

An Equivalence between the Lasso and Support Vector Machines

Martin Jaggi*

ETH Zürich, Switzerland

jaggi@inf.ethz.ch

Abstract. We investigate the relation of two fundamental tools in machine learning and signal processing, that is the support vector machine (SVM) for classification, and the Lasso technique used in regression. We show that the resulting optimization problems are equivalent, in the following sense. Given any instance of an ℓ_2 -loss soft-margin (or hard-margin) SVM, we construct a Lasso instance having the same optimal solutions, and vice versa.

As a consequence, many existing optimization algorithms for both SVMs and Lasso can also be applied to the respective other problem instances. Also, the equivalence allows for many known theoretical insights for SVM and Lasso to be translated between the two settings. One such implication gives a simple kernelized version of the Lasso, analogous to the kernels used in the SVM setting. Another consequence is that the sparsity of a Lasso solution is equal to the number of support vectors for the corresponding SVM instance, and that one can use screening rules to prune the set of support vectors. Furthermore, we can relate sublinear time algorithms for the two problems, and give a new such algorithm variant for the Lasso. We also study the regularization paths for both methods.

1. Introduction

Linear classifiers and kernel methods, and in particular the support vector machine (SVM) [CV95], are among the most popular standard tools for classification. On the other hand, ℓ_1 -regularized least squares regression, i.e. the Lasso estimator [Tib96], is one of the most widely used tools for robust regression and sparse estimation.

Along with the many successful practical applications of SVMs and the Lasso in various fields, there is a vast amount of existing literature¹ on the two methods themselves, considering both theory and also algorithms for each of the two. However, the two research topics developed largely independently and were not much set into context with each other so far.

In this chapter, we attempt to better relate the two problems, with two main goals in mind. On the algorithmic side, we show that many of the existing algorithms for each of the two problems can be set into comparison, and can be directly applied to the other respective problem. As a particular example of this idea, we can apply the recent sublinear time SVM algorithm by [CHW10] also to any Lasso problem, resulting in a new alternative sublinear time algorithm variant for the Lasso.

On the other hand, we can relate and transfer existing theoretical results between the literature for SVMs and the Lasso. In this spirit, a first example is the idea of the kernel trick.

*A shorter version of this article has appeared in [Jag13a].

¹As of January 2014, Google Scholar returned 300'000 publications containing the term "Support Vector Machine", and over 20'000 for Lasso regression.

Originally employed for SVMs, this powerful concept has allowed for lifting most insights from the linear classifier setting also to the more general setting of non-linear classifiers, by using an implicit higher dimensional space. Here, by using our equivalence, we propose a simple kernelized variant of the Lasso, being equivalent to the well-researched use of kernels in the SVM setting.

As another example, we can also transfer some insights in the other direction, from the Lasso to SVMs. The important datapoints, i.e. those that define the solution to the SVM classifier, are called the support vectors. Having a small set of support vectors is crucial for the practical performance of SVMs. Using our equivalence, we see that the set of support vectors for a given SVM instance is exactly the same as the sparsity pattern of the corresponding Lasso solution.

Screening rules are a way of pre-processing the input data for a Lasso problem, in order to identify inactive variables. We show that screening rules can also be applied to SVMs, in order to eliminate potential support vectors beforehand, and thereby speeding up the training process by reducing the problem size.

Finally, we study the complexity of the solution paths of Lasso and SVMs, as the regularization parameter changes. We discuss path algorithms that apply to both problems, and also translate a result on the Lasso path complexity to show that a single SVM instance, depending on the scaling of the data, can have very different patterns of support vectors.

Support Vector Machines. In this chapter, we focus on large margin linear classifiers, and more precisely on those SVM variants whose dual optimization problem is of the form

$$\min_{x \in \Delta} \|Ax\|^2 . \quad (1)$$

Here the matrix $A \in \mathbb{R}^{d \times n}$ contains all n datapoints as its columns, and Δ is the unit simplex in \mathbb{R}^n , i.e. the set of probability vectors, that is the non-negative vectors whose entries sum up to one. The formulation (1) includes the commonly used soft-margin SVM with ℓ_2 -loss. It includes both the one or two classes variants, both with or without using a kernel, and both using a (regularized) offset term (allowing hyperplanes not passing through the origin) or no offset. We will explain these variants and the large margin interpretation of this optimization problem in more detail in Section 2.

Lasso. On the other hand, the Lasso [Tib96], is given by the quadratic program

$$\min_{x \in \diamond} \|Ax - b\|^2 . \quad (2)$$

It is also known as the constrained variant of ℓ_1 -regularized least squares regression. Here the right hand side b is a fixed vector $b \in \mathbb{R}^d$, and \diamond is the ℓ_1 -unit-ball in \mathbb{R}^n . Note that the 1-norm $\|\cdot\|_1$ is the sum of the absolute values of the entries of a vector. Sometimes in practice, one would like to have the constraint $\|x\|_1 \leq r$ for some value $r > 0$, instead of the simple unit-norm case $\|x\|_1 \leq 1$. However, in that case it is enough to simply re-scale the input matrix A by a factor of r , in order to obtain our above formulation (2) for any general Lasso problem.

In applications of the Lasso, it is important to distinguish two alternative interpretations of the data matrix A , which defines the problem instance (2): on one hand, in the setting of *sparse regression*, the matrix A is usually called the dictionary matrix, with its columns $A_{\cdot j}$ being the dictionary elements, and the goal being to approximate the single vector b by a combination of few dictionary vectors. On the other hand if the Lasso problem is interpreted as *feature-selection*, then each row $A_{i \cdot}$ of the matrix A is interpreted as an input vector, and for each of those, the Lasso is approximating the response b_i to input row $A_{i \cdot}$. The book [BvdG11] gives a recent overview of Lasso-type methods and their broad applications. While the penalized variant of the Lasso (meaning that the term $\|x\|_1$ is added to the objective instead of imposed as a constraint) is also popular in applications, here we focus on the original constrained variant.

The Equivalence. We will prove that the two problems (1) and (2) are indeed equivalent, in the following sense. For any Lasso instance given by (A, b) , we construct an equivalent (hard-margin) SVM instance, having the same optimal solution. This will be a simple reduction preserving all objective values. On the other hand, the task of finding an equivalent Lasso instance for a given SVM appears to be a harder problem. Here we show that there always exists such an equivalent Lasso instance, and furthermore, if we are given a weakly-separating vector for the SVM (formal definition to follow soon below), then we can explicitly construct the equivalent Lasso instance. This reduction also applies to the ℓ_2 -loss soft-margin SVM, where we show that a weakly-separating vector is trivial to obtain, making the reduction efficient. The reduction does *not* require that the SVM input data is separable.

Our shown equivalence is on the level of the (SVM or Lasso) *training* formulations, we’re not making any claims on the performance of the two different kind of methods on unseen *test* data.

On the way to this goal, we will also explain the relation to the “non-negative” Lasso variant when the variable vector x is required to lie in the simplex, i.e.

$$\min_{x \in \Delta} \|Ax - b\|^2 . \quad (3)$$

It turns out the equivalence of the optimization problems (1) and (3) is straightforward to see. Our main contribution is to explain the relation of these two optimization problems to the original Lasso problem (2), and to study some of the implications of the equivalence.

Related Work. The early work of [Gir98] has already significantly deepened the joint understanding of kernel methods and the sparse coding setting of the Lasso. Despite its title, [Gir98] is not addressing SVM classifiers, but in fact the ε -insensitive loss variant of support vector regression (SVR), which the author proves to be equivalent to a Lasso problem where ε then becomes the ℓ_1 -regularization. Unfortunately, this reduction does not apply anymore when $\varepsilon = 0$, which is the case of interest for standard hinge-loss SVR, and also for SVMs in the classification setting, which are the focus of our work here.

Another reduction has been known if the SVR insensitivity parameter ε is chosen close enough to one. In that case, [PRE98] has shown that the SVR problem can be reduced to SVM classification with standard hinge-loss. Unfortunately, this reduction does not apply to Lasso regression.

In a different line of research, [LYX05] have studied the relation of a dual variant of the Lasso to the primal of the so called *potential SVM* originally proposed by [HO04, HO06], which is not a classifier but a specialized method of feature selection.

In the application paper [GC05] in the area of computational biology, the authors already suggested to make use of the “easier” direction of our reduction, reducing the Lasso to a very particular SVM instance. The idea is to employ the standard trick of using two non-negative vectors to represent a point in the ℓ_1 -ball [BS83, CDS98]. Alternatively, this can also be interpreted as considering an SVM defined by all Lasso dictionary vectors together with their negatives ($2n$ many points). We formalize this interpretation more precisely in Section 3.2. The work of [GC05] does not address the SVM regularization parameter.

