ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Summer 2015

Jakob Olbrich

Screening Rules for Convex Problems

Submission Date: September 11th 2015

Advisers: Dr. Martin Jaggi, Prof. Dr. Bernd Gärtner, Prof. Dr. Peter Bühlmann

Abstract

This thesis gives a general approach to deriving screening rules for convex optimization problems. It splits up in three steps. As the first step, the Karush-Kuhn-Tucker conditions are used to derive necessary conditions that allow to reduce the problem size. They depend on the optimal solution itself. The second step is to gather information on the optimal solution from a known approximation. In the third and final step the information is used to get conditions that do not depend on the optimal solution, which are then called screening rules. This thesis studies in particular the unit simplex, the unit box and polytopes as domain. The resulting screening rules can be applied to various problems, such as Support Vector Machines (SVM), the Minimum Enclosing Ball (MEB), LASSO problems and logistic regression. The resulting screening rules are compared to existing rules for those problems.

Contents

Notation

- Bold letters are used for vectors.
- Inequalities between vectors such as $\boldsymbol{a} < \boldsymbol{b}$ are a shorter notation for $\forall i : a_i < b_i$.
- For any function $f : \mathbb{R}^n \to \mathbb{R}$, its derivative is denoted by ∇f.
- If not stated otherwise $\|\cdot\|$ is used for the euclidean norm.
- The i-th unit vector is written as $\boldsymbol{e}_i$.

Chapter 1

Introduction

In the introduction the main ideas behind screening are explained, an overview of existing screening rules is given and the basic definitions and properties are stated for later use. The centre part concentrates on a generalized approach to finding screening rules. Rules are given for three types of domains, the unit simplex, box constraints and polytope constraints. In the final part, implications of those rules on various problems are described and compared to existing work on screening rules for those problems.

1.1 Goal of Screening

Assume a function $f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$ is given, where $\boldsymbol{x}$ is a variable vector with m components, $\boldsymbol{A}$ is a data matrix with columns $\boldsymbol{a}_1$ to $\boldsymbol{a}_m$ and $\boldsymbol{b}$ is a given vector. The goal is to minimize that function over some domain $\mathcal{D}(\boldsymbol{B})$ depending on the data matrix $\boldsymbol{B}$. The size of the problem and therefore also the computation time usually depends on the size of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The goal of screening is to eventually decrease computation cost by reducing the problem size. Screening rules can be even more relevant if the solutions are computed in an iterative manner. In that case it is tried to reduce the problem size in each step using information from the current approximation of the solution.

1.2 Existing Screening Rules

So far screening rules have been studied for several specific problems. A description of the problems mentioned here can be found in Chapter 3. The minimum enclosing ball was considered by Ahipaşaoğlu and Yildirim (2008) and Källberg and Larsson (2014). Regarding screening rules the most examined problem is the penalized version of the LASSO. The first paper on that problem was written by Ghaoui, Viallon, and Rabbani (2010), followed by extensions and alterations from Wang, Wonka, and Ye (2012), Bonnefoy, Emiya, Ralaivola, and Gribonval (2014) and Olivier Fercoq and Salmon (2015). Ndiaye, Fercoq, Gramfort, and Salmon (2015) were able to generalize their previous findings to a wider range of functions. As a third problem, support vector machines were studied by Ogawa, Suzuki, Suzumura, and Takeuchi (2014), Zhao, Liu, and Cox (2014) and Wysling (2015). For more details on the contribution of these papers read Section 4.1.

1.3 Basic Definitions of Convex Optimization

To be able to study convex problems, convexity has to be defined for sets and functions and some conditions for convexity will be needed. Also a convex optimization problem should be defined.

Definition 1.1. *A* set $X \subseteq \mathbb{R}^n$ *is called* convex *iff:*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in X \quad \forall 0 \leq s \leq 1 : \quad s\boldsymbol{x} + (1-s)\boldsymbol{y} \in X \tag{1.1}$$

Definition 1.2. *A* function $f : \mathbb{R}^n \to \mathbb{R}$ *is* convex *iff its domain* $\mathcal{D}$ *is a convex set and:*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \quad \forall 0 \leq s \leq 1 : \quad f(s\boldsymbol{x} + (1-s)\boldsymbol{y}) \leq sf(\boldsymbol{x}) + (1-s)f(\boldsymbol{y}) \tag{1.2}$$

Definition 1.3. *A differentiable function is called* strongly convex *with parameter* μ *if:*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{1.3}$$

Or equivalently:

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y}) \geq \mu\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{1.4}$$

Definition 1.4. *A function* $f : \mathbb{R}^n \to \mathbb{R}$ *is called* Lipschitz continuous *with parameter* L *if:*

$$L\|\boldsymbol{x} - \boldsymbol{y}\| \geq \|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \tag{1.5}$$

Proposition 1.5. *Let* f *be a differentiable function. Then* f *is convex iff its domain* $\mathcal{D}$ *is convex and:*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D} \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})^T\nabla f(\boldsymbol{x}) \tag{1.6}$$

(Boyd and Vandenberghe, 2004, Chapter 3.1.3)

Observation 1.6. *Proposition 1.5 is equivalent to the definition of strong convexity with parameter* $\mu = 0$.

Definition 1.7. *An* optimization problem *with constraints* in standard form *is given by:*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f_0(\boldsymbol{x}) \\ s.t. \quad & f_i(\boldsymbol{x}) \leq 0 \quad \forall i \in \{1 \dots m\} \\ & h_i(\boldsymbol{x}) = 0 \quad \forall i \in \{1 \dots q\} \end{aligned} \tag{1.7}$$

f_0 *is called* objective function *and* $f_0(\boldsymbol{x})$ objective value *for a given* $\boldsymbol{x}$. f_i *and* h_i *are the* constraint functions.

It is called a convex optimization problem if f_0 *and* f_i *are convex functions for all* i *and the functions* h_i *are linear for all* i.

Definition 1.8. *The* feasible set X *of an optimization problem in standard form is the set:*

$$X = \{\boldsymbol{x} | f_i(\boldsymbol{x}) \leq 0 \quad \forall i \in \{1 \dots m\}, h_j(\boldsymbol{x}) = 0 \quad \forall j \in \{1 \dots q\}\} \tag{1.8}$$

1.4 Basic Definitions of Lagrange Duality

The Karush Kuhn Tucker (KKT) conditions will be the tool that allows to reduce the problem size. Also weak duality plays a role in gathering information on the optimal solutions and the derivative of the objective function at optimal points. Here a Lagrange duality perspective is used to develop a notion for the duality gap and the KKT conditions.

Definition 1.9. *The associated* Lagrangian *for the objective function of an optimization problem in standard form is:*

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{q} \nu_i h_i(\boldsymbol{x}) \tag{1.9}$$

Definition 1.10. *The* Lagrange dual function *is defined as:*

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x}} \quad \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \tag{1.10}$$

Definition 1.11. *The* Lagrangian dual optimization problem *is given by:*

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ s.t. \quad & \boldsymbol{\lambda} \geq 0 \end{aligned} \tag{1.11}$$

Definition 1.12. *For feasible points $\boldsymbol{x}$ and $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ of an optimization problem of the form (1.7) and its corresponding dual problem (1.11), the difference of the objective values is called* duality gap.

$$gap_D(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\boldsymbol{x}) - g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \tag{1.12}$$

Proposition 1.13.

$$gap_D(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \geq 0 \tag{1.13}$$

(Boyd and Vandenberghe, 2004, Chapter 5.5.1)

Remark. Proposition 1.13 says that $f_0(\boldsymbol{x}) - g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is always bigger than zero. Hence the primal objective is always bigger than the dual objective, therefore the same holds for the optimal values. This property is usually referenced as *weak duality*. If primal and dual optimal value coincide one says that *strong duality* holds.

Definition 1.14. *Given an optimization problem with objective function $f_0(\boldsymbol{x})$ and an optimal solution $\boldsymbol{x}^*$. A value s is called suboptimality certificate, if the following holds:*

$$f_0(\boldsymbol{x}) - f_0(\boldsymbol{x}^*) \leq s \tag{1.14}$$

Remark. Since $f_0(\boldsymbol{x}) \geq f_0(\boldsymbol{x}^*) \geq g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ for all feasible points $\boldsymbol{x}$ and $(\boldsymbol{\lambda}, \boldsymbol{\nu})$, the duality gap is a *suboptimality certificate*. If strong duality holds it measures how good a pair of feasible points approximates the optimal solution. Otherwise it still gives a lower bound on the quality of the approximation.

1.5 Properties of Optimal Points in Convex Optimization

In this section the main tools necessary for the first two steps of finding screening rules are given. Propositions 1.15 and 1.18 are used to find restrictions on the optimal solutions. The KKT conditions will give necessary conditions that allow to reduce the problem size.

Proposition 1.15. *Consider a convex optimization problem with differentiable f_0 and feasible set X. Then a point $\boldsymbol{x}^* \in X$ is optimal iff:*

$$\forall \boldsymbol{y} \in X: \quad (\boldsymbol{y} - \boldsymbol{x}^*)^T \nabla f_0(\boldsymbol{x}^*) \geq 0 \tag{1.15}$$

(Boyd and Vandenberghe, 2004, Chapter 4.2.3)

Proposition 1.16. *Consider a convex optimization problem with compact feasible set X. Assume that the objective value $f_0(\boldsymbol{x})$ for a given point $\boldsymbol{x}$ is known. A lower bound for the optimal value is then given by:*

$$f_0(\boldsymbol{x}^*) \geq f_0(\boldsymbol{x}) + \min_{\boldsymbol{y} \in X} (\boldsymbol{y} - \boldsymbol{x})^T \nabla f_0(\boldsymbol{x}) \tag{1.16}$$

Definition 1.17. *For a convex optimization problem with compact feasible set the* gap function *is given by:*

$$gap_{FW}(\boldsymbol{x}) = \max_{\boldsymbol{y} \in X} (\boldsymbol{x} - \boldsymbol{y})^T \nabla f_0(\boldsymbol{x}) \tag{1.17}$$

Hearn (1982)

Remark. Since $f_0(\boldsymbol{x}^*) \geq f_0(\boldsymbol{x}) - gap_{FW}(\boldsymbol{x})$ the gap function is a suboptimality certificate.

Proposition 1.18. *Consider a convex optimization problem with gap function gap_{FW}. Then the following holds:*

$$gap_{FW}(\boldsymbol{x}) \geq (\boldsymbol{x} - \boldsymbol{x}^*) \nabla f_0(\boldsymbol{x}) \tag{1.18}$$

Proof.

$$\begin{aligned} gap_{FW}(\boldsymbol{x}) &= \max_{\boldsymbol{y} \in X} (\boldsymbol{x} - \boldsymbol{y})^T \nabla f_0(\boldsymbol{x}) && (1.19) \\ &\geq (\boldsymbol{x} - \boldsymbol{x}^*)^T \nabla f_0(\boldsymbol{x}) && (1.20) \end{aligned}$$

□

Definition 1.19. Slater's condition *on the constraint functions $f_0, \ldots, f_m, h_1, \ldots, h_q$ of an optimization problem as in Definition 1.7 is that:*

- *The functions $h_1, \ldots, h_q$ are linear.*
- *There exists a point $\boldsymbol{x}$ such that:*

$$\forall i \in \{1 \ldots m\}: \quad f_i \text{ is linear} \quad \vee \quad f_i(\boldsymbol{x}) < 0 \tag{1.21}$$

If there is a non-linear constraint function, such a point is called stricty feasible.

