ℓ_1 Regularization in Infinite Dimensional Feature Spaces

Saharon Rosset[1], Grzegorz Swirszcz[1], Nathan Srebro[2], and Ji Zhu[3]

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY 10549, USA
{srosset, swirszcz}@us.ibm.com
[2] IBM Haifa Research Lab, Haifa, Israel and Toyota Technological Institute, Chicago, IL 60637, USA
[3] University of Michigan, Ann Arbor, MI 48109, USA

Abstract. In this paper we discuss the problem of fitting ℓ_1 regularized prediction models in infinite (possibly non-countable) dimensional feature spaces. Our main contributions are: a. Deriving a generalization of ℓ_1 regularization based on measures which can be applied in non-countable feature spaces; b. Proving that the sparsity property of ℓ_1 regularization is maintained in infinite dimensions; c. Devising a path-following algorithm that can generate the set of regularized solutions in "nice" feature spaces; and d. Presenting an example of penalized spline models where this path following algorithm is computationally feasible, and gives encouraging empirical results.

1 Introduction

Given a data sample $(x_i, y_i)_{i=1}^n$ (with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for regression, $y_i \in \{\pm 1\}$ for classification), the "non-linear" regularized optimization problem calls for fitting models to the data, embedded into a high dimensional feature space, while controlling complexity, by solving a penalized fitting problem:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_i L(y_i, \beta^T \phi(x_i)) + \lambda J(\beta) \tag{1}$$

where L is a convex loss function; J is a convex model complexity penalty (typically taken to be the ℓ_q norm of β, with $q \geq 1$); $\phi(x_i) \in \mathbb{R}^\Omega$ is an embedding of x_i into the feature space indexed by Ω; and $\beta \in \mathbb{R}^\Omega$ is the parameter vector describing model fit. This formulation is at the heart of many successful modern data analysis tools.

Kernel Support Vector Machines (Schöelkopf and Smola 2002) and other kernel methods, fit ℓ_2 regularized models in high (often infinite) dimensional reproducing kernel Hilbert spaces (RKHS). The key observation which allows us to solve these problems is that the optimal solution in fact lies in an n-dimensional sub-space spanned by the embedded data. When we move away from ℓ_2 regularization, the nice algebra of kernel methods no longer applies, and the prevalent view is that exact very high dimensional fitting becomes practically impossible.

Boosting (Freund and Schapire 1997), is a popular and successful *committee* method, which builds prediction models as linear combinations of *weak learners* (usually small decision trees), which we can think of as features in a high dimensional feature space. As shown in Rosset et al. (2004) and references therein, boosting approximately and

incrementally fits ℓ_1 regularized models in high dimensional spaces — typically the space of all trees with a given number of terminal nodes.

Fitting ℓ_1-regularized models in very high (finite) dimension is known to be attractive because of their "primal" sparsity property:

> Every ℓ_1 regularized problem has an optimal solution with at most n non-zero coefficients, no matter how high the dimension of the feature space used. Under mild conditions, this solution is unique.

This result is a simple consequence of Caratheodory's convex hull theorem. It is proven, for example, in Rosset et al. (2004).

Thus, the success of boosting (approximate ℓ_1 regularization under embedding) and the attractiveness of the ℓ_1 sparsity property, lead us to the two main questions we address in this paper:

1. Can we generalize ℓ_1 regularization to infinite dimension in a consistent way, and will the sparsity property still hold?
2. Can we solve the resulting regularized problems despite the fact that they are infinite dimensional?

We answer the first question in Sections 2 and 3. In Section 2 we offer a formulation of ℓ_1 regularization based on measure rather than norm, which naturally generalized to infinite non-countable dimension. We then show (Section 3) that the sparsity property extends to infinite dimensional fitting, and even to non-countable dimensions, when using this definition. However, this property is contingent on the existence of the solution (which is not guaranteed in non-countable dimension), and we present sufficient conditions for this existence. We also formulate a simple, testable criterion for optimality of finite-dimensional solutions to infinite dimensional problems.

Armed with these results, in Section 4 we offer an answer to our second question, and present an algorithm that can provably generate these solutions, if they exist, which is based on a generalization of path-following algorithms previously devised for the Lasso and its extensions (Efron et al. 2004, Zhu et al. 2004).

We then describe in Section 5 an embedding problem — of fitting penalized splines to low dimensional data — where our algorithm is practical, and demonstrate its application on several datasets.

Throughout this paper we denote the index set of the functions in our feature space by Ω. The notation we use in this feature space is: $\phi(x) \in \mathbb{R}^{\Omega}$ is the embedding of x, $\phi_A(x) \in \mathbb{R}^A$ with $A \subset \Omega$ is the subset of coordinates of this embedding indexed by A (in particular, $\phi_\omega(x) \in \mathbb{R}$ is the "ω coordinate" of $\phi(x)$ for $\omega \in \Omega$), while $\phi_A(X) \in \mathbb{R}^{n \times A}$ is a matrix of the empirical partial embedding of all observations. We also assume throughout that $\sup_{\omega,x} |\phi_\omega(x)| < \infty$, i.e., that embedded coordinates are uniformly bounded.

Remark: Throughout the first part of this paper we also assume no intercept (or bias), i.e., that all features in Ω participate in the norm being penalized. This is done for simplicity of exposition, but we note that all our results hold and are easily generalized to the case that contains intercept (or even multi-dimensional intercept, like the spline basis in Section 5).