Notation. The following three sets of points will be central for our investigations. We denote the unit simplex, the filled simplex as well as the ℓ_1 -unit-ball in \mathbb{R}^n as follows.

$$\begin{aligned} \triangle & := \{x \in \mathbb{R}^n \mid x \geq \mathbf{0}, \sum_i x_i = 1\} , \\ \blacktriangle & := \{x \in \mathbb{R}^n \mid x \geq \mathbf{0}, \sum_i x_i \leq 1\} , \\ \blacklozenge & := \{x \in \mathbb{R}^n \mid \|x\|_1 \leq 1\} . \end{aligned}$$

The 1-norm $\|\cdot\|_1$ is the sum of the absolute values of the entries of a vector. The standard Euclidean norm is written as $\|\cdot\|$.

For a given matrix $A \in \mathbb{R}^{d \times n}$, we write $A_i \in \mathbb{R}^d$, $i \in [1..n]$ for its columns. We use the notation $AS := \{Ax \mid x \in S\}$ for subsets $S \subseteq \mathbb{R}^d$ and matrices A . The convex hull of a set S is written as $\text{conv}(S)$. By $\mathbf{0}$ and $\mathbf{1}$ we denote the all-zero and all-ones vectors in \mathbb{R}^n , and \mathbf{I}_n is the $n \times n$ identity matrix. We write $(A|B)$ for the horizontal concatenation of two matrices A, B .

2. Linear Classifiers and Support Vector Machines

Linear classifiers have become the standard workhorse for many machine learning problems. Suppose we are given n datapoints $X_i \in \mathbb{R}^d$, together with their binary labels $y_i \in \{\pm 1\}$, for $i \in [1..n]$.

As we illustrate in Figure 1, a linear classifier is a hyperplane that partitions the space \mathbb{R}^d into two parts, such that hopefully each point with a positive label will be on one side of the plane, and the points of negative label will be on the other site. Writing the classifier as the normal vector w of that hyperplane, we can formally write this separation as $X_i^T w > 0$ for those points i with $y_i = +1$, and $X_i^T w < 0$ if $y_i = -1$, assuming we consider hyperplanes that pass through the origin.

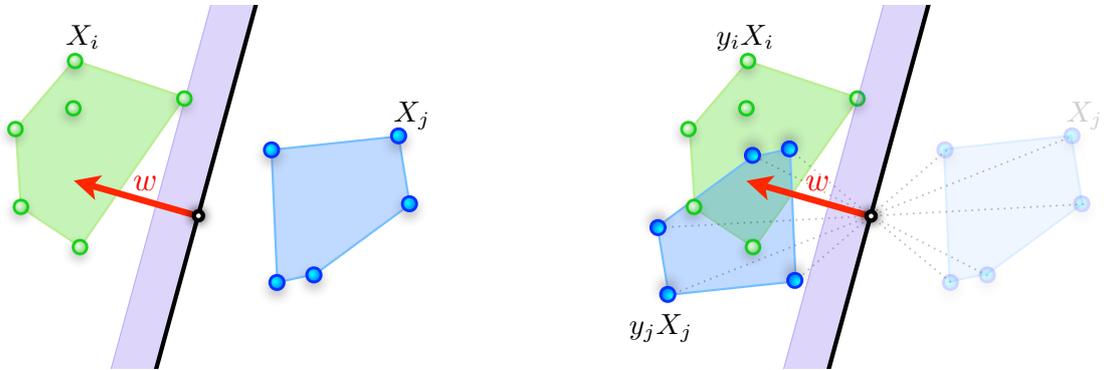


Figure 1: A linear classifier: illustration of the separation of two point classes, for a normal vector w . Here we highlight two points i, j of different labels, with $y_i = +1$ and $y_j = -1$.

2.1. Support Vector Machines

The popular concept of *support vector machines* (SVM) is precisely this linear classifier idea as we have introduced above, with one important addition: instead of being satisfied with any arbitrary one among all separating hyperplanes, we want to find the hyperplane which separates the two point classes by the best possible margin.

The *margin* is defined as the smallest distance to the hyperplane among all the datapoints. Maximizing this margin over all possible linear classifiers w can be simply formalized as the following optimization problem:

$$\max_{w \in \mathbb{R}^d} \min_i y_i X_i^T \frac{w}{\|w\|},$$

i.e. maximizing the smallest projection onto the direction w , among all datapoints.

The SVM optimization problem (1) that we defined in the introduction is very closely related to this form of optimization problem, as follows. If we take the Lagrange dual of problem (1), we exactly obtain a problem of this margin maximization type, namely

$$\max_{x \in \Delta} \min_i A_i^T \frac{Ax}{\|Ax\|}. \quad (4)$$

Here we think of the columns of A being the SVM datapoints with their signs $A_i = y_i X_i$. For an overview of Lagrange duality, we refer the reader to [BV04, Section 5], or see for example [GJ09, Appendix A] for the SVM case here. An alternative and slightly easier way to obtain

this dual problem is to compute the simple “linearization” dual function as in [Hea82, Jag13b], which avoids the notion of any dual variables. When starting from the non-squared version of the SVM problem (1), this also gives the same formulation (4).

No matter if we use problem formulation (1) or (4), any feasible weight vector x readily gives us a candidate classifier $w = Ax$, represented as a convex combination of the datapoints, because the weight vectors $x \in \Delta \subset \mathbb{R}^n$ lie in the simplex. The datapoints corresponding to non-zero entries in x are called the *support vectors*.

Several other SVM variants of practical interest also have the property that their dual optimization problem is of the form (1), as we will discuss in the next Subsection 3.2.

Kernelization. A crucial and widely used observation is that both optimization problems (1) and (4) are formulated purely in terms of the inner products between pairs of datapoints $A_i := y_i X_i$, meaning that they can directly be optimized in the *kernel* case [CV95], provided that we only have access to the entries of the matrix $A^T A \in \mathbb{R}^{n \times n}$, but not the explicit features $A \in \mathbb{R}^{d \times n}$. The matrix $K = A^T A$ is called the *kernel matrix* in the literature. Using this notation, the SVM dual optimization problem (1) becomes

$$\min_{x \in \Delta} x^T K x .$$

Separation, and Approximate Solutions. It is natural to measure the quality of an approximate solution x to the SVM problem as the attained margin, which is precisely the attained value in the above problem (4).

Definition 1. A vector $w \in \mathbb{R}^d$ is called σ -weakly-separating for the SVM instance (1) or (4) respectively, for a parameter $\sigma \geq 0$, if it holds that

$$A_i^T \frac{w}{\|w\|} \geq \sigma \quad \forall i ,$$

meaning that w attains a margin of separation of σ .

The margin value σ in this definition, or also the objective in (4), can be interpreted as a useful certificate for the attained optimization quality as follows. If we take some x (the separating vector now being interpreted as $w = Ax$), then the *duality gap* is given by the difference of the margin value from the corresponding objective $\|Ax\|$ in problem (1). This gap function is a certificate for the approximation quality (since the unknown optimum must lie within the gap), which makes it a very useful stopping criterion for optimizers, see e.g. [Hea82, GJ09, CHW10, Jag13b].

The simple perceptron algorithm [Ros58] is known to return a σ -weakly-separating solution to the SVM after $O(1/\varepsilon^2)$ iterations, for $\varepsilon := \sigma^* - \sigma$ being the additive error, if σ^* is the optimal solution to (1) and (4).

2.2. Soft-Margin SVMs

For the successful practical application of SVMs, the soft-margin concept of tolerating outliers is of central importance. Here we recall that also the soft-margin SVM variants using ℓ_2 -loss, with regularized offset or no offset, both in the one-class and the two-class case, can be formulated in the form (1). This fact is known in the SVM literature [SS02, KSBM00, TKC05], and can be formalized as follows.

