Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex Sets

Authored by:

Jieping Ye
Jie Wang

Abstract

Sparse-Group Lasso (SGL) has been shown to be a powerful regression technique for simultaneously discovering group and within-group sparse patterns by using a combination of the l1 and l2 norms. However, in large-scale applications, the complexity of the regularizers entails great computational challenges. In this paper, we propose a novel two-layer feature reduction method (TLFre) for SGL via a decomposition of its dual feasible set. The two-layer reduction is able to quickly identify the inactive groups and the inactive features, respectively, which are guaranteed to be absent from the sparse representation and can be removed from the optimization. Existing feature reduction methods are only applicable for sparse models with one sparsity-inducing regularizer. To our best knowledge, TLFre is the first one that is capable of dealing with multiple sparsity-inducing regularizers. Moreover, TLFre has a very low computational cost and can be integrated with any existing solvers. Experiments on both synthetic and real data sets show that TLFre improves the efficiency of SGL by orders of magnitude.

1 Paper Body

Sparse-Group Lasso (SGL) [5, 16] is a powerful regression technique in identifying important groups and features simultaneously. To yield sparsity at both group and individual feature levels, SGL combines the Lasso [18] and group Lasso [28] penalties. In recent years, SGL has found great success in a wide range of applications, including but not limited to machine learning [20, 27], signal processing [17], bioinformatics [14] etc. Many research efforts have been devoted to developing efficient solvers for SGL [5, 16, 10, 21]. However, when the feature dimension is extremely high, the complexity of the SGL regularizers imposes great computational challenges. Therefore, there is an increasingly urgent need for nontraditional techniques to address the challenges posed by the massive volume of the data sources. Recently, El Ghaoui et al. [4] proposed a promising feature reduction method, called SAFE screening, to screen out

the so-called inactive features, which have zero coefficients in the solution, from the optimization. Thus, the size of the data matrix needed for the training phase can be significantly reduced, which may lead to substantial improvement in the efficiency of solving sparse models. Inspired by SAFE, various exact and heuristic feature screening methods have been proposed for many sparse models such as Lasso [25, 11, 19, 26], group Lasso [25, 22, 19], etc. It is worthwhile to mention that the discarded features by exact feature screening methods such as SAFE [4], DOME [26] and EDPP [25] are guaranteed to have zero coefficients in the solution. However, heuristic feature screening methods like Strong Rule [19] may mistakenly discard features which have nonzero coefficients in the solution. More recently, the idea of exact feature screening has been extended to exact sample screening, which screens out the nonsupport vectors in SVM [13, 23] and LAD [23]. As a promising data reduction tool, exact feature/sample screening would be of great practical importance because they can effectively reduce the data size without sacrificing the optimality [12]. 1

However, all of the existing feature/sample screening methods are only applicable for the sparse models with one sparsity-inducing regularizer. In this paper, we propose an exact two-layer feature screening method, called TLFre, for the SGL problem. The two-layer reduction is able to quickly identify the inactive groups and the inactive features, respectively, which are guaranteed to have zero coefficients in the solution. To the best of our knowledge, TLFre is the first screening method which is capable of dealing with multiple sparsity-inducing regularizers. We note that most of the existing exact feature screening methods involve an estimation of the dual optimal solution. The difficulty in developing screening methods for sparse models with multiple sparsity-inducing regularizers like SGL is that the dual feasible set is the sum of simple convex sets. Thus, to determine the feasibility of a given point, we need to know if it is decomposable with respect to the summands, which is itself a nontrivial problem (see Section 2). One of our major contributions is that we derive an elegant decomposition method of any dual feasible solutions of SGL via the framework of Fenchel?s duality (see Section 3). Based on the Fenchel?s dual problem of SGL, we motivate TLFre by an in-depth exploration of its geometric properties and the optimality conditions. We derive the set of the regularization parameter values corresponding to zero solutions. To develop TLFre, we need to estimate the upper bounds involving the dual optimal solution. To this end, we first give an accurate estimation of the dual optimal solution via the normal cones. Then, we formulate the estimation of the upper bounds via nonconvex optimization problems. We show that these nonconvex problems admit closed form solutions. Experiments on both synthetic and real data sets demonstrate that the speedup gained by TLFre in solving SGL can be orders of magnitude. All proofs are provided in the long version of this paper [24]. Notation: Let k ? k1 , k ? k and k ? k? be the ‘1 , ‘2 and ‘? norms, respectively. Denote by B1n , B n , and n the unit ‘1 , ‘2 , and ‘? norm balls in Rn (we omit the superscript if it is clear from the context). B? For a set C, let int C be its interior. If C is closed and convex, we define the projection operator as PC (w) := argminu?C kw ? uk. We denote by IC (?) the indicator function of C, which is 0 on C and

? elsewhere. Let ?0 (Rn) be the class of proper closed convex functions on Rn . For f ? ?0 (Rn), let ?f be its subdifferential. The domain of f is the set dom f := {w : f (w) ¡ ?}. For w ? Rn , let [w]i be its ith component. For ? ? R, let sgn(?) = sign(?) if ? 6= 0, and sgn(0) = 0. We define

sign([w]i), if [w]i 6= 0; SGN(w) = s ? Rn : [s]i ? [?1, 1], if [w]i = 0. We denote by ?+ = max(?, 0). Then, the shrinkage operator S? (w) : Rn ? Rn with ? ? 0 is [S? (w)]i = (—[w]i — ? ?)+ sgn([w]i), i = 1, . . . , n.