2 ℓ_1 regularization in finite and infinite dimensions

The standard ℓ_1-penalized problem in Eq. (1) has $J(\beta) = \|\beta\|_1 = \sum_{\omega\in\Omega} |\beta_\omega|$. The alternative "constrained" formulation, which is equivalent under convexity of L , is:

$$\hat{\beta}(C) = \arg\min_{\beta} \sum_i L(y_i, \beta^\intercal \phi(x_i)) \text{ s.t. } \|\beta\|_1 \leq C \tag{2}$$

This definition works fine when $|\Omega| \leq \aleph_0$, i.e., when the feature space is finite or countably infinite. We now generalize it to the non-countable case. First, we replace the ℓ_1 norm by a sum with a positivity constraint, by the well known trick of "doubling" the dimension of the feature space. We define $\tilde{\Omega} = \Omega \times \{-1, 1\}$ and for every $\tilde{\omega} \in \tilde{\Omega}$, $\tilde{\omega} = \{\omega, s\}$ define $\tilde{\phi}_{\tilde{\omega}}(x) = s\phi_\omega(x)$. Our new feature space is: $\tilde{\phi}(x) \in \mathbb{R}^{|\tilde{\Omega}|}$. It is well known and very easy to prove that any optimal solution $\hat{\beta}$ of Eq. (2) corresponds to one (or more) optimal solutions of a positive constrained problem

$$\hat{\tilde{\beta}}(C) = \arg\min_{\tilde{\beta}} \sum_i L(y_i, \tilde{\beta}^\intercal \tilde{\phi}(x_i)) \text{ s.t. } \|\tilde{\beta}\|_1 \leq C \ , \ \tilde{\beta} \succeq 0. \tag{3}$$

Through the transformation,

$$\hat{\beta}_\omega = \hat{\tilde{\beta}}_{\omega,1} - \hat{\tilde{\beta}}_{\omega,-1}.$$

Thus without loss of generality we can limit ourselves to only formulation Eq. (3) with positive coefficients and drop $\sim$ from our notation.

Given the positivity constraint, we next replace the coefficient vector β by a positive measure on Ω. Let (Ω, Σ) be a measurable space, where we require $\Sigma \supset \{\{\omega\} : \omega \in \Omega\}$, i.e., the sigma algebra Σ contains all singletons (this is a very mild assumption, which holds for example for the "standard" Borel sigma algebra). Let $\mathcal{P}$ be the set of positive measures on this space. Then we generalize (3) as:

$$\hat{P}_C = \arg\min_{P\in\mathcal{P}} \sum_i L(y_i, \int_\Omega \phi_\omega(x_i) dP(\omega)) \text{ s.t. } P(\Omega) \leq C \tag{4}$$

For finite or infinite countable Ω we will always get $\Sigma = 2^\Omega$ (which is the only possible choice given our singleton-containment requirement above), and recover exactly the formulation of (3) since $P(\Omega) = \|\beta\|_1$, but the problem definition in (4) also covers the non-countable case.

3 Existence and sparsity of ℓ_1 regularized solutions in infinite dimensions

In this section we show that using the formulation (4), we can generalize the sparsity property of ℓ_1 regularized solutions to infinite dimensions, assuming an optimal solution exists. We then formulate a sufficient condition for existence of optimal solutions, and a testable criterion for optimality of a sparse solution.

3.1 Sparsity result

Theorem 1. *Assume that an optimal solution of the problem (4) exists, then there exists an optimal solution $\hat{P}_C$ supported on at most $n+1$ features in Ω.*

To understand this result and its proof let us define the set $D = \{\phi_\omega(X) : \omega \in \Omega\} \subset \mathbb{R}^n$ as the collection of *feature columns* in $\mathbb{R}^n$. Then the sparsity property simply states that any (scaled) convex combination of points in D can be described as an (identically scaled) convex combination of no more than $n+1$ points in D. For this finite case, this is simply Caratheodory's convex hull theorem. For the infinite case, we need to generalize this result, as follows:

Theorem 2. *Let μ be a positive measure supported on a bounded subset D of $\mathbb{R}^n$. Then there exists a measure ν whose support is a finite subset of D, $\{z_1, \ldots, z_k\}$, $k \leq n+1$, such that*

$$\int_D z d\mu(z) = \sum_{i=1}^{k} z_i d\nu(z_i).$$

We postpone the proof of Theorem 2 to Appendix A, and use it to prove Theorem (1). For simplicity we assume that $\mu(D) = C$, or equivalently, that $P_C(\Omega) = C$ in (4). If this is not the case and $\hat{P}_C(\Omega) = C' < C$ then we can simply apply Theorem 1 to $\hat{P}_{C'}$ for which equality holds, and the resulting sparse solution will also be optimal for constraint value C, i.e. $\hat{P}_C = \hat{P}_{C'}$.

Proof (of Theorem 1). Let $\hat{P}_C$ be an optimal solution of (4). We define a measure μ on $\mathbb{R}^n$ as a push–forward of $\hat{P}$, i.e. $\mu(B) = P(\{\omega : \phi_\omega(X) \in B\})$. Let D (as previously defined) be the image of Ω under mapping $\phi_\cdot(X)$. The measure μ is supported on D, and by our assumption from Section 1, D is bounded. We apply Theorem 2 to set D and measure μ. Each $z_i \in D$, so the preimage of z_i under the mapping $\phi_\cdot(X)$ is nonempty. For each i we pick any ω_i such that $\phi_{\omega_i}(X) = z_i$. Then $\sum_{i=1}^{k} \nu(z_i) \cdot \phi_{\omega_i}(\cdot)$ is an optimal solution of (4) supported on at most $n+1$ features.

3.2 Sufficient conditions for existence of solution of (4)

Theorem 3. *If the set $D = \{\phi_\omega(X) : \omega \in \Omega\} \subset \mathbb{R}^n$ is compact, then the problem (4) has an optimal solution*

Proof of Theorem 3 uses the following result:

Proposition 1. *The convex hull of a compact set in $\mathbb{R}^n$ is compact*

Proof of Proposition 1 is provided in Appendix A.

