Screening Rules for Convex Problems

Anant Raj
Max-Planck Institute for Intelligent Systems

Jakob Olbrich
ETH Zürich

Bernd Gärtner
ETH Zürich

Bernhard Schölkopf
Max-Planck Institute for Intelligent Systems

Martin Jaggi
EPFL

Abstract

We propose a new framework for deriving screening rules for convex optimization problems. Our approach covers a large class of constrained and penalized optimization formulations, and works in two steps. First, given any approximate point, the structure of the objective function and the duality gap is used to gather information on the optimal solution. In the second step, this information is used to produce screening rules, i.e. safely identifying unimportant weight variables of the optimal solution. Our general framework leads to a large variety of useful existing as well as new screening rules for many applications. For example, we provide new screening rules for general simplex and L_1-constrained problems, Elastic Net, squared-loss Support Vector Machines, minimum enclosing ball, as well as structured norm regularized problems, such as group lasso.

1 Introduction

Optimization techniques for high-dimensional problems have become the work-horses for most machine-learning methods. With the rapid increase of available data, major challenges occur as the number of optimization variables (weights) grows beyond capacity of current systems.

The idea of screening refers to eliminating optimization variables that are guaranteed to *not* contribute to any optimal solution, and can therefore safely be removed. Such screening techniques have received increased interest in several machine learning related applications in recent years, and have been shown to lead to very significant computational efficiency improvements in various cases, in particular for many types of sparse methods. Screening techniques can be used either as a pre-processing before passing the problem to the optimizer, or also interactively during any iterative solver (called dynamic screening), to gradually reduce the problem complexity during optimization.

While existing screening methods were mainly relying on problem-specific geometric properties, we in this paper take a different approach. We propose a new framework allowing screening on general convex optimization problems, using simple tools from convex duality instead of any geometric arguments. Our framework applies to a very large class of optimization problems both for constrained as well as penalized problems, including most machine learning methods of interest.

Our main contributions in this paper are summarized as follows [1] : (All proofs provided in appendix)

1. We propose a new framework for screening for a more general class of optimization problem with a simple primal-dual structure.
2. The framework leads to a large set of new screening rules that could not be screened before. Furthermore, it also recovers many existing screening rules as special cases.
3. We are able to express all screening rules using general optimization complexity notions such as smoothness or strong convexity, getting rid of problem-specific geometric properties.
4. Our proposed rules are dynamic (allowing any existing algorithm to be additionally equipped with screening) and safe (guaranteed to only eliminate truly unimportant variables).

[1] Parts of this work have appeared in the Master's Thesis [22].

Related Work. Coming from computational geometry, [1] has proposed a screening technique for the minimum enclosing ball problem for a given set of data points. Later [11] improve the threshold for this rule in the minimum enclosing ball setting.

Independently, the breakthrough work of [6] gave the first screening rules for the case of sparse regression, as given in the Lasso. Since then, there have been many extensions and alterations of the general concept. While [6] exploits geometric quantities to bound the the Lasso dual solution within a compact region, we recommend the survey paper by [28] for an overview of geometric methods for Lasso screening. Apart from being geometry specific, most existing approaches such as [26, 27, 15, 6, 21] are not agnostic to the regularization parameter used, but instead are restricted to perform screening along the entire regularization path. In contrast, our proposed framework here allows any internal optimization algorithms to be equipped with screening.

As opposed to penalized optimization problems, much less is known about screening for constrained optimization. For the dual of the hinge loss SVM, which is a box-constrained optimization problem, [20] proposed a geometric screening rule based on the intersection region of two spheres.More recently [29] provided new screening rules for that case in the dynamic setting using a method similar to our approach. However their method is restricted to the SVM case.

As a first step to allow screening for more general optimization objectives, [5, 17, 18, 19] have developed more systematic duality gap based screening rules for several problems, including group lasso, multi-task and multi-class problems (in the penalized setting) under a wider class of objectives f. While the earlier work of [5, 17] assumed separability of f over the group structure, later extensions of [23, 18, 19], have generalized the applicability, but still rely on geometric and application-specific quantities in order to perform screening. The approach of [23] allows screening rules for (sparse) SVM problems on both dimensions, the features as well as the datapoints, but is limited in terms of generality of this specific sparse problem structure. We here provide screening rules for a more general framework of box constrained optimization, while hinge-loss SVM happens to be a special case of this. Our approach here is most similar to the Blitz framework of [8, 9], which provides a general possibility to exploit piece-wise linear structure in an optimization problem in order to do screening. The method of [8, 9] is however tied to a relatively specific algorithm, leading to very efficient active set methods on this problem class. Our proposed approach aims at capturing the largest possible general class of optimization problems allowing for screening. It can be shown to recover many of the other existing rules including e.g. [5, 17, 18, 19] and [29], but significantly generalizing the method to general objectives and constraints as well as regularizers.

2 Primal-Dual Structure and Duality Gap Certificates

In this paper, we consider optimization problems of the following structure. A wide range of machine learning optimization problems can be formulated as (A) and (B), which are dual to each other:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad \left[\ \mathcal{O}_A(\boldsymbol{x}) := \quad f(A\boldsymbol{x}) \ + \ g(\boldsymbol{x}) \ \right] \tag{A}$$

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \left[\ \mathcal{O}_B(\boldsymbol{w}) := \quad f^*(\boldsymbol{w}) \ + \ g^*(-A^\top\boldsymbol{w}) \ \right] \tag{B}$$

The two problems are associated to a given data matrix $A \in \mathbb{R}^{d\times n}$, and the functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are allowed to be arbitrary closed convex functions. The functions f^*, g^* in formulation (B) are defined as the *convex conjugates* of their corresponding counterparts f, g in (A). Here $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{w} \in \mathbb{R}^d$ are the respective variable vectors. The association of problems (A) and (B) is a special case of Fenchel Duality, see e.g. [3, Theorem 4.4.2] or [2, Proposition 15.18], see the Appendix A for a self-contained derivation. The powerful features of this general duality structure is that the two problems are fully symmetric, when changing respective roles of f and g. As we will see later, the two roles can be swapped, depending on the application.

Optimality Conditions. The first-order optimality conditions for our pair of vectors $\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{x} \in \mathbb{R}^n$ in problems (A) and (B) are given as

$$\boldsymbol{w} \in \partial f(A\boldsymbol{x}) \ , \tag{1a}$$

$$A\boldsymbol{x} \in \partial f^*(\boldsymbol{w}) \ , \tag{1b}$$

$$-A^\top\boldsymbol{w} \in \partial g(\boldsymbol{x}) \ , \tag{2a}$$

$$\boldsymbol{x} \in \partial g^*(-A^\top\boldsymbol{w}) \tag{2b}$$

see e.g. [2, Proposition 19.18], see Appendix A for details. In similar way, the optimization problem as well as optimality condition can be written for the case when one of the function becomes separable.

Duality Gap and Certificates. The duality gap is the most important tool for us to provide guaranteed information about the optimal point, which will then be the foundation for the second step, to perform screening on the optimal point. For the problem structure (A) and (B) as given by Fenchel-Rockafellar duality, the *duality gap* for any pair of primal and dual variables $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{w} \in \mathbb{R}^d$ is defined as $G(\boldsymbol{w}, \boldsymbol{x}) := \mathcal{O}_B(\boldsymbol{w}) + \mathcal{O}_A(\boldsymbol{x})$. Most importantly, the duality gap acts as a certificate of approximation quality — the true optimum values $\mathcal{O}_B(\boldsymbol{w}^\star)$ and $-\mathcal{O}_A(\boldsymbol{x}^\star)$ (which are unknown) will always lie within the (known) duality gap. We discuss more details in Appendix B.

Obtaining Information about the Optimal Points. As we have mentioned, any type of screening will crucially rely on first deriving safe knowledge about the unknown optimal points of our given optimization problem. Here, we will use the duality gap to obtain such knowledge on the optimal points $\boldsymbol{x}^\star \in \mathbb{R}^n$ and $\boldsymbol{w}^\star \in \mathbb{R}^d$ of the respective optimization problems (A) and (B) respectively. Statements and proofs are provided in Appendix B.2.

3 Screening Rules for Constrained Problems

In the following, we will develop screening rules for constrained optimization problems of the form (3), by exploiting the structure of the constraint set for a variety of sparsity-inducing problems.

The Constrained Case. Any constrained convex optimization problem of the form

$$\min_{\boldsymbol{x}\in\mathcal{C}} f(A\boldsymbol{x}) \tag{3}$$

can be written in the form (A) by using the indicator function of the constraint set $\mathcal{C}$ set as g.

3.1 Simplex Constrained Problems

Optimization over unit simplex $\triangle := \{\boldsymbol{x} \in \mathbb{R}^n \mid x_i \geq 0, \quad \sum_{i=1}^n x_i = 1\}$ is a important class of constrained problems (3), as it includes optimization over any finite polytope. In this case, the columns of A describe the vertices, and $\boldsymbol{x}$ are barycentric coordinates representing the point $A\boldsymbol{x}$. Formally, $g(\boldsymbol{x})$ is the indicator function of the unit simplex $\mathcal{C} = \triangle$ in this case.

The following two theorems provide screening rules for simplex constrained problems.

Theorem 1. *For general simplex constrained optimization* $\min_{\boldsymbol{x}\in\triangle} f(A\boldsymbol{x})$*, the optimality condition* (2a) *gives rise to the following screening rule at the optimal point, for any* $i \in [n]$

$$(\boldsymbol{a}_i - A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star > 0 \ \Rightarrow\ x_i^\star = 0\,. \tag{4}$$

In the following Theorem 2 we now assume smoothness and strong convexity of function f to provide screening rules for simplex problems, in terms of an arbitrary iterate $\boldsymbol{x}$, without knowing $\boldsymbol{x}^\star$.

Theorem 2. *Let f be L-smooth and μ-strongly convex over the unit simplex $\mathcal{C} = \triangle$. Then for simplex constrained optimization* $\min_{\boldsymbol{x}\in\triangle} f(A\boldsymbol{x})$ *we have the following screening rule, for any* $i \in [n]$

$$(\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) > L\sqrt{\tfrac{G_{\mathcal{C}}(\boldsymbol{x})}{\mu}}\, \|\boldsymbol{a}_i - A\boldsymbol{x}\| \ \Rightarrow\ x_i^\star = 0\,. \tag{5}$$

Our general screening rules for simplex constrained problems as in Theorem 2 allows many practical implications. For example, new screening rules for squared loss SVM and minimum enclosing ball problem come as a direct consequence which we provide in Appendix C.1.

3.2 L_1*-Constrained Problems*

L_1-constrained formulations are very widely used in order to induce sparsity in the variables. Here below we provide results for screening on general L_1-constrained problems, that is $\min_{\boldsymbol{x}\in\mathcal{C}} f(A\boldsymbol{x})$ for $\mathcal{C} = L_1 \subset \mathbb{R}^n$ (or a scaled version of the L_1-ball).

Theorem 3. *For general L_1-constrained optimization* $\min_{\boldsymbol{x}\in L_1} f(A\boldsymbol{x})$*, the optimality condition* (2a) *gives rise to the following screening rule at the optimal point, for any* $i \in [n]$

$$\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| + (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star < 0 \ \Rightarrow\ x_i^\star = 0\,. \tag{6}$$

Using only a current iterate $\boldsymbol{x}$ instead of an optimal point, we obtain screening for general smooth and strongly convex function f:

Theorem 4. *Let f be L-smooth and μ-strongly convex over the L_1-ball. Then for L_1-constrained optimization* $\min_{\boldsymbol{x}\in L_1} f(A\boldsymbol{x})$ *we have the following screening rule, for any* $i \in [n]$

$$\left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + (A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) + L(\|\boldsymbol{a}_i\|_2 + \|A\boldsymbol{x}\|_2)\sqrt{\tfrac{G_{\mathcal{C}}(\boldsymbol{x})}{\mu}} < 0 \ \Rightarrow\ \boldsymbol{x}_i^\star = 0 \tag{7}$$

3.3 *Screening for Box Constrained Problems*

Box-constrained problems are important in several machine learning applications, including SVMs. After variable rescaling, w.l.o.g. we can assume the constraint set $\mathcal{C} = \square := \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \leq x_i \leq 1\}$. We derive screening rules for predicting both if a variable will take the upper or lower constraint.

Theorem 5. *Let f be L-smooth. Then for box-constrained optimization* $\min_{\boldsymbol{x}\in\square} f(A\boldsymbol{x})$*, we obtain the following screening rules, for any* $i \in [n]$

$$\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x}) - \|\boldsymbol{a}_i\|_2 \sqrt{2LG(\boldsymbol{x})} > 0 \ \Rightarrow\ x_i^\star = 0\,, \ \textit{and}$$
$$\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x}) + \|\boldsymbol{a}_i\|_2 \sqrt{2L\,G(\boldsymbol{x})} < 0 \ \Rightarrow\ x_i^\star = 1\,.$$

The above screening rules (5) is directly applicable but not only limited to screening for hinge loss SVM. More details are provided in the Appendix C.3.

4 Screening for Penalized Problems

In this section we discuss screening methods for general penalized convex optimization problems of the form (A) and (B). Our method can reproduce the screening rules of [17, 19] as special cases, whereas their method does not directly extend to general f as in the setting here. Beyond L_1 problems, we also describe new screening rules for elastic net regularized squared loss regression problems, as well as the important case of structured norm regularized optimization.

4.1 L_1-*Penalized Problems*

The next theorem describes a screening rule for general L_1-penalized problems, under a smoothness assumption on function f.

Theorem 6. *Consider an L_1-regularized optimization problem:* $\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(A\boldsymbol{x}) + \lambda \left\|\boldsymbol{x}\right\|_1$
If f is L-smooth, then the following screening rule holds for all $i \in [n]$:

$$\left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| < \lambda - \|\boldsymbol{a}_i\|_2 \sqrt{2L\, G(\boldsymbol{x})} \;\Rightarrow\; \boldsymbol{x}_i^\star = 0$$

In the Appendix D.1, we discuss the special cases of squared loss regression and logistic loss regression with L_1 penalization. These results are presented in Corollaries 20 and 21 as direct consequences of Theorem 6. Both can also be derived from the framework discussed in [17, 19].