2

(1)

Basics and Motivation

In this section, we briefly review some basics of SGL. Let y ? RN be the response vector and X ? RN ?p be the matrix of features. With the group information available, the SGL problem [5] is

2 XG XG ?

1

minp y ? X ? ng k?g k + ?2 k?k1 , (2) g g + ?1

g=1 g=1 ??R 2 where ng is the number of features in the g th group, Xg ? RN ?ng denotes the predictors in that group with the corresponding coefficient vector ?g , and ?1 , ?2 are positive regularization parameters. Without loss of generality, let ?1 = ?? and ?2 = ? with ? ¿ 0. Then, problem (2) becomes:

2

X XG

G ? 1

min y? Xg ? g + ? ? ng k?g k + k?k1 . (3) g=1 g=1 ??Rp 2 By the Lagrangian multipliers method [24], the dual problem of SGL is n o

2 ? sup 12 kyk2 ? 12 y? ? ? : XTg ? ? Dg? := ? ng B + B? , g = 1, . . . , G .

(4)

?

It is well-known that the dual feasible set of Lasso is the intersection of closed half spaces (thus a polytope); for group Lasso, the dual feasible set is the intersection of ellipsoids. Surprisingly, the geometric properties of these dual feasible sets play fundamentally important roles in most of the existing screening methods for sparse models with one sparsity-inducing regularizer [23, 11, 25, 4]. When we incorporate multiple sparse-inducing regularizers to the sparse models, problem (4) indicates that the dual feasible set can be much more complicated. Although (4) provides a geometric 2

description of the dual feasible set of SGL, it is not suitable for further analysis. Notice that, even the feasibility of a given point ? is not easy to determine, since it is nontrivial to tell if XTg ? can ? be decomposed into b1 + b2 with b1 ? ? ng B and b2 ? B? . Therefore, to develop screening methods for SGL, it is desirable to gain deeper understanding of the sum of simple convex sets. In the next section, we analyze the dual feasible set of SGL in depth via the Fenchel?s Duality Theorem. We show that for each XTg ? ? Dg? , Fenchel?s duality naturally leads to an explicit decom? position XTg ? = b1 + b2 , with

one belonging to ? ng B and the other one belonging to B? . This lays the foundation of the proposed screening method for SGL.

3

The Fenchel?s Dual Problem of SGL

In Section 3.1, we derive the Fenchel?s dual of SGL via Fenchel?s Duality Theorem. We then motivate TLFre and sketch our approach in Section 3.2. In Section 3.3, we discuss the geometric properties of the Fenchel?s dual of SGL and derive the set of (?, ?) leading to zero solutions. 3.1 The Fenchel?s Dual of SGL via Fenchel?s Duality Theorem To derive the Fenchel?s dual problem of SGL, we need the Fenchel?s Duality Theorem as stated in Theorem 1. The conjugate of f ? ?0 (Rn) is the function f ? ? ?0 (Rn) defined by f ? (z) = supw hw, zi ? f (w). Theorem 1. [Fenchel?s Duality Theorem] Let f ? ?0 (RN), ? ? ?0 (Rp), and T (?) = y ? X? be an affine mapping from Rp to RN . Let p? , d? ? [??, ?] be primal and dual values defined, respectively, by the Fenchel problems: p? = inf ??Rp f (y ? X?) + ??(?); d? = sup??RN ?f ? (??) ? ??? (XT ?) + ?hy, ?i. One has p? ? d? . If, furthermore, f and ? satisfy the condition 0 ? int (dom f ? y + Xdom ?), then the equality holds, i.e., p? = d? , and the supreme is attained in the dual problem if finite. We omit the proof of Theorem 1 since it is a slight modification of Theorem 3.3.5 in [2]. Let f (w) = 21 kwk2 , and ??(?) be the second term in (3). Then, SGL can be written as min? f (y ? X?) + ??(?). To derive the Fenchel?s dual problem of SGL, Theorem 1 implies that we need to find f ? and ?? . It is well-known that f ? (z) = 21 kzk2 . Therefore, we only need to find ?? , where the concept infimal convolution is needed. Let h, g ? ?0 (Rn). The infimal convolution of h and g is defined by (hg)(?) = inf ? h(?) + g(? ? ?), and it is exact at a point ? if there exists a ? ? (?) such that (hg)(?) = h(? ? (?)) + g(? ? ? ? (?)). hg is exact if it is exact at every point of its domain, in which case it is denoted by h g. PG ? ? Lemma 2. Let ?? 1 (?) = ? g=1 ng k?g k, ?2 (?) = k?k1 and ?(?) = ?1 (?) + ?2 (?). More? over, let Cg? = ? ng B ? Rng , g = 1, . . . , G. Then, the following hold: PG PG ? ? (i): (?? (?2)? (?) = g=1 IB? (?g), 1) (?) = g=1 ICg (?g) ,