Proof (of Theorem 3). We consider the set $C \cdot D = \{C \cdot \phi_\omega(X) : \omega \in \Omega\} \subset \mathbb{R}^n$, where C is the (scalar) constraint from (4). By Proposition 1, the convex hull $co(C \cdot D)$ is also a compact set. By Weierstraß Theorem the continuous function $\sum_i L(y_i, z_i)$, $(z_1, \ldots, z_n)^T \in \mathbb{R}^n$ obtains its minimum at some point $\hat{z} = (\hat{z}_1, \ldots, \hat{z}_n)^T \in co(C \cdot D)$. By Caratheodory's Convex Hull Theorem 6 there exist points $z^1, \ldots, z^k \in D, k \leq n+1$ and $b_i > 0$, $\sum_{i=1}^{k} b_i z^i = z$. For each z^i we pick any ω_i such that $C \cdot \phi_{\omega_i}(X) = z^i$. The measure $\mu = C \sum_i b_i \delta_{\omega_i}$ on Ω solves the problem (4).

The condition for existence of an optimal solution of the problem (4) provided in Theorem 3 can be difficult to check in practice. The following corollary provides us with much simpler and elegant criterion

Corollary 1. *If the set Ω is compact and the mapping $\phi_\cdot(X) : \Omega \to \mathbb{R}^n$ is continuous, then the problem (4) has an optimal solution.*

Proof. It is an immediate consequence of the fact that the continuous image of a compact set is compact.

3.3 Simple criterion for optimality

Given our results here, we can now devise a simple criterion for optimality of a finite solution to an infinite dimensional problem:

Theorem 4. *If an optimal solution to the regularized problem exists, and we are presented with a finite-support candidate solution of (3) $\tilde{P}$ such that $\exists A \subset \Omega,\ |A| < \infty,\ supp(\tilde{P}) = A$, we can test its optimality using the following criterion:*
$\tilde{P}$ is optimal solution of (3) $\Leftrightarrow$ $\forall B$ s.t. $A \subseteq B,\ |B| < \infty,$ $\tilde{P}$ is optimal solution for:

$$\min_{P \in \mathcal{P}_B} \sum_i C(y_i, \int_B \phi_w(x_i) dP(\omega)) \text{ s.t. } P(B) \leq C$$

Proof.
$\Rightarrow$:$\tilde{P}$ is the optimal solution in the whole (infinite) space, so it is the optimal solution in any subspace containing its support.
$\Leftarrow$: Assume by contradiction that $\tilde{P}$ is not optimal. We know a finite-support optimal solution exists from Theorem 1, mark this by $\hat{P}$. Set $B = \text{supp}(\tilde{P}) \cup \text{supp}(\hat{P})$. Then $|B| < \infty$ and $A \subseteq B$ obviously, and $\hat{P}$ is also better than $\tilde{P}$ in B.

This theorem implies that in order to prove that a finite solution is optimal for the *infinite* problem, it is sufficient to show that it is optimal for any *finite* sub-problem containing it. We will use it in the next section to prove that our proposed algorithm does indeed generate the optimal solutions to the infinite problem, if they exist.

4 Algorithms to generate the full solution paths

In this section, we are assuming that the optimal solution to the problem (4) exists for every C (possibly because the feature space complies with the required sufficient conditions of Theorem 3 or Corollary 1).

We now show how we can devise and implement a "path-following" algorithm, which generates this full solution path at a manageable computational cost. We describe this construction for the case of Lasso, i.e., when the loss is quadratic, and note that a similar algorithm can be devised for ℓ_1 regularized hinge loss (AKA ℓ_1-SVM) (Zhu et al. 2004).

Efron et al. (2004) have shown how an incremental homotopy algorithm can be used to generate the full regularized path at the cost of approximately one least square

calculation on the full data set, for a finite feature set. Their algorithm is geometrically motivated and derived. For our purposes, we prefer to derive and analyze it from an optimization perspective, through the Karush-Kuhn-Tucker (KKT) conditions for optimality of solutions to (3). See Rosset and Zhu (2006) for details of the KKT conditions and their implications. The resulting algorithm, in our parameterized basis notation, and with our space-doubling, non-negativity trick of Section 2:

Algorithm 1 *LAR-Lasso with parameterized feature space* [4]

1. *Initialize:*
   Set $\beta = \mathbf{0}$ *(Starting from empty model)*
   $\mathcal{A} = \arg\min_\omega \phi_\omega(X)^\intercal \mathbf{y}$ *(initial set of active variables)*
   $\mathbf{r} = \mathbf{y}$ *(residual vector)*
   $\gamma_\mathcal{A} = -(\phi_\mathcal{A}(X)^\intercal \phi_\mathcal{A}(X))^{-1} sgn(\phi_\mathcal{A}(X)^\intercal \mathbf{y}),\ \gamma_{\mathcal{A}^C} = 0$ *(direction of model change)*
2. *While* $(min_\omega \phi_\omega(X)^\intercal \mathbf{r} < 0)$
   (a) $d_1 = \min\{d > 0 : \phi_\omega(X)^\intercal(\mathbf{r} - d\phi_\mathcal{A}(X)\gamma_\mathcal{A}) = \phi_{\omega'}(X)^\intercal(\mathbf{r} - d\phi_\mathcal{A}(X)\gamma_\mathcal{A}),\ \omega \notin \mathcal{A},\ \omega' \in \mathcal{A}\}$
   (b) $d_2 = \min\{d > 0 : \beta_\omega + d\gamma_\omega = 0,\ \omega \in \mathcal{A}\}$ *(hit* 0*)*
   (c) $d = \min(d_1, d_2)$
   (d) Update:
   $\beta \leftarrow \beta + d\gamma$
   $\mathbf{r} = \mathbf{y} - \phi_\mathcal{A}(X)\beta_\mathcal{A}$
   If $d = d_1$ *then add feature attaining equality at* d *to* $\mathcal{A}$*.*
   If $d = d_2$ *then remove feature attaining* 0 *at* d *from* $\mathcal{A}$*.*
   $\gamma_\mathcal{A} = -(\phi_\mathcal{A}(X)^\intercal \phi_\mathcal{A}(X))^{-1} sgn(\phi_\mathcal{A}(X)^\intercal \mathbf{r})$
   $\gamma_{\mathcal{A}^C} = 0$

This algorithm generates the full regularized solution path for (3), i.e., for a

Theorem 5. *At any iteration of Algorithm 1, assume we are after step 2(c), and let* $l \leq d$*, where* d *is given by step 2(c). Then the finitely-supported measure* P_l *with atoms at* $\mathcal{A}$ *of size* $\beta_\mathcal{A} + l\gamma_\mathcal{A}$ *is an optimal solution to (3) with* $C = \|\beta_\mathcal{A}\|_1 + l$*.*

Proof. For finite Ω this algorithm is equivalent to LARS-Lasso of Efron et al.(2004), and hence is known to generate the solution path.