The two-class soft-margin SVM with squared loss (ℓ_2 -loss), without offset term, is given by the primal optimization problem

$$\begin{aligned} \min_{\substack{\bar{w} \in \mathbb{R}^d, \rho \in \mathbb{R}, \\ \xi \in \mathbb{R}^n}} \quad & \frac{1}{2} \|\bar{w}\|^2 - \rho + \frac{C}{2} \sum_i \xi_i^2 \\ \text{s.t.} \quad & y_i \cdot \bar{w}^T X_i \geq \rho - \xi_i \quad \forall i \in [1..n] . \end{aligned} \tag{5}$$

For each datapoint, we have introduced a slack-variable ξ_i which is penalized in the objective function in the case that the point does violate the margin. Here $C > 0$ is the regularization parameter, steering the tradeoff between large margin and punishing outliers. In the end, $\rho/\|\bar{w}\|$ will be the attained margin of separation. Note that in the classical SVM formulation, the margin parameter ρ is usually fixed to one instead, while ρ is explicitly used in the equivalent ν -SVM formulation known in the literature, see e.g. [SS02]. The equivalence of the soft-margin SVM dual problem to the optimization problem (1) is stated in the following Lemma:

Lemma 2. *The dual of the soft-margin SVM (5) is an instance of the classifier formulation (1), that is $\min_{x \in \Delta} \|Ax\|^2$, with*

$$A := \begin{pmatrix} Z \\ \frac{1}{\sqrt{C}} \mathbf{I}_n \end{pmatrix} \in \mathbb{R}^{(d+n) \times n}$$

where the data matrix $Z \in \mathbb{R}^{d \times n}$ consists of the n columns $Z_i := y_i X_i$.

Proof. Given in Appendix A for completeness, using standard Lagrange duality. \square

Not all SVM variants have our desired structure of the dual problem. For example, the hinge-loss (or ℓ_1 -loss) SVM refers to the setting where the outliers are penalized according to their margin-violation ξ_i , not the squared values ξ_i^2 as we use here. Changing the loss function in the primal optimization problem also affects the dual problem, so that the dual of the ℓ_1 -loss SVM is *not* of the form (1). However in practice, it is known that both SVM variants with ℓ_1 - or ℓ_2 -loss do perform similarly well for most applications [LM01, CHL08].

Obtaining a Weakly-Separating Vector for the ℓ_2 -loss Soft-Margin SVM. By the above lemma, we observe that a weakly-separating vector is trivial to obtain for the ℓ_2 -loss SVM. This holds without any assumptions on the original input data (X_i, y_i) . We set $w := \begin{pmatrix} \mathbf{0} \\ \frac{1}{\sqrt{n}} \mathbf{1} \end{pmatrix} \in \mathbb{R}^{d+n}$ to the all-one vector only on the second block of coordinates, rescaled to unit length. Clearly, this direction w attains a separation margin of

$$A_i^T \frac{w}{\|w\|} = \begin{pmatrix} y_i X_i \\ \frac{1}{\sqrt{C}} \mathbf{e}_i \end{pmatrix}^T \begin{pmatrix} \mathbf{0} \\ \frac{1}{\sqrt{n}} \mathbf{1} \end{pmatrix} = \frac{1}{\sqrt{nC}} > 0$$

for all points i in Definition 1.

Incorporating an Offset Term. Our above SVM formulation also allows the use of an *offset* (or bias) variable $b \in \mathbb{R}$ to obtain a classifier that does not necessarily pass through the origin. Formally, the separation constraints then become $y_i \cdot (w^T X_i + b) \geq \rho - \xi_i \quad \forall i \in [1..n]$. There is a well-known trick to efficiently emulate such an offset parameter while still using our formulation (5), by simply increasing the dimensionality of X_i and w by one, and adding a fixed value of one as the last coordinate to each of the datapoints X_i , see e.g. [KSBM00, TKC05]. As a side-effect, the offset b^2 is then also regularized in the new term $\|w\|^2$. Nevertheless, if desired, the effect of this additional regularization can be made arbitrary weak by re-scaling the fixed additional feature value from one to a larger value.

One-Class SVMs. All mentioned properties in this section also hold for the case of *one-class SVMs*, by setting all labels y_i to one, resulting in the same form of optimization problems (1) and (4). One class SVMs are popular for example for anomaly or novelty detection applications.

3. The Equivalence

3.1. Warm-Up: Equivalence between SVM and Non-Negative Lasso

Before we investigate the “real” Lasso problem (2) in the next two subsections, we will warm-up by considering the non-negative variant (3). It is a simple observation that the non-negative Lasso (3) is directly equivalent to the dual SVM problem (1) by a translation:

Equivalence by Translation. Given a non-negative Lasso instance (3), we can translate each column vector of the matrix A by the vector $-b$. Doing so, we precisely obtain an SVM instance (1), with the data matrix being

$$\tilde{A} := A - b\mathbf{1}^T \in \mathbb{R}^{d \times n} .$$

Here we have crucially used the simplex domain, ensuring that $b\mathbf{1}^T x = b$ for any $x \in \Delta$. To summarize, for those two optimization problems, the described translation precisely preserves all the objective values of all feasible points for both problems (3) and (1), that is for all $x \in \Delta$. This is why we say that the problems are equivalent.

The reduction in the other direction — i.e. reducing an SVM instance (1) to a non-negative Lasso instance (3) — is trivial by choosing $b := \mathbf{0}$.

Now to relate the SVM to the “real” Lasso, the same translation idea is of crucial importance. We explain the two reductions in the following subsections.

3.2. (Lasso \preceq SVM): Given a Lasso Instance, Constructing an Equivalent SVM Instance

(This reduction is significantly easier than the other direction.)

Parameterizing the ℓ_1 -Ball as a Convex Hull. One of the main properties of polytopes — if not the main one — is that every polytope can be represented as the convex hull of its vertices [Zie95]. When expressing an arbitrary point in the polytope as a convex combination of some vertices, this leads to the standard concept of *barycentric* coordinates.

In order to represent the ℓ_1 -ball \blacklozenge by a simplex Δ , this becomes particularly simple. The ℓ_1 -ball \blacklozenge is the convex hull of its $2n$ vertices, which are $\{\pm e_i \mid i \in [1..n]\}$, illustrating why \blacklozenge is also called the cross-polytope.

The barycentric representation of the ℓ_1 -ball therefore amounts to the simple trick of using two non-negative variables to represent each real variable, which is standard for example when writing linear programs in standard form. For ℓ_1 problems such as the Lasso, this representation was known very early [BS83, CDS98]. Formally, any n -vector $x_\diamond \in \blacklozenge$ can be written as

$$x_\diamond = (\mathbf{I}_n \mid -\mathbf{I}_n)x_\Delta \text{ for } x_\Delta \in \Delta \subset \mathbb{R}^{2n} .$$

Here x_Δ is a $2n$ -vector, and we have used the notation $(A|B)$ for the horizontal concatenation of two matrices A, B .

Note that the barycentric representation is *not* a bijection in general, as there can be several $x_\Delta \in \Delta \subset \mathbb{R}^{2n}$ representing the same point $x_\diamond \in \mathbb{R}^n$.

The Equivalent SVM Instance. Given a Lasso instance of the form (2), that is, $\min_{x \in \blacklozenge} \|Ax - b\|^2$, we can directly parameterize the ℓ_1 -ball by the $2n$ -dimensional simplex as described above. By writing $(\mathbf{I}_n \mid -\mathbf{I}_n)x_\Delta$ for any $x \in \blacklozenge$, the objective function becomes $\|(A \mid -A)x_\Delta - b\|^2$. This means we have obtained the equivalent non-negative regression problem of the form (3) over

the domain $x_\Delta \in \Delta$ which, by our above remark on translations, is equivalent to the SVM formulation (1), i.e.

$$\min_{x_\Delta \in \Delta} \left\| \tilde{A}x_\Delta \right\|^2 ,$$

where the data matrix is given by

$$\tilde{A} := (A \mid -A) - b\mathbf{1}^T \in \mathbb{R}^{d \times 2n} .$$

The additive rank-one term $b\mathbf{1}^T$ for $\mathbf{1} \in \mathbb{R}^{2n}$ again just means that the vector b is subtracted from each original column of A and $-A$, resulting in a translation of the problem. So we have obtained an equivalent SVM instance consisting of $2n$ points in \mathbb{R}^d .

Note that this equivalence not only means that the optimal solutions of the Lasso and the SVM coincide, but indeed gives us the correspondence of all feasible points, preserving the objective values: for any points solution $x \in \mathbb{R}^n$ to the Lasso, we have a feasible SVM point $x_\Delta \in \Delta \subset \mathbb{R}^{2n}$ of the *same* objective value, and vice versa.

3.3. (SVM \preceq Lasso): Given an SVM Instance, Constructing an Equivalent Lasso Instance

This reduction is harder to accomplish than the other direction we explained before. Given an instance of an SVM problem (1), we suppose that we have a (possibly non-optimal) σ -weakly-separating vector $w \in \mathbb{R}^d$ available, for some (small) value $\sigma > 0$. Given w , we will demonstrate in the following how to construct an equivalent Lasso instance (2).

