Project number IST-25582

CGL

Computational Geometric Learning

"Compressing Support Vector Machines"

STREP

Information Society Technologies

Period covered:	November 1, 2012–October 31, 2013
Date of preparation:	October 31, 2013
Date of revision:	October 31, 2013
Start date of project:	November 1, 2010
Duration:	3 years
Project coordinator name:	Joachim Giesen (FSU)
Project coordinator organisation:	Friedrich-Schiller-Universität Jena Jena, Germany

Compressing Support Vector Machines

Joachim Giesen Sören Laue Jens K. Müller

November 4, 2013

Abstract

A support vector machine is an optimization problem that takes as input n feature vectors of length m. Hence, to store the input memory of size $O(nm)$ is needed. We show how to construct another support vector machine whose input are n feature vectors of total size only $O(n \log n)$ that are derived from the original feature vectors by a random projection. Here, we do not assume that the data matrix is sparse or of low rank. Instead, we rely on the weaker assumption that the solution is sparse. We also provide an algorithm for recovering an approximate solution to the original problem from an optimal solution to the compressed problem, where the approximation factor depends on the compression ratio. The original feature vectors can be streamed from secondary memory, e.g., from tape or hard drive, to fast main memory and compressed on the fly into the new feature vectors. With this compression scheme the memory requirement for the support vector machine can be reduced from $O(nm)$ to $O\big((n+m)\log n\big)$. Reducing the memory requirement becomes beneficial when the original feature vectors do not fit into main memory. In this case state-of-the-art iterative solvers need to access the slower secondary memory repeatedly, which can be avoided if the compressed feature vectors fit into main memory.

1 Introduction

Set up. Given observations $(x_1, y_1), \ldots, (x_n, y_n)$ of labels $y_i \in \{-1, 1\}$ at data points x_i in some space Ω. The goal is to learn from the observations a predictor that maps the points in Ω to the label space $\{-1, 1\}$. In the popular support vector machine approach [7] this is accomplished through the following optimization/learning problem,

$$\min_{w \in \mathbb{R}^m} \left(L(\Phi w, y) + c \, \|w\|_2^2 \right),$$

where

$$L(\Phi w, y) \;=\; \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i \langle \Phi_i, w \rangle\}$$

is the loss function, c is a regularization parameter, and $\Phi \in \mathbb{R}^{n \times m}$ is a matrix of n row feature vectors whose i-th row Φ_i is given as

$$\Phi_i = \Phi(x_i) := \big(e_1(x_i), \ldots, e_m(x_i)\big),$$

with feature functions $e_j : \Omega \to \mathbb{R}$. In the case that $\Omega = \mathbb{R}^m$, the feature functions are often chosen to be the coordinate functions

$$e_j : x = \big(x^{(1)}, \ldots, x^{(m)}\big) \mapsto x^{(j)}.$$

The learning problem allows to derive a predictor from the observations as follows: if an optimal solution

$$\begin{aligned}\hat{w} &= (\hat{w}^{(1)}, \ldots, \hat{w}^{(m)}) \\ &= \operatorname{argmin}_{w \in \mathbb{R}^m} \left(L(\Phi w, y) + c\|w\|_2^2 \right)\end{aligned}$$

exists, then it provides a mapping

$$\Omega \ni x \mapsto \sum_{j=1}^{m} \hat{w}^{(j)} e_j(x) = \langle \Phi(x), \hat{w} \rangle \in \mathbb{R}.$$

The predicted label at $x \in \Omega$ is then given as $\operatorname{sign}(\langle \Phi(x), \hat{w} \rangle)$.

Contributions. The dimension of the support vector machine (optimization problem) as we have stated it here is m, i.e., there are m optimization variables. We assume that the problem is high-dimensional, i.e., the number of features m is in $\Omega(\sqrt{n})$, where n is the number of data points. In the following we will show how to approximate the original problem with reconstruction guarantees by a problem of dimension $O(\log n)$. The constant that is hidden in the big-O notation for the target dimension depends on the approximation guarantee and on the sparsity of the optimal solution. The original feature matrix needs space of the order of $O(nm)$, whereas the compressed feature matrix only needs space of the order $O(n \log n)$. The solution of the compressed problem can be expanded (uncompressed) and the expansion is an approximate solution of the original problem, where the approximation guarantee depends on the compression factor, i.e., the approximation gets better when we compress less. For compression and expansion we need to store another matrix whose size is of the order $O(m \log n)$. Hence, by compression we can reduce the required memory from $O(nm)$ to $O((m+n)\log n)$.