For infinite Ω, Theorem 4 and the finite Ω result combined complete the proof, since for any finite $\mathcal{B}$ such that $\mathcal{A} \subseteq \mathcal{B} \subset \Omega$, the finite feature set result implies optimality of the finite-support measure $\hat{P}(C)$, generated by the algorithm, in the feature set $\mathcal{B}$.

The key computational observation regarding Algorithm 1 is that the only step where the size of the feature space comes into play is step 2(a). All other steps only consider the set of (at most $n + 1$) features included in the current solution. So the key to applying this algorithm in very high dimension lies in being able to do the search in step 2(a) efficiently over the whole non active feature space. Denote:

$$\lambda(\beta) = -\phi_{\omega'}(X)^\intercal \mathbf{r}$$

[4] For simplicity, our description does not include a non-penalized constant. Including the constant (or constants, as we do in Section 5) complicates the notation but does not cause any difficulty.

where $\mathbf{r}$, β , and $\omega' \in \mathcal{A}$ are as in step 2(a). We can then re-write 2(a) as:

$$d_1 = \min\{d > 0 : -\phi_\omega(X)^\intercal(\mathbf{r} - d\phi_\mathcal{A}(X)\gamma_\mathcal{A}) = \lambda(\beta) - d, \text{ for some } \omega \notin \mathcal{A}\}$$

If we fix $\omega \notin \mathcal{A}$, we can find the value $l(\omega)$ at which we would attain equality. Denote:

$$l(\omega) = \frac{\phi_\omega(X)^\intercal \mathbf{r} + \lambda(\beta)}{\phi_\omega(X)^\intercal \phi_\mathcal{A}(X)\gamma_\mathcal{A} + 1} \tag{5}$$

and let:

$$d(\omega) = \begin{cases} l(\omega) \text{ if } l(\omega)) \geq 0 \\ \infty \quad \text{if } l(\omega)) < 0 \end{cases} \tag{6}$$

then our search problem in 2(a) becomes one of finding:

$$\omega^* = \arg\min_{\omega \notin \mathcal{A}} d(\omega) \tag{7}$$

Now, feature spaces in which our algorithm would be applicable are ones that allow a minimization of $d(\omega)$ over the infinite feature space, e.g., by analytically solving the problem (7) using a parametrization of Ω.

4.1 Computational cost

Efron et al. (2004) argue that for the algorithm we present, the number of pieces of the regularized path, and hence the number of iterations is "typically" $O(n)$, with a finite number of features. The switch to infinite dimension does not change the fundamental setting: the sparsity property we prove in Section 3 implies that, once we have $n + 1$ features included in our solution, we do not have to consider other features anymore (except if a "drop" event happens, which reduces the number of active features).

Assuming $O(n)$ iterations, the cost hinges on the complexity of the step length / next feature search. For the lasso spline example below, the step length calculation for each iteration is $O(n^2 p)$ (where p, the dimension of the original data, is typically very small), and the direction calculation is $O(n^2)$ (using an updating formula) for an overall iteration complexity of $O(n^2 p)$. The total complexity thus comes to $O(n^3 p)$ under the assumption on the number of iterations. In our experiments, this assumption seemed to hold.

5 Example: additive splines with total variation penalty

In this Section, we illustrate the power of infinite-dimensional ℓ_1-regularized learning by considering a regression problem on functions in $[0, 1] \to \mathbb{R}$. We will suggest a specific (infinite) feature space, and show that ℓ_1-regularization under this feature space corresponds closely to bounding the kth total variation for the predictor function, recovering at the optimum a kth order polynomial spline (i.e., a piecewise degree $k - 1$ polynomial function with $k - 2$ continuous derivatives). We focus here on quadratic loss, but our results can be easily generalized to other loss functions.

For a given order k, let $\Omega = \{(a, s)|a \in [0, 1], s \in \pm 1\}$ and consider the features:

$$\phi_{a,s}(x) = s(x - a)_+^{k-1}$$

We also allow k additional unregularized features ("intercepts"):

$$\phi_r(x) = x^r$$

for $r = 0, \ldots, k-1$. For observations $(x_i, y_i), i = 1, \ldots, n$, our optimization problem is then given by:

$$\text{minimize} \quad \sum_{i=1}^{n} (y_i - f_{P,\beta}(x_i))^2 \text{ s.t. } P(\Omega) \leq C \tag{8}$$

where P is a measure over Ω, $\beta \in \mathbb{R}^k$ and

$$f_{P,\beta}(x) = \int_{(a,s)} \phi_{a,s}(x) dP(a, s) + \sum_r \beta_r \phi_r(x) \tag{9}$$

is the fitted function corresponding to (P, β). From Theorem 1 and Corollary 1 we know that a sparse optimal solution to problem (8) exists. This will be a k-th order spline.

We note that with the above features we can approximate any function arbitrary well, and can exactly match any finite number of (consistent) observations. The key to this specific choice of basis for functions is the regularization cost (i.e. $P(\Omega)$) that applies to some predictor $f_{P,\beta}$. This is a familiar situation in learning with infinite-dimensional feature spaces, which we are used to encountering in kernel-based methods, where the choice of kernel (implicitly specifying a feature space) defines the regularization cost of predictor, rather than the space of available predictors.