PG ? ?P (?) ? (ii): ?? (?) = ((?? (?2)?) (?) = g=1 IB g ??Bn?g g , 1) where ?g ? Rng is the sub-vector of ? corresponding to the g th group. Note that PB? (?g) admits a closed form solution, i.e., [PB? (?g)]i = sgn ([?g]i) min (—[?g]i — , 1). Combining Theorem 1 and Lemma 2, the Fenchel?s dual of SGL can be derived as follows. Theorem 3. For the SGL problem in (3), the following hold: (i): The Fenchel?s dual of SGL is given by:

? inf 12 k y? ? ?k2 ? 12 kyk2 : XTg ? ? PB? (XTg ?) ? ? ng , g = 1, . . . , G . ?

(5)

(ii): Let ? ? (?, ?) and ?? (?, ?) be the optimal solutions of problems (3) and (5), respectively. Then, ??? (?, ?) =y ? X? ? (?, ?), ? XTg ?? (?, ?) ?? ng ?k?g? (?, ?)k + ?k?g? (?, ?)k1 , g = 1, . . . , G. 3

(6) (7)

Remark 1. We note that the shrinkage operator can also be expressed by S? (w) = w ? P?B? (w), ? ? 0.

(8)

Therefore, problem (5) can be written more compactly as

? inf 21 k y? ? ?k2 ? 12 kyk2 : S1 (XTg ?) ? ? ng , g = 1, . . . , G .

(9)

?

Remark 2. Eq. (6) and Eq. (7) can be obtained by the Fenchel-Young inequality [2, 24]. They are the so-called KKT conditions [3] and can also be obtained by the Lagrangian multiplier method [24]. Moreover, for the SGL problem, its Lagrangian dual in (4) and Fenchel?s dual in (5) are indeed equivalent to each other [24]. Remark 3. An appealing advantage of the Fenchel?s dual in (5) is that we have a natural decomposition of all points ?g ? Dg? : ?g = PB? (?g) + S1 (?g)) with PB? (?g) ? B? and S1 (?g) ? Cg? . As a result, this leads to a convenient way to determine the feasibility of any dual variable ? by checking if S1 (XTg ?) ? Cg? , g = 1, . . . , G. 3.2 Motivation of the Two-Layer Screening Rules We motive the two-layer screening rules via the KKT condition in Eq. (7). As implied by the name, there are two layers in our method. The first layer aims to identify the inactive groups, and the second layer is designed to detect the inactive features for the remaining groups. by Eq. (7), we have the following cases by noting ?kwk1 = SGN(w) and o (n w , if w 6= 0, kwk ?kwk = {u : kuk ? 1}, if w = 0. Case 1. If ?g? (?, ?) 6= 0, we have (? [? ? (?,?)] i ? ng k?g? (?,?)k + sign([?g? (?, ?)]i), if [?g? (?, ?)]i 6= 0, T ? g [Xg ? (?, ?)]i ? [?1, 1], if [?g? (?, ?)]i = 0.

(10)

In view of Eq. (10), we can see that ? ? ? ? (?1 ,?2) (a): S1 (XTg ?? (?, ?)) = ? ng k?g? (?1 ,?2)k and kS1 (XTg ?? (?, ?))k = ? ng , g

(b): If [XTg ?? (?, ?]i ? 1 then [?g? (?, ?)]i = 0. Case 2. If

?g? (?, ?)

(11) (12)

= 0, we have ? [XTg ?? (?, ?)]i ? ? ng [ug]i + [?1, 1], kug k ? 1.

(13)

The first layer (group-level) of TLFre From (11) in Case 1, we have

S1 (XTg ?? (?, ?)) ¡ ??ng ? ?g? (?, ?) = 0.