Perhaps surprisingly, such a weakly-separating vector w is trivial to obtain for the ℓ_2 -loss soft-margin SVM, as we have observed in Section 2.2 (even if the SVM input data is not separable). Also for hard-margin SVM variants, finding such a weakly-separating vector for a small σ is still significantly easier than the final goal of obtaining a near-perfect $(\sigma^* - \varepsilon)$ -separation for a small precision ε . It corresponds to running an SVM solver (such as the perceptron algorithm) for only a constant number of iterations. In contrast, obtaining a better ε -accurate solution by the same algorithm would require $O(1/\varepsilon^2)$ iterations, as mentioned in Section 2.

The Equivalent Lasso Instance. Formally, we define the Lasso instance (\tilde{A}, \tilde{b}) as the translated SVM datapoints

$$\tilde{A} := \left\{ A_i + \tilde{b} \mid i \in [1..n] \right\}$$

together with the right hand side

$$\tilde{b} := -\frac{w}{\|w\|} \cdot \frac{D^2}{\sigma} .$$

Here $D > 0$ is a strict upper bound on the length of the original SVM datapoints, i.e. $\|A_i\| < D \ \forall i$.

By definition of \tilde{A} , the resulting new Lasso objective function is

$$\left\| \tilde{A}x - \tilde{b} \right\| = \left\| (A + \tilde{b}\mathbf{1}^T)x - \tilde{b} \right\| = \left\| Ax + (\mathbf{1}^T x - 1)\tilde{b} \right\| . \quad (6)$$

Therefore, this objective coincides with the original SVM objective (1), for any $x \in \Delta$ (meaning that $\mathbf{1}^T x = 1$). However, this does not necessarily hold for the larger part of the Lasso domain when $x \in \blacklozenge \setminus \Delta$. In the following discussion and the main Theorem 3, we will prove that all those candidates $x \in \blacklozenge \setminus \Delta$ can be discarded from the Lasso problem, as they do not contribute to any optimal solutions.

As a side-remark, we note that the quantity $\frac{D}{\sigma}$ that determines the magnitude of our translation is a known parameter in the SVM literature. [Bur98, SS02] have shown that the VC-dimension of an SVM, a measure of “difficulty” for the classifier, is always lower than $\frac{D^2}{\sigma^2}$. Note that by the definition of separation, $\sigma \leq D$ always holds.

Geometric Intuition. Geometrically, the Lasso problem (2) is to compute the smallest Euclidean distance of the set $A\blacklozenge$ to the point $b \in \mathbb{R}^d$. On the other hand the SVM problem — after translating by b — is to minimize the distance of the smaller set $A\triangle \subset A\blacklozenge$ to the point b . Here we have used the notation $AS := \{Ax \mid x \in S\}$ for subsets $S \subseteq \mathbb{R}^d$ and linear maps A (it is easy to check that linear maps do preserve convexity of sets, so that $\text{conv}(AS) = A \text{conv}(S)$).

Intuitively, the main idea of our reduction is to mirror our SVM points A_i at the origin, so that both the points and their mirrored copies — and therefore the entire larger polytope $A\blacklozenge$ — do end up lying “behind” the separating SVM margin. The hope is that the resulting Lasso instance will have all its optimal solutions being non-negative, and lying in the simplex. Surprisingly, this can be done, and we will show that all SVM solutions are preserved (and no new solutions are introduced) when the feasible set \triangle is extended to \blacklozenge . In the following we will formalize this precisely, and demonstrate how to translate along our known weakly-separating vector w so that the resulting Lasso problem will have the same solution as the original SVM.

Properties of the Constructed Lasso Instance. The following theorem shows that for our constructed Lasso instance, all interesting feasible points are contained in the simplex \triangle . By our previous observation (6), we already know that all those candidates are feasible for both the Lasso (2) and the SVM (1), and obtain the same objective values in both problems.

In other words, we have a one-to-one correspondence between all feasible points for the SVM (1) on one hand, and the subset $\triangle \subset \blacklozenge$ of feasible points of our constructed Lasso instance (2), preserving all objective values. Furthermore, we have that in this Lasso instance, all points in $\blacklozenge \setminus \triangle$ are strictly worse than the ones in \triangle . Therefore, we have also shown that all optimal solutions must coincide.

Theorem 3. *For any candidate solution $x_\diamond \in \blacklozenge$ to the Lasso problem (2) defined by (\tilde{A}, \tilde{b}) , there is a feasible vector $x_\triangle \in \triangle$ in the simplex, of the same or better Lasso objective value γ . Furthermore, this $x_\triangle \in \triangle$ attains the same objective value γ in the original SVM problem (1).*

On the other hand, every $x_\triangle \in \triangle$ is of course also feasible for the Lasso, and attains the same objective value there, again by (6).

Proof. The proof follows directly from the two main facts given in Propositions 4 and 5 below, which state that “flipping negative signs improves the objective”, and that “scaling up improves for non-negative vectors”, respectively. We will see below as why these two facts hold, which is precisely by the choice of the translation \tilde{b} along a weakly separating vector w , in order to define our Lasso instance.

We assume that the given x_\diamond does not already lie in the simplex. Now by applying Propositions 4 and 5, we obtain $x_\triangle \in \triangle$, of a strictly better objective value γ for problem (3). By the observation (6) about the Lasso objective, we know that the original SVM objective value attained by this x_\triangle is equal to γ . \square

Proposition 4 (Flipping negative signs improves the objective). *Consider the Lasso problem (2) defined by (\tilde{A}, \tilde{b}) , and assume that $x_\diamond \in \blacklozenge$ has some negative entries.*

Then there is a strictly better solution $x_\blacktriangle \in \blacktriangle$ having only non-negative entries.

Proof. We are given $x_\diamond \neq 0$, having at least one negative coordinate. Define $x_\blacktriangle \neq 0$ as the vector you get by flipping all the negative coordinates in x_\diamond . We define $\delta \in \blacktriangle$ to be the difference vector corresponding to this flipping, i.e. $\delta_i := -(x_\diamond)_i$ if $(x_\diamond)_i < 0$, and $\delta_i := 0$ otherwise, so that $x_\blacktriangle := x_\diamond + 2\delta$ gives $x_\blacktriangle \in \blacktriangle$. We want to show that with respect to the quadratic objective function, x_\blacktriangle is strictly better than x_\diamond . We do this by showing that the following difference in

the objective values is strictly negative:

$$\begin{aligned}
\left\| \tilde{A}x_{\blacktriangle} - \tilde{b} \right\|^2 - \left\| \tilde{A}x_{\diamond} - \tilde{b} \right\|^2 &= \|c + d\|^2 - \|c\|^2 \\
&= c^T c + 2c^T d + d^T d - c^T c = (2c + d)^T d \\
&= 4(\tilde{A}x_{\diamond} - \tilde{b} + \tilde{A}\delta)^T \tilde{A}\delta \\
&= 4(\tilde{A}(x_{\diamond} + \delta) - \tilde{b})^T \tilde{A}\delta
\end{aligned}$$

where in the above calculations we have used that $\tilde{A}x_{\blacktriangle} = \tilde{A}x_{\diamond} + 2\tilde{A}\delta$, and we substituted $c := \tilde{A}x_{\diamond} - \tilde{b}$ and $d := 2\tilde{A}\delta$. Interestingly, $x_{\diamond} + \delta \in \blacktriangle$, since this addition just sets all previously negative coordinates to zero.

The proof then follows from Lemma 8 below. \square

Proposition 5 (Scaling up improves for non-negative vectors). *Consider the Lasso problem (2) defined by (\tilde{A}, \tilde{b}) , and assume that $x_{\blacktriangle} \in \blacktriangle$ has $\|x_{\blacktriangle}\|_1 < 1$.*

Then we obtain a strictly better solution $x_{\Delta} \in \Delta$ by linearly scaling x_{\blacktriangle} .