The compression scheme can be utilized as follows: the feature vectors $\Phi_i, i = 1, \ldots, n$, are streamed from secondary memory, e.g., from tape or hard drive, to fast main memory and get compressed on the fly into new feature vectors $\bar{\Phi}_i$ that have only $O(\log n)$ entries each. The support vector machine can then be solved for the compressed feature matrix while accessing only main memory. Finally, the solution $\bar{w}$ of the compressed support vector machine can be expanded and the expansion is an approximate solution to the original problem.

The compressed support vector machine like the original support vector machine can be solved very efficiently, namely in time linear in the size of the (compressed) feature matrix, using for example trust region Newton methods [15] or cutting plane methods [11] that have been adapted to solve support vector machines. These methods work in iterations. Thus, if the feature vectors do not fit into many memory, then they need to be read again from secondary memory in every iteration. Hence, compressing the feature vectors can help to avoid accessing the slower secondary memory repeatedly.

Related work. Balcan, Blum and Vempala [2, 3] were the first to realize that the intrinsic dimension of a binary classification problem depends on structural properties of the solution of the problem, i.e., in their analysis the margin of an optimal solution. They proved that if binary labeled data can be separated with margin γ, then the data can be projected (using a random projection matrix) from m dimensions (number of features) to $O\left(\frac{1}{\gamma^2}\log(1/\varepsilon)\right)$ dimensions such that the projected data still can be separated with error at most ε at margin $\gamma/4$ with high probability. Here we consider another structural property of learning problems, namely the sparsity of the solution, and show how the intrinsic dimension of the problem depends on the sparsity. We do not need to assume anything about the margin of the original solution. Still, our analysis like the analysis of Balcan et

al. makes use of the Johnson-Lindenstrauss lemma, see [12]. Since our goal is to reconstruct the solution in the original feature space, where the features often have a well defined meaning, the Johnson-Lindenstrauss lemma is not enough. For the approximate reconstruction we need the sparsity of the original solution and the restricted isometry property of the projection that has been introduced in the context of compressive sensing [5, 9].

Compressed learning has been studied by Calderbank et al. in [4], where it has been shown that learning in a compressed data domain works—specifically, classification with support vector machines, provided that the original data are sparse in some, even unknown but fixed, basis. The assumption of sparse data is justified for several data domains, e.g., natural images. Nevertheless, here we do not assume that the data are sparse. We only assume that the result of the learning problem is sparse. Sparsity of the solution can be always encouraged if one combines the support vector machine with some feature selection method. Feature selection is of independent interest since sparse solutions are favored in practice because they provide predictors that are easier to interpret and faster to evaluate. Note that our result is more general than that of Calderbank et al. [4]. A sparse data set will of course provide a sparse solution, however the contrary is not true. The assumption of a sparse optimal solution implies that some of the features of the data set are not relevant for the specific classification task. However, it does not imply that the data is necessarily sparse, even in some unknown but fixed, basis. Instead, the data can still be of full rank or not sparse in any basis.

Paul et al. [18] study random projections for linear support vector machines and provide generalization guarantees when the data matrix is of low rank. This can be seen as a special case of the assumption made by Calderbank et al [4]. For the same setup Zhang et al. [24] provide an algorithm for recovering the optimal solution to the original problem from the optimal solution to the compressed problem.

Shi et al. [20] study margin distortions under random projections for linearly separable data. If the data is linearly separable by some margin in the original space it is not guaranteed that the projected data is linearly separable. They provide conditions under which a margin is preserved under random projections. Here we do not assume that the data is linearly separable. We show that the data set can always be compressed by a random projection and that an almost optimal solution for the original problem can be reconstructed from the optimal solution of the compressed problem.