4.2 *Elastic Net Penalized Squared Loss Regression*

In the next corollary, we present a novel screening rule for the elastic net problem.

Corollary 7. *Consider the elastic net regression formulation*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; 1/2\|A\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \lambda_2\|\boldsymbol{x}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 \tag{8}$$

The following screening rule holds for all $i \in [n]$:

$$\left|(\boldsymbol{a}_i^\top A + 2\lambda_2\boldsymbol{e}_i^\top)\boldsymbol{x} - \boldsymbol{a}_i^\top\boldsymbol{b}\right| < \lambda_1 - \sqrt{2(\boldsymbol{a}_i^\top\boldsymbol{a}_i + 2\lambda_2)G(\boldsymbol{x})} \;\Rightarrow\; x_i^\star = 0.$$

The screening rules presented in Corollary 7 is different from the rule presented in [23].

4.3 *Structured Norm Penalized Problems*

In this section we present screening rules for non-overlapping group norm regularized problems. In particular, we discuss screening for general objectives with an ℓ_2/ℓ_1-regularization (the most prominent special case of that being the group lasso).

Group Norm - ℓ_2/ℓ_1 Regularization In the following, we use the notation $\{\boldsymbol{x}_1 \cdots \boldsymbol{x}_G\}$ to express a vector $\boldsymbol{x}$ as a partition of the groups of variables, such that $\boldsymbol{x}^\top = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top \cdots \boldsymbol{x}_G^\top\right]$. Correspondingly, the matrix A can be denoted as the concatenation of the respective columns $A = [A_1 \; A_2 \cdots A_G]$.

Theorem 8. *For ℓ_2/ℓ_1-regularized optimization problem of the form*

$$\min_{\boldsymbol{x}} f(A\boldsymbol{x}) + \textstyle\sum_{g=1}^G \sqrt{\rho_g}\|\boldsymbol{x}_g\|_2$$

Assuming f is L-smooth, then the following screening rule holds for all groups g:

$$\|A_g^\top \nabla f(A\boldsymbol{x})\|_2 + \sqrt{2L\, G(\boldsymbol{x})}\, \|A_g\|_{\mathit{Fro}} < \sqrt{\rho_g} \;\Rightarrow\; \boldsymbol{x}_g^\star = \boldsymbol{0} \in \mathbb{R}^{|g|}\;.$$

Screening for group lasso regression as discussed in Corollary 25 can be directly derived from Theorem 8 which is widely used in applications as an working example case of structured norm penalization. For the squared-loss special case, group lasso screening rules were recently developed by [18]. Similarly, [14] is also restricted to least-squares f objective.

5 Discussion

We have presented a unified way to derive screening rules for general constrained and penalized optimization problems. For both cases, our framework crucially utilizes the structure of piece-wise linearity of the problem at hand. For the constrained case, we showed that screening rules follow from the piece-wise linearity of the boundary of the constraint set. On the other hand for penalized optimization problems, we are able to derive screening rules from either piece-wise linearity of the penalty function, or as well from exploiting piece-wise linearity of the constraint set arising from the dual (conjugate) of the penalty function. In Appendix E we also provide illustrative experiments as well as discuss the connection with the sphere test method in Appendix D.3.

Acknowledgements. The authors would like to thank Julia Wysling, Rohit Babbar and Peter Bühlmann for valuable discussions.

References

[1] S. D. Ahipasaoğlu, P. Sun, and M. Todd. Linear Convergence of a Modified Frank–Wolfe Algorithm for Computing Minimum-Volume Enclosing Ellipsoids. *Optimization Methods and Software*, 23(1). 2008.

[2] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, New York, 2011.

[3] J. M. Borwein and Q. Zhu. *Techniques of Variational Analysis and Nonlinear Optimization*. Canadian Mathematical Society Books in Math, Springer, 2005.

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

[5] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the Lasso. *ICML*, 2015.

[6] L. E. Ghaoui, V. Viallon, and T. Rabbani. Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems. *arXiv.org*, Sept. 2010.

[7] M. Jaggi. An Equivalence between the Lasso and Support Vector Machines. In *Regularization, Optimization, Kernels, and Support Vector Machines*, Chapman and Hall/CRC, 2014.

[8] T. Johnson and C. Guestrin. Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization. *ICML*, 2015.

[9] T. B. Johnson and C. Guestrin. Unified methods for exploiting piecewise linear structure in convex optimization. *NIPS*, 2016.

[10] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. Technical report, Toyota Technological Institute, 2009.

[11] L. Källberg and T. Larsson. Improved Pruning of Large Data Sets for the Minimum Enclosing Ball Problem. *Graphical Models*, 2014.

[12] S. Lacoste-Julien and M. Jaggi. On the Global Linear Convergence of Frank-Wolfe Optimization Variants. *NIPS*, 2015.

[13] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. *ICML*, 2013.

[14] S. Lee and E. P. Xing. Screening Rules for Overlapping Group Lasso. *arXiv*, 2014.

[15] J. Liu, Z. Zhao, J. Wang, and J. Ye. Safe Screening with Variational Inequalities and Its Application to Lasso. *ICML*, 2014.

[16] J. Matoušek and B. Gärtner. *Understanding and using linear programming*. Springer, 2007.

[17] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP Safe screening rules for sparse multi-task and multi-class models. *NIPS*, 2015.

[18] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP Safe Screening Rules for Sparse-Group-Lasso. *arXiv*, Feb. 2016.

[19] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap Safe screening rules for sparsity enforcing penalties. *arXiv*, stat.ML, Nov. 2016.

[20] K. Ogawa, Y. Suzuki, S. Suzumura, and I. Takeuchi. Safe Sample Screening for Support Vector Machines. *arXiv.org*, 2014.

[21] K. Ogawa, Y. Suzuki, and I. Takeuchi. Safe Screening of Non-Support Vectors in Pathwise SVM Computation. *ICML*, 2013.

[22] J. Olbrich. Screening Rules for Convex Problems. Master's thesis, ETH Zürich, 2015.

[23] A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi. Simultaneous Safe Screening of Features and Samples in Doubly Sparse Modeling. *ICML*, 2016.

[24] V. Smith, S. Forte, M. Jaggi, and M. I. Jordan. L_1-Regularized Distributed Optimization: A Communication-Efficient Primal-Dual Framework. *arXiv cs.LG*, 2015.

[25] I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core Vector Machines: Fast SVM Training on Very Large Data Sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[26] J. Wang, B. Lin, P. Gong, P. Wonka, and J. Ye. Lasso Screening Rules via Dual Polytope Projection. *NIPS*, 2014.

[27] J. Wang, J. Zhou, J. Liu, P. Wonka, and J. Ye. A Safe Screening Rule for Sparse Logistic Regression. *NIPS*, 2014.

[28] Z. J. Xiang, Y. Wang, and P. J. Ramadge. Screening Tests for Lasso Problems. *arXiv.org*, 2014.

[29] J. Zimmert, C. S. de Witt, G. Kerg, and M. Kloft. Safe screening for support vector machines. In *NIPS Workshop on Optimization for Machine Learning*, 2015.

A Primal Dual Structure (Section 2)

The relation of our primal and dual problems (A) and (B) is standard in convex analysis, and is a special case of the concept of Fenchel Duality. Using the combination with the linear map A as in our case, the relationship is called *Fenchel-Rockafellar Duality*, see e.g. [3, Theorem 4.4.2] or [2, Proposition 15.18]. For completeness, we here illustrate this correspondence with a self-contained derivation of the duality.

Proof. Starting with the original formulation (A), we introduce a helper variable vector $\boldsymbol{v} \in \mathbb{R}^d$ representing $\boldsymbol{v} = A\boldsymbol{\alpha}$. Then optimization problem (A) becomes:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad f(\boldsymbol{v}) + g(\boldsymbol{\alpha}) \quad \text{such that } \boldsymbol{v} = A\boldsymbol{\alpha}\,. \tag{9}$$

Introducing Lagrange multipliers $\boldsymbol{w} \in \mathbb{R}^d$, the Lagrangian is given by:

$$L(\boldsymbol{\alpha}, \boldsymbol{v}; \boldsymbol{w}) := f(\boldsymbol{v}) + g(\boldsymbol{\alpha}) + \boldsymbol{w}^\top (A\boldsymbol{\alpha} - \boldsymbol{v})\ .$$

The dual problem of (A) follows by taking the infimum with respect to both $\boldsymbol{\alpha}$ and $\boldsymbol{v}$:

$$\begin{aligned} \inf_{\boldsymbol{\alpha}, \boldsymbol{v}} L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{v}) &= \inf_{\boldsymbol{v}} \left\{ f(\boldsymbol{v}) - \boldsymbol{w}^\top \boldsymbol{v} \right\} + \inf_{\boldsymbol{\alpha}} \left\{ g(\boldsymbol{\alpha}) + \boldsymbol{w}^\top A\boldsymbol{\alpha} \right\} \\ &= -\sup_{\boldsymbol{v}} \left\{ \boldsymbol{w}^\top \boldsymbol{v} - f(\boldsymbol{v}) \right\} - \sup_{\boldsymbol{\alpha}} \left\{ (-\boldsymbol{w}^\top A)\boldsymbol{\alpha} - g(\boldsymbol{\alpha}) \right\} \qquad (10) \\ &= -f^*(\boldsymbol{w}) - g^*(-A^\top \boldsymbol{w})\,. \qquad (11) \end{aligned}$$

We change signs and turn the maximization of the dual problem (11) into a minimization and thus we arrive at the dual formulation (B) as claimed:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \Big[\mathcal{O}_B(\boldsymbol{w}) := f^*(\boldsymbol{w}) + g^*(-A^\top \boldsymbol{w}) \Big]\,.$$

The Partially Separable Case. A very important special case arises when one part of the objective becomes separable. Formally, this is expressed as $g(\boldsymbol{x}) = \sum_{i=1}^n g_i(x_i)$ for univariate functions $g_i : \mathbb{R} \to \mathbb{R}$ for $i \in [n]$. Nicely in this case, the conjugate of g also separates as $g^*(\boldsymbol{y}) = \sum_i g_i^*(y_i)$. Therefore, the two optimization problems (A) and (B) write as

$$\mathcal{O}_A(\boldsymbol{x}) := f(A\boldsymbol{x}) + \textstyle\sum_i g_i(x_i) \tag{SA}$$

$$\mathcal{O}_B(\boldsymbol{w}) := f^*(\boldsymbol{w}) + \textstyle\sum_i g_i^*(-\boldsymbol{a}_i^\top \boldsymbol{w})\ , \tag{SB}$$

where $\boldsymbol{a}_i \in \mathbb{R}^d$ denotes the i-th column of A.

Crucially in this case, the optimality conditions (2a) and (2b) now become separable, that is

$$\boldsymbol{a}_i^\top \boldsymbol{w} \in \partial g_i(-x_i) \qquad \forall i\ , \tag{12a}$$

$$-x_i \in \partial g_i^*(\boldsymbol{a}_i^\top \boldsymbol{w}) \qquad \forall i\ . \tag{12b}$$

Note that the two other conditions (1a) and (1b) are unchanged in this case. For $g(\boldsymbol{x})$ is separable, i.e. $g(\boldsymbol{x}) = \sum_{i=1}^n g_i(x_i)$ for univariate functions $g_i : \mathbb{R} \to \mathbb{R}$ for $i \in [n]$, the primal-dual structure remains the separable. In this case, the conjugate of g also separates as $g^*(\boldsymbol{y}) = \sum_i g_i^*(y_i)$. Therefore, in terms of the the primal-dual structure (A) and (B) we obtain the separable special case (SA) and (SB). □

Optimality Conditions. The first-order optimality conditions follow from the standard definition of the conjugate functions in the Fenchel dual problem, see also e.g. [3, 2].

Proof. The first-order optimality conditions for our pair of vectors $\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{x} \in \mathbb{R}^n$ in problems (A) and (B) are given by equations (1a), (2a), (1b) and (2b). The proof directly comes from equation (10) by separately writing optimizing conditions for two expressions $\boldsymbol{w}^\top \boldsymbol{v} - f(\boldsymbol{v})$ and $(-\boldsymbol{w}^\top A)\boldsymbol{\alpha} - g(\boldsymbol{\alpha})$ in equation (10).

Crucially in the partially separable case, the optimality conditions (2a) and (2b) become separable. Comparing the expressions (SA) and (A), we see that $g(\boldsymbol{x}) = \sum_i g_i(x_i)$ and hence

$$g^*(\boldsymbol{x}) = \sum_i g_i^*(x_i)$$

Hence by applying (2a) and (2b) we obtain the separable optimality conditions (12a) and (12b). □

B Duality Gap and Objective Function Properties

The Gap Function. For the special case of differentiable function f, we can study a simpler duality gap

$$G(\boldsymbol{x}) := \mathcal{O}_B(\boldsymbol{w}(\boldsymbol{x})) + \mathcal{O}_A(\boldsymbol{x}) \tag{13}$$

purely defined as a function of $\boldsymbol{x}$, using the optimality relation (1a), i.e. $\boldsymbol{w}(\boldsymbol{x}) := \nabla f(A\boldsymbol{x})$.

The Wolfe-Gap Function. For any constrained optimization problem (3) defined over a bounded set $\mathcal{C}$ and $\boldsymbol{x} \in \mathcal{C}$, the Wolfe gap function (also known as Hearn gap or Frank-Wolfe gap) is defined as the difference of f to the minimum of its linearization over the same domain. Formally,

$$G_{\mathcal{C}}(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathcal{C}} \, (A\boldsymbol{x} - A\boldsymbol{y})^\top \nabla f(A\boldsymbol{x}). \tag{14}$$

It is not hard to see that the convenient Wolfe gap function is a special case of our above defined general duality gap $G(\boldsymbol{x}) := \mathcal{O}_B(\boldsymbol{w}(\boldsymbol{x})) + \mathcal{O}_A(\boldsymbol{x})$, for g being the indicator function of the constraint set $\mathcal{C}$, and $\boldsymbol{w}(\boldsymbol{x}) := \nabla f(A\boldsymbol{x})$. For more details, see Appendix B.1, or also [13, Appendix D].