Proof. The proof follows along similar lines as the above proposition. We are given $x_{\blacktriangle} \neq 0$ with $\|x_{\blacktriangle}\|_1 < 1$. Define x_{Δ} as the vector we get by scaling up $x_{\Delta} := \lambda x_{\blacktriangle}$ by $\lambda > 1$ such that $\|x_{\Delta}\|_1 = 1$. We want to show that with respect to the quadratic objective function, x_{Δ} is strictly better than x_{\blacktriangle} . As in the previous proof, we again do this by showing that the following difference in the objective values is strictly negative:

$$\begin{aligned}
\left\| \tilde{A}x_{\Delta} - \tilde{b} \right\|^2 - \left\| \tilde{A}x_{\blacktriangle} - \tilde{b} \right\|^2 &= \|c + d\|^2 - \|c\|^2 \\
&= c^T c + 2c^T d + d^T d - c^T c = (2c + d)^T d \\
&= \lambda'(2\tilde{A}x_{\blacktriangle} - 2b + \lambda'\tilde{A}x_{\blacktriangle})^T \tilde{A}x_{\blacktriangle} \\
&= 2\lambda'(\tilde{A}(1 + \frac{\lambda'}{2})x_{\blacktriangle} - \tilde{b})^T \tilde{A}x_{\blacktriangle}
\end{aligned}$$

where in the above calculations we have used that $\tilde{A}x_{\Delta} = \lambda\tilde{A}x_{\blacktriangle}$ for $\lambda > 1$, and we substituted $c := \tilde{A}x_{\blacktriangle} - \tilde{b}$ and $d := \tilde{A}x_{\Delta} - \tilde{A}x_{\blacktriangle} = (\lambda - 1)\tilde{A}x_{\blacktriangle} =: \lambda'\tilde{A}x_{\blacktriangle}$ for $\lambda' := \lambda - 1 > 0$. Note that $x_{\Delta} := (1 + \lambda')x_{\blacktriangle} \in \Delta$ so $(1 + \frac{\lambda'}{2})x_{\blacktriangle} \in \blacktriangle$.

The proof then follows from Lemma 8 below. \square

Definition 6. For a given axis vector $w \in \mathbb{R}^d$, the cone with axis w , angle $\alpha \in (0, \frac{\pi}{2})$ with tip at the origin is defined as $\text{cone}(w, \alpha) := \{x \in \mathbb{R}^d \mid \angle(x, w) \leq \alpha\}$, or equivalently $\frac{x^T w}{\|x\|\|w\|} \geq \cos \alpha$. By $\overset{\circ}{\text{cone}}(w, \alpha)$ we denote the interior of the convex set $\text{cone}(w, \alpha)$, including the tip $\mathbf{0}$.

Lemma 7 (Separation). *Let w be some σ -weakly-separating vector for the SVM (1) for $\sigma > 0$. Then*

- i) $A\blacktriangle \subseteq \overset{\circ}{\text{cone}}(w, \arccos(\frac{\sigma}{D}))$
- ii) Any vector in $\text{cone}(w, \arcsin(\frac{\sigma}{D}))$ is still σ' -weakly-separating for A for some $\sigma' > 0$.

Proof. i) Definition 1 of weakly separating, and using that $\|A_i\| < D$.

ii) For any unit length vector $v \in \text{cone}(w, \arcsin(\frac{\sigma}{D}))$, every other vector having a zero or negative inner product with this v must have angle at least $\frac{\pi}{2} - \arcsin(\frac{\sigma}{D}) = \arccos(\frac{\sigma}{D})$ with the cone axis w . However, by using i), we have $A\Delta \subseteq \overset{\circ}{\text{cone}}(w, \arccos(\frac{\sigma}{D}))$, so every column vector of A must have strictly positive inner product with v , or in other words v is σ' -weakly-separating for A (for some $\sigma' > 0$). \square

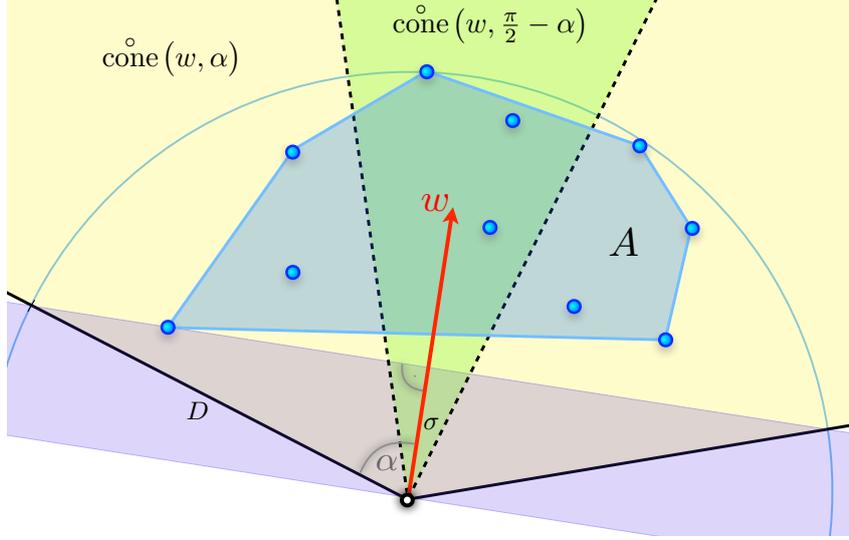


Figure 2: Illustration of the separation idea from Lemma 7, showing the cone of vectors that are still weakly-separating for the set of points A . Here we used the angle $\alpha := \arccos(\frac{\sigma}{D})$.

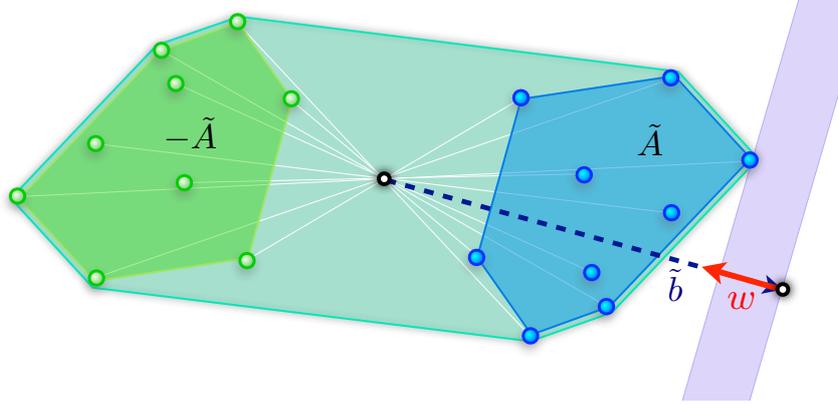


Figure 3: Illustration of Lemma 8. Recall that the translated points are defined by $\tilde{A} := \{A_i + \tilde{b} \mid i \in [1..n]\}$, where the translation is $\tilde{b} := -\frac{w}{\|w\|} \cdot \frac{D^2}{\sigma}$.

Lemma 8. *Let w be some σ -weakly-separating vector for the SVM for $\sigma > 0$. Then we claim that the translation by the vector $\tilde{b} := -\frac{w}{\|w\|} \cdot \frac{D^2}{\sigma}$ has the following properties. For any pair of vectors $x, \delta \in \blacktriangle, \delta \neq \mathbf{0}$, we have that $(\tilde{A}x - \tilde{b})^T(-\tilde{A}\delta) > 0$.*

Proof. (See also Figure 3). By definition of the translation \tilde{b} , we have that the entire Euclidean ball of radius D around the point $-\tilde{b}$ — and therefore also the point set $-\tilde{A}\blacktriangle$ and in particular $v := -\tilde{A}\delta$ — is contained in $\text{cone}(w, \arcsin(\frac{\sigma}{D}))$. Therefore by Lemma 7 ii), v is separating for A , and by translation v also separates \tilde{A} from \tilde{b} . This establishes the result $(\tilde{A}x - \tilde{b})^T v > 0$ for any $x \in \Delta$.

To extend this to the case $x \in \blacktriangle$, we observe that by definition of \tilde{b} , also the point $\mathbf{0} - \tilde{b}$ has strictly positive inner product with v . Therefore the entire convex hull of $\tilde{A}\Delta \cup \mathbf{0}$ and thus the set $\tilde{A}\blacktriangle$ has the desired property. \square

4. Implications & Remarks

In the following, we will explain a handful of implications of the shown equivalence, by relating both algorithms as well as known theoretical results for the Lasso or the SVM also to the respective other method.