In our case the ℓ_1 regularization cost, $P(\Omega)$, using our feature space, corresponds to the kth total variation (the total variation of the $(k-1)$th derivative). We can demonstrate that on our sparse spline solution

Proposition 2. *For an optimal solution that is a polynomial spline $f_{\hat{P},\hat{\beta}}$ with m knots at $(a_1, s_1), ...(a_m, s_m)$, and for which $\hat{P}(\Omega) = C$ (i.e., the constraint in (8) is tight) we have:*

$$TV(f_{\hat{P},\hat{\beta}}^{(k-1)}) = (k-1)!\hat{P}(\Omega)$$

Proof. We first observe:

$$f_{\hat{P},\hat{\beta}}^{(k-1)}(x) = (k-1)! \sum_{a_i < x} s_i \hat{P}(a_i, s_i)$$

Assume we have some i, j such that $a_i = a_j$ and $s_i \neq s_j$, and assume wlog that $s_i = 1$ and $\hat{P}(a_i, 1) > \hat{P}(a_i, -1)$. We can now define $\tilde{P}$ by $\tilde{P}(a_i, 1) = \hat{P}(a_i, 1) - \hat{P}(a_i, -1)$, $\hat{P}(a_i, -1) = 0$ and $\tilde{P} = \hat{P}$ everywhere else. Then $\tilde{P}(\Omega) < \hat{P}(\Omega)$ and $f_{\tilde{P},\hat{\beta}} = f_{\hat{P},\hat{\beta}}$ and we get a contradiction to optimality

Thus we have no knot with both positive and negative coefficient, and it follows that:

$$TV(f_{\hat{P},\hat{\beta}}^{(k-1)}) = (k-1)! \sum_i |s_i \hat{P}(a_i, s_i)| = (k-1)!\hat{P}(\Omega)$$

For motivation and discussion of total variation penalties, we refer the reader to Mammen and van de Geer (1997) and refrences therein. Intuitively, by imposing a total variation constraint on a (very) large family of functions, we are forcing the resulting solution to be *smooth* (by limiting wiggliness of the $(k-1)$th derivative).

It has previously been shown that minimizing a quadratic loss subject to a constraint on the kth total variation yields a kth order spline (Mammen and van de Geer 1997). It follows immediately that our sparse spline solution is indeed the optimal solution, not only of our ℓ_1 regularized problem, but also of the fully non-parametric regression problem, where a total-variation penalty is applied.

5.1 Practical implementation and the feature search problem

Looking back at Algorithm 1 and the next feature search problem, we observe that at each iteration of the path following algorithm we have a set $\mathcal{A}$ of indices of active functions with indexes in Ω_{pen}, characterized by their knots:
$\omega \in \mathcal{A} \Rightarrow (x-\omega)_+^{k-1}$ has non-0 coefficient in the solution.
In the search criterion for the next basis function in (5), $l(\omega)$ comprises a ratio of polynomials of degree $k-1$ in ω. The coefficients of these polynomials are fixed as long as ω does not cross a data point or a current knot in $\mathcal{A}$ (since both of these events change the parametric form of ϕ_ω, due to the positive-part function $(\cdot)_+$).

Investigating these polynomials we observe that for $k \in \{1,2\}$ we get in (5) ratios of constant or linear functions, respectively. It is easy to show that the extrema of such functions on closed intervals are always at the end points. Thus, the chosen knots will always be at the data points (this was first observed by Mammen and van de Geer 1997). Interestingly, we get here a situation that is analogous to the RKHS case: we have identified an $n+k$ dimensional sub-space of the feature space such that the solution path lies fully within this sub-space. If $k \geq 3$, however, then we get ratios of higher degree polynomials in (5), and their extrema are not guaranteed to be at the ends of the intervals. Hence, knots can fall outside data points and we are really facing an optimization problem in infinite dimensional space.

As a concrete example, we now concentrate on the case $k=3$ and the lasso modeling problem. The ratio of quadratics we get in (5) can be optimized analytically within each segment (flanked by two points which are either existing knots or data points), and once we do this for all such segments (there are at most $2n$ per dimension, or a maximum of $2np$ for the additive model), we can find ω^* — the global minimizer of $d(\omega)$ in (6) — which will be the next knot.

We demonstrate this on a 2-dimensional simulation example. For $x \in [0,1]$, let:
$g(x) = 0.125 - 0.125x - x^2 + 2(x-0.25)_+^2 - 2(x-0.5)_+^2 + 2(x-0.75)_+^2$.
a quadratic spline with knots at $0.25, 0.5, 0.75$. Our target function, drawn in the upper left box of Figure 1, is $f(x_1,x_2) = g(x_1) + g(x_2)$.

We draw 100 training samples uniformly in $[0,1] \times [0,1]$ with gaussian noise:
$y_i = f(x_{i1}, x_{i2}) + \epsilon_i, \;\; \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, 0.03)$
We then apply our quadratic spline algorithm. The results can be seen in Figure 1. Initially the data is clearly under-fitted, but in about 40 iterations of Algorithm 1 we get a reasonable estimate of the true surface. After that, the fit deteriorates as over-fitting occurs and we are mostly fitting the noise.