B.1 Wolfe Gap as a Special Case of Duality Gap

Proof. To see this as a special case of general duality gap of the problem formulation, we consider the constraint as indicator function of set $\mathcal{C}$ such that $g(\boldsymbol{x}) = \boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x})$. Now from the definition of the Wolfe gap function

$$G_{\mathcal{C}}(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathcal{C}} \, (A\boldsymbol{x} - A\boldsymbol{y})^\top \partial f(A\boldsymbol{x})$$

Here $\partial f(A\boldsymbol{x})$ is an arbitrary subgradient of f at the candidate position $\boldsymbol{x}$, and $\boldsymbol{\iota}_{\mathcal{C}}^*(\boldsymbol{y}) := \sup_{\boldsymbol{s} \in \mathcal{C}} \langle \boldsymbol{s}, \boldsymbol{y} \rangle$ is the support function of $\mathcal{C}$. Now writing the general duality gap $G(\boldsymbol{x})$ as

$$\begin{aligned} G(\boldsymbol{x}) &:= \mathcal{O}_B(\boldsymbol{w}(\boldsymbol{x})) + \mathcal{O}_A(\boldsymbol{x}) \\ &:= f^*(\boldsymbol{w}(\boldsymbol{x})) + \boldsymbol{\iota}_{\mathcal{C}}^*(-(A^\top \boldsymbol{w}(\boldsymbol{x}))) + f(A\boldsymbol{x}) + \boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x}) \end{aligned}$$

the last term disappears since we assumed $\boldsymbol{x} \in \mathcal{C}$. Using the definition of the Fenchel conjugate, one has the Fenchel-Young inequality, i.e.

$$f^*(\boldsymbol{w}) := \max_{\boldsymbol{u} \in \mathbb{R}^d} \boldsymbol{w}^\top \boldsymbol{u} - f(\boldsymbol{u}) \;\Rightarrow\; f^*(\boldsymbol{w}) + f(\boldsymbol{u}) \geq \boldsymbol{w}^\top \boldsymbol{u}$$

The above holds with equality if $\boldsymbol{w}$ is chosen as a subgradient of f at $\boldsymbol{u} = A\boldsymbol{x}$. Therefore, using our first-order optimality mapping $\boldsymbol{w}(\boldsymbol{x}) := \partial f(A\boldsymbol{x})$, we have

$$G(\boldsymbol{x}) = (A\boldsymbol{x})^\top \partial f(A\boldsymbol{x}) + \boldsymbol{\iota}_{\mathcal{C}}^*(-(A^\top \boldsymbol{w}(\boldsymbol{x}))) = G_{\mathcal{C}}(\boldsymbol{x})$$

This derivation is adapted from [13, Appendix D]. □

B.2 Obtaining Information about the Optimal Points

Lemma 9 (Conjugates of Indicator Functions and Norms)**.**

i) The conjugate of the indicator function $\boldsymbol{\iota}_{\mathcal{C}}$ of a set $\mathcal{C} \subset \mathbb{R}^n$ (not necessarily convex) is the support function of the set $\mathcal{C}$, that is $\boldsymbol{\iota}_{\mathcal{C}}^*(\boldsymbol{x}) = \sup_{\boldsymbol{s} \in \mathcal{C}} \langle \boldsymbol{s}, \boldsymbol{x} \rangle$

ii) The conjugate of a norm is the indicator function of the unit ball of the dual norm.

Proof. [4, Example 3.24 and 3.26] □

Lemma 10. *Assume that f is a closed and convex function then f^* is μ-strongly convex with respect to a norm $\|\cdot\|$ if and only if f is $1/\mu$-Lipschitz gradient with respect to dual norm $\|\cdot\|_*$.*

Proof. [10, Theorem 3] □

The following lemma shows how to bound the distance between any (feasible) current dual iterate and the solution $\boldsymbol{w}^\star$ using standard assumptions on the objective functions.

Lemma 11. *Consider the problem* (B) *with optimal solution $\boldsymbol{w}^\star \in \mathbb{R}^d$. For f being μ-smooth, we have*

$$\|\boldsymbol{w} - \boldsymbol{w}^\star\|^2 \leq \frac{2}{\mu}(f^*(\boldsymbol{w}) - f^*(\boldsymbol{w}^\star)) \tag{15}$$

Proof of Lemma 11. From the definition of μ-strongly convex function, we know that

$$\begin{aligned} f^*(\boldsymbol{w}) &\geq f^*(\boldsymbol{w}^\star) + (\boldsymbol{w} - \boldsymbol{w}^\star)^\top \nabla f^*(\boldsymbol{w}^\star) + \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 \\ &\geq f^*(\boldsymbol{w}^\star) + \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 \end{aligned}$$

The first inequality follows directly by using the first order optimality condition for $\boldsymbol{w}^\star$ being optimal. For any optimal point $\boldsymbol{w}^\star$ and another feasible point $\boldsymbol{w}$,

$$(\boldsymbol{w} - \boldsymbol{w}^\star)^\top \nabla f^*(\boldsymbol{w}^\star) \geq 0.$$

Hence, $\quad \|\boldsymbol{w}^\star - \boldsymbol{w}\|_2^2 \leq \frac{2}{\mu}(f^*(\boldsymbol{w}) - f^*(\boldsymbol{w}^\star))$ □

The following corollary will be important to derive screening rules for penalized problems in Section 4, as well as box-constrained problems (Section 3.3).

Corollary 12. *We consider the problem setup* (A) *and* (B)*, and assume f is μ-smooth. Then*

$$\|\boldsymbol{w} - \boldsymbol{w}^\star\|^2 \leq \frac{2}{\mu} G(\boldsymbol{x}). \tag{16}$$

Here $G(\boldsymbol{x})$ is the duality gap function as defined in equation (13)*.*

Proof of Corollary 12. This statement directly comes from (11) and the definition of the duality gap. By definition we know that the true optimum values $-\mathcal{O}_B(\boldsymbol{w}^*)$ and $\mathcal{O}_A(\boldsymbol{x}^*)$ respectively for primal (A) and dual formulation (B) will always lie within the duality gap which implies

$$G(\boldsymbol{x}) \geq \mathcal{O}_B(\boldsymbol{w}) - \mathcal{O}_B(\boldsymbol{w}^\star)$$

By equation (B), we know that $\quad \mathcal{O}_B(\boldsymbol{w}) = f^*(\boldsymbol{w}) + g^*(-A^\top \boldsymbol{w}^\star)$

Now since f^* is μ-strongly convex function and g^* is convex hence,

$$f^*(\boldsymbol{w}) \geq f^*(\boldsymbol{w}^\star) + \nabla f^*(\boldsymbol{w}^\star)^\top (\boldsymbol{w} - \boldsymbol{w}^\star) + \frac{\mu}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^\star\right\|_2^2 \tag{17}$$

$$g^*(-A^\top \boldsymbol{w}) \geq g^*(-A^\top \boldsymbol{w}^\star) + \nabla g^*(-A^\top \boldsymbol{w}^\star)^\top (-A^\top \boldsymbol{w} + A^\top \boldsymbol{w}^\star) \tag{18}$$

Hence by adding equation (17) and (18), we get

$$\begin{aligned} \mathcal{O}_B(\boldsymbol{w}) &\geq \mathcal{O}_B(\boldsymbol{w}^\star) + (\nabla f^*(\boldsymbol{w}^\star) - A\nabla g^*(-A^\top \boldsymbol{w}^\star))^\top (\boldsymbol{w} - \boldsymbol{w}^\star) + \frac{\mu}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^\star\right\|_2^2 \\ \Rightarrow \mathcal{O}_B(\boldsymbol{w}) &\geq \mathcal{O}_B(\boldsymbol{w}^\star) + \nabla \mathcal{O}_B(\boldsymbol{w}^\star)^\top (\boldsymbol{w} - \boldsymbol{w}^\star) + \frac{\mu}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^\star\right\|_2^2 \end{aligned}$$

At optimal point $\boldsymbol{w}^\star$, $\nabla \mathcal{O}_B(\boldsymbol{w}^\star)^\top (\boldsymbol{w} - \boldsymbol{w}^\star) \geq 0$.

Hence,

$$G(\boldsymbol{x}) \geq \mathcal{O}_B(\boldsymbol{w}) - \mathcal{O}_B(\boldsymbol{w}^\star) \geq \frac{\mu}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^\star\right\|_2^2$$

□

The following two results hold for general constrained optimization problems of the form (3), where g is the indicator function of a constraint set $\mathcal{C} \subset \mathbb{R}^n$ and hence are useful for deriving screening rules for such problems.

Lemma 13. *Consider problem* (A) *and assume that f is μ-strongly convex over a bounded set $\mathcal{C}$. Then it holds that*

$$\|A\boldsymbol{x} - A\boldsymbol{x}^\star\|_2^2 \leq \frac{1}{\mu} G_{\mathcal{C}}(\boldsymbol{x}), \tag{19}$$

where $\boldsymbol{x}^\star$ is an optimal solution and $G_{\mathcal{C}}$ is the Wolfe-Gap function of f over the bounded set $\mathcal{C}$.

Proof of Lemma 13. From the definition of μ-strong convexity of f and using optimality condition,

$$\mu\left\|A\boldsymbol{x} - A\boldsymbol{x}^\star\right\|^2 \leq (A\boldsymbol{x} - A\boldsymbol{x}^\star)^\top (\nabla f(A\boldsymbol{x}) - \nabla f(A\boldsymbol{x}^\star)) \tag{20}$$

$$\leq (A\boldsymbol{x} - A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}) \tag{21}$$

$$\leq G_{\mathcal{C}}(\boldsymbol{x}) \tag{22}$$

Equation (20) comes from the definition of μ-strong convexity.

Equation (21) is first order optimality condition for $\boldsymbol{x}^\star$ being optimal which implies

$$(A\boldsymbol{x} - A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \geq 0$$

The inequality (22) follows by the definition of the gap function given in (14). □

Corollary 14. *Assuming f is L-smooth as well as μ-strongly convex over a bounded set $\mathcal{C}$, we have*

$$\|\nabla f(A\boldsymbol{x}) - \nabla f(A\boldsymbol{x}^\star)\| \leq \frac{L}{\sqrt{\mu}}\sqrt{G_{\mathcal{C}}(\boldsymbol{x})} \tag{23}$$

Proof of Corollary 14. This comes by definition of L-smooth functions and Lemma 13. From the definition,

$$\begin{aligned}\|\nabla f(A\boldsymbol{x}) - \nabla f(A\boldsymbol{x}^\star)\| &\leq L\,\|A\boldsymbol{x} - A\boldsymbol{x}^\star\| \\ &\leq \frac{L}{\sqrt{\mu}}\sqrt{G_{\mathcal{C}}(\boldsymbol{x})}\end{aligned}$$

Second inequality directly comes from Lemma 13. □

C Screening on Constrained Problems

Lemma 15. *Let $\mathcal{C}$ be a convex set, and $\boldsymbol{\iota}_{\mathcal{C}}$ be its indicator function, then*

1. *For $\boldsymbol{x} \notin \mathcal{C}$, $\partial\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x}) = \emptyset$*
2. *For $\boldsymbol{x} \in \mathcal{C}$, we have that $\boldsymbol{w} \in \partial\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x})$ if $\boldsymbol{w}^\top(\boldsymbol{z} - \boldsymbol{x}) \leq 0 \quad \forall \boldsymbol{z} \in \mathcal{C}$*

Proof. Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a closed convex set. Then subgradient of indicator function $\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x})$ at $\boldsymbol{x}$ will be vectors $\boldsymbol{u}$ which satisfy

$$\begin{aligned}&\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) \geq \boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x}) + \boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \quad \forall \boldsymbol{z} \in dom(\boldsymbol{\iota}_{\mathcal{C}}) \\ &\Rightarrow \boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) \geq \boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x}) + \boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \quad \forall \boldsymbol{z} \in \mathbb{R}^n\end{aligned} \tag{24}$$

If $Int(\mathcal{C})$ represents interior of the set such that it contains n-dimensional ball of radius $r > 0$ and $Bd(\mathcal{C})$ represents boundary of the set $\mathcal{C}$. Now we have to assume various cases for proving Lemma 15.

Case 1 We evaluate Equation (24) when $\boldsymbol{x} \in Int(\mathcal{C})$. Equation (24) becomes

$$\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) \geq \boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \quad \forall \boldsymbol{z} \in \mathbb{R}^n$$

Now since the above equation is satisfied for all $\boldsymbol{z} \in \mathbb{R}^n$, we assume $\boldsymbol{z} \in Int(\mathcal{C})$ such that $(\boldsymbol{z} - \boldsymbol{x})$ can be anywhere in the ball. Hence $\boldsymbol{u}$ needs to be 0 in this case.