4.1. Sublinear Time Algorithms for Lasso and SVMs

The recent breakthrough SVM algorithm of [CHW10, HKS11] in time $O(\varepsilon^{-2}(n+d) \log n)$ returns an ε -accurate solution to problem (1). Here ε -accurate means $(\sigma^* - \varepsilon)$ -weakly-separating. The running time of the algorithm is remarkable since for large data, it is significantly smaller than even the size of the input matrix, being $d \cdot n$. Therefore, the algorithm does not read the full input matrix \tilde{A} . More precisely, [CHW10, Corollary III.2] proves that the algorithm provides (with high probability) a solution $p^* \in \Delta$ of additive error at most ε to

$$\min_{p \in \Delta} \max_{\substack{w \in \mathbb{R}^d, \\ \|w\| \leq 1}} w^T \tilde{A} p .$$

This is a reformulation of $\min_{p \in \Delta} p^T \tilde{A}^T \tilde{A} p$, which is exactly our SVM problem (1), since for given p , the inner maximum is attained when $w = \tilde{A} p$. Therefore, using our simple trick from Section 3.2 of reducing any Lasso instance (2) to an SVM (1) (with its matrix \tilde{A} having twice the number of columns as A), we directly obtain a sublinear time algorithm for the Lasso. Note that since the algorithm of [CHW10, HKS11] only accesses the matrix \tilde{A} by simple entry-wise queries, it is not necessary to explicitly compute and store \tilde{A} (which is a preprocessing that would need linear time and storage). Instead, every entry \tilde{A}_{ij} that is queried by the algorithm can be provided on the fly, by returning the corresponding (signed) entry of the Lasso matrix A , minus b_i .

It will be interesting to compare this alternative algorithm to the recent more specialized sublinear time Lasso solvers in the line of work of [CBSS11, HK12], which are only allowed to access a constant fraction of the entries (or features) of each row of A . If we use our proposed reduction here instead, the resulting algorithm from [CHW10] has more freedom: it can (randomly) pick arbitrary entries of A , without necessarily accessing an equal number of entries from each row.

On the other hand, it is an open research question if a sublinear SVM algorithm exists which only accesses a constant small fraction of each datapoint, or of each feature of the input data.

4.2. A Kernelized Lasso

Traditional kernel regression techniques [SS04, SGV98, Rot04] try to learn a real-valued function f from the space \mathbb{R}^d of the datapoints, such that the resulting real value for each datapoint approximates some observed value. The regression model is chosen as a linear combination of the (kernel) inner products with few existing landmark datapoints (the support vectors).

Here, as we discuss a kernelization of the Lasso that is in complete analogy to the classical kernel trick for SVMs, our goal is different. We are not trying to approximate n many individual real values (one for each datapoint, or row of A), but instead we are searching for a linear combination of our points in the kernel space, such that the resulting combination is *close* to the lifted point b , measured in the kernel space norm. Formally, suppose our kernel space \mathcal{H} is given by an inner product $\kappa(y, z) = \langle \Phi(y), \Phi(z) \rangle$ for some implicit mapping $\Phi : \mathbb{R}^d \rightarrow \mathcal{H}$. Then we define our kernelized variant of the Lasso as

$$\min_{x \in \diamond} \left\| \sum_i \Phi(A_i) x_i - \Phi(b) \right\|_{\mathcal{H}}^2 . \quad (7)$$

Nicely, analogous to the SVM case, also this objective function here is determined purely in terms of the pairwise (kernel) inner products $\kappa(\cdot, \cdot)$.

An alternative way to see this is to observe that our simple “mirror-and-translate” trick from Section 3.2 also works the very same way in any kernel space \mathcal{H} . Here, the equivalent SVM instance is given by the $2n$ new points $\{\pm\Phi(A_i) - \Phi(b) \mid i \in [1..n]\} \subset \mathcal{H}$. The crucial observation is that the (kernel) inner product of any two such points is

$$\begin{aligned} & \langle s_i\Phi(A_i) - \Phi(b), s_j\Phi(A_j) - \Phi(b) \rangle \\ &= s_i s_j \kappa(A_i, A_j) - s_i \kappa(A_i, b) - s_j \kappa(A_j, b) + \kappa(b, b) . \end{aligned}$$

Here $s_i, s_j \in \pm 1$ are the signs corresponding to each point. Therefore we have completely determined the resulting $2n \times 2n$ kernel matrix K that defines the kernelized SVM (1), namely $\min_{x \in \Delta} x^T K x$, which solves our equivalent Lasso problem (7) in the kernel space \mathcal{H} .

Discussion. While traditional kernel regression corresponds to a lifting of the *rows* of the Lasso matrix A into the kernel space, our approach (7) by contrast is lifting the *columns* of A (and the r.h.s. b). We note that it seems indeed counter-intuitive to make the regression “more difficult” by artificially increasing the dimension of b . Using e.g. a polynomial kernel, this means that we also want the higher moments of b to be well approximated by our estimated x . On the other hand, increasing the dimension of b naturally corresponds to adding more data rows (or measurements) to a classical Lasso instance (2).

In the light of the success of the kernel idea for the classification case with its existing well-developed theory, we think it will be interesting to relate these results to the above proposed kernelized version of the Lasso, and to study how different kernels will affect the solution x for applications of the Lasso. Using a different connection to SVMs, the early work of [Gir98] has studied a similar kernelization of the penalized version of the Lasso, see also [EPP00]. For applications in image retrieval, [TRS12] has recently applied a similar Lasso kernelization idea.

4.3. The Pattern of Support Vectors, in the View of Lasso Sparsity

Using our construction of the equivalent Lasso instance for a given SVM, we can translate sparsity results for the Lasso to understand the pattern of support vectors of SVMs.

The motivation here is that a small number of support vectors is crucial for the efficient application of SVMs in practice, in particular in the kernelized case, because the cost to evaluate the resulting classifier is directly proportional to the number of support vectors. Furthermore, the support vectors are the most informative points for the classification task, while the non-support vectors could be safely discarded from the problem.

Using Sparse Recovery Results. There has been a vast amount of literature studying the sparsity of solutions to the Lasso and related ℓ_1 -regularized methods, in particular the study of the sparsity of x when A and b are from distributions with certain properties. For example, in the setting known as *sparse recovery*, the goal is to approximately recover a sparse solution x using a Lasso instance A, b (consisting of only a small number of rows). Here b is interpreted as a noisy or corrupted linear measurement $A\hat{x}$, and the unknown original \hat{x} is sparse. Classical recovery results then show that under weak assumptions on A, b and the sparsity of \hat{x} , the optimal Lasso solution x must be identical to the (unknown) sparse \hat{x} , see e.g. [CDS98, PS12].

Now our construction of the equivalent Lasso instance for a given SVM allows us to translate such sparsity results to the pattern of SVM support vectors. More precisely, any result that characterizes the Lasso sparsity for some distribution of matrices A and suitable b , will also characterize the patterns of support vectors for the equivalent SVM instance (and in particular the number of support vectors). This assumes that a Lasso sparsity result is applicable for the type of translation b that we have used to construct the equivalent Lasso instance. However, this is not hopeless. For example, existence of a weakly separating vector that is a sparse convex combination of the SVM datapoints is sufficient, since this results in a translation b that satisfies $b \propto A\hat{x}$ for a sparse weight vector \hat{x} . It remains to investigate which distributions and corresponding sparsity results are of most practical interest from the SVM perspective, in order to guarantee a small number of support vectors.

Lasso Sparsity in the View of SVMs. In the very other direction, sparsity has also been studied for SVMs in the literature, for example in the work of [Ste03], which analyzes the asymptotic regime $n \rightarrow \infty$. Using the simpler one of our reductions, the same results also hold for the Lasso, when the number of variables n grows.

4.4. Screening Rules for Support Vector Machines

For the Lasso, *screening rules* have been developed recently. This approach consists of a pre-processing of the data A , in order to immediately discard those predictors A_i that can be guaranteed to be inactive for the optimal solution. Provable guarantees for such rules were first obtained by [GVR10], and studied also in the later work [WLG⁺13], or the heuristic paper [TBF⁺11].

Translated to the SVM setting by our reduction, any such existing Lasso screening rule can be used to permanently discard input points before the SVM optimization is started. The screening rule then guarantees that any discarded point will not be a support vector, so the resulting optimal classifier remains unchanged. We are not aware of screening rules in the SVM literature so far, with the exception of the more recent paper of [OST13].

While the two previous subsections have mainly made use of the more complicated direction of our reduction (SVM \preceq Lasso) from Section 3.3, we can also gain some insights into the pattern of support vectors of SVMs by using the other (simpler) direction of reduction, as we will do next.

4.5. Regularization Paths and Homotopy Methods

For most machine learning methods — including SVMs and Lasso — one of the main hurdles in practice is the selection of the right free parameters. For SVMs and Lasso, the main question boils down on to how to select the best value for the regularization parameter, which determines the trade-off between the best fit of the model, and the model complexity. For the SVM, this is the soft-margin parameter C , while in the Lasso, the regularization parameter is the value r for the required bound $\|x\|_1 \leq r$.

Since naive grid-search for the best parameter is error-prone and comes without guarantees between the grid-values, algorithms that follow the solution path — as the regularization parameter changes — have been developed for both SVMs [HRTZ04] and Lasso [OPT00, EHJT04], and have become popular in particular on the Lasso side.