In very recent work Clarkson and Woodruff [6] designed a distribution over random matrices for fast random "Johnson-Lindenstrauss"-type projections. Specifically, they improved the running time of a low-dimensional embedding of a sparse data matrix to linear time in the number of non-zeros of the data matrix. Meng and Mahoney [17] generalize the result of Clarkson and Woodruff [6] to L_p-norms for $p \in [1, 2]$. Both papers also apply their results to linear regression problems.

Zhou et al. [25] consider L_1-regularized linear least squares regression (Lasso) for low-rank data matrices. Instead of compressing the number of features they compress the number of data points by a random projection. They show that the solution to the compressed problem can predict the non-zero coefficients of the true predictor of the original problem and that the predictor of the compressed problem is persistent with the predictive risk over the uncompressed data. Maillard and Munos [16] also consider regularized least squares regression problems. They map the high-dimensional data into low-dimensional space and provide bounds on the excess risk, i.e., the sum of the estimation error and the approximation error of the linear estimator of the compressed data.

Rahimi and Recht [19] have applied random projections to approximate shift-invariant kernels for classification and regression problems. Specifically, they map the high-dimensional data from the corresponding reproducing Hilbert space of a

kernel to a low-dimensional space. Achlioptas et al. [1] use the Johnson-Lindenstrauss lemma for speeding up the computation of kernel matrices for kernels that only depend on the inner product and/or Euclidean distances between two points. They use a random projection to embed the data into a lower dimensional space and compute the kernel matrix there.

On the practical side, Vedaldi and Zisserman [21] construct high-dimensional sparse feature vectors from arbitrary kernels using among others a technique called product quantization. One motivation for their work is to compress the given data—in their case large image descriptors, such that it fits into fast main memory, exploiting the fact that sparse vectors can be stored efficiently.

Yu et al. [23] address the support vector machine problem for large data that do not fit into main memory in a block minimization framework. In the framework the data are divided into blocks that fit into main memory, and at each step one block is loaded into main memory and handled by either a primal or dual support vector machine solver. The framework has been tested on data sets that are 20 times larger than the available main memory. It is possible to combine the framework of Yu et al. with our approach for handling gigantic data sets, i.e., a large number of data points (large n) with a large number of features (large m).

All previous work assumes a sparse data matrix, or a data matrix of low rank. Here, we require a weaker assumption, i.e., sparsity of the optimal solution. In addition, we also provide a method for recovering a solution to the original problem with approximation guarantees.

2 Compression

In our compression scheme we are working with random combinations of the given feature functions, i.e., from the feature functions $e_1, \dots, e_m$ we define another set of feature functions as follows

$$\tilde{e}_i = \sum_{j=1}^{m} \lambda_{ij} e_j,\ i = 1, \dots, m,$$

where the λ_{ij} are independent, identically distributed Gaussian random variables with expectation 0 and variance $1/m$, i.e., $\lambda_{ij} \sim \mathcal{N}\left(0, \frac{1}{m}\right)$. That is, the combinations are orthogonal in expectation.

Instead of $\Phi_i = \big(e_1(x_i), \dots, e_m(x_i)\big)$, we now consider the feature vectors $\tilde{\Phi}_i = \big(\tilde{e}_1(x_i), \dots, \tilde{e}_m(x_i)\big)$, $i = 1, \dots, n$, using the new feature functions. The support vector machine for these feature vectors reads

$$\min_{\tilde{w} \in \mathbb{R}^m} \left(L(\tilde{\Phi}\tilde{w}, y) + c \left\| \tilde{w} \right\|_2^2 \right),$$

where $\tilde{\Phi} \in \mathbb{R}^{n \times m}$ is the matrix whose i-th row is $\tilde{\Phi}_i$. Note that this problem is related but not strictly equivalent to the problem with feature matrix Φ because the new feature functions $\tilde{e}_i$ are orthonormal only in expectation.