(Boyd and Vandenberghe, 2004, Chapter 5.2.3)

Remark. For convex problems Slater's condition implies strong duality. (Boyd and Vandenberghe, 2004, Chapter 5.3.2)

Theorem 1.20 (Karush-Kuhn-Tucker (KKT) conditions)**.** *Consider a convex optimization problem of the form (1.7). Let $f_0, \dots, f_m, h_1, \dots, h_q$ be differentiable and $\boldsymbol{x}^*, (\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ be primal, respectively dual feasible points. Assume that the constraint functions fulfil Slater's condition. Then $\boldsymbol{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are optimal iff they fulfil the KKT conditions:*

$$f_i(\boldsymbol{x}^*) \leq 0 \quad \forall i \in \{1 \dots m\} \tag{1.22}$$

$$h_i(\boldsymbol{x}^*) = 0 \quad \forall i \in \{1 \dots q\} \tag{1.23}$$

$$\lambda_i^* \geq 0 \quad \forall i \in \{1 \dots m\} \tag{1.24}$$

$$\lambda_i^* f_i(\boldsymbol{x}^*) = 0 \quad \forall i \in \{1 \dots m\} \tag{1.25}$$

$$\nabla f_0(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla f_i(\boldsymbol{x}^*) + \sum_{i=1}^{q} \nu_i^* \nabla h_i(\boldsymbol{x}^*) = 0 \tag{1.26}$$

(Boyd and Vandenberghe, 2004, Chapter 5.5.3)
The conditions that allow us to reduce the problem size emerge from equation (1.25), which is known as *complementary slackness* on its own. At this point a short description is given how this works for two types of constraints f_i. Equation (1.25) implies that $\lambda_i^* > 0 \Rightarrow f_i(\boldsymbol{x}^*) = 0$. In cases where the function f_i only depends on x_i and is linear, i.e. $f_i(\boldsymbol{x}) = c_i x_i + d_i$, this gives $\lambda_i^* > 0 \Rightarrow x_i = -d_i/c_i$. Assuming that the objective of the considered problem can be written as $f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, it can be rewritten as $f(\boldsymbol{A}'\boldsymbol{x}' + \boldsymbol{b}')$ with a smaller matrix $\boldsymbol{A}'$. Where $\boldsymbol{b}' = \boldsymbol{b} - \frac{d_i}{c_i}\boldsymbol{a}_i$ and $\boldsymbol{A}'$ is obtained by deleting the i-th column of $\boldsymbol{A}$.

Definition 1.21. *Consider a convex optimization problem of the form (1.7), whose constraints $f_i(\boldsymbol{x})$ are linear functions only depending on x_i, with an objective f_0 that can be written as $f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})+\boldsymbol{c}^T\boldsymbol{x}$ and Lagrange dual problem as in Definition 1.11. Then a data point $\boldsymbol{a}_i$ s called* non-influential *if $\lambda_i^* > 0$ for all dual optimal solutions $\boldsymbol{\lambda}^*$.*

Remark. As a synonym for a data point $\boldsymbol{a}_i$ to be non-influential it is sometimes said that it does not contribute to the optimal solution. In some occasions the entry x_i^* or the corresponding dimension may as well be called non-influential or not contributing.

The second type of constraints regards the case that the domain depends on a matrix $\boldsymbol{B}$ in the following way. Assuming that the constraints have the form $f_i(\boldsymbol{x}) = \boldsymbol{b}_i^T\boldsymbol{x} \leq c_i$, complementary slackness is now used in the other direction, i.e. $\boldsymbol{b}_i^T\boldsymbol{x}^* < c_i \Rightarrow \lambda_i^* = 0$. This gives information about the dual solution. Another observation is that for convex objectives the statement $\boldsymbol{b}_i^T\boldsymbol{x}^* < c_i$ directly implies that deleting this constraint from the problem would not effect the set of optimal solutions. This observation does not even rely on complementary slackness.

Definition 1.22. *For convex optimization problems of the form (1.7), with constraint functions of the form $f_i(\boldsymbol{x}) = \boldsymbol{b}_i^T\boldsymbol{x} \leq c_i$, a data point $\boldsymbol{b}_i$ is called* non-influential *if $\boldsymbol{b}_i^T\boldsymbol{x}^* < c_i$ holds for all optimal solutions $\boldsymbol{x}^*$.*

Chapter 2

Developing Screening Rules

2.1 Conditions for Contribution to the Optimal Solution

This chapter is about the necessary conditions for a data point to be non-influential for an optimization problem. They usually are derived from the KKT conditions and only depend on the domain of the optimization problem. When optimizing a function $f(\boldsymbol{A}\boldsymbol{x})$, proving that an entry x_i^* of the optimal solution has a fixed value is enough to show that the vector $\boldsymbol{a}_i$ is non-influential. Three specific domains are considered here: first the unit simplex, second box constraints and third polytope constraints.

2.1.1 Convex Optimization over the Unit Simplex

Optimization problems over the unit simplex can be interpreted as finding the optimal weights for functions of a weighted sum of objects. Therefore it is an interesting domain to study. Examples such as Support vector machines with squared hinge loss, the minimum enclosing ball problem and the constrained variant of the LASSO are discussed in Chapter 3. Using barycentric coordinates, optimization over a polytope given by its vertices can be described as well.

Definition 2.1. *The* unit simplex $\triangle$ *in* $\Re^m$ *is defined as*

$$\triangle = \{\boldsymbol{x} \in \mathbb{R}^m | \quad \sum_{i=1}^{m} x_i = 1, \quad x_j \geq 0 \quad \forall j \in \{1 \dots m\}\} \tag{2.1}$$

Writing an optimization problem over the unit simplex in standard form (1.7), the following constraint functions are obtained:

$$f_i(\boldsymbol{x}) = -x_i \qquad \forall i \in \{1 \dots m\} \tag{2.2}$$

$$h_1(\boldsymbol{x}) = 1 - \sum_{i=1}^{m} x_i \tag{2.3}$$

The objective function f_0 remains unchanged. These are all linear constraints, hence Slater's condition is fulfilled. Therefore strong duality holds for all convex optimization problems over the unit simplex, which means in particular that the KKT conditions hold

for optimal solutions. Now the Lagrangian and the KKT conditions of the problem are given and used to get a sufficient condition for $x_i^* = 0$.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \nu) &= f_0(\boldsymbol{x}) - \boldsymbol{\lambda}^T \boldsymbol{x} + \nu - \nu \mathbf{1}^T \boldsymbol{x} && (2.4)\\
-x_i^* &\leq 0 \qquad \forall i \in \{1 \dots m\} && (2.5)\\
1 - \sum_{i=1}^{m} x_i^* &= 0 && (2.6)\\
\lambda_i^* &\geq 0 \qquad \forall i \in \{1 \dots m\} && (2.7)\\
-\lambda_i^* x_i^* &= 0 \qquad \forall i \in \{1 \dots m\} && (2.8)\\
\nabla f_0(\boldsymbol{x}^*) - \boldsymbol{\lambda}^* - \nu^* \mathbf{1} &= 0 && (2.9)
\end{aligned}$$

From this the desired condition for an entry of the optimal solution to have a fixed value can be obtained as follows.

Lemma 2.2. *Consider an optimization problem of the form (1.7) over the unit simplex, i.e. f_i and h_i as in the equations (2.2). Then the following characterization on the entries of an optimal solution $\boldsymbol{x}^*$ holds:*

$$\forall i \in \{1 \dots m\} : \quad ((\boldsymbol{e}_i - \boldsymbol{x}^*)^T \nabla f_0(\boldsymbol{x}^*) > 0 \Rightarrow x_i^* = 0) \tag{2.10}$$

Proof. Equation (2.9) gives:

$$\forall i \in \{1 \dots m\} \quad \boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) - \lambda_i^* = \nu^* \tag{2.11}$$

Multiplying with x_i^* returns:

$$\forall i \in \{1 \dots m\} \quad x_i^* \boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) - x_i^* \lambda_i^* = x_i^* \nu^* \tag{2.12}$$

Summing all those equations and using equation (2.11) gives:

$$\boldsymbol{x}^{*T} \nabla f_0(\boldsymbol{x}^*) - \boldsymbol{x}^{*T} \boldsymbol{\lambda}^* = \nu^* = \boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) - \lambda_i^* \tag{2.13}$$

From equation (2.8) it is known that $\boldsymbol{x}^{*T} \boldsymbol{\lambda}^* = 0$, hence it holds that:

$$\lambda_i^* = (\boldsymbol{e}_i - \boldsymbol{x}^*)^T \nabla f_0(\boldsymbol{x}^*) \tag{2.14}$$

Which again by equation (2.8) proves the result. □

2.1.2 Convex Optimization with Box Constraints

In this subsection optimization problems of the following form are examined:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f_0(\boldsymbol{x}) \\ \text{s.t.} \quad & 0 \leq \boldsymbol{x} \leq C\mathbf{1} \end{aligned} \tag{2.15}$$

Where f_0 is a convex function and C is a constant. One example of such a problem is the dual version of a support vector machine problem with hinge loss and no bias. Clearly the constraints are linear, so the KKT conditions must be fulfilled. The Lagrangian and KKT conditions for optimal points $\boldsymbol{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are given as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \nu) &= f_0(\boldsymbol{x}) - \boldsymbol{\lambda}^T\boldsymbol{x} + \boldsymbol{\nu}^T\boldsymbol{x} - C\boldsymbol{\nu}^T\mathbf{1} && (2.16)\\ -x_i^* &\leq 0 \quad \forall i \in \{1 \dots m\} && (2.17)\\ x_i^* - C &\leq 0 \quad \forall i \in \{1 \dots m\} && (2.18)\\ \lambda_i^* &\geq 0 \quad \forall i \in \{1 \dots m\} && (2.19)\\ \nu_i^* &\geq 0 \quad \forall i \in \{1 \dots m\} && (2.20)\\ -\lambda_i^* x_i^* &= 0 \quad \forall i \in \{1 \dots m\} && (2.21)\\ \nu_i(x_i - C) &= 0 \quad \forall i \in \{1 \dots m\} && (2.22)\\ \nabla f_0(\boldsymbol{x}^*) - \boldsymbol{\lambda}^* + \boldsymbol{\nu}^* &= 0 && (2.23) \end{aligned}$$

As before, those are used to find a condition on the entries of the optimal solutions.

Lemma 2.3. *Consider an optimization problem of the form (2.15). Then the following characterization on the entries of an optimal solution $\boldsymbol{x}^*$ holds:*

$$\begin{aligned} \forall i \in \{1 \dots m\}: \quad & (\boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) > 0 \Rightarrow x_i^* = 0) && (2.24)\\ \forall i \in \{1 \dots m\}: \quad & (\boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) < 0 \Rightarrow x_i^* = C) && (2.25) \end{aligned}$$

Proof. Equation (2.23) and inequality (2.19) give:

$$\boldsymbol{\lambda}^* = \boldsymbol{\nu}^* + \nabla f_0(\boldsymbol{x}^*) \geq \nabla f_0(\boldsymbol{x}^*) \tag{2.26}$$

In the next step inequality (2.21) is used for the second implication to prove the first part of the claim:

$$\boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) > 0 \Rightarrow \lambda_i > 0 \Rightarrow x_i^* = 0 \tag{2.27}$$

The second part works analogous. Equation (2.23) and inequality (2.20) imply:

$$\boldsymbol{\nu}^* = \boldsymbol{\lambda}^* - \nabla f_0(\boldsymbol{x}^*) \geq -\nabla f_0(\boldsymbol{x}^*) \tag{2.28}$$

Now inequality (2.22) is used for the second implication:

$$\boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) < 0 \Rightarrow \nu_i > 0 \Rightarrow x_i^* = C \tag{2.29}$$

□

2.1.3 Convex Optimization over Polytopes

In this section, problems of the following form are discussed:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f_0(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{A}^T\boldsymbol{x} \leq \boldsymbol{b} \end{aligned} \tag{2.30}$$

Where $f_0(\boldsymbol{x})$ is a convex function and $\boldsymbol{A}$ is a $n \times m$ matrix whose columns are the m data points $\boldsymbol{a}_i$. The domain described by $\boldsymbol{A}^T\boldsymbol{x} \leq \boldsymbol{b}$ is a polytope. Each inequality represents a facet. Therefore the polytope is described by its n facets. Problems of this type occur often considering ℓ_1-regularized problems, since their dual can be formulated as an optimization problem with polytope constraints. Polytopes described as convex combinations of their vertices on the other hand can be considered using barycentric coordinates as problems over the unit simplex. The KKT-conditions for problem (2.30) read:

$$\lambda_i^*(\boldsymbol{a}_i^T\boldsymbol{x}^* - b_i) = 0 \quad \forall i \in \{1 \dots m\} \tag{2.31}$$

There are two ways to interpret it. On the one hand considering the KKT condition given above, knowing that $\boldsymbol{a}_i^T\boldsymbol{x}^* < b_i$ gives $\lambda_i^* = 0$. Which means that we fix an entry of the dual solution $\boldsymbol{\lambda}^*$. So if the dual objective has the form $g(\tilde{\boldsymbol{A}}\boldsymbol{\lambda}, \boldsymbol{\nu})$, the vector $\tilde{\boldsymbol{a}}_i$ is non-influential for the dual problem. On the other hand, $\boldsymbol{a}_i^T\boldsymbol{x}^* < b_i$ implies that the problem emerging from deleting the constraint $\boldsymbol{a}_i^T\boldsymbol{x} < b_i$ has the same solution as the original one. Hence, $\boldsymbol{a}_i$ is non-influential on the primal problem. This would be screening a data point defining the domain of the optimization problem.