(R1)

Clearly, (R1) can be used to identify the inactive groups and thus a group-level screening rule. The second layer (feature-level) of TLFre Let xgi be the ith column of Xg . [XTg ?? (?, ?)]i = xTgi ?? (?, ?). In view of (12) and (13), we can see that T ?

xg ? (?, ?) ? 1 ? [?g? (?, ?)]i = 0. i

We have (R2)

Different from (R1), (R2) detects the inactive features and thus it is a feature-level screening rule. However, we cannot directly apply (R1) and (R2) to identify the inactive groups/features because both need to know ?? (?, ?). Inspired by the SAFE rules [4], we can first estimate a region ? containing ?? (?, ?). Let XTg ? = {XTg ? : ? ? ?}. Then, (R1) and (R2) can be relaxed as follows:

? (R1?) sup?g kS1 (?g)k : ?g ? ?g ? XTg ? ¡ ? ng ? ?g? (?, ?) = 0, T

sup? xg ? : ? ? ? ? 1 ? [?g? (?, ?)]i = 0. (R2?) i

Inspired by (R1?) and (R2?), we develop TLFre via the following three steps: Step 1. Given ? and ?, we estimate a region ? that contains ?? (?, ?). Step 2. We solve for the supreme values in (R1?) and (R2?). Step 3. By plugging in the supreme values from Step 2, (R1?) and (R2?) result in the desired two-layer screening rules for SGL. 4

3.3

The Set of Parameter Values Leading to Zero Solution ? For notational convenience, let Fg? = {? : kS1 (XTg ?)k ? ? ng }, g = 1, . . . , G; and thus the ? feasible set of the Fenchel?s dual of SGL is F = ?g=1,...,G Fg? . In view of problem (5) [or (9)], we can see that ?? (?, ?) is the projection of y/? on F ? , i.e., ?? (?, ?) = PF ? (y/?). Thus, if y/? ? F ? , we have ?? (?, ?) = y/?. Moreover, by (R1), we can see that ? ? (?, ?) = 0 if y/? is an interior point of F ? . Indeed, we have the following stronger result.

? T

Theorem 4. For the SGL problem, let ?? max = maxg {?g : S1 (Xg y/?g) = ? ng }. Then, y y ? ? ? ? ? ? F ? ? (?, ?) = ? ? ? (?, ?) = 0 ? ? ? ?max . ?g in the definition of ?? max admits a closed form solution [24]. Theorem 4 implies that the optimal solution ? ? (?, ?) is 0 as long as y/? ? F ? . This geometric property also leads to an explicit characterization of the set of (?1 , ?2) such that the corresponding solution of problem (2) is 0. We denote by ??? (?1 , ?2) the optimal solution of problem (2). Corollary 5. For the SGL problem in (2), let ?max (?2) = maxg ?1ng kS?2 (XTg y)k. Then, 1 (i): ??? (?1 , ?2) = 0 ? ?1 ? ?max (?2). 1

(ii): If ?1 ? ?max := maxg 1

4

?1 kXT g yk ng

or ?2 ? ?max := kXT yk? , then ??? (?1 , ?2) = 0. 2

The Two-Layer Screening Rules for SGL

We follow the three steps in Section 3.2 to develop TLFre. In Section 4.1, we give an accurate estimation of ?? (?, ?) via normal cones [15]. Then, we compute the supreme values in (R1?) and (R2?) by solving nonconvex problems in Section 4.2. We present the TLFre rules in Section 4.3. 4.1 Estimation of the Dual Optimal Solution Because of the geometric property of the dual problem in (5), i.e., ?? (?, ?) = PF ? (y/?), we have a very useful characterization of the dual optimal solution via the so-called normal cones [15]. Definition 1. [15] For a closed convex set C ? Rn and a point w ? C, the normal cone to C at w is NC (w) = {v : hv, w0 ? wi ? 0, ?w0 ? C}. (14) ? ?). ? ?) is known if ? ? = ?? . Thus, we can estimate ?? (?, ?) in terms of ?? (?, By Theorem 4, ?? (?, max ? for ? (?, ?) to be estimated. Due to the same reason, we only consider the cases with ? ¡ ?? max Remark 4. In many applications, the parameter values that perform the best are usually unknown. To determine appropriate parameter values, commonly used approaches such as cross validation and stability selection involve solving SGL many times over a grip of parameter values. Thus, given {?(i) }Ii=1 and ?(1) ? ? ? ? ? ?(J) , we can fix the value of ? each time and solve SGL by varying the value of ?.

We repeat the process until we solve SGL for all of the parameter values. ? ?) is known with ? ? ? ?? . Let ?g , Theorem 6. For the SGL problem in (3), suppose that ?? (?, max ? g = 1, . . . , G, be defined by Theorem 4. For any ? ? (0, ?), we define

? ? ? ? ¡ ?? , if ? max ? = y/? ? ? (?, ?), where X? = argmaxXg ?g , n? (?) T ? ? X? S1 (X? y/?max), if ? = ?? max , ? = y ? ?? (?, ? ?), v? (?, ?) ? ? ? = v? (?, ?) ? ? v? (?, ?)

? ? hv? (?,?),n ? (?)i ? n? (?). ? 2 kn? (?)k

Then, the following hold: ? ? NF ? (?? (?, ? ?)), (i): n? (?) ? ? ?) + 1 v? (?, ?))k ? ? (ii): k?? (?, ?) ? (?? (?, ? 21 kv? (?, ?)k. 2 ?