In our compression scheme we now simply truncate the vectors $\tilde{w}$ and $\tilde{\Phi}_i$ after the first $\ell < m$ entries, i.e., we reduce the dimension of the problem from m to ℓ by considering only the first ℓ feature functions $\tilde{e}_i$. That is, we obtain the following optimization problem,

$$\min_{\bar{w} \in \mathbb{R}^\ell} \left(L(\bar{\Phi}\bar{w}, y) + c \left\| \bar{w} \right\|_2^2 \right),$$

where $\bar{\Phi} \in \mathbb{R}^{n \times \ell}$ is the matrix whose rows are the ℓ-dimensional feature vectors

$$\begin{aligned}\bar{\Phi}_i &= \big(\tilde{e}_1(x_i), \ldots, \tilde{e}_\ell(x_i)\big) \\ &= \left(\sum_{j=1}^{m} \lambda_{1j} e_j(x_i), \ldots, \sum_{j=1}^{m} \lambda_{\ell j} e_j(x_i)\right)\end{aligned}$$

for $i = 1, \ldots, n$. Thus, $\bar{\Phi} = \Phi \Lambda^T$, where

$$\Lambda = \big(\lambda_{ij}\big) \in \mathbb{R}^{\ell \times m} \text{ with } \lambda_{ij} \sim \mathcal{N}\left(0, \frac{1}{m}\right).$$

In the remainder of this paper we relate an optimal solution of the compressed problem to an optimal solution of the original problem, i.e., the problem with the feature vectors Φ_i. We show that up to a small error the optimal solution of the original problem can be reconstructed from an optimal solution of the compressed problem.

3 Reconstruction

Let Λ be an instance of the $\ell \times m$ Gaussian coefficient matrix. Remember that the feature vectors $\Phi_i = \big(e_1(x_i), \ldots, e_m(x_i)\big)$ for $i = 1, \ldots, n$, are determined by the fixed data points $x_1, \ldots, x_n$, and $\bar{\Phi}_i \in \mathbb{R}^\ell$ is the image of $\Phi_i \in \mathbb{R}^m$ under the random projection matrix Λ. The matrix Λ is a "Johnson-Lindenstrauss"-transform, see [12], for the set $P = \{\Phi_1, \ldots, \Phi_n, w\} \subset \mathbb{R}^m$ of $n+1$ points, where $w \in \mathbb{R}^m$ is arbitrary but fixed, i.e.,

$$(1-\delta)\|p\|_2^2 \leq \|\Lambda p\|_2^2 \leq (1+\delta)\|p\|_2^2$$

for all $p \in P$ and $\delta \in (0,1)$ with probability at least $1 - 2(n+1)\exp(-(\delta^2 - \delta^3)\ell/4)$, see for example [22]. Furthermore, also the scalar products $\langle \Phi_i, w \rangle$ are preserved up to an additive error δ with probability at least $1 - 2(n+1)\exp(-(\delta^2 - \delta^3)\ell/4)$, see also [22], i.e.,

$$|\langle \Phi_i, w \rangle - \langle \bar{\Phi}_i, \Lambda w \rangle| < \delta,$$

where by definition $\langle \bar{\Phi}_i, \Lambda w \rangle = \langle \Lambda \Phi_i, \Lambda w \rangle$. The "Johnson-Lindenstrauss"-property of Λ allows us to prove the following lemma.

Lemma 1. *For any fixed $w \in \mathbb{R}^m$, $\delta \in (0,1)$ and*

$$\ell \geq \frac{8}{\delta^2 - \delta^3} \log(4(n+1))$$

it holds with probability at least $1 - 1/n$ that

$$|L(\bar{\Phi}\Lambda w, y) - L(\Phi w, y)| < \delta.$$

Proof. By the "Johnson-Lindenstrauss"-property of Λ, if we choose $\ell \geq \frac{8}{\delta^2 - \delta^3} \log(4(n+1))$, then the norms of all the points in P are preserved up to a factor $(1 \pm \delta)$ and the scalar products $|\langle \bar{\Phi}_i, \Lambda w \rangle|$ are simultaneously preserved up to an additive error of δ with probability at least $1 - 1/n$ (by taking a union bound).