2.2 Restrictions on Optimal Solutions

The previous section showed that the conditions for a data point to be non-influential depend on an optimal solution or the derivative of the objective function at an optimal point. In this chapter, basic properties of convex functions are used to get restrictions on those quantities. As it can be seen those restrictions usually depend on properties of the objective function. The simplest restriction one could think of is given directly by the definition of strong convexity:

Lemma 2.4. *Let f be a strongly convex function with parameter μ. Consider the convex optimization problem:*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})+\boldsymbol{c}^T\boldsymbol{x} \\ s.t. \quad & \boldsymbol{x}\in\mathcal{D} \end{aligned} \tag{2.32}$$

With convex domain $\mathcal{D}$. Then for an optimal point $\boldsymbol{x}^*$ it holds:

$$\|\boldsymbol{A}\boldsymbol{x}^*-\boldsymbol{A}\boldsymbol{x}\|^2 \leq \frac{2}{\mu}(f(\boldsymbol{A}\boldsymbol{x}^*+\boldsymbol{b})-f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})) \tag{2.33}$$

Proof.

$$\begin{aligned} f(\boldsymbol{A}\boldsymbol{x}^*+\boldsymbol{b}) &\geq f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})+\nabla f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x}^*+\boldsymbol{b}-(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})) && (2.34)\\ &\quad +\frac{\mu}{2}\|\boldsymbol{A}\boldsymbol{x}^*+\boldsymbol{b}-(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})\|^2 && (2.35)\\ &\geq f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})+\frac{\mu}{2}\|\boldsymbol{A}\boldsymbol{x}^*-\boldsymbol{A}\boldsymbol{x}\|^2 && (2.36) \end{aligned}$$

The first inequality is the definition of strong convexity.
The second inequality is Proposition 1.15.
Reordering the terms gives the property. □

Observation 2.5. *This result implies that a certificate for the suboptimality of a point $\boldsymbol{x}$ immediately gives a restriction on the position of an optimal point $\boldsymbol{x}^*$.*

The gap function could be used as a suboptimality certificate in Lemma 2.4, but it can be used directly to get an even better restriction.

Theorem 2.6. *Consider a convex optimization problem of the form (1.7) with feasible set X. Assume that a function f, a matrix $\boldsymbol{A}$ and two vectors $\boldsymbol{b}$, $\boldsymbol{c}$ are given such that the objective function f_0 can be written as $f_0(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})+\boldsymbol{c}^T\boldsymbol{x}$. Furthermore assume that f is strongly convex with parameter μ. Then for an optimal solution $\boldsymbol{x}^*$ it holds that:*

$$\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\boldsymbol{x}^*\|^2 \leq \frac{1}{\mu}gap_{FW}(\boldsymbol{x}) \tag{2.37}$$

Proof.

$$\begin{aligned} \mu\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\boldsymbol{x}^*\|^2 &\leq (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\boldsymbol{x}^*)^T(\nabla f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})-\nabla f(\boldsymbol{A}\boldsymbol{x}^*+\boldsymbol{b})) && (2.38)\\ &= (\boldsymbol{x}-\boldsymbol{x}^*)^T(\nabla f_0(\boldsymbol{x})-\nabla f_0(\boldsymbol{x}^*)) && (2.39)\\ &\leq (\boldsymbol{x}-\boldsymbol{x}^*)^T\nabla f_0(\boldsymbol{x}) && (2.40)\\ &\leq gap_{FW}(\boldsymbol{x}) && (2.41) \end{aligned}$$

The first inequality is Definition 1.3.
Equation (2.39) holds, since $\nabla f_0(\boldsymbol{x}) = \nabla_x f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})+\boldsymbol{c} = \boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})+\boldsymbol{c}$.
The second inequality uses Proposition 1.15.
The third inequality is Proposition 1.18. □

Observation 2.7. *This theorem improves the restriction one would get taking the result of Lemma 2.4 and using the gap function as suboptimality certificate, by a factor of two, i.e. the upper bound on the distance between the approximate solution to each of the optimal solutions is cut in half.*

To get a restriction for the derivatives of the objective function at the optimal solutions one can use another property of the objective, namely smoothness, i.e. Lipschitz continuity of the gradient.

Theorem 2.8. *Consider a convex optimization problem of the form (1.7) with feasible set X. Assume that a function f, a matrix $\boldsymbol{A}$ and two vectors $\boldsymbol{b}$, $\boldsymbol{c}$ are given such that the objective function f_0 can be written as $f_0(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x}$. Furthermore assume that the derivative ∇f is Lipschitz continuous with parameter L and that there is an upper bound $\delta(\boldsymbol{x})$ for the distance between $\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{A}\boldsymbol{x}^*$, which may depend on $\boldsymbol{x}$, i.e. $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*\| \leq \delta(\boldsymbol{x})$. Then it holds that:*

$$\|\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b})\| \leq L\,\delta(\boldsymbol{x}) \tag{2.42}$$

Proof. The statement follows directly by Definition 1.4 and Theorem 2.6. □

These are the restrictions that will eventually be used later to obtain screening rules in this thesis. There are two additional restrictions mentioned here. First one can use Lipschitz continuity of the objective to bound the distance from an approximate solution to the optimal ones from below given a lower bound for the suboptimality of the solution.

Lemma 2.9. *Consider an optimization problem of the form:*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x} \\ s.t. \quad & \boldsymbol{x} \in \mathcal{D} \end{aligned} \tag{2.43}$$

Let f be Lipschitz continuous with parameter L, then

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*\| \geq \frac{1}{L}(f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) - f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b})) \tag{2.44}$$

The second one is more interesting and uses duality considerations.

Lemma 2.10. *Consider an optimization problem of the form (1.7) with $f_0(\boldsymbol{x}) := f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x}$. Call the primal objective $p(\boldsymbol{x})$ and the dual $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$. Then given primal and dual feasible points $\boldsymbol{x}$ and $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ it holds:*

$$d(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\boldsymbol{x}^*) \leq p(\boldsymbol{x}) \tag{2.45}$$

If the level sets of the objective function are easy to determine, this lemma can be useful. Lemma 2.10 basically says that the optimal solutions have to lie in between the level sets of the objective function with values $p(\boldsymbol{x})$ and $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$. This thesis focusses on rather general objective functions and therefore does not assume it to be easy to determine its level sets. It has been used for the penalized version of the LASSO in a paper by Olivier Fercoq and Salmon (2015). For the dual problem of the penalized LASSO the level sets are simply spheres. They took the set in between those spheres and intersected it with the ball one can obtain from Lemma 2.4 to define a GAP SAFE DOME screening rule. However their simulations did not show much improvement in terms of computation time using the dome instead of directly using the ball obtained from Lemma 2.4.

2.3 Resulting Screening Rules

Here, the results of the previous two chapters are used to state screening rules. Section 2.1 gave sufficient conditions for a data point to be non-influential and Chapter 2.2 supplied information about the optimal points $\boldsymbol{x}^*$ necessary to conclude screening rules. Note that the information about the optimal points does not occur directly in the stated theorems. Instead for modularity, just an upper bound $\delta(\boldsymbol{x})$ on $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*\|$ is required. As such either Lemma 2.4 or Theorem 2.6 is used in Chapter 3.

2.3.1 Screening Rule for the Unit Simplex

Theorem 2.11. *Consider an optimization problem of the form (1.7) over the unit simplex. Assume that a function f, a matrix $\boldsymbol{A}$ and two vectors $\boldsymbol{b}$, $\boldsymbol{c}$ are given such that the objective function f_0 can be written as $f_0(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x}$. Furthermore assume that the derivative ∇f is Lipschitz continuous with parameter L and that there is an upper bound $\delta(\boldsymbol{x})$ for the distance between $\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{A}\boldsymbol{x}^*$, which may depend on $\boldsymbol{x}$, i.e. $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*\| \leq \delta(\boldsymbol{x})$. Let $\boldsymbol{x}^*$ be an optimal solution. Then the following statement holds:*

$$(\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}) > L\,\delta(\boldsymbol{x})\,\|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\| \Rightarrow x_i^* = 0 \Rightarrow \boldsymbol{a}_i \text{ is non-influential} \tag{2.46}$$

Proof.

$$\begin{aligned}
& (\boldsymbol{e}_i - \boldsymbol{x}^*)^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + \boldsymbol{c}) && (2.47)\\
= \; & (\boldsymbol{e}_i - \boldsymbol{x} + \boldsymbol{x} - \boldsymbol{x}^*)^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + \boldsymbol{c}) && (2.48)\\
= \; & (\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + \boldsymbol{c}) + (\boldsymbol{x} - \boldsymbol{x}^*)^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + \boldsymbol{c}) && (2.49)\\
\geq \; & (\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + \boldsymbol{c}) && (2.50)\\
= \; & (\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + \boldsymbol{c} - \boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})) && (2.51)\\
= \; & (\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}) && (2.52)\\
& +(\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x})^T(\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})) && (2.53)\\
\geq \; & (\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}) && (2.54)\\
& - \|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\|\,\|\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})\| && (2.55)\\
\geq \; & (\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}) - \|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\|\,L\,\delta(\boldsymbol{x}) && (2.56)
\end{aligned}$$

Step (2.50) uses Proposition 1.15.
Step (2.55) uses that $\boldsymbol{a}^T\boldsymbol{b} = \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|\cos(\angle(\boldsymbol{a},\boldsymbol{b})) \geq -\|\boldsymbol{a}\|\,\|\boldsymbol{b}\|$.
Step (2.56) uses Theorem 2.8.
Lemma 2.2 now implies the result, since

$$(\boldsymbol{e}_i - \boldsymbol{x}^*)^T\nabla f_0(\boldsymbol{x}^*) = (\boldsymbol{e}_i - \boldsymbol{x}^*)^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + \boldsymbol{c}) \tag{2.57}$$

. □

2.3.2 Screening Rule for Box Constraints

Theorem 2.12. *Consider an optimization problem of the form (2.15). Assume that we are given a function f, a matrix $\boldsymbol{A}$ with columns $\boldsymbol{a}_i$ and two vectors $\boldsymbol{b}$, $\boldsymbol{c}$ such that the objective function f_0 can be written as $f_0(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x}$. Furthermore assume that the derivative ∇f is Lipschitz continuous with parameter L and that there is have an upper bound $\delta(\boldsymbol{x})$ for the distance between $\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{A}\boldsymbol{x}^*$, which may depend on $\boldsymbol{x}$, i.e. $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*\| \leq \delta(\boldsymbol{x})$. Let $\boldsymbol{x}^*$ be an optimal solution. Then the following statement holds:*

$$\boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i > L \|\boldsymbol{a}_i\| \, \delta(\boldsymbol{x}) \quad \Rightarrow x_i^* = 0 \quad \Rightarrow \boldsymbol{a}_i \text{ is non-influential} \tag{2.58}$$

$$\boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i < -L \|\boldsymbol{a}_i\| \, \delta(\boldsymbol{x}) \quad \Rightarrow x_i^* = C \quad \Rightarrow \boldsymbol{a}_i \text{ is non-influential} \tag{2.59}$$