? Theorem 6 shows that ?? (?, ?) ? = ?? (?, ? ?) + 1 v? (?, ?). For notational convenience, let o? (?, ?) 2 ? ? ? centered at o? (?, ?). ? lies inside the ball of radius 21 kv? (?, ?)k 4.2 Solving for the supreme values via Nonconvex Optimization We solve the optimization problems in (R1?) and (R2?). To simplify notations, let ? ? 1 kv? (?, ?)k}, ? ? = {? : k? ? o? (?, ?)k ? 2

? ? 1 kv? (?, ?)kkX ? ?g = ?g : k?g ? XT o? (?, ?)k g k2 , g = 1, . . . , G. g

2

5

?

(15) (16)

Theorem 6 indicates that ?? (?, ?) ? ?. Moreover, we can see that XTg ? ? ?g , g = 1, . . . , G. To develop the TLFre rule by (R1?) and (R2?), we need to solve the following optimization problems: ? ?) = sup {kS1 (?g)k : ?g ? ?g }, g = 1, . . . , G, s?g (?, ?; (17) ?g ? T ? t (?, ?; ?) = sup {—x ?— : ? ? ?}, i = 1, . . . , ng , g = 1, . . . , G. (18) ?

gi

gi

Solving problem (17) We consider the following equivalent problem of (17):

1 1 ? ?) 2 = sup kS1 (?g)k2 : ?g ? ?g . s? (?, ?;

(19) We can see that the objective function of problem (19) is continuously differentiable and the feasible set is a ball. Thus, (19) is a nonconvex problem because we need to maximize a convex function subject to a convex set. We derive the closed form solutions of problems (17) and (19) as follows. ? ? r = 1 kv? (?, ?)kkX ? Theorem 7. For problems (17) and (19), let c = XTg o? (?, ?), g k2 and ?g be ? 2 the set of the optimal solutions. (i) Suppose that c ? / B? , i.e., kck? ¿ 1. Let u = rS1 (c)/kS1 (c)k. Then, ? ?) = kS1 (c)k + r and ?? = {c + u}. s?g (?, ?; (20) g (ii) Suppose that c is a boundary point of B? , i.e., kck? = 1. Then, ? ?) = r and ?? = {c + u : u ? NB (c), kuk = r} . s?g (?, ?; (21) g ? ? ? (iii) Suppose that c ? int B? , i.e., kck? ¡ 1. Let i ? I = {i : —[c]i — = kck? }. Then, ? ?) = (kck? + r ? 1) , s?g (?, ?; (22) + ? if ?g ? B? , ??g , ? ? ? ? ? ?g = {c + r ? sgn([c]i)ei : i ? I } , if ?g 6? B? and c 6= 0, ? {r ? ei? , ?r ? ei? : i? ? I ? } , if ?g 6? B? and c = 0, 2

g

?g

2

where ei is the ith standard basis vector. Solving problem in (18) Problem (18) can be solved directly via the Cauchy-Schwarz inequality. ? + 1 kv? (?, ?)kkx ? ?) = —xT o? (?, ?)— ? Theorem 8. For problem (18), we have t?gi (?, ?; gi k. gi ? 2 4.3 The Proposed Two-Layer Screening Rules To develop the two-layer screening rules for SGL, we only need to plug the supreme values ? 2 ; ?1) in (R1?) and (R2?). We present the TLFre rule as follows. ? 2 ; ?1) and t? (?2 , ? s?g (?2 , ? gi Theorem 9. For the SGL problem in (3), suppose that we are given ? and a sequence of parameter (0) values ?? ¿ ?(1) ¿ . . . ¿ ?(J) . For each integer 0 ? j ¡ J , we assume that ? ? (?(j) , ?) max = ? ? (j) ? (j+1) (? , ?(j)) and s?g (?(j+1) , ?(j) ; ?) be given by Eq. (6), Theorems 6 is known. Let ? (? , ?), v? and 7, respectively. Then, for g = 1, . . . , G, the following holds ? s?g (?(j+1) , ?(j) ; ?) ¡ ? ng ? ?g? (?(j+1) , ?) = 0. (L1) For the g?th group that does not pass the rule in (L1), we have [?g?? (?(j+1) , ?)]i = 0 if

T y?X? ? (?(j) ,?) 1 ? (j+1) (j) 1 ? (j+1) (j) + v (? , ?) , ?)kkxg?i k ? 1. xg?i + 2 kv? (? (j) ? 2 ?

(L2)

(L1) and (L2) are the first layer and second layer screening rules of TLFre, respectively.