Case 2 In this case we assume $\boldsymbol{x} \in Bd(\mathcal{C})$. That gives

$$\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) \geq \boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \quad \forall \boldsymbol{z} \in \mathbb{R}^n$$

If we take $\boldsymbol{z} \in \mathcal{C}$ then $\boldsymbol{u}$ satisfies $\boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \leq 0 \quad \forall \boldsymbol{z} \in \mathcal{C}$
If $\boldsymbol{z} \notin \mathcal{C}$ then $\boldsymbol{u}$ can take all the value. Hence taking intersection, $\boldsymbol{u}$ satisfies

$$\boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \leq 0 \quad \forall \boldsymbol{z} \in \mathcal{C}$$

Case 3 When we assume $\boldsymbol{x} \notin \mathcal{C}$, we get

$$\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) \geq +\infty + \boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \quad \forall \boldsymbol{z} \in \mathbb{R}^n$$

If we again take $\boldsymbol{z} \in \mathcal{C}$ then no finite $\boldsymbol{u}$ can satisfy the equation $\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) \geq +\infty + \boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{x}) \quad \forall \boldsymbol{z} \in \mathcal{C}$ because $\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) = 0$ if $\boldsymbol{z} \in \mathcal{C}$.
And if $\boldsymbol{z} \notin \mathcal{C} \Rightarrow \boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{z}) = +\infty$ then again nothing can be said about the vector $\boldsymbol{u}$. Hence by convention it is assumed that $\boldsymbol{x} \notin \mathcal{C} \Rightarrow \boldsymbol{u} \in \emptyset$

By the above arguments we conclude that,

1. For $\boldsymbol{x} \notin \mathcal{C}$, $\partial\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x}) = \emptyset$
2. For $\boldsymbol{x} \in \mathcal{C}$, we have that $\boldsymbol{w} \in \partial\boldsymbol{\iota}_{\mathcal{C}}(\boldsymbol{x})$ if $\boldsymbol{w}^\top(\boldsymbol{z} - \boldsymbol{x}) \leq 0 \quad \forall \boldsymbol{z} \in \mathcal{C}$

Hence the claim made in Lemma 15 is proved. □

C.1 Screening on Simplex Constrained Problems (Section 3.1)

General Simplex Constrained Screening

Proof of Theorem 1. In the simplex case, we have $g(\boldsymbol{x}) = \boldsymbol{\iota}_{\triangle}(\boldsymbol{x})$ and by Lemma 15

$$\partial g(\boldsymbol{x}^\star) = \{ \boldsymbol{s} \mid \forall \boldsymbol{z} \in \triangle \;\; \boldsymbol{s}^\top(\boldsymbol{z} - \boldsymbol{x}^\star) \le 0 \;\; \}$$
$$= \{ \boldsymbol{s} \mid \forall \boldsymbol{z} \in \triangle \;\; \boldsymbol{s}^\top \boldsymbol{z} \le \boldsymbol{s}^\top \boldsymbol{x}^\star \;\; \} \tag{25}$$

Now, by the optimality condition (2a), $-A^\top \boldsymbol{w}^\star \in \partial g(\boldsymbol{x}^\star)$ and since this holds, hence $-A^\top \boldsymbol{w}^\star$ should satisfy the required constrained which is needed to be in the set of subgradients of $\partial g(\boldsymbol{x}^\star)$ according to conditions in equation (25). Hence,

$$(-A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z} \le (-A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \quad \forall \boldsymbol{z} \in \triangle \tag{26}$$
$$\Rightarrow (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \le (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z} \quad \forall \boldsymbol{z} \in \triangle \tag{27}$$
$$\Rightarrow (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \le \min_z (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z} \quad s.t \;\; \boldsymbol{z} \in \triangle \tag{28}$$
$$\Rightarrow (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \le \min_i \boldsymbol{a}_i^\top \boldsymbol{w}^\star \tag{29}$$
$$\Rightarrow (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star \le \min_i \boldsymbol{a}_i^\top \boldsymbol{w}^\star \tag{30}$$
$$\Rightarrow (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star = \min_i \boldsymbol{a}_i^\top \boldsymbol{w}^\star \tag{31}$$

Equation (29) is due to the fact that $\boldsymbol{z}$ lie in the simplex, hence minimum value of $(A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z}$ is $\min_i \boldsymbol{a}_i^\top \boldsymbol{w}^\star$ and equation (31) also comes from the same fact that $\boldsymbol{x}^\star$ lie in the simplex and hence $(A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star$ can not be smaller than $\min_i \boldsymbol{a}_i^\top \boldsymbol{w}^\star$. That implies these two quantities need to be equal and all the i's where this equality doesn't hold refers to $x_i^\star = 0$ for all such i's.

$$\boldsymbol{a}_i^\top \boldsymbol{w}^\star > \; (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \;\Rightarrow\; x_i = 0$$
$$(\boldsymbol{a}_i - A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star > 0 \;\Rightarrow\; x_i = 0$$

□

Proof of Theorem 2. From the optimality condition (1a), we have $\boldsymbol{w}^\star = \nabla f(A\boldsymbol{x}^\star)$ since f is differentiable. Hence,

$$(\boldsymbol{a}_i - A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star = (\boldsymbol{a}_i - A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \tag{32}$$
$$= (\boldsymbol{a}_i - A\boldsymbol{x}^\star + A\boldsymbol{x} - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}^\star) \tag{33}$$
$$= (\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}^\star) + (A\boldsymbol{x} - A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \tag{34}$$
$$\ge (\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}^\star) \quad \{\text{From the optimality of } f(A\boldsymbol{x})\} \tag{35}$$
$$= (\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) - (\boldsymbol{a}_i - A\boldsymbol{x})^\top (\nabla f(A\boldsymbol{x}) - \nabla f(A\boldsymbol{x}^\star)) \tag{36}$$
$$\ge (\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) - \|\boldsymbol{a}_i - A\boldsymbol{x}\| \, \|\nabla f(A\boldsymbol{x}) - \nabla f(A\boldsymbol{x}^\star)\| \tag{37}$$
$$\ge (\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) - L\sqrt{\frac{G_{\mathcal{C}}(\boldsymbol{x})}{\mu}} \, \|\boldsymbol{a}_i - A\boldsymbol{x}\| \tag{38}$$

Eq. (35) comes from the fact that at the optimal point $\boldsymbol{x}^\star$, the inequality $(A\boldsymbol{x} - A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \ge 0$ holds $\forall\, \boldsymbol{x}$. Equation (38) comes from Corollary 14 for smooth function f over a constrained set $\mathcal{C}$. Hence from Theorem 1, we obtain the screening rule

$$(\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) > L\sqrt{\frac{G_{\mathcal{C}}(\boldsymbol{x})}{\mu}} \, \|\boldsymbol{a}_i - A\boldsymbol{x}\| \;\Rightarrow\; x_i^\star = 0$$

□

Squared Hinge Loss SVM. The squared hinge-loss SVM problem in its dual form is formulated as

$$\min_{\boldsymbol{x}} \left[f(A\boldsymbol{x}) := \tfrac{1}{2}\boldsymbol{x}^\top A^\top A\boldsymbol{x}\right] \quad \text{such that} \quad \boldsymbol{x} \in \triangle\ . \tag{39}$$

Here for given points $\bar{\boldsymbol{a}}_1, \ldots, \bar{\boldsymbol{a}}_n \in \mathbb{R}^d$ and corresponding labels $y_i \in \pm 1$, the matrix A collects the columns $\boldsymbol{a}_i = y_i \bar{\boldsymbol{a}}_i$, see e.g. [25]. We obtain the following novel screening rule for square loss SVM:

Corollary 16. *For the squared hinge loss SVM* (39) *we have the screening rule*

$$(\boldsymbol{a}_i - A\boldsymbol{x})^\top A\boldsymbol{x} > \sqrt{\max_i\ (A\boldsymbol{x} - \boldsymbol{a}_i)^\top A\boldsymbol{x}}\|\boldsymbol{a}_i - A\boldsymbol{x}\| \ \Rightarrow\ x_i^\star = 0. \tag{40}$$

Proof of Corollary 16. Theorem 2 is directly applicable to problems of the form (39). The objective function $f(\boldsymbol{y}) = f(A\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A^\top A\boldsymbol{x}$ is strongly convex with parameter $\mu = 1$. Also the derivative ∇f is Lipschitz-continuous with parameter $L = 1$. To obtain an upper bound on the distance between any approximate solution and the optimal solution $\|A\boldsymbol{x} - A\boldsymbol{x}^\star\|$, we employ Lemma 13. Since the constrained of the optimization problem is unit simplex and hence the value of Wolfe gap function $G_{\mathcal{C}}(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{C}}\ (A\boldsymbol{x} - A\boldsymbol{y})^\top \nabla f(A\boldsymbol{x})$ will be attained on one of the vertices. So, $G_{\mathcal{C}}(\boldsymbol{x}) = \max_{i\in 1\ldots m}\ (A\boldsymbol{x} - \boldsymbol{a}_i)^\top A\boldsymbol{x}$. Finally, Theorem 2 gives us the screening rule for squared hinge loss SVM:

$$(\boldsymbol{a}_i - A\boldsymbol{x})^\top A\boldsymbol{x} > \sqrt{\max_{i\in 1\ldots m}\ (A\boldsymbol{x} - \boldsymbol{a}_i)^\top A\boldsymbol{x}}\ \|\boldsymbol{a}_i - A\boldsymbol{x}\| \ \Rightarrow\ x_i^\star = 0 \tag{41}$$

□

Screening on Minimum Enclosing Ball.

Minimum Enclosing Ball - Given a set of n points, $\boldsymbol{a}_1$ to $\boldsymbol{a}_n$ in $\mathbb{R}^d$, the minimum enclosing ball is defined as the smallest ball $B_{\boldsymbol{c},r}$ with center $\boldsymbol{c}$ and radius r, i.e.: $B_{\boldsymbol{c},r} := \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{c} - \boldsymbol{x}\| \le r\}$, such that all points $\boldsymbol{a}_i$ lie in its interior. In this set-up, screening means to identify points $\boldsymbol{a}_i$ lying in the interior of the optimal ball $B_{\boldsymbol{c}^\star,r^\star}$. Removing those points from the problem does not change the optimal ball. The primal and dual for the minimum enclosing ball problem is given as the following pair of optimization formulations (42) and (43) respectively.

$$\min_{\boldsymbol{c}\in\mathbb{R}^d, r\in\mathbb{R}} \quad r^2 \quad \text{s.t.} \quad \|\boldsymbol{c} - \boldsymbol{a}_i\|_2^2 \le r^2 \quad \forall i \in [n] \tag{42}$$

$$\min_{\boldsymbol{x}\in\triangle\subset\mathbb{R}^n} \quad \boldsymbol{x}^\top A^\top A\boldsymbol{x} + \boldsymbol{c}^\top \boldsymbol{x}\ , \tag{43}$$

where $\boldsymbol{c}$ is a vector whose i^{th} element c_i is $-\boldsymbol{a}_i^\top \boldsymbol{a}_i$, see for example [16] or our Appendix C.1.

Corollary 17. *For the minimum enclosing ball problem* (42) *we have the screening rule*

$$(\boldsymbol{e}_i - \boldsymbol{x})^\top (2A^\top A\boldsymbol{x} + \boldsymbol{c}) > 2\sqrt{\frac{1}{2}\max_i (\boldsymbol{x} - \boldsymbol{e}_i)^\top (2A^\top A\boldsymbol{x} + \boldsymbol{c})}\ \|\boldsymbol{a}_i - A\boldsymbol{x}\| \ \Rightarrow\ x_i^\star = 0\ . \tag{44}$$

Our result improves upon the known rules by [11, 1] by providing a broader selection criterion (44).

Proof of Corollary 17. The minimum enclosing ball problem can be formulated as an optimization problem of the form given in Equation (42):

$$\min_{\boldsymbol{c},r} \quad r^2 \quad \text{s.t.}\ \|\boldsymbol{c} - \boldsymbol{a}_i\|_2^2 \le r^2 \quad \forall i \in [n]$$

As we have seen, the dual formulation can be written in the form of Equation (43) as given in [16, Chapter 8.7]:

$$\min_{\boldsymbol{x}}\ \boldsymbol{x}^\top A^\top A\boldsymbol{x} - \sum_{j=1}^{p} \boldsymbol{a}_j^\top \boldsymbol{a}_j x_j \quad \text{s.t. } \boldsymbol{x} \in \triangle$$

Now the function $\boldsymbol{x}^\top A^\top A\boldsymbol{x} - \sum_{j=1}^{p} \boldsymbol{a}_j^\top \boldsymbol{a}_j x_j$ is strongly convex in $A\boldsymbol{x}$ with parameter $\mu = 2$. Since the constrained of the optimization problem is unit simplex and hence the value of the Wolfe gap function $G_{\mathcal{C}}(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{C}}\ (A\boldsymbol{x} - A\boldsymbol{y})^\top \nabla f(A\boldsymbol{x})$ as defined in Appendix B will be attained at one of the vertices of unit simplex. Hence Corollary 14 gives $G_{\mathcal{C}}(\boldsymbol{x}) = \sqrt{\frac{1}{2}\max_i (\boldsymbol{x} - \boldsymbol{e}_i)^\top (2A^\top A\boldsymbol{x} + \boldsymbol{c}')}$. Now applying the findings of Theorem 2, we get a sufficient condition for $\boldsymbol{a}_i$ to be non-influential, i.e. $\boldsymbol{a}_i$ lies in the interior of the MEB. But before that we will simplify the left hand side of the theorem

2 a bit. $(\boldsymbol{a}_i - A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x})$ can we written as $(\boldsymbol{e}_i - \boldsymbol{x})^\top A^\top \nabla f(A\boldsymbol{x})$. Hence we get our result claimed in Corollary 17.

$$(\boldsymbol{e}_i - \boldsymbol{x})^\top (2A^\top A\boldsymbol{x} + \boldsymbol{c}') > 2\sqrt{\frac{1}{2}\max_j (\boldsymbol{x} - \boldsymbol{e}_i)^\top (2A^\top A\boldsymbol{x} + \boldsymbol{c}')}\, \|\boldsymbol{a}_i - A\boldsymbol{x}\| \ \Rightarrow\ x_i^\star = 0 \tag{45}$$

That means $\boldsymbol{a}_i$ is non influential. □

C.2 Screening on L_1-ball Constrained Problems

Proof of Theorem 3. In the constrained Lasso case, we have $g(\boldsymbol{x}) = \iota_{\mathcal{B}_{L_1}}(\boldsymbol{x})$ and by Lemma 15

$$\begin{aligned}\partial g(\boldsymbol{x}^\star) &= \{\boldsymbol{s} \mid \forall \boldsymbol{z} \in \mathcal{B}_{L_1}\ \ \boldsymbol{s}^\top(\boldsymbol{z} - \boldsymbol{x}^\star) \le 0\ \ \}\\ &= \{\boldsymbol{s} \mid \forall \boldsymbol{z} \in \mathcal{B}_{L_1}\ \ \boldsymbol{s}^\top \boldsymbol{z} \le \boldsymbol{s}^\top \boldsymbol{x}^\star\ \ \}\end{aligned} \tag{46}$$