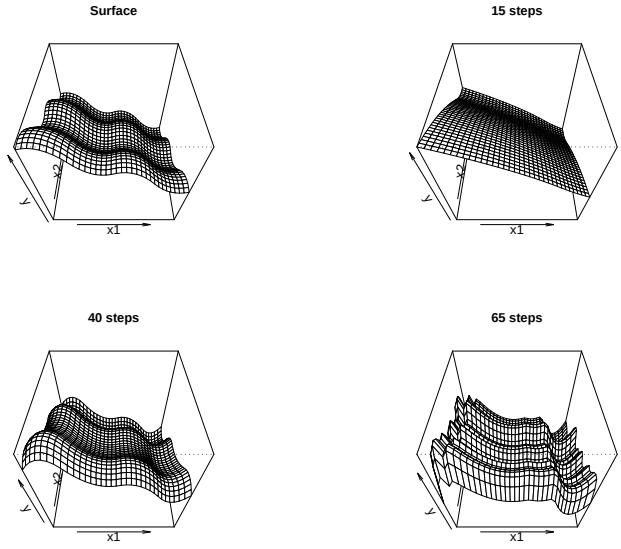

Fig. 1. True model (top left) and models generated in 15,40 and 65 steps of Algorithm 1

5.2 Real data examples: Boston and California Housing datasets

We briefly describe application of our additive spline algorithm with $k = 3$ to the Boston Housing dataset (Blake and Merz 1998) (13 features, 506 observations, of them 455 used for fitting, the rest held out for evaluation) and the California Housing dataset (Pace and Barry 1997) (8 features, 20640 observations, of them 1000 used for fitting). Figure 2 shows the performance on holdout data, as a function of the number of iterations of the path following algorithm (an indication of model complexity). We observe that for both datasets, increased complexity through additive spline fitting does seem to significantly improve the predictive performance (although the small size of the holdout set for the Boston Housing dataset implies we should take these results with some caution). For both datasets, the performance still seems to be improving after about 200 iterations, when the additive spline model already contains 10 knots across all original variables for the Boston Housing dataset and 15 knots for the California Housing dataset. Overall performance improvement due to the use of splines was 10% (California) and 15% (Boston) in MSE compared to quadratic regression and 17% (California) and 45% (Boston) compared to simple linear regression.

Remark 1 *We were not able to run the algorithm beyond about 200 iterations for Boston Housing and about 250 iterations for California Housing due to accumulation of numerical inaccuracies in our* R *implementation (caused by operations like squared root performed in finding ω^*). So the knots selected are not exactly where they should be and these tiny errors accumulate as the algorithm proceeds, eventually leading it astray.*

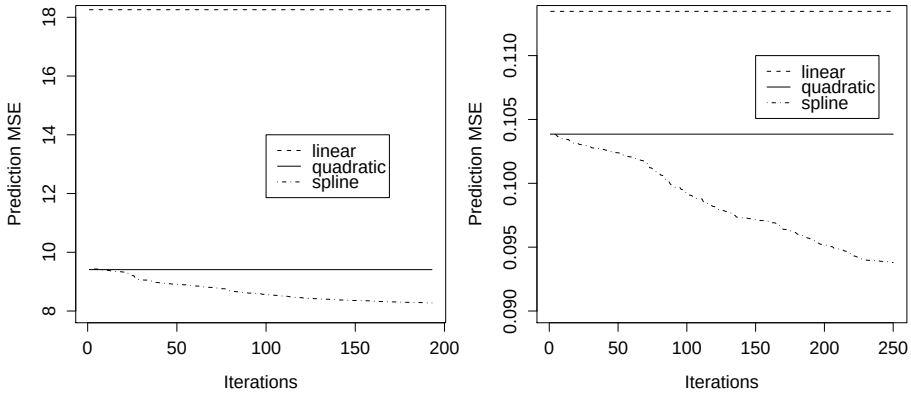

Fig. 2. Results of running additive spline algorithm on Boston (left) and California (right) Housing datasets. For comparison, both plots show the holdout MSE of regular linear regression (dashed) and quadratic regression (solid), compared to the improved performance from the additive splines. See text for details.

6 Discussion

In this paper we have addressed some of the theoretical and practical aspects of fitting ℓ_1 regularized models in infinite dimensional embedding feature spaces. In Section 3 we described some of the important mathematical and statistical properties of the solutions: existence, sparsity, and optimality testing. In Section 4 we developed an algorithm which can practically find the solutions, if the feature spaces facilitate the next feature search problem we defined in Eq. (7). We demonstrated in Section 5 that this indeed leads to a practical and useful modeling tool in the case of penalized regression splines.

While our results combine together to give a coherent picture of a theoretically attractive and practically applicable methodology, there are clearly a few additional questions that should be addressed to fully understand the potential of ℓ_1 regularization in infinite dimensions.

First and foremost is the question of learning performance, both practical and theoretical — can ℓ_1 regularized embeddings really offer a useful learning tool? From the practical side, we have evidence in the success of boosting, basis pursuit and other ℓ_1-type methods in high dimension. We can also add our spline example and the promising performance it demonstrates.

From the learning theory perspective, learning with ℓ_2 regularization in infinite-dimensional spaces enjoys strong learning guarantees which depend only on the ℓ_2 norm of the classifier and the ℓ_2 norm of the feature vector (i.e. the the kernel values). Unfortunately, the situation is not as favorable in the case of ℓ_1 regularized learning in infinite dimensional spaces. Learning guarantees based on the ℓ_1-norm of the classifier and on $\sup_{\omega,x} |\phi_\omega(x)|$ (i.e. the ℓ_∞-norm of the feature vectors) also depend logarithmically on the dimensionality (Zhang 2002). In fact, it can easily be seen that bounding

$\sup_{\omega,x} |\phi_\omega(x)|$ alone is not enough to guarantee learning in infinite dimensional spaces: consider a space with a feature w_B for each finite set $B \subset \Omega$ such that $\phi_B(x) = 1$ iff $x \in B$. Any finite sample can be perfectly fitted with a classifier of ℓ_1-norm 1 without attaining any generalization.

However, learning can be assured if the space of feature mappings $\{\phi_w : x \to \mathbb{R} | w \in \Omega\}$ is of low-complexity, e.g. low VC-dimension, low Rademacher complexity, or having a low covering numbers. In this case, we can view a bounded ℓ_1-classifier as a convex combination of (scaled) base-predictors, and apply results for combined classifiers (Koltchinski and Panchenko 2004).