5

Experiments

We evaluate TLFre on both synthetic and real data sets. To measure the performance of TLFre, we compute the rejection ratios of (L1) and (L2), respectively. Specifically, let m be the number of features that have 0 coefficients in the solution, G be the index set of groups that are discarded by (L1) and p be the number of inactive features that are detected by (L2). The rejection ratios of (L1) P ng and (L2) are defined by r1 = g?G and r2 = —p— m m , respectively. Moreover, we report the speedup gained by TLFre, i.e., the ratio of the running time of solver without screening to the running time of solver with TLFre. The solver used in this paper is from SLEP [9]. To determine appropriate values of ? and ? by cross validation or stability selection, we can run TLFre with as many parameter values as we need. Given a data set, for illustrative purposes only, we select seven values of ? from {tan(?) : ? = 5? , 15? , 30? , 45? , 60? , 75? , 85? }. Then, for each value of ?, we run TLFre along a sequence of 100 values of ? equally spaced on the logarithmic scale of ?/?? max from 1 to 0.01. Thus, 700 pairs of parameter values of (?, ?) are sampled in total. 6

0.3 0.1

200

400 ?2

600

0.01 0.02 0.04

800

Rejection Ratio

Rejection Ratio

0.1 0.2 0.4 ?/??max
1
0.7 0.5 0.3 0.1 0.1 0.2 0.4 ?/??max
1
0.7 0.5 0.3
0.1 0.2 0.4 ?/??max
1
?
(e) ? = tan(45)
1
0.5 0.3
0.01 0.02 0.04
0.7 0.5 0.3
0.1 0.2 0.4 ?/??max
1
1 0.9 0.7 0.5 0.3 0.1
0.1 0.2 0.4 ?/??max
1
0.01 0.02 0.04
?
(f) ? = tan(60)
0.7
(d) ? = tan(30?)
1 0.9
0.01 0.02 0.04
1 0.9
0.1 0.1 0.2 0.4 ?/??max
0.1
0.01 0.02 0.04
?
0.3
(c) ? = tan(15?)
1 0.9
0.1
0.01 0.02 0.04
0.5
0.01 0.02 0.04
(b) ? = tan(5?)
(a) 1 0.9
0.7
0.1
Rejection Ratio
0 0
0.5
Rejection Ratio
100

0.7
1 0.9
Rejection Ratio
200
1 0.9
Rejection Ratio
?1
300
Rejection Ratio
?max 1 (?2) ? = tan(5?) ? = tan(15?) ? = tan(30?) ? = tan(45?) ? = tan(60?) ? = tan(75?) ? = tan(85?)
400
0.1 0.2 0.4 ?/??max
1
?
(g) ? = tan(75)
(h) ? = tan(85)
Figure 1: Rejection ratios of TLFre on the Synthetic 1 data set.
0.5 0.3 0.1
500 ?2
0.01 0.02 0.04
1000
Rejection Ratio
Rejection Ratio
0.7 0.5 0.3 0.1 0.01 0.02 0.04
0.1 0.2 0.4 ?/??max
1
?
(e) ? = tan(45)
1
0.3
0.7 0.5 0.3
1
0.1 0.2 0.4 ?/??max
1
0.7 0.5 0.3
0.01 0.02 0.04
?
0.5 0.3
0.01 0.02 0.04
0.1 0.2 0.4 ?/??max
1
1 0.9 0.7 0.5 0.3 0.1
0.1 0.2 0.4 ?/??max
?
(f) ? = tan(60)

0.7
(d) ? = tan(30?)
1 0.9
0.1
0.01 0.02 0.04
1 0.9
0.1 0.1 0.2 0.4 ?/??max
(c) ? = tan(15?)
1 0.9
0.1 0.1 0.2 0.4 ?/??max
0.5
0.01 0.02 0.04
(b) ? = tan(5?)
(a) 1 0.9
0.7
0.1
Rejection Ratio
0 0
Rejection Ratio
0.7
1 0.9
Rejection Ratio
200
1 0.9
Rejection Ratio
?1
400
Rejection Ratio
?max 1 (?2) ? = tan(5?) ? = tan(15?) ? = tan(30?) ? = tan(45?) ? = tan(60?) ? = tan(75?) ? = tan(85?)
600
(g) ? = tan(75)
1
0.01 0.02 0.04
0.1 0.2 0.4 ?/??max
1
?
(h) ? = tan(85)

Figure 2: Rejection ratios of TLFre on the Synthetic 2 data set. 5.1 Simulation Studies We perform experiments on two synthetic data sets that are commonly used in the literature [19, 29]. The true model is y = X? ? + 0.01, ? N (0, 1). We generate two data sets with 250 ? 10000 entries: Synthetic 1 and Synthetic 2. We randomly break the 10000 features into 1000 groups. For Synthetic 1, the entries of the data matrix X are i.i.d. standard Gaussian with pairwise correlation zero, i.e., corr(xi , xi) = 0. For Synthetic 2, the entries of the data matrix X are drawn from i.i.d. standard Gaussian with pairwise