Now, by the optimality condition (2a), $-A^\top \boldsymbol{w}^\star \in\ \partial g(\boldsymbol{x}^\star)$ and since this holds, hence $-A^\top \boldsymbol{w}^\star$ should satisfy the required constrained which is needed to be in the set of subgradients of $\partial g(\boldsymbol{x}^\star)$ according to conditions in equation (46). Hence,

$$(-A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z} \le (-A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \quad \forall \boldsymbol{z} \in \mathcal{B}_{L_1} \tag{47}$$

$$\Rightarrow (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \le (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z} \quad \forall \boldsymbol{z} \in \mathcal{B}_{L_1} \tag{48}$$

$$\Rightarrow (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \le \min_z (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z} \quad s.t\ \ \boldsymbol{z} \in \mathcal{B}_{L_1} \tag{49}$$

$$\Rightarrow\ (A^\top \boldsymbol{w}^\star)^\top \boldsymbol{x}^\star \le -\max_i\ \left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| \tag{50}$$

$$\Rightarrow (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star \le -\max_i\ \left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| \tag{51}$$

$$\Rightarrow (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star = -\max_i\ \left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| \tag{52}$$

Equation (50) is due to the fact that $\boldsymbol{z}$ lie in the L_1-ball and hence minimum value of $(A^\top \boldsymbol{w}^\star)^\top \boldsymbol{z}$ is $-\max_i\ \left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right|$ and Equation (52) also comes from the same fact that $\boldsymbol{x}^\star$ lie in the L_1-ball and hence $(A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star$ can not be smaller than $-\max_i\ \left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right|$. That implies these two quantities need to be equal and all the i's where this equality doesn't hold refers to $x_i^\star = 0$ for all such i's. Hence whenever these two quantities are not equal this holds:

$$\begin{aligned}-\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| > (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star \ &\Rightarrow\ x_i^\star = 0\\ \Rightarrow\ \left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| + (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star < 0 \ &\Rightarrow\ x_i^\star = 0\end{aligned}$$

□

Proof of Theorem 4. Using optimality condition (1a), we know that $\boldsymbol{w}^\star \in \partial f(A\boldsymbol{x})$

$$\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| + (A\boldsymbol{x}^\star)^\top \boldsymbol{w}^\star = \left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x}^\star)\right| + (A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \tag{53}$$

$$= \left|\boldsymbol{a}_i^\top (\nabla f(A\boldsymbol{x}) - \nabla f(A\boldsymbol{x}) + \nabla f(A\boldsymbol{x}^\star))\right| + (A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \tag{54}$$

$$\le \left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + \left|\boldsymbol{a}_i^\top (\nabla f(A\boldsymbol{x}^\star) - \nabla f(A\boldsymbol{x}))\right| + (A\boldsymbol{x}^\star - A\boldsymbol{x} + A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}^\star) \tag{55}$$

$$= \left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + \left|\boldsymbol{a}_i^\top (\nabla f(A\boldsymbol{x}^\star) - \nabla f(A\boldsymbol{x}))\right| + (A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}^\star) - (A\boldsymbol{x} - A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \tag{56}$$

$$\le \left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + \left|\boldsymbol{a}_i^\top (\nabla f(A\boldsymbol{x}^\star) - \nabla f(A\boldsymbol{x}))\right| + (A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}^\star) \tag{57}$$

$$\le \left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + \left|\boldsymbol{a}_i^\top (\nabla f(A\boldsymbol{x}^\star) - \nabla f(A\boldsymbol{x}))\right| + (A\boldsymbol{x})^\top (\nabla f(A\boldsymbol{x}^\star) - \nabla f(A\boldsymbol{x}) + \nabla f(A\boldsymbol{x})) \tag{58}$$

$$\le \left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + \left|\boldsymbol{a}_i^\top (\nabla f(A\boldsymbol{x}^\star) - \nabla f(A\boldsymbol{x}))\right| + (A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) + (A\boldsymbol{x})^\top (\nabla f(A\boldsymbol{x}^\star) - \nabla f(A\boldsymbol{x})) \tag{59}$$

$$\le \left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + (A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) + L(\|\boldsymbol{a}_i\| + \|A\boldsymbol{x}\|)\sqrt{\frac{G_{\mathcal{C}}(\boldsymbol{x})}{\mu}} \tag{60}$$

Eq. (56) comes from the fact that at the optimal point $\boldsymbol{x}^\star$, the inequality $(A\boldsymbol{x} - A\boldsymbol{x}^\star)^\top \nabla f(A\boldsymbol{x}^\star) \geq 0$ holds $\forall \boldsymbol{x}$. Hence using Theorem 3, Lemma 13 and Corollary 14, we get the screening rule for L_1 constrained as whenever,

$$\left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| + (A\boldsymbol{x})^\top \nabla f(A\boldsymbol{x}) + L(\|\boldsymbol{a}_i\| + \|A\boldsymbol{x}\|)\sqrt{\frac{G_{\mathcal{C}}(\boldsymbol{x})}{\mu}} < 0 \;\Rightarrow\; \boldsymbol{x}_i^\star = 0 \qquad \square$$

C.3 Screening for Box Constrained Problems

Screening for General Box Constrained Problems (Section 3.3)

Proof of Theorem 5. The box-constrained case can be seen in the form of the partially separable optimization problem pair (SA) and (SB). According to optimality condition (12a) for this case, we have

$$-\boldsymbol{a}_i^\top \boldsymbol{w}^\star \in \partial g_i(x_i^\star) \qquad \forall i \tag{61}$$

Now from the definition of subgradient for an indicator function as given in Lemma 15. Also since x_i is a number now, we will get rid of the transpose here.

$$\begin{aligned}\partial g(x_i^\star) &= \{s \mid 0 \leq z \leq C,\;\; s(z - x_i^\star) \leq 0 \quad \} \\ &= \{s \mid 0 \leq z \leq C,\;\; sz \leq s x_i^\star \quad \}\end{aligned} \tag{62}$$

Now, by the optimality condition (12a), $-\boldsymbol{a}_i^\top \boldsymbol{w}^\star \in \;\partial g(x_i^\star)$ and since this holds, hence $-\boldsymbol{a}_i^\top \boldsymbol{w}^\star$ should satisfy the required constrained which is needed to be in the set of subgradients of $\partial g(x_i^\star)$ according to conditions in Equation (62). Hence,

$$\begin{aligned}(-\boldsymbol{a}_i^\top \boldsymbol{w}^\star)z &\leq (-\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star \quad \forall z \;\text{ s.t }\; 0 \leq z \leq C, \\ \Rightarrow \min_z(\boldsymbol{a}_i^\top \boldsymbol{w}^\star)z &\geq (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star \quad s.t \;\; 0 \leq z \leq C\end{aligned} \tag{63}$$

Now (63) can be manipulated in two ways

Case 1

$$\begin{aligned}\boldsymbol{a}_i^\top \boldsymbol{w}^\star > 0 \Rightarrow \min_z(\boldsymbol{a}_i^\top \boldsymbol{w}^\star)^\top z &\geq (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star \quad s.t \;\; 0 \leq z \leq C \\ \Rightarrow 0 &\geq (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)^\top x_i^\star\end{aligned}$$

But since $\boldsymbol{a}_i^\top \boldsymbol{w}^\star > 0$ and also $x_i^\star \geq 0$ hence $(\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star \nless 0$. This implies $(\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star = 0$ and hence if $\boldsymbol{a}_i^\top \boldsymbol{w}^\star > 0 \;\Rightarrow\; x_i^\star = 0$

Case 2

$$\begin{aligned}\boldsymbol{a}_i^\top \boldsymbol{w}^\star < 0 \Rightarrow \min_z(\boldsymbol{a}_i^\top \boldsymbol{w}^\star)z &\geq (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star \quad s.t \;\; 0 \leq z \leq C \\ \Rightarrow (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)C &\geq (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star\end{aligned}$$

But since $\boldsymbol{a}_i^\top \boldsymbol{w}^\star < 0$ and also $x_i^\star \leq C$ hence $(\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star \nless (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)C$. This implies $(\boldsymbol{a}_i^\top \boldsymbol{w}^\star)x_i^\star = (\boldsymbol{a}_i^\top \boldsymbol{w}^\star)C$ and hence if $\boldsymbol{a}_i^\top \boldsymbol{w}^\star < 0 \;\Rightarrow\; x_i^\star = C$

Final optimality arguments can be given as

$$\begin{aligned}\boldsymbol{a}_i^\top \boldsymbol{w}^\star > 0 \;&\Rightarrow\; x_i^\star = 0 \\ \boldsymbol{a}_i^\top \boldsymbol{w}^\star < 0 \;&\Rightarrow\; x_i^\star = C\end{aligned} \tag{64}$$

Now

$$\begin{aligned}\boldsymbol{a}_i^\top \boldsymbol{w}^\star = \boldsymbol{a}_i^\top(\boldsymbol{w}^\star + \boldsymbol{w} - \boldsymbol{w}) &= \boldsymbol{a}_i^\top \boldsymbol{w} + \boldsymbol{a}_i^\top(\boldsymbol{w}^\star - \boldsymbol{w}) \\ \boldsymbol{a}_i^\top \boldsymbol{w} - \|\boldsymbol{a}_i\|_2 \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \leq \boldsymbol{a}_i^\top \boldsymbol{w}^\star &\leq \boldsymbol{a}_i^\top \boldsymbol{w} + \|\boldsymbol{a}_i\|_2 \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2\end{aligned} \tag{65}$$

Since f is L-Lipschitz gradient hence f^* is $1/L$-strongly convex, hence using Lemmas 11 and 10, Equation (64) becomes

$$\boldsymbol{a}_i^\top \boldsymbol{w} - \|\boldsymbol{a}_i\|_2 \sqrt{2L\, G(\boldsymbol{x})} \leq \boldsymbol{a}_i^\top \boldsymbol{w}^\star \leq \boldsymbol{a}_i^\top \boldsymbol{w} + \|\boldsymbol{a}_i\|_2 \sqrt{2L\, G(\boldsymbol{x})} \tag{66}$$

Hence using equation (66) and earlier arguments we get,

$$\boldsymbol{a}_i^\top \boldsymbol{w}^\star > 0 \;\Rightarrow\; x_i^\star = 0$$
$$\Rightarrow \boldsymbol{a}_i^\top \boldsymbol{w} - \|\boldsymbol{a}_i\|_2 \sqrt{2L\,G(\boldsymbol{x})} > 0 \;\Rightarrow\; x_i^\star = 0$$

And if

$$\boldsymbol{a}_i^\top \boldsymbol{w}^\star < 0 \;\Rightarrow\; x_i^\star = C$$
$$\Rightarrow \boldsymbol{a}_i^\top \boldsymbol{w} + \|\boldsymbol{a}_i\|_2 \sqrt{2L\,G(\boldsymbol{x})} > 0 \;\Rightarrow\; x_i^\star = C$$

□

Screening on SVM with hinge loss and no bias

Hinge Loss SVM. The dual of the classical support vector machine with hinge loss, when not using a bias value, is a box-constrained problem. As a direct consequence of Theorem 5 we therefore obtain screening rules for SVM with hinge loss and no bias. The primal formulation of the SVM in this setting, for a regularization parameter $C > 0$, is

$$\min_{\boldsymbol{w}\in\mathbb{R}^d, \boldsymbol{\epsilon}\in\mathbb{R}^n} \quad \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\mathbf{1}^\top \boldsymbol{\epsilon} \quad \text{s.t.} \quad \boldsymbol{w}^\top \boldsymbol{a}_i \geq 1 - \varepsilon_i \;\forall i \in [n] \qquad \varepsilon_i \geq 0 \quad \forall i \in [n] \tag{67}$$

Corollary 18. *For SVM with hinge loss and no bias as given in* (67)*, we have the screening rules*

$$\boldsymbol{a}_i^\top A\boldsymbol{x} - \|\boldsymbol{a}_i\|_2 \sqrt{2G(\boldsymbol{x})} > 0 \;\Rightarrow\; x_i^\star = 0\,, \text{ and}$$
$$\boldsymbol{a}_i^\top A\boldsymbol{x} + \|\boldsymbol{a}_i\|_2 \sqrt{2G(\boldsymbol{x})} < 0 \;\Rightarrow\; x_i^\star = C\,.$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is any feasible dual point.

The closest known result to our Corollary 18 for screening in hinge loss SVM is given in [29]. The work of [29] also covers the kernelized SVM case, and improves the threshold given in our Corollary 18 by a constant of $\sqrt{2}$. In Appendix C.3, we show that our more general approach here can also be adjusted to gain this constant factor.

Proof of Corollary 18. Here the primal problem is given by:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\epsilon}} \quad & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\mathbf{1}^\top \boldsymbol{\epsilon} \\ \text{s.t.} \quad & \boldsymbol{w}^\top \boldsymbol{a}_i \geq 1 - \epsilon_i \qquad \forall i \in \{1:p\} \\ & \epsilon_i \geq 0 \qquad\qquad\quad \forall i \in \{1:p\} \end{aligned} \tag{68}$$

A dual formulation of the problem can be written as:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & -\boldsymbol{x}^\top \mathbf{1} + \frac{1}{2}\boldsymbol{x}^\top A^\top A\boldsymbol{x} \\ \text{s.t.} \quad & 0 \leq \boldsymbol{x} \leq C\mathbf{1} \end{aligned} \tag{69}$$

Theorem 5 is applied on the dual formulation. The objective function $\frac{1}{2}\boldsymbol{x}^\top A^\top A\boldsymbol{x} - \boldsymbol{x}^\top \mathbf{1}$ is strongly convex with parameter 1 and its derivative Lipschitz continuous with parameter 1. The duality gap between primal and dual feasible points $G(\boldsymbol{w}, \boldsymbol{\epsilon}, \boldsymbol{x})$ is now used as suboptimality certificate which can play the role of the upper bound $\|\boldsymbol{w} - \boldsymbol{w}^\star\|$ using Lemma 12. For a given $\boldsymbol{x}$ a primal feasible point can be obtained by setting $\boldsymbol{w} = A\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ minimal such that the first constraint of the primal problem is satisfied. Using the obtained point for the duality gap, it only depends on the point $\boldsymbol{x}$. All together this gives the screening rule:

$$\boldsymbol{a}_i^\top A\boldsymbol{x} + 1 > \|\boldsymbol{a}_i\| \sqrt{2G(\boldsymbol{x})} \quad \Rightarrow \quad x_i^\star = 0 \tag{70}$$
$$\boldsymbol{a}_i^\top A\boldsymbol{x} + 1 < -\|\boldsymbol{a}_i\| \sqrt{2G(\boldsymbol{x})} \quad \Rightarrow \quad x_i^\star = C \tag{71}$$