Proof.

$$\begin{aligned}
\boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + c_i &= \boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i && (2.60)\\
&\quad + \boldsymbol{a}_i^T (\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})) && (2.61)\\
&\geq \boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i && (2.62)\\
&\quad - \|\boldsymbol{a}_i\| \, \|\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})\| && (2.63)\\
&\geq \boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i - \|\boldsymbol{a}_i\| \, L \, \delta(\boldsymbol{x}) && (2.64)
\end{aligned}$$

The first inequality holds, since $\boldsymbol{a}^T\boldsymbol{b} = \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \cos(\angle(\boldsymbol{a}, \boldsymbol{b})) \geq - \|\boldsymbol{a}\| \, \|\boldsymbol{b}\|$.
The second inequality follows from Theorem 2.8.
Lemma 2.3 now implies the first part.

$$\begin{aligned}
\boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) + c_i &= \boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i && (2.65)\\
&\quad + \boldsymbol{a}_i^T (\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})) && (2.66)\\
&\leq \boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i && (2.67)\\
&\quad + \|\boldsymbol{a}_i\| \, \|\nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})\| && (2.68)\\
&\leq \boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i + \|\boldsymbol{a}_i\| \, L \, \delta(\boldsymbol{x}) && (2.69)
\end{aligned}$$

The first inequality holds, since $\boldsymbol{a}^T\boldsymbol{b} = \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \cos(\angle(\boldsymbol{a}, \boldsymbol{b})) \leq \|\boldsymbol{a}\| \, \|\boldsymbol{b}\|$.
The second inequality follows from Theorem 2.8.
Lemma 2.3 now implies the second part.

□

2.3.3 Screening Rule for Polytope Constraints

Theorem 2.13. *Consider an optimization problem of the form (2.30). Assume that there is an upper bound $\delta(\boldsymbol{x})$ for the distance between $\boldsymbol{x}$ and $\boldsymbol{x}^*$, which may depend on $\boldsymbol{x}$, i.e. $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \delta(\boldsymbol{x})$. Let $\boldsymbol{x}^*$ be an optimal solution. Then the following statement holds:*

$$\boldsymbol{a}_i^T \boldsymbol{x} < b_i - \|\boldsymbol{a}_i\| \, \delta(\boldsymbol{x}) \Rightarrow \boldsymbol{a}_i^T \boldsymbol{x}^* < b_i \Rightarrow \boldsymbol{a}_i \text{ is non-influential} \tag{2.70}$$

Proof.

$$\begin{aligned}
\boldsymbol{a}_i^T \boldsymbol{x}^* &= \boldsymbol{a}_i^T \boldsymbol{x} + \boldsymbol{a}_i^T (\boldsymbol{x}^* - \boldsymbol{x}) && (2.71)\\
&\leq \boldsymbol{a}_i^T \boldsymbol{x} + \|\boldsymbol{a}_i\| \, \|\boldsymbol{x}^* - \boldsymbol{x}\| && (2.72)\\
&\leq \boldsymbol{a}_i^T \boldsymbol{x} + \|\boldsymbol{a}_i\| \, \delta(\boldsymbol{x}) && (2.73)
\end{aligned}$$

The first inequality holds, since $\boldsymbol{a}^T\boldsymbol{b} = \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|\cos(\angle(\boldsymbol{a},\boldsymbol{b})) \leq \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|$.
The second inequality follows from Theorem 2.6.
This proves the result. □

2.3.4 Quality of these Screening Rules

The screening rules developed here use an approximation $\boldsymbol{x}$ to identify non-influential data points in an optimization problem. That leads to the question: how many of those can be identified in the limit $\|\boldsymbol{x} - \boldsymbol{x}^*\| \to 0$?

Proposition 2.14. *Consider a convex optimization problem with differentiable objective and constraint functions. Let $\mathcal{X}$ be its feasible set, $\boldsymbol{x}$ an approximate solution and $\boldsymbol{x}^*$ the optimal solution and $gap_{FW}(\boldsymbol{x})$ its gap function. Then in the limit $\|\boldsymbol{x} - \boldsymbol{x}^*\| \to 0$ it holds that:*

$$gap_{FW}(\boldsymbol{x}) \quad \to \quad 0 \tag{2.74}$$

Proof.

$$\lim_{\|\boldsymbol{x}-\boldsymbol{x}^*\|\to 0} gap_{FW}(\boldsymbol{x}) \quad = \quad \lim_{\|\boldsymbol{x}-\boldsymbol{x}^*\|\to 0} \max_{\boldsymbol{y}\in X} (\boldsymbol{x}-\boldsymbol{y})^T \nabla f_0(\boldsymbol{x}) \tag{2.75}$$

$$= \quad \max_{\boldsymbol{y}\in X} (\boldsymbol{x}^*-\boldsymbol{y})^T \nabla f_0(\boldsymbol{x}^*) \tag{2.76}$$

$$= \quad 0 \tag{2.77}$$

The limit can be evaluated since, due to the convexity of f_0, $\nabla f_0(\boldsymbol{x})$ is bounded from below and f_0 is differentiable. The last step follows by Proposition 1.15. □

The gap function can be used as the upper bound $\delta(\boldsymbol{x})$ in Theorem 2.11, 2.12 and 2.13. Which implies, regarding the KKT condition (1.25), that in the limit $\|\boldsymbol{x} - \boldsymbol{x}^*\| \to 0$ the screening rules are equivalent to either $\lambda_i^* = 0$ or $\nabla f_i(\boldsymbol{x}^*) = 0$, depending on the Theorem considered. The KKT-conditions do not identify all the non-influential data points of the optimal solution, but almost all of them. To be more precise the only way a non-influential data point is not identified is if both $\lambda_i^* = 0$ and $f_i(\boldsymbol{x}^*) = 0$. If the duality gap $gap_D(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is used as suboptimality certificate to get an upper bound δ via Lemma 2.4, it is more complicated. There are actually two approximate solutions involved. In the limit of both being close to the optimum and assuming strong duality the same reasoning as before can be done.

Chapter 3

Applications

In this chaper the results of this thesis are applied on typical convex optimizatio problems and discussed to what extend they can reproduce known results.

3.1 Support Vector Machine (SVM)

The problem setting here is that there is a set of data points $\boldsymbol{a}_i$ which belong to two categories. The goal is to find an optimal classifying hyperplane separating the two groups of data points. If they are perfectly separable, i.e. there exists a hyperplane such that the point of one category are on one side of it and all other points on the other side, the optimal hyperplane maximizes the minimal distance to a data point, which is called margin. Otherwise a loss for points being on the wrong side of the hyperplane has to be defined and minimized. The problem is then called soft-margin SVM.

3.1.1 SVM with Squared Loss

It can be shown that soft-margin SVM variants with squared loss can be formulated as an optimization problem with objective function $f_0(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}$, which is minimized over the unit simplex $\triangle$ (Jaggi, 2013; Scholkopf and Smola, 2001; Keerthi, Shevade, Bhattacharyya, and Murthy, 2000; Tsang, Kwok, ming Cheung, and Cristianini, 2005). That includes one-class and two-class SVM as well as versions with regularized offset or no offset. Theorem 2.11 is directly applicable to this problem. To do so the objective function is written as $f_0(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x}$, where $\boldsymbol{b}$ and $\boldsymbol{c}$ are zero. The function f is then strongly convex with parameter $\mu = 1$. Also the derivative ∇f is Lipschitz-continuous with parameter $L = 1$. To get an upper bound $\delta(\boldsymbol{x})$ Theorem 2.6 is used, where $gap_{FW}(\boldsymbol{x}) = \max\limits_{i \in 1 \dots m} (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{a}_i)^T\boldsymbol{A}\boldsymbol{x}$. In the end Theorem 2.11 becomes:

$$(\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x} > \sqrt{gap_{FW}(\boldsymbol{x})}\,\|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\| \Rightarrow x_i^* = 0 \tag{3.1}$$

Which is a screening rule that also can be obtained by geometric considerations. It also improves the result of Wysling (2015) in the sense that the threshold for the quantity $(\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x}$, that has to be exceeded to assure $x_i^* = 0$, is smaller.

3.1.2 SVM with Hinge Loss and No Bias

Here the primal problem can be formulated as:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\epsilon}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\mathbf{1}^T\boldsymbol{\epsilon} \\ \text{s.t.} \quad & \boldsymbol{w}^T\boldsymbol{a}_i \geq 1 - \epsilon_i \qquad \forall i \in \{1:p\} \\ & \epsilon_i \geq 0 \qquad\qquad\quad \forall i \in \{1:p\} \end{aligned} \tag{3.2}$$

A dual formulation of that problem is given by:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \boldsymbol{x}^T\mathbf{1} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} \\ \text{s.t.} \quad & 0 \leq \boldsymbol{x} \leq C\mathbf{1} \end{aligned} \tag{3.3}$$

Lemma 2.4 is applied on the negative of the dual formulation. Writing the objective function as $f_0(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x}$ we have to set $\boldsymbol{b} = 0$ and $\boldsymbol{c} = -\mathbf{1}$. Also f is then strongly convex with parameter 1 and its derivative Lipschitz continuous with parameter 1. The duality gap between primal and dual feasible points $gap_D(\boldsymbol{w}, \boldsymbol{\epsilon}, \boldsymbol{x})$ is now used as suboptimality certificate which can play the role of the upper bound $\delta(\boldsymbol{x})$ using Lemma 2.4. For a given $\boldsymbol{x}$ a primal feasible point can be obtained by setting $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{x}$ 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:

$$\boldsymbol{a}_i^T\boldsymbol{A}\boldsymbol{x} + 1 > \|\boldsymbol{a}_i\|\sqrt{2gap_D(\boldsymbol{x})} \quad \Rightarrow \quad x_i^* = 0 \tag{3.4}$$

$$\boldsymbol{a}_i^T\boldsymbol{A}\boldsymbol{x} + 1 < -\|\boldsymbol{a}_i\|\sqrt{2gap_D(\boldsymbol{x})} \quad \Rightarrow \quad x_i^* = C \tag{3.5}$$

One could also try to use Theorem 2.6 as upper bound $\delta(\boldsymbol{x})$, which would give a similar rule, where $2gap_D(\boldsymbol{x})$ is replaced by $gap_{FW}(\boldsymbol{x})$. To do so a linear program must be solved, namely:

$$\begin{aligned} \max_{\boldsymbol{y}} \quad & -\boldsymbol{y}^T(\mathbf{1} - \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}) \\ \text{s.t.} \quad & 0 \leq \boldsymbol{y} \leq C \end{aligned} \tag{3.6}$$

It is not clear how it compares in terms of computation time or performance to the result presented here to use this upper bound for screening.
Ogawa et al. (2014) also obtained a screening rule for this problem. They combined two approaches to screening. A dynamic one using the exact solution for a different parameter $\tilde{C}$ and a sequential approach using previously computed approximate solutions for the currently considered parameter C. Both are then used to find two balls that necessarily contain the optimal solution. Afterwards they intersect those balls to get a smaller region. By applying Lemma 2.4 this thesis also gives a ball containing the optimal solution depending on an approximate solution. Due to different notation it is not directly clear how this ball compares to the one Ogawa et al. (2014) used.