correlation 0.5—i?j— , i.e., corr(xi , xj) = 0.5—i?j— . To construct ? ? , we first randomly select ?1 percent of groups. Then, for each selected group, we randomly select ?2 percent of features. The selected components of ? ? are populated from a standard Gaussian and the remaining ones are set to 0. We set ?1 = ?2 = 10 for Synthetic 1 and ?1 = ?2 = 20 for Synthetic 2. The figures in the upper left corner of Fig. 1 and Fig. 2 show the plots of ?max (?2) (see Corollary 1 5) and the sampled parameter values of ? and ? (recall that ?1 = ?? and ?2 = ?). For the other figures, the blue and red regions represent the rejection ratios of (L1) and (L2), respectively. We can see that TLFre is very effective in discarding inactive groups/features; that is, more than 90% of inactive features can be detected. Moreover, we can observe that the first layer screening (L1) becomes more effective with a larger ?. Intuitively, this is because the group Lasso penalty plays a more important role in enforcing the sparsity with a larger value of ? (recall that ?1 = ??). The top and middle parts of Table 1 indicate that the speedup gained by TLFre is very significant (up to 30 times) and TLFre is very efficient. Compared to the running time of the solver without screening, the running time of TLFre is negligible. The running time of TLFre includes that of computing kXg k2 , g = 1, . . . , G, which can be efficiently computed by the power method [6]. Indeed, this can be shared for TLFre with different parameter values. 5.2 Experiments on Real Data Set We perform experiments on the Alzheimer?s Disease Neuroimaging Initiative (ADNI) data set (http://adni.loni.usc.edu/). The data matrix consists of 747 samples with 426040 single 7

Table 1: Running time (in seconds) for solving SGL along a sequence of 100 tuning parameter values of ? equally spaced on the logarithmic scale of ?/?? max from 1.0 to 0.01 by (a): the solver [9] without screening; (b): the solver combined with TLFre. The top and middle parts report the results of TLFre on Synthetic 1 and Synthetic 2. The bottom part reports the results of TLFre on the ADNI data set with the GMV data as response. tan(5?) tan(15?) tan(30?) tan(45?) tan(60?) tan(75?) tan(85?)

? solver TLFre Synthetic 1 TLFre+solver speedup
298.36 0.77 10.26 29.09
301.74 0.78 12.47 24.19
308.69 0.79 15.73 19.63
307.71 0.79 17.69 17.40
311.33 0.81 19.71 15.79
307.53 0.79 21.95 14.01
291.24 0.77 22.53 12.93
solver TLFre Synthetic 2 TLFre+solver speedup
294.64 0.79 11.05 26.66
294.92 0.80 12.89 22.88
297.29 0.80 16.08 18.49
297.50 0.81 18.90 15.74
297.59 0.81 20.45 14.55
295.51 0.81 21.58 13.69
292.24 0.82 22.80 12.82

30838.29 64.96 386.80 79.73
31096.10 65.00 402.72 77.22
30850.78 64.89 391.63 78.78
30728.27 65.17 385.98 79.61
30572.35 65.05 382.62 79.90
0.5 0.3 0.1
50
100 ?2
150
1
0.7 0.5 0.3 0.1 0.1 0.2 0.4 ?/??max
(e) ? = tan(45?)
1
0.3
1
0.7 0.5 0.3
0.1 0.2 0.4 ?/??max
1
(f) ? = tan(60?)
0.3
0.1 0.2 0.4 ?/??max
1
(d) ? = tan(30)
1 0.9 0.7 0.5 0.3
0.01 0.02 0.04
0.5
?
0.1
0.01 0.02 0.04
0.7
0.01 0.02 0.04
(c) ? = tan(15)
1 0.9
1 0.9
0.1 0.1 0.2 0.4 ?/??max
?
0.1
0.01 0.02 0.04
0.5
0.01 0.02 0.04
(b) ? = tan(5) Rejection Ratio
Rejection Ratio
0.1 0.2 0.4 ?/??max
?
(a)
0.7

0.1
0.01 0.02 0.04
1 0.9
1 0.9
Rejection Ratio
0.7
Rejection Ratio
0 0
1 0.9
Rejection Ratio
50
Rejection Ratio
?max 1 (?2) ? = tan(5?) ? = tan(15?) ? = tan(30?) ? = tan(45?) ? = tan(60?) ? = tan(75?) ? = tan(85?)
100
?1
solver 30652.56 30755.63 TLFre 64.08 64.56 TLFre+solver 372.04 383.17 speedup 82.39 80.27
Rejection Ratio
ADNI+GMV
1 0.9 0.7 0.5 0.3 0.1
0.1 0.2 0.4 ?/??max
(g) ? = tan(75?)
1
0.01 0.02 0.04
0.1 0.2 0.4 ?/??max
1
(h) ? = tan(85?)