□

Note - Since the primal and dual of hinge loss SVM have very nice structure with smooth quadratic function with an addition to piece-wise linear convex function, hence it is not hard to show that both primal and dual function is 1 strongly convex as shown in [29]. For more detailed proof, we recommend to go through [29]. Now for an instance, if we write duality gap function as a function of $\boldsymbol{w}$ then

$$G(\boldsymbol{w}) \geq G(\boldsymbol{w}^{\star}) + \nabla G(\boldsymbol{w}^{\star})^{\top}(\boldsymbol{w} - \boldsymbol{w}^{\star}) + \|\boldsymbol{w} - \boldsymbol{w}^{\star}\|_2^2$$

Since strong duality hold in SVM case, hence at optimal point $\boldsymbol{w}^{\star}$, $G(\boldsymbol{w}^{\star}) = 0$. Finally we get,

$$G(\boldsymbol{w}) \geq \|\boldsymbol{w} - \boldsymbol{w}^{\star}\|_2^2$$

Hence the screening rule comes out as given in [29]:

$$\boldsymbol{a}_i^{\top} A\boldsymbol{x} + 1 > \|\boldsymbol{a}_i\| \sqrt{G(\boldsymbol{x})} \quad \Rightarrow \quad x_i^{\star} = 0 \tag{72}$$

$$\boldsymbol{a}_i^{\top} A\boldsymbol{x} + 1 < -\|\boldsymbol{a}_i\| \sqrt{G(\boldsymbol{x})} \quad \Rightarrow \quad x_i^{\star} = C \tag{73}$$

D Screening on Penalized Problems

D.1 Screening L_1-regularized Problems

Lemma 19. *Considering general L_1-regularized optimization problems*

$$\min_{\boldsymbol{x} \in \mathbb{R}^p} \quad f(A\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_1 \tag{74}$$

At optimum points $\boldsymbol{x}^{\star}$ and dual optimal point $\mathbf{w}^{\star}$, the following rule is satisfied for the above problem formulation (74) *:*

$$\left|\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}\right| < \lambda \;\Rightarrow\; x_i^{\star} = 0$$

Proof . Since the optimization problem (74) comes under the partially separable framework and we can use the first order optimality condition (12a) as well as (12b) to derive screening rules for the problem. Also we know that, the conjugate of the norm function is the indicator function of its dual norm ball. By optimality condition (12b), we know that

$$x_i \in \; \partial g_i^*(-\boldsymbol{a}_i^{\top}\boldsymbol{w})$$

g_i^* here is the indicator function written as $\boldsymbol{\iota}_{L_\infty}(-\boldsymbol{a}_i^{\top}\boldsymbol{w})$. Hence for the indicator function g^* by Lemma 15

$$\begin{aligned}\partial g_i^*(-\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) &= \left\{ s| \;\; \forall \boldsymbol{z} \; s.t \;\; \left|\frac{\boldsymbol{a}_i^{\top}\boldsymbol{z}}{\lambda}\right| \leq 1\,;\; s(-\boldsymbol{a}_i^{\top}\boldsymbol{z} + \boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \leq 0 \quad \right\} \\ &= \left\{ s| \;\; \forall \boldsymbol{z} \; s.t \;\; \left|\boldsymbol{a}_i^{\top}\boldsymbol{z}\right| \leq \lambda\,;\; s(\boldsymbol{a}_i^{\top}\boldsymbol{z}) \geq s(\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \quad \right\}\end{aligned}$$

Since the optimality condition (12b) holds hence $-x_i^{\star}$ should satisfy the required constrained which is needed to be in the set of subgradients of $\partial g_i^*(-\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star})$ according to conditions given above. That is

$$-\,x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{z}) \leq -x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \quad \forall \boldsymbol{z} \; s.t \;\; \left|\boldsymbol{a}_i^{\top}\boldsymbol{z}\right| \leq \lambda \tag{75}$$

$$x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{z}) \geq x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \quad \forall \boldsymbol{z} \; s.t \;\; \left|\boldsymbol{a}_i^{\top}\boldsymbol{z}\right| \leq \lambda \tag{76}$$

$$\Rightarrow x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \leq \min_{z}(x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{z})) \quad s.t \quad \left|\boldsymbol{a}_i^{\top}\boldsymbol{z}\right| \leq \lambda \tag{77}$$

Case 1: $x_i^{\star} > 0$

$$\begin{aligned}&x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \leq \min_{z}(x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{z})) \quad s.t \quad \left|\boldsymbol{a}_i^{\top}\boldsymbol{z}\right| \leq \lambda \\ &\Rightarrow\; x_i^{\star}(\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \leq -\lambda x_i^{\star} \\ &\Rightarrow (\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) \leq -\lambda \\ &\Rightarrow (\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}) = -\lambda \end{aligned} \tag{78}$$

Equation (78) comes from the fact that $\left|\boldsymbol{a}_i^{\top}\boldsymbol{w}^{\star}\right| \leq \lambda$

Case 2: $x_i^\star < 0$

$$\begin{aligned}
& x_i^\star(\boldsymbol{a}_i^\top \boldsymbol{w}^\star) \le \min_z (x_i^\star(\boldsymbol{a}_i^\top \boldsymbol{z})) \quad s.t \quad \left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le \lambda \\
& \Rightarrow\ x_i^\star(\boldsymbol{a}_i^\top \boldsymbol{w}^\star) \le \lambda x_i^\star \\
& \Rightarrow (\boldsymbol{a}_i^\top \boldsymbol{w}^\star) \ge \lambda \\
& \Rightarrow (\boldsymbol{a}_i^\top \boldsymbol{w}^\star) = \lambda
\end{aligned} \tag{79}$$

Equation (78) comes from the fact that $\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| \le \lambda$

Case 3: $x_i^\star = 0$ Since if we assume f as a continuous smooth function then $\boldsymbol{a}_i^\top \boldsymbol{w}^\star$ is also continuous. Now if we consider arguments given for $x_i^\star < 0$ and $x_i^\star > 0$ we conclude that $\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| = \lambda$ in all of the above two cases. Since $x_i^\star = 0$ is in the domain of the function (A), hence at $x_i^\star = 0$, $\boldsymbol{a}_i^\top \boldsymbol{w}^\star$ will lie in the open range of $-\lambda$ to λ. Which implies whenever $\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| < \lambda$, then $x_i^\star = 0$

Another view on the proof can be derived from the optimality condition (12a).

The optimization problem (74) can be taken as partially separable problem and from the optimality condition (12a) kk

$$-\,\boldsymbol{a}_i^\top \boldsymbol{w}^\star \in \partial g_i(x_i^\star) \tag{80}$$

$$\partial g_i(x_i^\star) \in \begin{cases} \lambda \frac{x_i^\star}{|x_i^\star|} & \text{if } x_i \neq 0 \\ [-\lambda, \lambda] & \text{if } x_i = 0 \end{cases} \tag{81}$$

From equations (80) and (81) we conclude that if

$$\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| < \lambda \ \Rightarrow\ x_i^\star = 0$$

□

Proof of Theorem 6. From Equation (1a), we know that $\boldsymbol{w}^\star \in \partial f(A\boldsymbol{x}^\star)$. Hence from Lemma 19,

$$\begin{aligned}
\left|\boldsymbol{a}_i^\top \boldsymbol{w}^\star\right| &= \left|\boldsymbol{a}_i^\top (\boldsymbol{w}^\star - \boldsymbol{w} + \boldsymbol{w})\right| \\
&\le \left|\boldsymbol{a}_i^\top \boldsymbol{w}\right| + \left|\boldsymbol{a}_i^\top (\boldsymbol{w}^\star - \boldsymbol{w})\right| \\
&\le \left|\boldsymbol{a}_i^\top \boldsymbol{w}\right| + \|\boldsymbol{a}_i\|_2 \|\boldsymbol{w}^\star - \boldsymbol{w}\|_2 \\
&\le \left|\boldsymbol{a}_i^\top \boldsymbol{w}\right| + \|\boldsymbol{a}_i\|_2 \sqrt{2L\, G(\boldsymbol{x})}
\end{aligned} \tag{82}$$

Eq. (82) comes from Corollary 12. Now using Lemma 19 and equation (82), we get

$$\left|\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x})\right| < \lambda - \|\boldsymbol{a}_i\|_2 \sqrt{2L\, G(\boldsymbol{x})} \ \Rightarrow\ \boldsymbol{x}_i^\star = 0$$

□

Penalized LASSO

Corollary 20. ***Penalized LASSO (Can be derived from GAP SAFE paper [17])-*** *Consider the optimization problem of the form:*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|A\boldsymbol{x} - b\|_2^2 + \lambda\|\boldsymbol{x}\|_1$$

Then the screening rule is given by: $\left|\boldsymbol{a}_i^\top (A\boldsymbol{x} - \boldsymbol{b})\right| < \lambda - \|\boldsymbol{a}_i\|_2 \sqrt{2G(\boldsymbol{x})} \ \Rightarrow\ \boldsymbol{x}_i^\star = 0.$

Proof of Corollary 20. By observing the cost function for penalized lasso it can be concluded that

$$f(A\boldsymbol{x}) = \frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|^2, \quad \boldsymbol{w} = A\boldsymbol{x} - \boldsymbol{b}, \quad \text{and } L = 1$$

Now results from Theorem 6 can be directly applied here and hence the screening rule becomes

$$\left|\boldsymbol{a}_i^\top (A\boldsymbol{x} - \boldsymbol{b})\right| < \lambda - \|\boldsymbol{a}_i\| \sqrt{2G(\boldsymbol{x})} \ \Rightarrow\ \boldsymbol{x}_i^\star = 0.$$

□

This result is known in the literature in [17] and we recover it using our prosed general approach in this paper by using Theorem 6.

Also, by applying same trick as mentioned after the end of proof of corollary 18, we can show that we can get rid of the factor 2 here also. Here also it is not hard to see that primal and dual ((A) and (B)) both are 1 strongly convex in the dual variable $\boldsymbol{w}$. Hence by the same argument as made in the proof of Corollary 18, we get that

$$G(\boldsymbol{w}) \geq \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2$$

And the improved screening rule comes out to be

$$\left|\boldsymbol{a}_i^\top (A\boldsymbol{x} - \boldsymbol{b})\right| < \lambda - \|\boldsymbol{a}_i\| \sqrt{G(\boldsymbol{x})} \;\Rightarrow\; \boldsymbol{x}_i^\star = 0.$$

Logistic Regression with L_1-regularization

Corollary 21. ***Logistic Regression with L_1-norm Penalization-*** *The optimization problem for logistic regression with L_1 regularizer can be written in the form of:*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^{n} \log(\exp([A\boldsymbol{x}]_i) + 1) + \lambda\|\boldsymbol{x}\|_1 \tag{83}$$

And screening rule for above problem can be written as :

$$\left|\boldsymbol{a}_i^\top \left(\frac{\exp(A\boldsymbol{x})}{\exp(A\boldsymbol{x}) + 1}\right)\right| < \lambda - \|\boldsymbol{a}_i\|_2 \sqrt{2G(\boldsymbol{x})} \;\Rightarrow\; \boldsymbol{x}_i^\star = 0$$

where $\left(\frac{\exp(A\boldsymbol{x})}{\exp(A\boldsymbol{x})+1}\right)$ *is element wise vector whose i_{th} element is* $\left(\frac{\exp([A\boldsymbol{x}]_i)}{\exp([A\boldsymbol{x}]_i)+1}\right)$

Proof. By observation we know that in equation (83)

$f(A\boldsymbol{x}) = \sum_{i=1}^{n} log(exp([A\boldsymbol{x}]_i) + 1)$ and $\boldsymbol{w}$ is elementwise vector of w_i s.t $\quad w_i = \dfrac{exp([A\boldsymbol{x}]_i)}{exp([A\boldsymbol{x}]_i) + 1}$

According to [24, Lemma 5], we get that the function $f(A\boldsymbol{x})$ is 1-smooth. Hence $L = 1$

Now from theorem 6, we derive the screening rule for logistic regression with L_1-regularization which is

$$\left|\boldsymbol{a}_i^\top \left(\frac{exp(A\boldsymbol{x})}{exp(A\boldsymbol{x}) + 1}\right)\right| < \lambda - \|\boldsymbol{a}_i\|_2 \sqrt{2G(\boldsymbol{x})} \;\Rightarrow\; \boldsymbol{x}_i^\star = 0$$

where $\left(\frac{exp(A\boldsymbol{x})}{exp(A\boldsymbol{x})+1}\right)$ is element wise vector whose i_{th} element is $\left(\frac{exp([A\boldsymbol{x}]_i)}{exp([A\boldsymbol{x}]_i)+1}\right)$ This result is also known in the literature in [17] (or see also [27] for a similar approach) and we recover it using our prosed general approach in this paper by using Theorem 6.