3.2 Minimum Enclosing Ball (MEB)

The setting is that there are p points $\boldsymbol{a}_1$ to $\boldsymbol{a}_p$ in $\mathbb{R}^n$. Now the goal is to find the smallest Ball $B_{\boldsymbol{c},r}$ with center $\boldsymbol{c}$ and radius r, i.e.: $B_{\boldsymbol{c},r} = \{\boldsymbol{x} \in \mathbb{R}^n | \, \|\boldsymbol{c} - \boldsymbol{x}\| \leq r\}$, 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 $B_{\boldsymbol{c}^*,r^*}$. Removing those points from the problem does not change the optimal ball. This can be formulated as an optimization problem:

$$\begin{aligned} \min_{\boldsymbol{c},r} \quad & r^2 \\ \text{s.t.} \quad & \|\boldsymbol{c} - \boldsymbol{a}_i\|_2^2 \leq r^2 \quad \forall i \in \{1:p\} \end{aligned} \tag{3.7}$$

A dual formulation is given by:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & -\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} + \sum_{j=1}^{p} \boldsymbol{a}_j^T \boldsymbol{a}_j x_j \\ \text{s.t.} \quad & \boldsymbol{x} \in \triangle \end{aligned} \tag{3.8}$$

(Matousek and Gärtner, 2006, Chapter 8.7)

On the dual formulation of the MEB problem as in equation (3.8) Theorem 2.11 can be applied by considering the minimization of the negative objective value. Where $\boldsymbol{A}' = \boldsymbol{A}$, $\boldsymbol{b}' = 0$ and $\boldsymbol{c}'_j = -\boldsymbol{a}_j^T \boldsymbol{a}_j$. The upper bound $\delta(\boldsymbol{x})$ is obtained by Theorem 2.6. The corresponding function $f(\boldsymbol{A}'\boldsymbol{x}+\boldsymbol{b}')$ is strongly convex with parameter $\mu = 2$, hence the theorem gives $\delta(\boldsymbol{x}) = \sqrt{\frac{1}{2}\max_i (\boldsymbol{x} - \boldsymbol{e}_i)^T (2\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{c}')}$. Theorem 2.11 now gives a sufficient condition for $\boldsymbol{a}_i$ to be non-influential, i.e. $\boldsymbol{a}_i$ lies in the interior of the MEB:

$$(\boldsymbol{e}_i - \boldsymbol{x})^T (2\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{c}') > 2\sqrt{\frac{1}{2}\max_j (\boldsymbol{x} - \boldsymbol{e}_i)^T (2\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{c}')} \, \|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\| \Rightarrow x_i^* = 0 \tag{3.9}$$

To compare this to the results of other groups the statement has to be rewritten in terms of quantities typical used for the MEB problem. Given a dual optimal solution $\boldsymbol{x}^*$ the MEB $B_{\boldsymbol{c}^*,r^*}$ is given by setting its centre $\boldsymbol{c}^* = \boldsymbol{A}\boldsymbol{x}^*$ and its squared radius to the objective value $r^* = \sqrt{-\boldsymbol{x}^{*T} \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}^* + \sum_{j=1}^{p} \boldsymbol{a}_j^T \boldsymbol{a}_j x_j}$ (Matousek and Gärtner, 2006, Chapter 8.7). For a dual feasible point $\boldsymbol{x}$ the ball with centre $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{x}$ and squared radius r^2 equal to the objective value in equation (3.8) can be considered an approximate solution. It won't contain all points $\boldsymbol{a}_i$, since the radius is smaller than the optimal one. Computing $\max_j \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{a}_j\|$ gives an upper bound R for the optimal radius. The result is now rewritten in terms of r, R and $\boldsymbol{c}$.

Proposition 3.1. *For the dual formulation of the MEB problem as in (3.8) we define:*

$$r = \sqrt{-\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} + \sum_{j=1}^{p} \boldsymbol{a}_j^T \boldsymbol{a}_j x_j} \tag{3.10}$$

$$R = \max_j \|\boldsymbol{c} - \boldsymbol{a}_j\| \tag{3.11}$$

$$\boldsymbol{c} = \boldsymbol{A}\boldsymbol{x} \tag{3.12}$$

Then the point $\boldsymbol{a}_i$ lies in the interior of the MEB and hence can be discarded if:

$$\|\boldsymbol{a}_i - \boldsymbol{c}\| < \frac{r^2}{\|\boldsymbol{a}_i - \boldsymbol{c}\|} - 2\sqrt{\frac{R^2 - r^2}{2}} \tag{3.13}$$

Proof. Consider the inequality in statement (3.9). The left-hand-side is:

$$\begin{aligned}(\boldsymbol{e}_i - \boldsymbol{x})^T(2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}') &= 2(\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x} + c_i' - \boldsymbol{x}^T\boldsymbol{c}' && (3.14)\\ &= 2\boldsymbol{a}_i^T\boldsymbol{c} - 2\boldsymbol{c}^T\boldsymbol{c} - \boldsymbol{a}_i^T\boldsymbol{a}_i + r^2 + \boldsymbol{c}^T\boldsymbol{c} && (3.15)\\ &= -\|\boldsymbol{a}_i - \boldsymbol{c}\|^2 + r^2 && (3.16)\end{aligned}$$

The right-hand-side becomes:

$$\begin{aligned}& 2\sqrt{\frac{1}{2}\max_j(\boldsymbol{x} - \boldsymbol{e}_j)^T(2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}')}\,\|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\| && (3.17)\\ = \;& 2\sqrt{\frac{1}{2}\max\|\boldsymbol{a}_i - \boldsymbol{c}\|^2 - r^2}\,\|\boldsymbol{a}_i - \boldsymbol{c}\| && (3.18)\\ = \;& 2\sqrt{\frac{R^2 - r^2}{2}}\,\|\boldsymbol{a}_i - \boldsymbol{c}\| && (3.19)\end{aligned}$$

In the first step the calculation for the left-hand-side is used. Together this gives:

$$-\|\boldsymbol{a}_i - \boldsymbol{c}\|^2 + r^2 > 2\sqrt{\frac{R^2 - r^2}{2}}\,\|\boldsymbol{a}_i - \boldsymbol{c}\| \Rightarrow x_i^* = 0 \tag{3.20}$$

Reordering the terms we proves claim. □

There are two papers on screening for the MEB problem.The first one was written by Ahipaşaoğlu and Yildirim (2008) using an approximate solution $B_{\boldsymbol{c},r}$ and calculating an upper bound for $\|\boldsymbol{c} - \boldsymbol{c}^*\|$. Källberg and Larsson (2014) improved this upper bound using duality considerations. Their result also considers situations where more than one approximate solution is available. In this thesis only one known approximation is used, hence only their result for that case is considered here:

Theorem 3.2. *Suppose the ball $B_{\boldsymbol{c},r}$ is dual feasible, i.e. $\boldsymbol{c}$ is a convex combination of the input points. Let $R = \max_j\|\boldsymbol{a}_j - \boldsymbol{c}\|$. A point $\boldsymbol{a}_i$ lies in the interior of $B_{\boldsymbol{c}^*,r^*}$ if*

$$\|\boldsymbol{a}_i - \boldsymbol{c}\| < \sqrt{\frac{R^2 + r^2}{2}} - \sqrt{\frac{R^2 - r^2}{2}} \tag{3.21}$$

Källberg and Larsson (2014)

This looks very similar to the inequality in Proposition 3.1:

$$\|\boldsymbol{a}_i - \boldsymbol{c}\| < \frac{r^2}{\|\boldsymbol{a}_i - \boldsymbol{c}\|} - 2\sqrt{\frac{R^2 - r^2}{2}} \tag{3.22}$$

The following proposition shows that the newly obtained screening rule is at least as good as the one from Källberg and Larsson (2014). This means that every point $\boldsymbol{a}_i$ discarded by their rule is also discarded by the new rule Proposition 3.1.

Proposition 3.3. *From*

$$\|\boldsymbol{a}_i - \boldsymbol{c}\| < \sqrt{\frac{R^2 + r^2}{2}} - \sqrt{\frac{R^2 - r^2}{2}} \tag{3.23}$$

It follows that

$$\|\boldsymbol{a}_i - \boldsymbol{c}\| < \frac{r^2}{\|\boldsymbol{a}_i - \boldsymbol{c}\|} - 2\sqrt{\frac{R^2 - r^2}{2}} \tag{3.24}$$

Proof. Assume that inequality (3.23) holds. Then it follows that:

$$\begin{aligned}(\sqrt{\frac{R^2+r^2}{2}}+\sqrt{\frac{R^2-r^2}{2}})\,\|\boldsymbol{a}_i-\boldsymbol{c}\| &< (\sqrt{\frac{R^2+r^2}{2}}+\sqrt{\frac{R^2-r^2}{2}})(\sqrt{\frac{R^2+r^2}{2}}-\sqrt{\frac{R^2-r^2}{2}})\\ &= \frac{R^2+r^2}{2}-\frac{R^2-r^2}{2}\\ &= r^2\end{aligned}$$

Which, reordering the terms, gives directly:

$$\|\boldsymbol{a}_i-\boldsymbol{c}\| < \sqrt{\frac{R^2+r^2}{2}}-\sqrt{\frac{R^2-r^2}{2}} \tag{3.25}$$

$$< \frac{r^2}{\|\boldsymbol{a}_i-\boldsymbol{c}\|}-2\sqrt{\frac{R^2-r^2}{2}} \tag{3.26}$$

□

3.3 The LASSO

The LASSO (least absolute shrinkage and selection operator) is a slightly changed least squares problem, designed to make the optimal solutions sparse, i.e. most entries will be zero. Therefore it is a promising set-up for the use of screening rules. There are two under certain restrictions equivalent formulations of the Lasso:

Definition 3.4. *The* ℓ_1-unit-ball *in* $\mathbb{R}^n$ *is defined as*

$$\blacklozenge := \{\boldsymbol{x}\in\mathbb{R}^n \,|\, \|\boldsymbol{x}\|_1 \le 1\} \tag{3.27}$$

Definition 3.5. *Given a Matrix* $\boldsymbol{A}$ *and a vector* b *the following optimization problem is denoted an instance of the* penalized version of the LASSO*:*

$$\min_{\boldsymbol{x}} \quad \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2+\lambda\|\boldsymbol{x}\|_1 \tag{3.28}$$

Definition 3.6. *Given a Matrix* $\boldsymbol{A}$ *and a vector* b *the following optimization problem is denoted an instance of the constrained variant of the LASSO:*

$$\begin{aligned}\min_{\boldsymbol{x}} \quad & \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2\\ s.t. \quad & \|\boldsymbol{x}\|_1 \le s\end{aligned} \tag{3.29}$$

Up to now only the penalized version has been studied using a dual formulation of the problem:

$$\begin{aligned}-\min_{\boldsymbol{\theta}} \quad & \frac{\lambda^2}{2}\left\|\boldsymbol{\theta}-\frac{\boldsymbol{b}}{\lambda}\right\|^2-\frac{1}{2}\|\boldsymbol{b}\|^2\\ \text{s.t.} \quad & \left|\boldsymbol{a}_i^T\boldsymbol{\theta}\right| \le 1 \qquad\qquad \forall i\in\{1:p\}\end{aligned} \tag{3.30}$$

The basic idea was formulated by Ghaoui et al. (2010). They computed a region in which the dual optimal point $\boldsymbol{\theta}^*$ had to be included and derived from there which primal variables x_i have to be equal to zero. Which implies that the corresponding $\boldsymbol{a}_i$ can safely

be omitted. There is a good summary how this approach developed in recent years in the paper from Olivier Fercoq and Salmon (2015). They distinguish into static rules as in Ghaoui et al. (2010) which only depend on the problem input and can be used prior to any computations on the one hand. On the other hand there are dynamic rules like the one from Wang et al. (2012) relying on already computed exact solutions for different penalty parameter λ, which usually shrinks the region containing the dual optimal solution. The third category they mention and propose a rule themselves are sequential rules relying on previously computed approximate solutions, as done in this thesis. In addition to previous results Fercoq et al. used considerations on the duality gap to get a region where $\boldsymbol{\theta}^*$ has to be included.
The dual formulation in Definition 3.5 can be interpreted as a projection of $\frac{\boldsymbol{b}}{\lambda}$ on a polyhedron whose facets are defined by the data points $\boldsymbol{a}_i$. The constrained variant of the LASSO 3.6 can be rewritten similarly:

$$\begin{aligned} \min_{\boldsymbol{y}} \quad & \frac{s^2}{2} \left\| \boldsymbol{y} - \frac{\boldsymbol{b}}{s} \right\|^2 \\ \text{s.t.} \quad & \boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{y} \\ & \|\tilde{\boldsymbol{x}}\|_1 \leq 1 \end{aligned} \tag{3.31}$$

In Comparison to the previous interpretation of the dual penalized LASSO as a projection it can be seen that this is a projection of $\frac{\boldsymbol{b}}{s}$ onto a polytope whose vertices are defined by the data points $\boldsymbol{a}_i$. In fact the two polytopes projected on are the polar of each other. This thesis gives the equipment to give screening rules for both types of the Lasso. The constrained variant is discussed in the following subsection. The penalized version is a special case of ℓ_1-regularized problems discussed in Section 3.4.