In the light of our joint investigation of Lasso and SVMs here, we can gain the following insights on path methods for the two problems:

General Solution Path Algorithms. We have observed that both the primal Lasso (2) and the dual ℓ_2 -SVM (1) are in fact convex optimization problems over the simplex. This enables us to apply the same solution path methods to both problems. More precisely, for problems of the form $\min_{x \in \Delta} f(x, t)$, general path following methods are available, that can maintain an approximation guarantee along the entire path in the parameter t , as shown in [GJL12] and more recently strengthened by [GMLS12]. These methods do apply for objective functions f that are convex in x and continuous in the parameter t (which in our case is r for the Lasso and C for the SVM).

Path Complexity. The exact solution path for the Lasso is known to be piecewise linear. However, the number of pieces, i.e. the complexity of the path, is not easy to determine. The recent work of [MY12] has constructed a Lasso instance $A \in \mathbb{R}^{d \times n}$, $b \in \mathbb{R}^n$, such that the complexity of the solution path (as the parameter r increases) is exponential in n . This is inspired by a similar result by [GJM12] which holds for the ℓ_1 -SVM.

Making use of the easier one of the reductions we have shown above, we ask if similar complexity worst-case results could also be obtained for the ℓ_2 -SVM. For every constraint value $r > 0$ for the Lasso problem, we have seen that the corresponding equivalent SVM instance (1) as constructed in Section 3.2 is $\min_{x \in \Delta} \|\tilde{A}_{(r)}x\|^2$, with

$$\tilde{A}_{(r)} := r(A|-A) - b\mathbf{1}^T \in \mathbb{R}^{d \times 2n}. \quad (8)$$

Therefore, we have obtained a hard-margin SVM, with the datapoints moving in the space \mathbb{R}^d as the Lasso regularization parameter r changes. The movement is a simple linear rescaling by r , relative to the reference point $b \in \mathbb{R}^d$. The result of [MY12] shows that essentially all sparsity patterns do occur in the Lasso solution as r changes², i.e. that the number of patterns is exponential in n . For each pattern, we know that the SVM solution $x_\diamond = (\mathbf{I}_n|-\mathbf{I}_n)x_\Delta$ is identical to the Lasso solution, and in particular also has the same sparsity pattern. Therefore, we also have the same (exponential) number of different sparsity patterns in the simplex parameterization x_Δ for the SVM (one could even choose more in those cases where the mapping is not unique).

To summarize, we have shown that a simple rescaling of the SVM data can have a very drastic effect, in that *every* pattern of support vectors can potentially occur as this scaling changes. While our construction is still a worst-case result, note that the operation of rescaling is not unrealistic in practice, as it is similar to popular data preprocessing by re-normalizing the data, e.g. for zero mean and variance one.

However, note that the constructed instance (8) is a hard-margin SVM, with the datapoints moving as the parameter r changes. It does not directly correspond to a soft-margin SVM, because the movement of points is different from changing the regularization parameter C in an ℓ_2 -SVM. As we have seen in Lemma 2, changing C in the SVM formulation (5) has the effect that the datapoints $\tilde{A}_{(C)}$ in the dual problem $\min_{x \in \Delta} \|\tilde{A}_{(C)}x\|^2$ move as follows (after re-scaling the entire problem by the constant factor C):

$$\tilde{A}_{(C)} := \begin{pmatrix} \sqrt{C}X \\ \mathbf{I}_n \end{pmatrix} \in \mathbb{R}^{(d+n) \times n}$$

In conclusion, the reduction technique here does unfortunately not yet directly translate the regularization path of the Lasso to a regularization path for an ℓ_2 -SVM. Still, we have gained some more insight as to how “badly” the set of SVM support vectors can change when the SVM data is simply re-scaled. We hope that the correspondence will be a first step to better relate the two kinds of regularization paths in the future. Similar methods could also extend to the case when other types of parameters are varied, such as e.g. a kernel parameter.

² The result of [MY12] applies to the penalized formulation of the Lasso, that is the solution path of $\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 + \lambda \|x\|_1$ as the parameter $\lambda \in \mathbb{R}_+$ varies. However, here we are interested in the constrained Lasso formulation, that is $\min_{x, \|x\|_1 \leq r} \|Ax - b\|^2$ where the regularization parameter is $r \in \mathbb{R}_+$.

Luckily, the penalized and the constrained problem formulations are known to be equivalent by standard convex optimization theory, and they have the same regularization paths, in the following sense. For every choice of $\lambda \in \mathbb{R}_+$, there is a value for the constraint parameter r such that the solutions to the two problems are identical (choose for example $r_{(\lambda)} := \|x_{(\lambda)}^*\|_1$ for some optimal $x_{(\lambda)}^*$, then the same vector is obviously also optimal for the constrained problem). This mapping from λ to $r_{(\lambda)}$ is monotone. As the regularization parameter λ is weakened (i.e. decreased), the corresponding constraint value $r_{(\lambda)}$ only grows larger.

The mapping in the other direction is similar: for every r , we know there is a value $\lambda_{(r)}$ (in fact this is the Lagrange multiplier of the constraint $\|x\|_1 \leq r$), such that the penalized formulation has the same solution, and the mapping from r to $\lambda_{(r)}$ is monotone as well.

Having both connections, it is clear that the two kinds of regularization paths must be the same, and that the support patterns that occur in the solutions — as we go along the path — are also appearing in the very same order as we track the respective other path.

5. Conclusions

We have investigated the relation between the Lasso and SVMs, and constructed equivalent instances of the respective other problem. While obtaining an equivalent SVM instance for a given Lasso is straightforward, the other direction is slightly more involved in terms of proof, but still simple to implement, in particular e.g. for ℓ_2 -loss SVMs.

The two reductions allow us to better relate and compare many existing algorithms for both problems. Also, it can be used to translate a lot of the known theory for each method also to the respective other method. In the future, we hope that the understanding of both types of methods can be further deepened by using this correspondence.

A. Some Soft-Margin SVM Variants that are Equivalent to (1)

We include the derivation of the dual formulation to the ℓ_2 -loss soft-margin SVM (5) for n datapoints $X_i \in \mathbb{R}^d$, together with their binary class labels $y_i \in \{\pm 1\}$, for $i \in [1..n]$, as defined above in Section 2.2. The equivalence to (1) directly extends to the one- and two-class case, without or with (regularized) offset term, and as well for the hard-margin SVM. These equivalent formulations have been known in the SVM literature, see e.g. [SS02, KSBM00, TKC05, GJ09], and the references therein.

Lemma' 2. *The dual of the soft-margin SVM (5) is an instance of the classifier formulation (1), that is $\min_{x \in \Delta} \|Ax\|^2$, with*

$$A := \begin{pmatrix} Z \\ \frac{1}{\sqrt{C}} \mathbf{I}_n \end{pmatrix} \in \mathbb{R}^{(d+n) \times n}$$

where the data matrix $Z \in \mathbb{R}^{d \times n}$ consists of the n columns $Z_i := y_i X_i$.

Proof. The Lagrangian [BV04, Section 5] of the soft-margin SVM formulation (5) with its n constraints can be written as

$$L(w, \rho, \xi, \alpha) := \frac{1}{2} \|w\|^2 - \rho + \frac{C}{2} \sum_i \xi_i^2 + \sum_i \alpha_i (-w^T Z_i + \rho - \xi_i) .$$

Here we introduced a non-negative Lagrange multiplier $\alpha_i \geq 0$ for each of the n constraints. Differentiating L with respect to the primal variables, we obtain the KKT optimality conditions

$$\begin{aligned} \mathbf{0} &\stackrel{!}{=} \frac{\partial}{\partial w} = w - \sum_i \alpha_i Z_i \\ \mathbf{0} &\stackrel{!}{=} \frac{\partial}{\partial \rho} = 1 - \sum_i \alpha_i \\ \mathbf{0} &\stackrel{!}{=} \frac{\partial}{\partial \xi} = C\xi - \alpha . \end{aligned}$$

When plugged into the Lagrange dual problem $\max_{\alpha} \min_{w, \rho, \xi} L(w, \rho, \xi, \alpha)$, these give us the equivalent formulation (sometimes called the *Wolfe-dual*)

$$\begin{aligned} \max_{\alpha} \quad & \frac{1}{2} \alpha^T Z^T Z \alpha - \rho + \frac{C}{2} \frac{1}{C^2} \alpha^T \alpha \\ & - \alpha^T Z^T Z \alpha + \rho - \frac{1}{C} \alpha^T \alpha \\ = \quad & -\frac{1}{2} \alpha^T Z^T Z \alpha - \frac{1}{2C} \alpha^T \alpha . \end{aligned}$$

In other words the dual is

$$\begin{aligned} \min_{\alpha} \quad & \alpha^T (Z^T Z + \frac{1}{C} \mathbf{I}_n) \alpha \\ \text{s.t.} \quad & \alpha \geq 0 \\ & \alpha^T \mathbf{1} = 1 . \end{aligned}$$

This is directly an instance of our first SVM formulation (1) used in the introduction, if we use the extended matrix

$$A := \begin{pmatrix} Z \\ \frac{1}{\sqrt{C}} \mathbf{I}_n \end{pmatrix} \in \mathbb{R}^{(d+n) \times n} .$$

□

Note that the (unique) optimal primal solution w can directly be obtained from any dual optimal α by using the optimality condition $w = A\alpha$.