□

Elastic norm regularization regression

Proof of Corollary 7.

$$\begin{aligned}
&\frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda_2\|\boldsymbol{x}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 \\
&= \frac{1}{2}[\boldsymbol{x}^\top A^\top A\boldsymbol{x} - 2\boldsymbol{b}^\top A\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{b}] + \lambda_2\boldsymbol{x}^\top\boldsymbol{x} + \lambda_1\|\boldsymbol{x}\|_1 \\
&= \frac{1}{2}[\boldsymbol{x}^\top(A^\top A + 2\lambda_2 I)\boldsymbol{x} - 2\boldsymbol{b}^\top A\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{b}] + \lambda_1\|\boldsymbol{x}\|_1
\end{aligned} \tag{84}$$

Now consider $A^\top A + 2\lambda_2 I = Q^\top Q$ and choose vector $\boldsymbol{m}$ such that $A^\top\boldsymbol{b} = Q^\top\boldsymbol{m}$. Hence line (84) can be written as

$$\begin{aligned}
&\tfrac{1}{2}\left[\boldsymbol{x}^\top(A^\top A + 2\lambda_2 I)\boldsymbol{x} - 2\boldsymbol{b}^\top A\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{b}\right] + \lambda_1\|\boldsymbol{x}\|_1 \\
&= \tfrac{1}{2}\left[\boldsymbol{x}^\top Q^\top Q\boldsymbol{x} - 2\boldsymbol{m}^\top Q\boldsymbol{x} + \boldsymbol{m}^\top\boldsymbol{m} - \boldsymbol{m}^\top\boldsymbol{m} + \boldsymbol{b}^\top\boldsymbol{b}\right] + \lambda_1\|\boldsymbol{x}\|_1 \\
&= \tfrac{1}{2}\|Q\boldsymbol{x} - \boldsymbol{m}\|_2^2 + \tfrac{1}{2}\left[\boldsymbol{b}^\top\boldsymbol{b} - \boldsymbol{m}^\top\boldsymbol{m}\right] + \lambda_1\|\boldsymbol{x}\|_1
\end{aligned}$$

Now the optimization problem (8) can be written as

$$\min_{\boldsymbol{x}} \; \frac{1}{2}\|Q\boldsymbol{x} - \boldsymbol{m}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 \tag{85}$$

Now results from Corollary 20 can be directly applied to (85).
From observation, we know that $f(Q\boldsymbol{x}) = \frac{1}{2}\|Q\boldsymbol{x} - \boldsymbol{m}\|^2$, $\quad \boldsymbol{w} = Q\boldsymbol{x} - \boldsymbol{m}, \quad$ and $L = 1$
Simplification,

$$\begin{aligned}
\left|\boldsymbol{q}_i^\top(Q\boldsymbol{x} - \boldsymbol{m})\right| &= \left|\boldsymbol{q}_i^\top Q\boldsymbol{x} - \boldsymbol{q}_i^\top \boldsymbol{m}\right| \\
&= \left|\boldsymbol{q}_i^\top Q\boldsymbol{x} - \boldsymbol{a}_i^\top \boldsymbol{b}\right| \\
&= \left|(\boldsymbol{a}_i^\top A + 2\lambda_2 \boldsymbol{e}_i^\top)\boldsymbol{x} - \boldsymbol{a}_i^\top \boldsymbol{b}\right| \qquad (86) \\
|\boldsymbol{q}_i| \sqrt{2G(\boldsymbol{x})} &= \sqrt{\boldsymbol{a}_i^\top \boldsymbol{a}_i + 2\lambda_2}\sqrt{2G(\boldsymbol{x})} \\
&= \sqrt{2(\boldsymbol{a}_i^\top \boldsymbol{a}_i + 2\lambda_2)G(\boldsymbol{x})} \qquad (87)
\end{aligned}$$

Now using results from Corollary 20, equations (86) and (87), we get screening rules for elastic norm regularization regression problem as:

$$\left|(\boldsymbol{a}_i^\top A + 2\lambda_2 \boldsymbol{e}_i^\top)\boldsymbol{x} - \boldsymbol{a}_i^\top \boldsymbol{b}\right| < \lambda_1 - \sqrt{2(\boldsymbol{a}_i^\top \boldsymbol{a}_i + 2\lambda_2)(G(\boldsymbol{x}))} \;\Rightarrow\; x_i^\star = 0.$$

□

Lemma 22. *If the data matrix A is full rank and there exist a square symmetric matrix M such that $A^\top M A = I$. Now if a function $f(A\boldsymbol{x})$ is L-Lipschitz gradient then the function defined as $f(A\boldsymbol{x}) + \lambda_2\|\boldsymbol{x}\|_2^2$ is $(L + 2\lambda_2\sigma_1)$-Lipschitz gradient in $A\boldsymbol{x}$ where σ_1 is the largest eigenvalue of a matrix M.*

Proof. Since $f(A\boldsymbol{x})$ is L-Lipschitz gradient hence $\nabla^2 f(A\boldsymbol{x}) \le LI$. Now,

$$\begin{aligned}
\lambda_2\|\boldsymbol{x}\|_2^2 &= \lambda_2 \boldsymbol{x}^\top \boldsymbol{x} \\
&= \lambda_2 \boldsymbol{x}^\top A^\top M A\boldsymbol{x} \quad \text{such that } A^\top M A = I \\
&= \lambda_2(\boldsymbol{x}^\top A^\top) M (A\boldsymbol{x}) \qquad (88)
\end{aligned}$$

Suppose $\tilde{f}(A\boldsymbol{x}) = f(A\boldsymbol{x}) + \lambda_2(A\boldsymbol{x})^\top M(A\boldsymbol{x})$
Hence from equation (88) and our assumption on f, we conclude that

$$\nabla^2 \tilde{f}(A\boldsymbol{x}) \le (L + 2\lambda_2\sigma_1)I$$

where σ_1 is the largest eigenvalue of the matrix M. □

Theorem 23. *If we consider the general elastic net formulation of the form*

$$\min_{\boldsymbol{x}} \; f(A\boldsymbol{x}) + \lambda_2\|\boldsymbol{x}\|_2^2 + \lambda_1\|\boldsymbol{x}\|_1 \tag{89}$$

If we define the matrix M similarly as in Lemma 22 and make similar assumptions as made in Lemma 22, then a screening rule for this problem class is given by

$$\left|(\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x}) + 2\lambda_2 x_i)\right| < \lambda_1 - |\boldsymbol{a}_i| \sqrt{2(L + 2\lambda_2\sigma_1)G(\boldsymbol{x})} \;\Rightarrow\; \boldsymbol{x}_i^\star = 0$$

Proof. If we assume $\tilde{f}(A\boldsymbol{x}) = f(A\boldsymbol{x}) + \lambda_2(A\boldsymbol{x})^\top M(A\boldsymbol{x})$, then optimization problem (89) can be written as

$$\min_{\boldsymbol{x}} \; \tilde{f}(A\boldsymbol{x}) + \lambda_1\|\boldsymbol{x}\|_1$$

From Lemma 22, $\tilde{f}(A\boldsymbol{x})$ is smooth with parameter $(L + 2\lambda_2\sigma_1)$. Using findings from Theorem 6, we get

$$\left|\boldsymbol{a}_i^\top \nabla \tilde{f}(A\boldsymbol{x})\right| < \lambda_1 - |\boldsymbol{a}_i| \sqrt{2(L + 2\lambda_2\sigma_1)G(\boldsymbol{x})} \;\Rightarrow\; \boldsymbol{x}_i^\star = 0$$

Now,

$$\begin{aligned}
\boldsymbol{a}_i^\top \nabla \tilde{f}(A\boldsymbol{x}) &= (A\boldsymbol{e}_i)^\top \nabla \tilde{f}(A\boldsymbol{x}) \\
&= (\boldsymbol{e}_i)^\top A^\top \nabla \tilde{f}(A\boldsymbol{x}) \\
&= (\boldsymbol{e}_i)^\top (A^\top \nabla f(A\boldsymbol{x}) + 2\lambda_2 \boldsymbol{x}) \\
&= (\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x}) + 2\lambda_2 x_i)
\end{aligned}$$

Hence the screening rules becomes

$$\left|(\boldsymbol{a}_i^\top \nabla f(A\boldsymbol{x}) + 2\lambda_2 x_i)\right| < \lambda_1 - |\boldsymbol{a}_i| \sqrt{2(L + 2\lambda_2\sigma_1)G(\boldsymbol{x})} \Rightarrow \boldsymbol{x}_i^\star = 0$$

□

D.2 Screening for Structured Norms

We use the same notation as mentioned in Section 4.3 i.e., we use the notation $\{\boldsymbol{x}_1 \cdots \boldsymbol{x}_G\}$ to express a vector $\boldsymbol{x}$ as a partition of non-overlapping groups $g \in \mathcal{G}$ of variables, such that $\boldsymbol{x}^\top = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top \cdots \boldsymbol{x}_G^\top\right]$. Correspondingly, the matrix A can be denoted as the concatenation of the respective column groups $A = [A_1\ A_2 \cdots A_G]$, and $\sum_{g \in \mathcal{G}} |g| = n$.

Lemma 24. *Now if we consider an optimization problem of the form*

$$\arg\min_{\boldsymbol{x}} f(A\boldsymbol{x}) + \sum_{g=1}^{G} \sqrt{\rho_g} \|\boldsymbol{x}_g\|_2$$

At the optimal point $\boldsymbol{x}^\star$ and dual optimal points $\boldsymbol{w}^\star$, we get rules according to the following equation:

$$\|A_g^\top \boldsymbol{w}^\star\|_2 < \sqrt{\rho_g} \Rightarrow \boldsymbol{x}_g^\star = 0$$

Proof. The dual of the problem is given by

$$\mathcal{O}_B(\boldsymbol{w}) = f^*(\boldsymbol{w}) + \sum_g \sqrt{\rho_g} \boldsymbol{\iota}_{L_\infty}\left(\frac{\left\|A_g^\top \boldsymbol{w}\right\|_2}{\sqrt{\rho_g}}\right) \tag{90}$$

Hence for the indicator function g_g^* by Lemma 15

$$\begin{aligned}
\partial g_g^*(-A_g^\top \boldsymbol{w}^\star) &= \left\{ \boldsymbol{s} |\ \forall \boldsymbol{z}\ s.t\ \left\| \frac{A_g^\top \boldsymbol{z}}{\sqrt{\rho_g}} \right\|_2 \le 1\ ;\ \boldsymbol{s}^\top(-A_g^\top \boldsymbol{z} + A_g^\top \boldsymbol{w}^\star) \le 0 \right\} \\
&= \left\{ \boldsymbol{s} |\ \forall \boldsymbol{z}\ s.t\ \left\|A_g^\top \boldsymbol{z}\right\|_2 \le \sqrt{\rho_g};\ \boldsymbol{s}^\top(A_g^\top \boldsymbol{z}) \ge \boldsymbol{s}^\top(A_g^\top \boldsymbol{w}^\star) \right\}
\end{aligned}$$

Now, by the optimality condition (12b) $\boldsymbol{x}_g \in \partial g_g^*(-A_g^\top \boldsymbol{w}^\star)$, and since this holds, hence $xv_g^\star$ should satisfy the required constrained which is needed to be in the set of subgradients of $\partial g^*(-A_g^\top \boldsymbol{w}^\star)$ according to conditions given above. Hence,

$$\begin{aligned}
& - {\boldsymbol{x}_g^\star}^\top(A_g^\top \boldsymbol{z}) \le -{\boldsymbol{x}_g^\star}^\top(A_g^\top \boldsymbol{w}^\star) \quad \forall \boldsymbol{z}\ s.t\ \left\|A_g^\top \boldsymbol{z}\right\|_2 \le \sqrt{\rho_g} \\
& \Rightarrow {\boldsymbol{x}_g^\star}^\top(A_g^\top \boldsymbol{z}) \ge {\boldsymbol{x}_g^\star}^\top(A_g^\top \boldsymbol{w}^\star) \quad \forall \boldsymbol{z}\ s.t\ \left\|A_g^\top \boldsymbol{z}\right\|_2 \le \sqrt{\rho_g} \\
& \Rightarrow {\boldsymbol{x}_g^\star}^\top(A_g^\top \boldsymbol{w}^\star) \le \min_z {\boldsymbol{x}_g^\star}^\top(A_g^\top \boldsymbol{z}) \quad s.t\ \left\|A_g^\top \boldsymbol{z}\right\|_2 \le \sqrt{\rho_g} \\
& \Rightarrow {\boldsymbol{x}_g^\star}^\top(A_g^\top \boldsymbol{w}^\star) \le \min_z \|\boldsymbol{x}_g\|_2 \|A_g^\top \boldsymbol{z}\|_2 \quad s.t\ \left\|A_g^\top \boldsymbol{z}\right\|_2 \le \sqrt{\rho_g} \\
& \Rightarrow \boldsymbol{x}_g^\top(A_g^\top \boldsymbol{w}^\star) \le -\|\boldsymbol{x}_g^\star\|_2 \sqrt{\rho_g} \\
& \Rightarrow \|A_g^\top \boldsymbol{w}^\star\|_2 = \sqrt{\rho_g}
\end{aligned} \tag{91}$$

Equation (91) comes from the cauchy inequality and true $\forall \boldsymbol{x}_g^\star\ :\ \boldsymbol{x}_g^\star \neq 0$. Whenever $\|A_g^\top \boldsymbol{w}^\star\|_2 < \sqrt{\rho_g}$ then $\boldsymbol{x}_g^\star = 0$

Another view on the screening of above optimization problem can be seen from the optimality condition (12a). The optimization problem in Lemma 24 can be taken as partially separable problem and from the optimality condition (12a)

$$- A_g^\top \boldsymbol{w}^\star \in \partial g(\boldsymbol{x}_g^\star) \tag{92}$$

$$\partial g(\boldsymbol{x}_g^\star) \in \begin{cases} \sqrt{\rho_g} \frac{\boldsymbol{x}_g}{\|x_g\|_2} & \text{if } \boldsymbol{x}_g \neq 0 \\ \mathcal{B}_2 & \text{if } \boldsymbol{x}_g = 0 \textit{ and } \mathcal{B}_2 \text{ is norm ball of radius } \sqrt{\rho_g} \end{cases} \tag{93}$$

From Equations (92) and (93), we conclude that if

$$\|A_g^\top \boldsymbol{w}^\star\|_2 < \sqrt{\rho_g} \Rightarrow \boldsymbol{x}_g^\star = 0$$

□

Theorem' 8. *For ℓ_2/ℓ_1-regularized optimization problem of the form*

$$\min_{\boldsymbol{x}} f(A\boldsymbol{x}) + \sum_{g=1}^{G} \sqrt{\rho_g}\|\boldsymbol{x}_g\|_2$$

Assuming f is L-smooth, then the following (group-level) screening rule holds for all groups g:

$$\|A_g^\top \nabla f(A\boldsymbol{x})\|_2 + \sqrt{2L\,G(\boldsymbol{x})}\,\|A_g\|_{\mathsf{Fro}} < \sqrt{\rho_g} \;\Rightarrow\; \boldsymbol{x}_g^\star = \mathbf{0} \in \mathbb{R}^{|g|}\ .$$