3.3.1 Constrained Variant of the LASSO

Consider an instance of the constrained variant of the LASSO as in Definition 3.6. For simplicity let $s = 1$, other values of s can be considered by scaling the data matrix $\boldsymbol{A}$ accordingly. This problem is usually only interesting as long as the optimal value is above zero and hence the optimal solution lies on the border of the ℓ_1-unit-ball ⧫. Assuming this to be the case the problem can be rewritten by setting $\tilde{\boldsymbol{A}} = [\boldsymbol{A}| - \boldsymbol{A}]$ as:

$$\begin{array}{lll} \min_{\boldsymbol{x}} & \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 & = \quad \min_{\boldsymbol{x}} \quad \frac{1}{2} \left\| \tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b} \right\|^2 \\ \text{s.t.} & x \in ⧫ & \text{s.t.} \qquad x \in \Delta \end{array} \tag{3.32}$$

Jaggi (2013)

Here Theorem 2.11 can be applied by writing $f_0(\boldsymbol{x}) = f_0(\boldsymbol{x}) = f(\boldsymbol{A}'\boldsymbol{x} + \boldsymbol{b}') + \boldsymbol{c}^T\boldsymbol{x}$ with $\boldsymbol{A}' = \tilde{\boldsymbol{A}}$, $\boldsymbol{b}' = -\boldsymbol{b}$ and $\boldsymbol{c} = 0$. The corresponding f is strongly convex with parameter 1 and its gradient is Lipschitz continuous with parameter 1. To obtain the upper bound $\delta(\boldsymbol{x})$ Theorem 2.6 is used with $gap_{FW}(\boldsymbol{x}) = \max_i (\tilde{\boldsymbol{A}}\boldsymbol{x} - \tilde{\boldsymbol{a}}_i)^T(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b})$. Theorem 2.11 then becomes:

$$(\tilde{\boldsymbol{a}}_i - \tilde{\boldsymbol{A}}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b}) > \sqrt{\max_i (\tilde{\boldsymbol{A}}\boldsymbol{x} - \tilde{\boldsymbol{a}}_i)^T(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b})} \left\| \tilde{\boldsymbol{a}}_i - \tilde{\boldsymbol{A}}\boldsymbol{x} \right\| \Rightarrow x_i^* = 0 \tag{3.33}$$

Where $\tilde{\boldsymbol{a}}_i$ are the columns of $\tilde{\boldsymbol{A}}$. Since $\tilde{\boldsymbol{A}}$ contains the column vectors of $\boldsymbol{A}$ and $-\boldsymbol{A}$ a point $\boldsymbol{a}_i$ can be dropped from the original problem if the following two inequalities hold:

$$(\boldsymbol{a}_i - \tilde{\boldsymbol{A}}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b}) > \sqrt{\max_i (\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{a}_i)^T(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b})}\left\|\boldsymbol{a}_i - \tilde{\boldsymbol{A}}\boldsymbol{x}\right\| \tag{3.34}$$

$$(-\boldsymbol{a}_i - \tilde{\boldsymbol{A}}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b}) > \sqrt{\max_i (\tilde{\boldsymbol{A}}\boldsymbol{x} + \boldsymbol{a}_i)^T(\tilde{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b})}\left\|\boldsymbol{a}_i + \tilde{\boldsymbol{A}}\boldsymbol{x}\right\| \tag{3.35}$$

Observation 3.7. *As an example an approximate solution with $\tilde{\boldsymbol{A}}\boldsymbol{x} = 0$ can be tried, which gives:*

$$\pm\boldsymbol{a}_i^T\boldsymbol{b} > \sqrt{\max_i \pm\boldsymbol{a}_i^T\boldsymbol{b}}\,\|\boldsymbol{a}_i\| \Rightarrow x_i^* = 0 \tag{3.36}$$

Obviously this rule dismisses at most either of the points $\boldsymbol{a}_i$ and $-\boldsymbol{a}_i$, which would only give the sign of the corresponding variable in the original problem. It seems that this approximate solution is not of much help because it cannot discard any data point completely.

3.4 ℓ_1-regularized Optimization Problems

With the penalized version of the LASSO one example of problems with ℓ_1-regularization was mentioned in the previous section. Here, ℓ_1-regularized problems are discussed in general. Consider an optimization problem of the form:

$$\min_{\boldsymbol{x}\in\mathbb{R}^p} \quad f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x} + \lambda\|\boldsymbol{x}\|_1 \tag{3.37}$$

There are no constraints in this problem. That arises a problem, since as described in Chapter 2.1 the constraints are utilized to get conditions whether data points, in that case the columns of $\boldsymbol{A}$, influence the optimal solution. This issue is overcome by considering a dual formulation of the problem, which has constraints.

Lemma 3.8. *Assume that f is strongly convex, then a dual formulation of the problem given in (3.37) is:*

$$\begin{aligned} &\max_{\boldsymbol{u}} \inf_{\boldsymbol{z}} \quad f(\boldsymbol{z}) + \boldsymbol{u}^T(\boldsymbol{b} - \boldsymbol{z}) \\ &s.t. \qquad \left|\frac{\boldsymbol{a}_i^T\boldsymbol{u} + c_i}{\lambda}\right| \leq 1 \qquad \forall i \in \{1:p\} \end{aligned} \tag{3.38}$$

Proof. The proof follows the idea Wang et al. (2012) used in the appendix of their paper to formulate a dual of the penalized version of the LASSO. To be able to construct the Lagrangian dual restrictions are required. By substituting $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ the problem reads:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^p} \quad & f(\boldsymbol{z}) + \boldsymbol{c}^T\boldsymbol{x} + \lambda\|\boldsymbol{x}\|_1 \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{z} \end{aligned} \tag{3.39}$$

By Definition 1.11 the dual problem is then given by:

$$\begin{aligned} \max_{\boldsymbol{u}} \quad \inf_{\boldsymbol{x},\boldsymbol{z}} \quad & f(\boldsymbol{z}) + \boldsymbol{c}^T\boldsymbol{x} + \lambda\|\boldsymbol{x}\|_1 + \boldsymbol{u}^T(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} - \boldsymbol{z}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{z} \end{aligned} \tag{3.40}$$

At this point, the infimum with respect to $\boldsymbol{x}$ is evaluated. It is attained at $\boldsymbol{x}'$ iff 0 is an element of the subdifferential $\{\boldsymbol{c} + \lambda \boldsymbol{v} + \boldsymbol{A}^T \boldsymbol{u} | \, \|\boldsymbol{v}\|_\infty \leq 1 \ \wedge \ \boldsymbol{v}^T \boldsymbol{x}' = \|\boldsymbol{x}'\|_1\}$, i.e.:

$$\begin{aligned} & 0 = \boldsymbol{c} + \lambda \boldsymbol{v} + \boldsymbol{A}^T \boldsymbol{u} \\ \text{s.t.} \quad & \|\boldsymbol{v}\|_\infty \leq 1 \\ & \boldsymbol{v}^T \boldsymbol{x}' = \|\boldsymbol{x}'\|_1 \end{aligned} \tag{3.41}$$

Using the equation to diminish v, this gives the following conditions on $\boldsymbol{x}'$:

$$\|\boldsymbol{x}'\|_1 = \boldsymbol{v}^T \boldsymbol{x}' = \frac{-\boldsymbol{u}^T \boldsymbol{A} - \boldsymbol{c}^T}{\lambda} \boldsymbol{x}' \tag{3.42}$$

$$\|\boldsymbol{v}\|_\infty = \left\| \frac{\boldsymbol{A}^T \boldsymbol{u} + \boldsymbol{c}}{\lambda} \right\|_\infty \leq 1 \tag{3.43}$$

From this it follows that $\inf_{\boldsymbol{x}} \ \boldsymbol{c}^T \boldsymbol{x} + \lambda \|\boldsymbol{x}\|_1 + \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{x} = \boldsymbol{c}^T \boldsymbol{x}' + \lambda \|\boldsymbol{x}'\|_1 + \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{x}' = 0$. The dual formulation reads now:

$$\begin{aligned} \max_{\boldsymbol{u}} \inf_{\boldsymbol{z}} \quad & f(\boldsymbol{z}) - \boldsymbol{u}^T \boldsymbol{z} + \boldsymbol{u}^T \boldsymbol{b} \\ \text{s.t.} \quad & \left\| \frac{\boldsymbol{A}^T \boldsymbol{u} + \boldsymbol{c}}{\lambda} \right\|_\infty \leq 1 \end{aligned} \tag{3.44}$$

Which gives the result. □

Considering the dual formulation (3.38) the j-th constricting inequality can be dropped or equivalently the data point $\boldsymbol{a}_j$ is non-influential if the inequality is strict at all optimal solutions:

$$\left| \frac{\boldsymbol{a}_j^T \boldsymbol{u}^* + c_j}{\lambda} \right| < 1 \Rightarrow \boldsymbol{a}_j \text{ is non-influential} \tag{3.45}$$

To get a screening rule Theorem 2.13 can be used. This requires an upper bound $\delta(\boldsymbol{u})$ on $\|\boldsymbol{u} - \boldsymbol{u}^*\|$, which can be obtained by either Lemma 2.4 or Theorem 2.6 assuming strong convexity with parameter μ for the dual objective as a function of $\boldsymbol{u}$. The latter requires to compute the gap function by solving a linear program. The screening rule then reads:

$$\left| \frac{\boldsymbol{a}_i^T \boldsymbol{u} + c_i}{\lambda} \right| < 1 - \frac{1}{\lambda} \|\boldsymbol{a}_i\| \, \delta(\boldsymbol{u}) \Rightarrow \boldsymbol{a}_i \text{ is non-influential} \tag{3.46}$$

Remark. Note that for this type of problem the screening uses a dual feasible point.

In the case of using the duality gap as bound δ a similar result was obtained by Ndiaye et al. (2015). Instead of Lagrange duality they used Fenchel duality to obtain the dual formulation of the problem. The only difference lies in the form of functions allowed. They additionally allow to optimize over matrices, which corresponds to applications like the group-LASSO. On the other hand they limit themselves to objectives of the form $\sum_1^m f_i(\boldsymbol{a}_i^T \boldsymbol{x})$.