References

- [BS83] Peter Bloomfield and William L Steiger. *Least absolute deviations: theory, applications, and algorithms*. Progress in probability and statistics. Birkhäuser, 1983.
- [Bur98] Christopher J C Burges. [A Tutorial on Support Vector Machines for Pattern Recognition](#). *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
- [BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
- [BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data - Methods, Theory and Applications*. Springer Series in Statistics 0172-7397. Springer, 2011.
- [CBSS11] Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. [Efficient Learning with Partially Observed Attributes](#). *The Journal of Machine Learning Research*, 12:2857–2878, 2011.
- [CDS98] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. [Atomic Decomposition by Basis Pursuit](#). *SIAM Journal on Scientific Computing*, 20(1):33, 1998.
- [CHL08] Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. [Coordinate Descent Method for Large-scale L2-loss Linear Support Vector Machines](#). *JMLR*, 9:1369–1398, 2008.
- [CHW10] Kenneth L Clarkson, Elad Hazan, and David P Woodruff. [Sublinear Optimization for Machine Learning](#). *FOCS 2010 - 51st Annual IEEE Symposium on Foundations of Computer Science*, 2010.
- [CV95] Corinna Cortes and Vladimir Vapnik. [Support-Vector Networks](#). *Machine Learning*, 20(3):273–297, 1995.
- [EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. [Least angle regression](#). *Annals of Statistics*, 32(2):407–499, 2004.
- [EPP00] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. [Regularization Networks and Support Vector Machines](#). *Advances in Computational Mathematics*, 13(1):1–50, 2000.
- [GC05] Debashis Ghosh and Arul M Chinnaiyan. [Classification and Selection of Biomarkers in Genomic Data Using LASSO](#). *Journal of Biomedicine and Biotechnology*, 2005(2):147–154, 2005.
- [Gir98] Federico Girosi. [An Equivalence Between Sparse Approximation and Support Vector Machines](#). *Neural Computation*, 10(6):1455–1480, 1998.
- [GJ09] Bernd Gärtner and Martin Jaggi. [Coresets for polytope distance](#). *SCG '09: Proceedings of the 25th annual Symposium on Computational Geometry*, 2009.
- [GJL12] Joachim Giesen, Martin Jaggi, and Sören Laue. [Approximating parameterized convex optimization problems](#). *ACM Transactions on Algorithms*, 9(10):1–17, 2012.
- [GJM12] Bernd Gärtner, Martin Jaggi, and Clément Maria. [An Exponential Lower Bound On The Complexity Of Regularization Paths](#). *Journal of Computational Geometry*, 3(1):168–195, 2012.
- [GMLS12] Joachim Giesen, Jens Müller, Soeren Laue, and Sascha Swiercy. [Approximating Concavely Parameterized Optimization Problems](#). In *NIPS*, 2012.
- [GVR10] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. [Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems](#). *arXiv.org*, 2010.
- [Hea82] Donald W Hearn. [The gap function of a convex program](#). *Operations Research Letters*, 1(2):67–71, 1982.
- [HK12] Elad Hazan and Tomer Koren. [Linear Regression with Limited Observation](#). In *ICML 2012 - Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [HKS11] Elad Hazan, Tomer Koren, and Nathan Srebro. [Beating SGD: Learning SVMs in Sublinear Time](#). In *NIPS*, 2011.
- [HO04] Sepp Hochreiter and Klaus Obermayer. [Gene Selection for Microarray Data](#). In Bernhard Schölkopf, Jean-Philippe Vert, and Koji Tsuda, editors, *Kernel Methods in Computational Biology*, page 319. MIT Press, 2004.

- [HO06] Sepp Hochreiter and Klaus Obermayer. [Support Vector Machines for Dyadic Data](#). *Neural Computation*, 18(6):1472–1510, 2006.
- [HRTZ04] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. [The Entire Regularization Path for the Support Vector Machine](#). *The Journal of Machine Learning Research*, 5:1391–1415, 2004.
- [Jag13a] Martin Jaggi. [An Equivalence between the Lasso and Support Vector Machines](#). In *ROKS 2013: International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: Theory and Applications*, 2013.
- [Jag13b] Martin Jaggi. [Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization](#). In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.
- [KSBM00] S Sathiya Keerthi, Shirish K Shevade, Chiranjib Bhattacharyya, and K R K Murthy. [A fast iterative nearest point algorithm for support vector machine classifier design](#). *IEEE Transactions on Neural Networks*, 11(1):124–136, 2000.
- [LM01] Yuh-Jye Lee and Olvi L Mangasarian. [RSVM: Reduced Support Vector Machines](#). In *SDM 2001 - Proceedings of the first SIAM International Conference on Data Mining*, 2001.
- [LYX05] Fan Li, Yiming Yang, and Eric P Xing. [From Lasso regression to Feature vector machine](#). In *NIPS*, 2005.
- [MY12] Julien Mairal and Bin Yu. [Complexity Analysis of the Lasso Regularization Path](#). In *ICML 2012 - Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [OPT00] Michael R Osborne, Brett Presnell, and Berwin A Turlach. [A new approach to variable selection in least squares problems](#). *IMA Journal of Numerical Analysis*, 20(3):389–403, 2000.
- [OST13] Kohei Ogawa, Yoshiki Suzuki, and Ichiro Takeuchi. [Safe Screening of Non-Support Vectors in Pathwise SVM Computation](#). In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, pages 1382–1390, 2013.
- [PRE98] Massimiliano Pontil, Ryan Rifkin, and Theodoros Evgeniou. [From Regression to Classification in Support Vector Machines](#). Technical Report A.I. Memo No. 1649, MIT, 1998.
- [PS12] Ely Porat and Martin J Strauss. [Sublinear time, measurement-optimal, sparse recovery for all](#). In *SODA '12: Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2012.
- [Ros58] Frank Rosenblatt. [The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain](#). *Psychological Review*, 65(6):386–408, 1958.
- [Rot04] Volker Roth. [The Generalized LASSO](#). *IEEE Transactions on Neural Networks*, 15(1):16–28, 2004.
- [SGV98] Craig Saunders, Alexander Gammerman, and Volodya Vovk. [Ridge Regression Learning Algorithm in Dual Variables](#). In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*. 1998.
- [SS02] Bernhard Schölkopf and Alex J Smola. [Learning with kernels](#). support vector machines, regularization, optimization, and beyond. The MIT Press, 2002.
- [SS04] Alex J Smola and Bernhard Schölkopf. [A tutorial on support vector regression](#). *Statistics and Computing*, 14:199–222, 2004.
- [Ste03] Ingo Steinwart. [Sparseness of Support Vector Machines—Some Asymptotically Sharp Bounds](#). In *NIPS*, 2003.
- [TBF⁺11] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. [Strong rules for discarding predictors in lasso-type problems](#). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2011.
- [Tib96] Robert Tibshirani. [Regression Shrinkage and Selection via the Lasso](#). *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

- [TKC05] Ivor W Tsang, James T Kwok, and Pak-Ming Cheung. [Core Vector Machines: Fast SVM Training on Very Large Data Sets](#). *Journal of Machine Learning Research*, 6:363–392, 2005.
- [TRS12] Jayaraman J Thiagarajan, Karthikeyan Natesan Ramamurthy, and Andreas Spanias. [Local Sparse Coding for Image Classification and Retrieval](#). Technical report, ASU, 2012.
- [WLG⁺13] Jie Wang, Binbin Lin, Pinghua Gong, Peter Wonka, and Jieping Ye. [Lasso Screening Rules via Dual Polytope Projection](#). In *NIPS*, 2013.
- [Zie95] Günter M Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer Verlag, 1995.