It is interesting whether more general learning guarantees can also be obtained based on analytic properties of the features $\phi_\omega(x)$. Zhang (2002) has already provided guarantees on infinite-dimensional learning with a bound on $\sup_{\omega,x} |\phi_\omega(x)|$ and on the ℓ_1-norm of the classifier, by also requiring the *entropy* of the classifier to be high. This requirement precludes sparse classifiers and so is disappointing from our perspective. However, perhaps it is possible to require some sort of "dual" constraint on the features instead, precluding them from being overly sparse or disjoint. Another possibility is obtaining guarantees in terms of smoothness or topological proprieties of the features, especially when there is a natural parametrization of the features, as in spline example of Section 5.

A second important question relates to deriving a general characterization of the "interesting" feature spaces where the feature search problem can be solved. We have seen in Section 5 that for the spline example, for $k < 2$ the problem is trivial, and for $k > 3$ it cannot easily be solved analytically. The case where the full power of our methodology was deployed was when $k = 3$ (quadratic splines): on the one hand, solving the truly infinite dimensional problem would not be possible without Algorithm 1, and on the other the feature search problem admits an analytic solution. We have some preliminary results about the properties of the feature spaces and their parametrization through Ω that facilitate such analytic solutions, but that is a topic for future work.

Our spline regression example has interesting connections to recent work on use of the ℓ_1 penalty for multiple kernel and multiple component learning (Bach et al. 2004, Zhang and Lin 2006). These works employ the ℓ_1 penalty *between components or kernels* to get sparsity in these objects (note that if they have less than n objects no sparsity is guaranteed). Within kernels or components the ℓ_2 penalty is still used. Our approach gives sparsity in the original feature space, and when it has several components (like the two dimensions x_1, x_2 in the simulated multivariate spline example), our methods control the total number of features used across all components combined. Another important difference is that our approach leads to simple, efficient, algorithms for generating the full regularized path, while Bach et al. (2004) and Zhang and Lin (2006) require complex optimization approaches to solve for a single regularization setting.

A Proofs of convexity and measure results

Appendix A is organized as follows. First we present necessary definitions and facts about measures and convex sets. Then we prove Theorem 2.

Definition 1 *We define $co(A)$ as the intersection of all convex sets B containing A,*

$$co(A) = \bigcap_{\substack{A \subset B \\ B - \text{convex}}} B.$$

Analogously $\overline{co}(A)$ will denote the closure of $co(A)$.

Another natural way to define a $co(A)$ is to define it as the set of all convex combinations of finite subsets of A. Next lemma states that both those definitions are equivalent.

Lemma 1. *For a set A let $co'(A) = \{x : x = \sum_{i=1}^{n} a_i x_i, x_i \in A, \sum_{i=1}^{n} a_i = 1, a_i > 0\}$. Then $co(A) = co'(A)$.*

This is a very well known fact, but the proof is easy so we provide it.

Proof (of Lemma 1). The inclusion $co'(A) \subset co(A)$ is obvious, as every convex set containing A contains all convex combinations of points of A.

It remains to prove that $co'(A) \supset co(A)$. We shall show that $co'(A)$ is a convex set. Indeed, let $x, y \in co'(A)$. By definition $x = \sum_{i=1}^{n} \alpha_i x_i, y = \sum_{i=1}^{k} \beta_i y_i, x_i, y_i \in A, \sum_{i=1}^{n} \alpha_i = 1, \sum_{i=1}^{k} \beta_i = 1, \alpha_i, \beta_i > 0$. Then for every $t \in [0, 1]$

$$tx + (1-t)y = t \sum_{i=1}^{n} \alpha_i x_i + (1-t) \sum_{i=1}^{k} \beta_i y_i$$

is a convex combination of points $x_1, \ldots, x_n, y_1, \ldots, y_k$, so it is an element of $co'(A)$. Trivially $A \subset co'(A)$, so $co'(A)$ is a convex set containing A, and thus it contains $co(A)$.

We are going to need the following classical result:

Theorem 6 (Caratheodory's Convex Hull Theorem). *Let A be a finite set of points in $\mathbb{R}^n$. Then every $x \in co(A)$ can be expressed as a convex combination of at most $n+1$ points of A.*

A corollary of Caratheodory's Convex Hull Theorem is Proposition 1, the classical fact which is essential for our considerations (see also Rudin (1991), Theorem 3.20.)

Proof (of Proposition 1 from Section 3.2). By Caratheodory's Convex Hull Theorem and Lemma 1 $co(A)$ is the image of a mapping $\{a_1, \ldots, a_{n+1}, z_1, \ldots, z_{n+1}\} \mapsto \sum_{i=1}^{n+1} a_i z_i$. This is a continuous mapping on a compact domain $\{\sum_{i=1}^{n+1} a_i, a_i \geq 0\} \times A^n$, so its image is compact.

Now we need to connect the theory of convex sets with measures on bounded subsets of $\mathbb{R}^n$. Lemma 2 provides such a link.

Lemma 2. *Let A be a bounded subset of $\mathbb{R}^n$. Then for any probability measure μ with $supp(\mu) \subset A$ there holds*

$$\int_A x d\mu(x) \in co(A).$$

Remark 2 *This result does not generalize to the non-Euclidean case .*
If A is a subset of a topological vector space V such that V^* [5] separates the points of V and if $\overline{co}(A)$ is a compact set then it is always true that $\int_A x d\mu(x) \in \overline{co}(A)$. Compare (Rudin 1991, Theorem 3.27).
However, even if A is a bounded subset of a a Hilbert space and coA is not closed, $\int_A x d\mu(x)$ might not be contained in coA.

For every bounded subset A of $\mathbb{R}^n$ and $C \in \mathbb{R}$ we define $\mathcal{D}_C(A)$ to be a set of all $\varphi \in (\mathbb{R}^n)^*$, $\|\varphi\| = 1$ such that $\varphi(x) \leq C$ for every $x \in A$.

For the proof of Lemma 2 we shall need the following two propositions. Proposition 3 states that $\overline{co}(A)$ is an intersection of all halfspaces containing A.

