|^2$ because J is assumed strictly[6] convex and $\mathcal{C}^2$ so that its spectrum is lower bounded. On a subsequence converging to W^*, since it decreases monotonically, we therefore have:

$$G(W^{0,k-1}) - G(W^*) \geq \sum_i G(W^{t_i,k-1}) - G(W^{t_i,k}) \geq \frac{1}{2}\mu_{\min} \sum_i \alpha_{t_i,k}^2 \|g^{t_i,k}\|^2. \qquad (*)$$

For any $W^{t,k-1}$ such that the closed segment joining w_k^{t-1} and w_k^t doesn't contain the zero vector, two full Taylor expansions yield:

$$\begin{aligned}
\exists \bar{W}^{t,k}, \qquad G(W^{t,k}) - G(W^{t,k-1}) &= -\alpha_{t,k}\|g^{t,k}\| + \frac{1}{2}\alpha_{t,k}^2 {g^{t,k}}^\top H_G^{kk}(\bar{W}^{t,k}) g^{t,k}. \\
\exists \tilde{W}^{t,k}, \qquad G(W^{t,k-1}) - G(W^{t,k}) &= \frac{1}{2}\alpha_{t,k}^2 {g^{t,k}}^\top H_G^{kk}(\tilde{W}^{t,k}) g^{t,k}.
\end{aligned}$$

For the latter, since $w_k^t \neq 0$, G has a partial differential at $W^{t,k}$ in the k^{th} subspace which is orthogonal to $g^{t,k}$, hence the missing linear term. We can assume $\alpha_{t_i,k} \neq 0$ (otherwise, either we can extract a further subsequence with that property, or the subsequence is finitely convergent to a stationary point). Combining the Taylor expansions and using $H_G^{kk}(W) = H_J^{kk}(W) + \frac{1}{\|w_k\|}(Id - \frac{w_k w_k^\top}{\|w_k\|^2})$ we get

$$1 = \frac{1}{2}\alpha_{t_i,k} \frac{{g^{t_i,k}}^\top \left[H_G(\tilde{W}^{t_i,k}) - H_G(\bar{W}^{t_i,k})\right] g^{t_i,k}}{\|g^{t_i,k}\|^2} \leq \frac{1}{2}\alpha_{t_i,k} \left[\left(\mu_{\max} + \frac{1}{\|\tilde{w}_k^{t_i}\|}\right) - \mu_{\min}\right].$$

Since in the limit $(w_k^{t_i})_i$ is lower bounded and, since $\|w_k^t - w_k^{t+1}\| \underset{t}{\to} 0$, so is $(\tilde{w}_k^{t_i})_i$ asymptotically. The previous inequality then implies that $\alpha_{t_i,k}$ is lower bounded as well and therefore, from Eqn. $(*)$, $\|g^{t_i,k}\| \underset{i}{\to} 0$. So either $w_k^{t_i} \underset{i}{\to} 0$ or $\|g^{t_i,k}\| \underset{i}{\to} 0$. This characterizes any accumulation point of the sequence $(W^{t,k})_t$. Note that, for any k_0 and any converging subsequence $(W^{t_i,k_i})_i$, we have $\|W^{t_i,k_i} - W^{t_i,k_0}\| \underset{i}{\to} 0$ since for all k, $\|w_k^t - w_k^{t+1}\| \underset{t}{\to} 0$, which means that the accumulation points of the overall sequence are the same as the accumulation points of any sequence $(W^{t,k})_t$ with k fixed. Furthermore, given our assumptions on J, G is strictly convex and therefore, once a set A^c of components for which $w_k^{t_i} \underset{i}{\to} 0$ is specified, only one accumulation point is possible because $g^{t_i,k} \underset{i}{\to} 0$ implies that $(W_A^{t_i})_i$ converges to a point which is the unique minimum of $W_A \mapsto G(0_{A^c}, W_A)$ where $(0_{A^c}, W_A)$ is a full matrix W in the domain of G where all components of subspaces indexed by $k \in A^c$ are set to 0 (note that this is even true if $(W_A^{t_i})_i$ converges to a point of non-differentiability

[5] We use subgradient notation for the steepest descent direction since the directional derivative of f in direction d can be calculated from the subgradients as $f'(x,d) = sup_{s \in \nabla f(x)} s^\top d$.

[6] Note that we actually only use the strict convexity in the direction of the gradient.

of $G(0_{A^c}, \cdot)$). This implies that there are at most a finite number of accumulation points possible (in particular at most 2^K).
We now argue that the only accumulation point possible is the minimum of G. Since G is strictly convex, at most one of the potential accumulation point has no descent direction. But by Lemma 5 any potential accumulation point W^0 with a descent direction has a neighborhood such that, for any iterate in that neighborhood, the next iterate V satisfies $G(V) < G(W^0)$, which implies that W^0 cannot be an accumulation point since the sequence is decreasing in G. The only accumulation point possible is therefore the minimum.