3.4.1 Penalized Variant of the LASSO

In the case of the penalized version of the LASSO the primal has the form (3.5), this needs to be brought to the form (3.37). Primed variables are the ones occurring in the latter. Which gives $\boldsymbol{A}' = \boldsymbol{A}$, $\boldsymbol{b}' = -\boldsymbol{b}$, $\boldsymbol{c}' = 0$ and $f(\boldsymbol{A}'\boldsymbol{x}' + \boldsymbol{b}') = \frac{1}{2}\|\boldsymbol{A}'\boldsymbol{x}' + \boldsymbol{b}'\|^2$. Now the dual formulation after Lemma 3.38 is:

$$\begin{aligned} \max_{\boldsymbol{u}} \inf_{\boldsymbol{z}} \quad & \frac{1}{2}\|\boldsymbol{z}\|^2 + \boldsymbol{u}^T(\boldsymbol{b} - \boldsymbol{z}) \\ \text{s.t.} \quad & \left|\frac{\boldsymbol{a}_i^T\boldsymbol{u}}{\lambda}\right| \leq 1 \qquad \forall i \in \{1:p\} \end{aligned} \tag{3.47}$$

The infimum with respect to $\boldsymbol{z}$ is attained at $\boldsymbol{z}'$ iff $\boldsymbol{u}' = \boldsymbol{z}$. Hence the dual problem becomes:

$$\begin{aligned} \max_{\boldsymbol{u}} \quad & \frac{1}{2}\|\boldsymbol{u}\|^2 + \boldsymbol{u}^T(\boldsymbol{b} - \boldsymbol{u}) \\ \text{s.t.} \quad & \left|\frac{\boldsymbol{a}_i^T\boldsymbol{u}}{\lambda}\right| \leq 1 \qquad \forall i \in \{1:p\} \end{aligned} \tag{3.48}$$

The negative dual objective as a function of $\boldsymbol{u}$ is strongly convex with parameter 1. Using Lemma 2.4 on the dual problem to get an upper bound $\delta(\boldsymbol{u})$ and the duality gap as suboptimality certificate the screening rule (3.46) reads:

$$\left|\frac{\boldsymbol{a}_i^T\boldsymbol{u}}{\lambda}\right| < 1 - \|\boldsymbol{a}_i\|\sqrt{\frac{2}{\lambda^2} gap_D(\boldsymbol{x}, \boldsymbol{u})} \Rightarrow \boldsymbol{a}_i \text{ is non-influential} \Rightarrow \boldsymbol{x}_i^* = 0 \tag{3.49}$$

The second implication occurs because discarding a data point from the problem is in this case the same as setting the corresponding primal variable to zero. This is the exact rule Olivier Fercoq and Salmon (2015) obtained as GAP SAFE sphere rule.
It is also possible to use Theorem 2.6 to get an upper bound $\delta(\boldsymbol{u})$. To do so it would be necessary to compute the gap function of the dual problem, which means that the following linear program has to be solved:

$$\begin{aligned} \max_{\boldsymbol{y}} \quad & -\boldsymbol{y}^T(\boldsymbol{u} - \frac{\boldsymbol{b}}{\lambda}) \\ \text{s.t.} \quad & \left|\boldsymbol{a}_i^T\boldsymbol{y}\right| \leq 1 \qquad \forall i \in \{1:p\} \end{aligned} \tag{3.50}$$

As for the SVM with hinge loss and no bias it is not clear how using the gap function as upper bound $\delta(\boldsymbol{x})$ compares to the other presented result.

3.4.2 ℓ_1-regularized Logistic Regression

Logistic regression is a classification method. There are m predictor variables and two labels, 1 and -1. The idea is to construct a model predicting the probability for a new point $\boldsymbol{p}$ to have label -1. A simple linear model $P(\boldsymbol{p} \text{ has label -1}) = \boldsymbol{p}^T\boldsymbol{x}$ takes values in $\mathbb{R}$, a probability should be in the $[0, 1]$ interval. To still be able to use a linear model the transformation $x \to \frac{\exp(x)}{\exp(x)+1}$ is used, which maps $\mathbb{R}$ to the $[0, 1]$ interval. The model reads:

$$P(\boldsymbol{p} \text{ has label -1}) = \frac{\exp(\boldsymbol{p}^T\boldsymbol{x})}{\exp(\boldsymbol{p}^T\boldsymbol{x}) + 1}. \tag{3.51}$$

The task is now to find the parameter $\boldsymbol{x}$ returning the "best" fit to given data. Assume there are n data points $\boldsymbol{p}_1$ to $\boldsymbol{p}_n$, with the corresponding labels stored in the vector $\boldsymbol{y}$. Denote the matrix containing the vectors $y_1\boldsymbol{p}_1^T$ to $y_n\boldsymbol{p}_n^T$ as rows by $\boldsymbol{A}$. The columns of $\boldsymbol{A}$ are $\boldsymbol{a}_1$ to $\boldsymbol{a}_m$, each of them represents the values for one predictor variable. For that problem screening corresponds to excluding predictor variables. The "best" fit is now defined as the one that would have the highest probability to get the given data, if it was the true data generating process. This gives the optimization problem:

$$\underset{\boldsymbol{x}}{\operatorname{argmax}} \quad \prod_{y_i=-1} \frac{1}{\exp(-\boldsymbol{p}_i^T\boldsymbol{x})+1} \prod_{y_i=1} \frac{1}{\exp(\boldsymbol{p}_i^T\boldsymbol{x})+1}. \tag{3.52}$$

It is easier to optimize over sums and since log is monotonous, it is possible to solve the following problem instead:

$$\underset{\boldsymbol{x}}{\operatorname{argmax}} \quad \sum_{1}^{n} -\log(\exp(y_i\boldsymbol{p}_i^T\boldsymbol{x})+1). \tag{3.53}$$

Or equivalently:

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \quad \sum_{i=1}^{n} \log(\exp(\boldsymbol{y}_i\boldsymbol{p}_i^T\boldsymbol{x})+1). \tag{3.54}$$

For high dimensional data typically a regularization term is added to avoid over-fitting. This gives the problem this section is about:

$$\min_{\boldsymbol{x}} \quad \lambda\|\boldsymbol{x}\|_1 + \sum_{i=1}^{n} \log(\exp(\boldsymbol{y}_i\boldsymbol{p}_i^T\boldsymbol{x})+1). \tag{3.55}$$

To apply the results of Section 3.4 this needs to be brought to the form (3.37). Which gives $\boldsymbol{b} = \boldsymbol{c} = 0$ and setting $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{w}$ one gets $f(\boldsymbol{w}) = \sum_{i=1}^{n} \log(\exp(w_i)+1)$, since the rows of $\boldsymbol{A}$ were $y_i\boldsymbol{p}_i^T$. The screening rule (3.46) has the same form as before:

$$\left|\frac{\boldsymbol{a}_i^T\boldsymbol{u}}{\lambda}\right| < 1 - \frac{1}{\lambda}\|\boldsymbol{a}_i\|\,\delta(\boldsymbol{u}) \Rightarrow \boldsymbol{a}_i \text{ is non-influential.} \tag{3.56}$$

$\boldsymbol{u}$ is the dual variable, hence the important part here is to write down the dual formulation after Lemma 3.8 and determining its parameter of strong convexity with respect to $\boldsymbol{u}$. $\delta(\boldsymbol{u})$ is an upper bound on $\|\boldsymbol{u}-\boldsymbol{u}^*\|$ which can be obtained by Lemma 2.4 or Theorem 2.6. Using the duality gap as suboptimality certificate we get the same as in the penalized variant of the LASSO:

$$\left|\frac{\boldsymbol{a}_i^T\boldsymbol{u}}{\lambda}\right| < 1 - \|\boldsymbol{a}_i\|\sqrt{\frac{2}{\lambda^2} gap_D(\boldsymbol{x},\boldsymbol{u})} \Rightarrow \boldsymbol{a}_i \text{ is non-influential} \Rightarrow \boldsymbol{x}_i^* = 0 \tag{3.57}$$

The dual formulation after Lemma 3.8 is given by:

$$\begin{aligned} \max_{\boldsymbol{u}\in(0,1)^n} \quad & -\sum_{i=1}^{n}\log(1-u_i) - \sum_{i=1}^{n} u_i \log\frac{u_i}{1-u_i} \\ \text{s.t.} \quad & \left|\frac{\boldsymbol{a}_i^T\boldsymbol{u}}{\lambda}\right| \le 1 \qquad\qquad \forall i \in \{1:p\} \end{aligned} \tag{3.58}$$

The objective is strongly convex with parameter 4, since its second derivative is always bigger than 4 for $\boldsymbol{u}$ in $(0,1)^n$. Running the framework described by Ndiaye et al. (2015) would lead to the same result. As before it would also be possible to use the gap function as upper bound $\delta(\boldsymbol{u})$, which would require to solve a linear problem.

Chapter 4

Summary

4.1 Related Work

Screening rules experienced increasing attention in the past couple of years. One of the first papers trying to reduce the problem size by removing non-influential data points addressed the minimum enclosing ball problem. Ahipaşaoğlu and Yildirim (2008) argued that knowing that a point lies in the interior of the minimum enclosing ball, it can be removed from the problem without changing the result. Given an approximate solution they found an threshold for the distance of a point to the current centre by geometric considerations. If the distance is lower than this threshold, the point lies in the interior of the minimum enclosing ball. Källberg and Larsson (2014) were able to find an even higher threshold. They used weak duality to get tighter bounds on the distance between current and optimal centre.
For sparse least squares linear regression, known as the LASSO, screening rules have been an even more active topic. It is a widely used technique, because it automatically does variable selection, i.e. most entries of the optimal solution are zero. After the first screening rule on the penalized version of the LASSO proposed by Ghaoui et al. (2010), there have been a couple of extensions and alterations. The goal there is to identify zero-entries of the optimal solution. They all have in common that they look at a dual formulation of the problem. Then a region $\mathcal{S}$ as small as possible containing the dual optimal solution is constructed. From there it has to be checked for each dimension of the primal solution whether there exists a dual feasible points in that region fulfilling an optimality condition. In Ghaoui et al. (2010) the region $\mathcal{S}$ is constructed from optimality conditions and the knowledge that – for high enough regularization parameter λ – the primal optimal solution is zero. There were several ideas on how to shrink that region. Usually the problem has to be solved for several regularization parameters. Therefore Wang et al. (2012) decided to take the known solution for a higher regularization parameter into account. Rules based on this approach are usually referenced as sequential screening rules. A good overview up to this point is given in Xiang, Wang, and Ramadge (2014). In Wang, Zhou, Liu, Wonka, and Ye (2014) they follow the same strategy to obtain a screening rule for ℓ_1-regularized logistic regression. Another promising way of shrinking the region $\mathcal{S}$ is to use information already calculated in the form of approximate solutions, as done in the minimum enclosing ball cases. This allows for iterative computations to screen in each step of the iteration. Bonnefoy et al. (2014) and Bonnefoy, Emiya, Ralaivola, and Gribonval (2015) introduced this approach. Olivier Fercoq and Salmon (2015) describe how to use the duality gap

at the current approximation to construct $\mathcal{S}$. In Ndiaye et al. (2015) they were able to generalize the result for ℓ_1-regularized objectives with Lipschitz gradient.
A third area where screening rules were discussed are support vector machines, which are large margin linear classifiers. Ogawa et al. (2014) considered the version with hinge loss following a similar approach as Ghaoui et al. (2010). They constructed two balls containing the optimal solution, so it has to be in their intersection. One of them is constructed from an approximate solution, the other one from the known solution for a different parameter. Then they also checked which optimality conditions can be fulfilled by points in the intersection. Wysling (2015) considered the squared loss case and studied a screening rule by pure geometric considerations.
All rules mentioned so far are safe rules in the sense that they only remove non-influential points from a problem. In contrast to that so called "strong" rules were invented by Tibshirani, Bien, Friedman, Hastie, Simon, Taylor, and Tibshirani (2012). They discard more dimensions but are known to sometimes dismiss dimensions which would have contributed to the optimal solution, which makes post-processing necessary.
The variety of problems studied already gives rise to the question if there is a general strategy behind which can be used for all of them. This thesis tackles this question.

4.2 Results

Given a convex optimization problem, a screening rule can be obtained following three steps:

- first, the KKT conditions are written down and complementary slackness is used to get a condition for a data point to be non-influential. For the domains considered in this thesis those conditions are listed in Table 4.1.
- Second, an approximate solution is used to gather information about the optimal points. The properties used to do to are summarized in Table 4.2.
- Third, the condition for a data point to be non-influential is formulated independent of the optimum using the obtained information. This is the screening rule. Table 4.3 shows general screening rules for specific domains.

The obtained screening rules were applied to specific problems. The implications are summarized in Table 4.4.

4.3 Impact of the Results

This thesis provides a general framework to think about screening rules for arbitrary convex problems. The relation between existing work on screening rules for the minimum enclosing ball, support vector machines and the LASSO is shown. Additionally a pure algebraic view on screening rules without geometric considerations was developed. In the case of the minimum enclosing ball, it was possible to get a screening rule which dismisses at least the same and probably more points than the screening rule by Källberg and Larsson (2014). For ℓ_1-regularized problems, including the penalized version of the LASSO and ℓ_1-regularized logistic regression, it was possible to reproduce the results that Ndiaye et al. (2015) obtained, which is the most recent paper on the topic.