Proposition 3. *Let A be a bounded subset of $\mathbb{R}^n$. Then $\overline{co}A$ is an intersection of all sets of the form $\{x : \varphi(x) \leq C, \varphi \in \mathcal{D}_C(A)\}$.*

Proof. This is an immediate corollary of a much stronger results - the Separation Theorem for topological vector space, see Rudin (1991, Theorem 3.21).

The next proposition states that every point on a boundary of a convex hull of A has a "supporting plane".

Proposition 4. *For every $z \in \overline{co}(A) \setminus co(A)$ there exist $C \in \mathbb{R}$ and $\Lambda \in \mathcal{D}_C(A)$ such that $\Lambda(z) - C = 0$.*

Proof. Let $z \in \overline{co}(A) \setminus co(A)$. Then there exists a convex set W containing A such that $z \notin W$ and $z \in \overline{W}$. Let $c = \sup_{C \in \mathbb{R}, \Lambda \in \mathcal{D}_C(W)} \Lambda(z) - C$. As $z \in \overline{W}$, for every $\Lambda \in \mathcal{D}_C(\omega)$ there holds $\Lambda(z) - c \leq 0$. By continuity of linear operators in $\mathbb{R}^n$ there holds $c \leq 0$. Due to compactness arguments there exist C_0 and $\Lambda_0 \in \mathcal{D}_{C_0}(W)$ such that $c = \Lambda_0(z) - C_0$. Thus for every point $z' \in B(z, -c)$ and every $\Lambda \in \mathcal{D}_C(W)$ for some C there holds $\Lambda(z') - C = \Lambda(z' - z) + \Lambda(z) - C \leq \Lambda(z' - z) - c \leq 0$ as $|\Lambda(z' - z)| < -c$ because $\|\Lambda\| = 1$. Thus $B(z, -c) \subset \overline{W}$. Let us suppose that $c \neq 0$. Let I be any diameter of $B(z, -c)$. The intersection $I \cap W$ is a convex set, it is a subinterval of I. Moreover, as $\overline{W \cap I} = \overline{W} \cap I$, only the endpoints of I can be not contained in W. As z is a midpoint of I, this is a contradiction with an assumption $z \notin W$, so there must be $c = 0$. Thus $\Lambda_0(z) - C_0 = 0$. As $\mathcal{D}_C(W) \subset \mathcal{D}_C(A)$, the proposition follows.

[5] For a topological space V we are using a V^* symbol to denote a dual space of V–the space of all continuous linear functionals on V. In case of $\mathbb{R}^n$ this space is of course isometric to $\mathbb{R}^n$ itself. In particular $\varphi \in (\mathbb{R}^n)^*$, $\|\varphi\| = 1$ can be identified with a set of all vectors of length 1.

Proof. (*of Lemma 2)* The proof is by induction on n, the dimension of the space. For $n = 0$, $\mathbb{R}^n$ consists of a single point and the theorem is trivially true.
Let us assume that the assertion holds for n and let A be a bounded subset of $\mathbb{R}^{n+1}$. We will denote $y = \int_A x d\mu(x)$. Let Λ be a linear functional on $\mathbb{R}^{n+1}$. We have (by linearity of an integral)

$$\Lambda(y) = \int_A \Lambda(x) d\mu(x)$$

and therefore if $\Lambda \in \mathcal{D}_C(A)$, then $\Lambda(y) \leq C$. By Proposition 3 $y \in \overline{co}(A)$. By Proposition 4 either $y \in co(A)$ and our assertion is true, or there exist C and $\Lambda \in \mathcal{D}_C(A)$ such that $\Lambda(y) = C$. In the second case $\mu(A \setminus \{x : \Lambda(x) = C\}) = 0$, and therefore $supp(\mu) \subset \mu(A \cap \{x : \Lambda(x) = C\})$. The later set is a convex subset of n-dimensional hyperplane, and by inductive assumption $y \in A \cap \{x : \Lambda(x) = C\} \subset A$.

Now we are ready to prove Theorem 2.

Proof (of Theorem 2 from Section 3.1). Is is an immediate consequence of Lemma 2 and Caratheodory's convex hull theorem .

References

Bach, F., Lanckriet, G. & Jordan, M.: Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. *ICML-04*

Blake, C.L. & Merz, C.J.: UCI Repository of machine learning databases. (1998) `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R.: Least Angle Regression (with discussion). The Annals of Statistics, **32** (2004) 2:407–499

Freund, Y. & Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55** (1997) 1:119–139

Hastie, T. & Tibshirani, R.: Generalized Additive Models. Chapman and Hall (1990)

Koltchinski, V. & Panchenko, D. : Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics **30** (2002), 1

Mammen, E. & Van de Geer, S.: Locally Adaptive Regression Splines. The Annals of Statistics **25** (1997) 1:387–413

Pace, R.K. & Barry, R. : Sparse Spatial Autoregressions. Stat. & Prob. Let. **33** (1997) 291–297

Rosset, S., Zhu, J. & Hastie,T. : Boosting as a regularized path to a maximum margin classifier. Journal of Machine Learning Research **5** (2004) (Aug):941-973

Rosset, S. & Zhu, J. : Piecewise linear regularized solution paths. Annals of Statistics, to appear (2006) `www-stat.stanford.edu/~saharon/papers/piecewise-revised.pdf`

Rudin, W. Functional Analysis, second edition, McGraw-Hill, Inc. (1991)

Schöelkopf, B. & Smola, A. J.: Learning with Kernels. MIT Press (2002)

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

Zhang, H. H. & Lin, Y. : Component Selection and Smoothing for Nonparametric Regression in Exponential Families. Statistica Sinica **16** (2006) 1021–1041

Zhang, T. : Covering number bounds of certain regularized linear function classes. Journal of Machine Learning Research **2** (2002) 527–550

Zhu, J., Rosset, S., Hastie, T. & Tibshirani, R. : 1-norm support vector machines. Neural Information Processing Systems **16** (2004)