Proof. From Equation (1a), we know that $\boldsymbol{w} \in \nabla f(A\boldsymbol{x})$. Now

$$\begin{aligned}
\|A_g^\top \boldsymbol{w}^\star\|_2 &= \|A_g^\top(\boldsymbol{w} + \boldsymbol{w}^\star - \boldsymbol{w})\|_2 \le \|A_g^\top \boldsymbol{w}\|_2 + \|A_g^\top(\boldsymbol{w}^\star - \boldsymbol{w})\|_2 \\
&= \|A_g^\top \boldsymbol{w}\|_2 + \sqrt{tr((A_g^\top(\boldsymbol{w}^\star - \boldsymbol{w}))((\boldsymbol{w}^\star - \boldsymbol{w})^\top)A_g)^\top} \\
&\le \|A_g^\top \boldsymbol{w}\|_2 + \sqrt{tr((\boldsymbol{w}^\star - \boldsymbol{w})^\top(\boldsymbol{w}^\star - \boldsymbol{w}))}\sqrt{tr(A_g^\top A_g)} \\
&= \|A_g^\top \boldsymbol{w}\|_2 + \|\boldsymbol{w}^\star - \boldsymbol{w}\|_2\,\|A_g\|_{\mathsf{Fro}}
\end{aligned} \tag{94}$$

Using Corollary 12 with Equation (94), we get

$$\|A_g^\top \boldsymbol{w}^\star\|_2 \le \|A_g^\top \nabla f(A\boldsymbol{x})\|_2 + \sqrt{2L\,G(\boldsymbol{x})}\,\|A_g\|_{\mathsf{Fro}}$$

Hence using the previous Lemma 24,

$$\|A_g^\top \nabla f(A\boldsymbol{x})\|_2 + \sqrt{2L\,G(\boldsymbol{x})}\,\|A_g\|_{\mathsf{Fro}} < \sqrt{\rho_g} \;\Rightarrow\; \boldsymbol{x}_g^\star = \mathbf{0} \in \mathbb{R}^{|g|}$$

□

Corollary 25. Group Lasso Regression with Squared Loss - *For the group lasso formulation*

$$\min_{\boldsymbol{x}}\ \frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda\sum_{g=1}^{G}\sqrt{\rho_g}\|\boldsymbol{x}_g\|_2$$

we have the following screening rule for all groups g:

$$\|A_g^\top(A\boldsymbol{x} - \boldsymbol{b})\|_2 + \sqrt{2G(\boldsymbol{x})}\,\|A_g\|_{\mathit{Fro}} < \lambda\sqrt{\rho_g} \;\Rightarrow\; \boldsymbol{x}_g^\star = 0\ .$$

Proof. This is an explicit case of the optimization problem mentioned in Lemma 24, see also [18]. By observation we know that,

$$f(A\boldsymbol{x}) = \tfrac{1}{2}\|A\boldsymbol{x} - b\|^2, \quad \boldsymbol{w} = A\boldsymbol{x} - b \quad and \quad L = 1$$

Now applying the findings of Theorem 8, we get

$$\|A_g^\top(A\boldsymbol{x} - \boldsymbol{b})\|_2 + \sqrt{2G(\boldsymbol{x})}\,\|A_g\|_{\mathsf{Fro}} < \lambda\sqrt{\rho_g} \;\Rightarrow\; \boldsymbol{x}_g^\star = 0$$

□

In Lemma 26 mentioned below, we show that the structured norm setting of [17] can be derived from our more general (A) and (B) structure.

Lemma 26. *Sparse Multi-Task and Multi Class Model [17] - If we consider general problem of the form*

$$\min_{X\in\mathbb{R}^{p\times q}} \sum_{i=1}^{n} f_i(\boldsymbol{a}_i^\top X) + \lambda\Omega(X) \tag{95}$$

where the regularization function $\Omega : \mathbb{R}^{p\times q} \to \mathbb{R}_+$ is such that $\Omega(X) = \sum_{g=1}^{p}\|\boldsymbol{x}_g\|_2$ and $X = [\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_G]$. We write $W = [\boldsymbol{w}_1, \boldsymbol{w}_2 \cdots \boldsymbol{w}_G]$ for variable of the dual problem. Then the screening rule becomes

$$\|{\boldsymbol{a}^{(g)}}^\top W\|_2 < \lambda - \|\boldsymbol{a}^{(g)}\|_2\,\|W - W^\star\|_2 \;\Rightarrow\; \boldsymbol{x}_g^\star = 0$$

Here $\boldsymbol{a}^{(g)}$ is the vector of the g^{th} element group of each vector $\boldsymbol{a}_i$.

Proof. Equations pair (A) and (B) can be used interchangeably by replacing primal with dual and f with g. Hence the partial separable primal-dual pair (SA) and (SB) can also be used interchangeably. By comparing Equation (95) with (SA) and (SB), we observe that separable function $\sum_{i=1}^{n} f_i(\boldsymbol{a}_i^\top X)$

takes the place of separable g^* in (SB) and $\lambda\Omega(X)$ takes the place of f^*. Hence we apply the optimality condition (1b) to get (with exchanged primal dual variable)

$$AW^\star \in \partial\lambda\Omega(X^\star)$$

Hence if,

$$\|\boldsymbol{a}^{(g)^\top} W^\star\|_2 < \lambda \;\Rightarrow\; \boldsymbol{x}_g = 0 \tag{96}$$

Now,

$$\begin{aligned}
\|\boldsymbol{a}^{(g)^\top} W^\star\|_2 &= \|\boldsymbol{a}^{(g)^\top}(W^\star - W + W)\|_2 \\
&\le \|\boldsymbol{a}^{(g)^\top} W\|_2 + \|\boldsymbol{a}^{(g)^\top}(W^\star - W)\|_2 \\
&\le \|\boldsymbol{a}^{(g)^\top} W\|_2 + \|\boldsymbol{a}^{(g)}\|_2 \|(W^\star - W)\|_2
\end{aligned} \tag{97}$$

Using equations (96) and (97), the screening rule comes out to be

$$\|\boldsymbol{a}^{(g)^\top} W\|_2 < \lambda - \|\boldsymbol{a}^{(g)}\|_2 \, \|W - W^\star\|_2 \;\Rightarrow\; \boldsymbol{x}_g^\star = 0$$

□

Corollary 27. *If for all $i \in [n]$, f_i is L-Lipschitz gradient then screening rule for equation* (95) *is*

$$\|\boldsymbol{a}^{(g)^\top} W\|_2 < \lambda - \|\boldsymbol{a}^{(g)}\|_2 \sqrt{2L\,G(X)} \;\Rightarrow\; \boldsymbol{x}_g^\star = 0$$

Proof. Using Lemma 26 and Corollary 12, we get the desired expression. □

D.3 Connection with Sphere Test Method

The general idea behind the sphere test method [28] is to consider the maximum value of desired function in a spherical region which contains the optimal dual variable. In context of our general framework (A) and (B), we obtain this case when considering an ℓ_1 penalty or ℓ_2/ℓ_1 penalty. That means g is a norm and hence from Lemma 9, g^* becomes the indicator function of the dual norm ball of $A^\top \boldsymbol{w}$. The dual norm function for ℓ_1 norm is of the form $\max_i |\boldsymbol{a}_i^\top \boldsymbol{w}|$ and for ℓ_2/ℓ_1 norm, it is $\max_g \|A_g^\top \boldsymbol{w}\|$. Hence, we try to find maximum value of the function of the forms $\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{q},r)} \boldsymbol{a}_i^\top \boldsymbol{\theta}$ where $\mathcal{S}(\boldsymbol{q}, r) = \{z : \|\boldsymbol{z} - \boldsymbol{q}\|_2 \le r\}$ the ball $\mathcal{S}$ also contains the optimal dual point $\boldsymbol{w}^\star$. If the maximum value of $\boldsymbol{a}_i^\top \boldsymbol{\theta}$ is less than some particular value for all the $\boldsymbol{\theta}$ in the ball hence $\boldsymbol{a}_i^\top \boldsymbol{w}$ will also be less than that particular value and that is the main reason we try to find maximum of $\boldsymbol{a}_i^\top \boldsymbol{\theta}$ over the ball $\mathcal{S}$.

$$\begin{aligned}
\max_{\boldsymbol{\theta}\in\mathcal{S}(\boldsymbol{q},r)} \boldsymbol{a}_i^\top \boldsymbol{\theta} &= \boldsymbol{a}_i^\top(\boldsymbol{\theta} - \boldsymbol{q} + \boldsymbol{q}) = \boldsymbol{a}_i^\top(\boldsymbol{\theta} - \boldsymbol{q}) + \boldsymbol{a}_i^\top \boldsymbol{q} \\
&\le \|\boldsymbol{a}_i\|_2 \|\boldsymbol{\theta} - \boldsymbol{q}\| + \boldsymbol{a}_i^\top \boldsymbol{q} \le r\|\boldsymbol{a}_i\|_2 + \boldsymbol{a}_i^\top \boldsymbol{q}
\end{aligned}$$

Similar arguments can be given in the ℓ_2/ℓ_1-norm case. A variety of existing screening test for lasso and group lasso are of this flavor of sphere tests. The difference between these approaches mainly lie in the way of choosing the center and bounding the radius of the sphere, such that the optimal dual variables lie inside the sphere. Our method can be seen as a general framework for such a sphere test based screening with dynamic screening rules. Our method can be interpreted as a sphere test with the current iterate of the dual variable $\boldsymbol{w}$ as a center of the ball, and we obtain the bound on the radius in terms of duality gap function.

E Illustrative Experiments

While the contribution of our paper is on the theoretical generality and the collection of new screening applications, we will still briefly illustrate the performance of some of the proposed screening algorithms, for the classical examples of simplex constrained and L_1-constrained problems. We compare the fraction of active variables and the Wolfe-Gap function as optimization algorithm progress.

We consider the optimization problem of the form $\min_{\boldsymbol{x} \in \mathcal{B}_{L_1}} \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2$. $\mathcal{B}_{L_1}$ is a scaled L_1-ball with radius 35. $A \in \mathbb{R}^{3000 \times 600}$ is a random Gaussian matrix and a noisy measurement $\boldsymbol{b} = A\boldsymbol{x}^\star$ where $\boldsymbol{x}^\star$ is a sparse vector of $+1$ and -1 with only 70 non zeros entries. We solve the above optimization problem using the Frank-Wolfe algorithm (pair-wise variant, see [12]). Before putting this optimization problem into the solver we convert this problem into the barycentric representation which is $\min_{\boldsymbol{x}_\triangle \in \triangle} \|A_\triangle \boldsymbol{x}_\triangle - \boldsymbol{b}\|_2^2$. The relation between the transformed variable and original variable can be given by $A_\triangle = [A| - A]$ and $\boldsymbol{x} = [I_n| - I_n]\boldsymbol{x}_\triangle$. For more details see [7]. Now we apply our Theorems 4 and 2 on variable of $\boldsymbol{x}$ and $\boldsymbol{x}_\triangle$ respectively to screen, in order to compare the two alternative screening approaches on the same problem. Note that the Wolfe gap is identical in both parameterizations, for any $\boldsymbol{x}$. One important point to note here is that dimension of $\boldsymbol{x}_\triangle$ is the double of the dimension of $\boldsymbol{x}$, and any L_1-coordinate value x_i is zero if and only if both "duplicate" variables $x_{\triangle,i}$ and $x_{\triangle,n+i}$ are zero, where n is the dimensionality of $\boldsymbol{x}$.

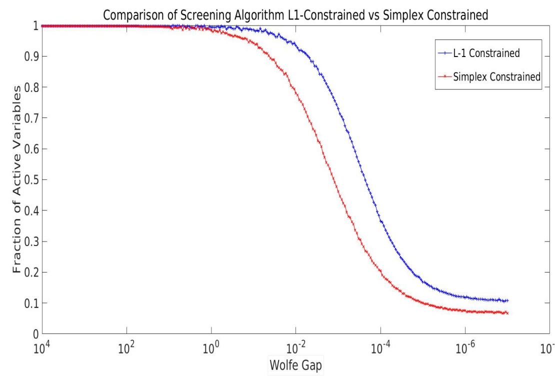

Figure 1: Simplex vs L-1 Constrained Screening

Therefore, the simplex variant (with more variables) performs a more fine-grained variant of screening, where we can screen each of the sign patterns separately for each variable. In Fig 1, the blue curve illustrates the screening efficiency for the L_1-constrained screening case, while the red curve illustrate simplex constrained screening. Our theorems 4 and 2 are well in line with the phenomena in Fig 1. For the L_1-constrained case, the screening starts relatively at later stage than simplex case due to the fact that in Equation (7), two out of three terms are absolute values of some quantity and hence it is very tough to compensate both of them by the third quantity to make the whole some less than 0. Hence in the beginning this rule can often be ineffective. As algorithm progresses, the duality gap becomes smaller and screening starts but at the same time the gradient (and therefore gap) also starts to decay which brings the trade-off shown in the plot. For both variants, screening becomes slow towards the end.

We also report the time taken to reach a duality gap of 10^{-7} with both the approaches mentioned above (simplex constrained and L_1-constrained) on for different datasets. The first two datasets (*Synth1* and *Synth2*) are generated under the same setting described earlier but *Synth1* with 5000 samples and *Synth2* with 10000 samples. *RCV1* is a real world dataset having $20,242$ samples and $47,236$ data dimensions. *news20Binary* is also a real world dataset having $19,996$ entries and $1,355,191$ dimensions. Below in Tables 1 and 2, we describe the running time of the optimization methods to reach a duality gap threshold of 10^{-7} with or without screening. On *RCV1* dataset we try the feature learning with L_1-norm ball constraint of 200 and on *news20Binary* we use L_1-norm ball constraint of 35. In the case of *RCV1* and *news20Binary*, A is the data matrix and $\boldsymbol{b}$ is the label of each instance in the dataset. From Tables 1 and 2 it is also evident that simplex screening rule is more tighter than the L_1-constrained screening rule.

Dataset/ No. of Samples	**No Screening (Simplex)**	**Screening (Simplex)**
Synth1 5000	13.1 sec	11.7sec
Synth2 10000	28.3 sec	23.1 sec
RCV1 20242	18.6 min	13.5 min
news20B 19996	33.4 min	25.2 min

Table 1: Simplex constrained screening time taken

Dataset/ No. of Samples	**No Screening (ℓ_1-constr.)**	**Screening (ℓ_1-constr.)**
Synth1 5000	13.1	12.2 sec
Synth2 10000	28.3 sec	24.7 sec
RCV1 20242	18.6 min	14.9 min
news20B 19996	33.4 min	27.1 min

Table 2: L_1-constrained screening time taken