Domain	Problem	Condition
Unit Simplex	$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x})$ s.t. $\quad -\boldsymbol{x} \leq 0$ $1 - \mathbf{1}^T\boldsymbol{x} = 0$	$(\boldsymbol{e}_i - \boldsymbol{x}^*)^T \nabla f_0(\boldsymbol{x}^*) > 0 \Rightarrow x_i^* = 0$
Box Constraints	$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x})$ s.t. $\quad 0 \leq \boldsymbol{x} \leq C\mathbf{1}$	$\boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) > 0 \Rightarrow x_i^* = 0$ $\boldsymbol{e}_i^T \nabla f_0(\boldsymbol{x}^*) < 0 \Rightarrow x_i^* = C$
Polytope Constraints	$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x})$ s.t. $\quad \boldsymbol{A}^T\boldsymbol{x} \leq \boldsymbol{b}$	$\boldsymbol{a}_i^T\boldsymbol{x}^* < b_i \Rightarrow \boldsymbol{a}_i$ is non-influential

Table 4.1: Summary of the conditions that an be used to reduce the problem size for the given domains. At this point they still depend on the optimal solution $\boldsymbol{x}^*$ and therefore are no screening rules yet.

Source	Restriction
Strong Convexity	$\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{A}\boldsymbol{x}\|^2 \leq \frac{2}{\mu}(f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b}) - f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}))$
Gap Function	$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*\|^2 \leq \frac{1}{\mu} gap_{FW}(\boldsymbol{x})$
Lipschitz Gradient	$\|\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) - \nabla f(\boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{b})\| \leq L\,\delta(\boldsymbol{x})$

Table 4.2: Summary of the information on the optimal solution $\boldsymbol{x}^*$ used in this thesis to get screening rules.

Domain	Problem	Screening Rule
Unit Simplex	$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x})$ s.t. $\quad -\boldsymbol{x} \leq 0$ $1 - \mathbf{1}^T\boldsymbol{x} = 0$	$(\boldsymbol{e}_i - \boldsymbol{x})^T(\boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}) > L\,\delta(\boldsymbol{x})\,\|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\|$ $\Rightarrow x_i^* = 0$
Box Constraints	$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x})$ s.t. $\quad 0 \leq \boldsymbol{x} \leq C\mathbf{1}$	$\boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i > L\,\|\boldsymbol{a}_i\|\,\delta(\boldsymbol{x}) \Rightarrow x_i^* = 0$ $\boldsymbol{a}_i^T \nabla f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + c_i < -L\,\|\boldsymbol{a}_i\|\,\delta(\boldsymbol{x}) \Rightarrow x_i^* = C$
Polytope Constraints	$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x})$ s.t. $\quad \boldsymbol{A}^T\boldsymbol{x} \leq \boldsymbol{b}$	$\boldsymbol{a}_i^T\boldsymbol{x} < b_i - \|\boldsymbol{a}_i\|\,\delta(\boldsymbol{x}) \Rightarrow \boldsymbol{a}_i$ is non-influential

Table 4.3: Summary of the general screening rules developed in this thesis.

Name	Problem	Screening Rule
Squared Loss SVM	$\min_{\boldsymbol{x}} \quad \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}$ s.t. $\quad -\boldsymbol{x} \leq 0$ $1 - \mathbf{1}^T\boldsymbol{x} = 0$	$(\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x} > \sqrt{gap_{FW}(\boldsymbol{x})}\,\|\boldsymbol{a}_i - \boldsymbol{A}\boldsymbol{x}\|$ $\Rightarrow x_i^* = 0$
Hinge Loss SVM	$\max_{\boldsymbol{x}} \quad \boldsymbol{x}^T\mathbf{1} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}$ s.t. $\quad 0 \leq \boldsymbol{x} \leq C\mathbf{1}$	$\boldsymbol{a}_i^T\boldsymbol{A}\boldsymbol{x} + 1 > \|\boldsymbol{a}_i\|\sqrt{2gap_D(\boldsymbol{x})} \Rightarrow x_i^* = 0$ $\boldsymbol{a}_i^T\boldsymbol{A}\boldsymbol{x} + 1 < -\|\boldsymbol{a}_i\|\sqrt{2gap_D(\boldsymbol{x})} \Rightarrow x_i^* = C$
MEB	$\max_{\boldsymbol{x}} \quad -\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \sum_{j=1}^p \boldsymbol{a}_j^T\boldsymbol{a}_j x_j$ s.t. $\quad -\boldsymbol{x} \leq 0$ $1 - \mathbf{1}^T\boldsymbol{x} = 0$	$\|\boldsymbol{a}_i - \boldsymbol{c}\| < \frac{r^2}{\|\boldsymbol{a}_i - \boldsymbol{c}\|} - 2\sqrt{\frac{R^2 - r^2}{2}}$ $\Rightarrow \boldsymbol{a}_i$lies in interior
LASSO Constrained Variant	$\min_{\boldsymbol{x}} \quad \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ s.t. $\quad \|\boldsymbol{x}\|_1 \leq s$	$(\pm\boldsymbol{a}_i - \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}})^T(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} - \boldsymbol{b}) >$ $\sqrt{\max_i (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} - \boldsymbol{a}_i)^T(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} - \boldsymbol{b})}\left\|\boldsymbol{a}_i - \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}\right\|$ $\Rightarrow \tilde{x}_i^* = 0$
LASSO Penalized Variant	$\min_{\boldsymbol{x}} \quad \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{x}\|_1$	$\left\|\frac{\boldsymbol{a}_i^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})}{\lambda}\right\| < 1 - \|\boldsymbol{a}_i\|\sqrt{\frac{2}{\lambda^2}gap_D(\boldsymbol{x})}$ $\Rightarrow \boldsymbol{x}_i^* = 0$
Logistic Regression	$\min_{\boldsymbol{x}} \quad \lambda\|\boldsymbol{x}\|_1 +$ $\sum_{i=1}^n(-y_i\boldsymbol{p}_i^T\boldsymbol{x} + \log(\exp(\boldsymbol{p}_i^T\boldsymbol{x}) + 1))$	$\left\|\frac{\boldsymbol{a}_i^T\boldsymbol{u}}{\lambda}\right\| < 1 - \|\boldsymbol{a}_i\|\sqrt{\frac{2}{\lambda^2}gap_D(\boldsymbol{x}, \boldsymbol{u})}$ $\Rightarrow \boldsymbol{a}_i$ is non-influential
ℓ_1-regularized Optimization Problems	$\min_{\boldsymbol{x}\in\mathbb{R}^p} \quad f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{c}^T\boldsymbol{x} + \lambda\|\boldsymbol{x}\|_1$	$\left\|\frac{\boldsymbol{a}_i^T\boldsymbol{u} + c_i}{\lambda}\right\| < 1 - \frac{1}{\lambda}\|\boldsymbol{a}_i\|\ \delta(\boldsymbol{x})$ $\Rightarrow \boldsymbol{a}_i$ is non-influential

Table 4.4: Summary of the screening rules on specific problems, derived from the general rules developed in this thesis. For the MEB r is the radius of the ball corresponding to the current approximate solution, i.e. $r = \sqrt{-\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \sum_{j=1}^p \boldsymbol{a}_j^T\boldsymbol{a}_j x_j}$. R is given by $\max_j \|\boldsymbol{c} - \boldsymbol{a}_j\|$ and $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{x}$. For the constrained variant of the LASSO $\tilde{\boldsymbol{A}}$ is the matrix $[\boldsymbol{A}| - \boldsymbol{A}]$. In the logistic regression case and for the ℓ_1-regularized optimization problems, the vector $\boldsymbol{u}$ corresponds to a dual feasible solution.

4.4 Future Work

On the basis of this thesis there are several directions one could continue studying.
By looking at other types of constraints and objective functions it should be possible to get further screening rules. One particular direction would be to look whether it can be useful to determine entries of optimal solutions for problems not depending on some data matrices.
It is also possible to include more information about the optimal solution into the analysis. In this thesis, either the gap function or a suboptimality certificate combined with weak duality are used to bound the distance of optimal solutions to the current approximation. In Section 2.2 further information on the optimal solution not yet used in this thesis is mentioned. For example Olivier Fercoq and Salmon (2015) used already something comparable to Lemma 2.10 for the penalized version of the LASSO.
Another direction would be to run simulations checking the performance of the screening rules given in this thesis. This is particularly interesting for the MEB problem and how it compares to the screening rule in Källberg and Larsson (2014). As well it could be interesting to check the performance of the rules for hinge loss SVM, the penalized version of the LASSO and ℓ_1-regularized problems using the gap function as upper bound for the distance between current approximation and optimal solutions.

Bibliography

Ahipaşaoğlu, S. D. and E. A. Yildirim (2008, November). Identification and elimination of interior points for the minimum enclosing ball problem. *SIAM J. on Optimization 19*(3), 1392–1396.

Bonnefoy, A., V. Emiya, L. Ralaivola, and R. Gribonval (2014, September). A Dynamic Screening Principle for the Lasso. In *European Signal Processing Conference EUSIPCO 2014*, Lisboa, Portugal, pp. 1–5.

Bonnefoy, A., V. Emiya, L. Ralaivola, and R. Gribonval (2015). Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. *IEEE Transactions on Signal Processing*, 20.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press.

Ghaoui, L. E., V. Viallon, and T. Rabbani (2010). Safe feature elimination in sparse supervised learning. *CoRR abs/1009.4219*.

Hearn, D. W. (1982). The gap function of a convex program. *Operations Research Letters 1*(2), 67 – 71.

Jaggi, M. (2013). An equivalence between the lasso and support vector machines. *CoRR abs/1303.1152*.

Källberg, L. and T. Larsson (2014, July). Improved pruning of large data sets for the minimum enclosing ball problem. *Graphical Models 76*(6), 609–619.

Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy (2000). A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Trans. Neural Netw. Learning Syst. 11*(1), 124–136.

Matousek, J. and B. Gärtner (2006). *Understanding and Using Linear Programming (Universitext)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Ndiaye, E., O. Fercoq, A. Gramfort, and J. Salmon (2015). GAP safe screening rules for sparse multi-task and multi-class models. In *NIPS*.

Ogawa, K., Y. Suzuki, S. Suzumura, and I. Takeuchi (2014, January). Safe Sample Screening for Support Vector Machines. *ArXiv e-prints*.

Olivier Fercoq, A. G. and J. Salmon (2015). Mind the duality gap: safer rules for the lasso.

Scholkopf, B. and A. J. Smola (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press.

Tibshirani, R., J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74* (2), 245–266.

Tsang, I. W., J. T. Kwok, P. ming Cheung, and N. Cristianini (2005). Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research 6*, 363–392.

Wang, J., P. Wonka, and J. Ye (2012). Lasso screening rules via dual polytope projection. *CoRR abs/1211.3966*.

Wang, J., J. Zhou, J. Liu, P. Wonka, and J. Ye (2014). A safe screening rule for sparse logistic regression. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 1053–1061. Curran Associates, Inc.

Wysling, J. (2015). Screening rules for the support vector machine and the minimum enclosing ball. *Bachelor thesis ETH Zürich*.

Xiang, Z. J., Y. Wang, and P. J. Ramadge (2014). Screening tests for lasso problems. *CoRR abs/1405.4897*.

Zhao, Z., J. Liu, and J. Cox (2014). Safe and efficient screening for sparse support vector machine. In *the 20th ACM SIGKDD international conference*, New York, USA, pp. 542–551. ACM Press.

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

Screening Rules for Convex Problems

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Olbrich	Jakob

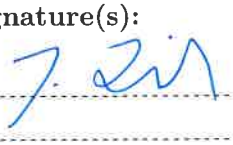
With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics*.

Place, date:

Zürich, 10.09.2015

Signature(s):



For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.