Figure 3: Rejection ratios of TLFre on the ADNI data set with grey matter volume as response. nucleotide polymorphisms (SNPs), which are divided into 94765 groups. The response vector is the grey matter volume (GMV). The figure in the upper left corner of Fig. 3 shows the plots of ?max (?2) (see Corollary 5) and the 1 sampled parameter values of ? and ?. The other figures present the rejection ratios of (L1) and (L2) by blue and red regions, respectively. We can see that almost all of the inactive groups/features are discarded by TLFre. The rejection ratios of r1 + r2 are very close to 1 in all cases. The bottom part of Table 1 shows that TLFre leads to a very significant speedup (about 80 times). In other words, the solver without screening needs about eight and a half hours to solve the 100 SGL problems for each value of ?. However, combined with TLFre, the solver needs only six to eight minutes. Moreover, we can observe that the computational cost of TLFre is negligible compared to that of the solver without screening. This demonstrates the efficiency of TLFre.

6

Conclusion

In this paper, we propose a novel feature reduction method for SGL via decomposition of convex sets. We also derive the set of parameter values that

lead to zero solutions of SGL. To the best of our knowledge, TLFre is the first method which is applicable to sparse models with multiple sparsity-inducing regularizers. More importantly, the proposed approach provides novel framework for developing screening methods for complex sparse models with multiple sparsity-inducing regularizers, e.g., ‘1 SVM that performs both sample and feature selection, fused Lasso and tree Lasso with more than two regularizers. Experiments on both synthetic and real data sets demonstrate the effectiveness and efficiency of TLFre. We plan to generalize the idea of TLFre to ‘1 SVM, fused Lasso and tree Lasso, which are expected to consist of multiple layers of screening.

8

2 References

[1] H. H. Bauschke and P. L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, 2011. [2] J. Borwein and A. Lewis. Convex Analysis and Nonlinear Optimization, Second Edition. Canadian Mathematical Society, 2006. [3] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004. [4] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. Pacific Journal of Optimization, 8:667?698, 2012. [5] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. arXiv:1001.0736. [6] N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review, 53:217?288, 2011. [7] J.-B. Hiriart-Urruty. From convex optimization to nonconvex optimization. necessary and sufficient conditions for global optimality. In Nonsmooth optimization and related topics. Springer, 1988. [8] J.-B. Hiriart-Urruty. A note on the Legendre-Fenchel transform of convex composite functions. In Nonsmooth Mechanics and Analysis. Springer, 2006. [9] J. Liu, S. Ji, and J. Ye. SLEP: Sparse Learning with Efficient Projections. Arizona State University, 2009. [10] J. Liu and J. Ye. Moreau-Yosida regularization for grouped tree structure learning. In Advances in neural information processing systems, 2010. [11] J. Liu, Z. Zhao, J. Wang, and J. Ye. Safe screening with variational inequalities and its application to lasso. In International Conference on Machine Learning, 2014. [12] K. Ogawa, Y. Suzuki, S. Suzumura, and I. Takeuchi. Safe sample screening for Support Vector Machine. arXiv:1401.6740, 2014. [13] K. Ogawa, Y. Suzuki, and I. Takeuchi. Safe screening of non-support vectors in pathwise SVM computation. In ICML, 2013. [14] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Noh, J. Pollack, and P. Wang. Regularized multivariate regression for indentifying master predictors with application to integrative genomics study of breast cancer. The Annals of Appliced Statistics, 4:53?77, 2010. [15] A. Ruszczy?nski. Nonlinear Optimization. Princeton University Press, 2006. [16] N. Simon, J. Friedman., T. Hastie., and R. Tibshirani. A Sparse-Group Lasso. Journal of Computational and Graphical Statistics, 22:231?245, 2013. [17] P. Sprechmann, I. Ram??rez, G. Sapiro., and Y. El-

dar. C-HiLasso: a collaborative hierarchical sparse modeling framework. IEEE Transactions on Signal Processing, 59:4183?4198, 2011. [18] R. Tibshirani. Regression shringkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58:267?288, 1996. [19] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. Tibshirani. Strong rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society Series B, 74:245? 266, 2012. [20] M. Vidyasagar. Machine learning methods in the cocomputation biology of cancer. In Proceedings of the Royal Society A, 2014. [21] M. Vincent and N. Hansen. Sparse group lasso and high dimensional multinomial classification. Computational Statistics and Data Analysis, 71:771?786, 2014. [22] J. Wang, J. Jun, and J. Ye. Efficient mixed-norm regularization: Algorithms and safe screening methods. arXiv:1307.4156v1. [23] J. Wang, P. Wonka, and J. Ye. Scaling svm and least absolute deviations via exact data reduction. In International Conference on Machine Learning, 2014. [24] J. Wang and J. Ye. Two-Layer feature reduction for sparse-group lasso via decomposition of convex sets. arXiv:1410.4210v1, 2014. [25] J. Wang, J. Zhou, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. In Advances in neural information processing systems, 2013. [26] Z. J. Xiang and P. J. Ramadge. Fast lasso screening tests based on correlations. In IEEE ICASSP, 2012. [27] D. Yogatama and N. Smith. Linguistic structured sparsity in text categorization. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2014. [28] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B, 68:49?67, 2006. [29] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67:301?320, 2005.

9