Hence, we have $|\langle \bar{\Phi}_i, \Lambda w \rangle - \langle \Phi_i, w \rangle| < \delta$ which implies

$$|1 - y_i \langle \bar{\Phi}_i, \Lambda w \rangle - (1 - y_i \langle \Phi_i, w \rangle)| < \delta$$

and thus

$$|\max\big\{0, 1 - y_i \langle \bar{\Phi}_i, \Lambda w \rangle\big\} - \max\{0, 1 - y_i \langle \Phi_i, w \rangle\}| < \delta$$

from which we derive $|L(\bar{\Phi}\Lambda w, y) - L(\Phi w, y)| < \delta$ with probability at least $1 - 1/n$ by using the definition $L(\Phi w, y) = \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i \langle \Phi_i, w \rangle\}$. □

Assume now that the optimal solution w of the original problem is always l-sparse in some fixed (that means independent of the data points $x_1, \ldots, x_n$) basis, i.e., only $l < m$ of the optimal coefficients are non-zero in this basis. To exploit sparsity we use that Λ satisfies the restricted isometry property, see for example [5, 9, 10], with high probability. A matrix $\Lambda \in \mathbb{R}^{\ell \times m}$ satisfies the restricted isometry property (RIP) with constant $\delta_l \in (0,1)$ if

$$(1-\delta_l)\|p\|_2^2 \;\leq\; \|\Lambda p\|_2^2 \;\leq\; (1+\delta_l)\|p\|_2^2$$

for all l-sparse vectors $p \in \mathbb{R}^m$, and δ_l is minimal with this property.

In fact, it holds for Λ, see for example [10], that the restricted isometry constant δ_l is upper bounded by $\delta > 0$ with probability at least $1-\varepsilon$ if $\ell \geq \frac{C}{\delta^2}\left(l \log(m/\ell) - \log(\varepsilon)\right)$ for some constant $C > 0$.

Even if w is not l-sparse in the orthonormal basis that corresponds to the feature functions $\{e_i\}$, it is by our assumption l-sparse with respect to some fixed orthonormal basis, i.e., a basis that is independent of the specific problem instance and its optimal solution w. The fixed basis can be derived from $\{e_i\}$ by applying an orthonormal transform U, i.e., Uw is l-sparse. Note that

$$\|w\|_2^2 = \|Uw\|_2^2 \quad \text{and} \quad \|\Lambda w\|_2^2 = \|\Lambda U^T U w\|_2^2,$$

and the matrix ΛU^T still satisfies the restricted isometry property. Hence, we can use the following lemma for the optimal solution of the original problem.

Lemma 2. *Let $w \in \mathbb{R}^m$ be l-sparse. Under our assumptions it holds that*

$$\|\Lambda w\|_2^2 \leq (1+\delta)\|w\|_2^2$$

with probability at least $1 - 1/m$, provided $\ell \geq (C \cdot l \log m)/\delta^2$ where $C > 0$ is some constant.

Proof. The proof follows immediately from the RIP for the projection matrix Λ if we set $\varepsilon = 1/m$. □

Observe that Lemma 2 together with Lemma 1 implies that

$$L(\bar{\Phi}\Lambda w, y) + c\,\|\Lambda w\|_2^2 - \delta \;\leq\; L(\Phi w, y) + (1+\delta)\,c\,\|w\|_2^2.$$

Multiplying both sides with $(1-\delta)$ and using the fact that $\delta > 0$ gives

$$(1-\delta)\left(L(\bar{\Phi}\Lambda w, y) + c\,\|\Lambda w\|_2^2\right) - \delta \;\leq\; L(\Phi w, y) + c\,\|w\|_2^2$$

for any l-sparse vector $w \in \mathbb{R}^m$.

Next we give an upper bound on the optimal value of the original problem.

Lemma 3. *Let $w \in \mathbb{R}^m$ be the optimal solution of the original support vector machine. Under our assumptions it holds for any $\bar{w} \in \mathbb{R}^\ell$ that*

$$\begin{aligned} L(\Phi w, y) + c\,\|w\|_2^2 \;&\leq\; L(\Phi\Lambda^T\bar{w}, y) + c\,\|\Lambda^T\bar{w}\|_2^2 \\ &=\; L(\bar{\Phi}\bar{w}, y) + c\,\|\Lambda^T\bar{w}\|_2^2. \end{aligned}$$

Proof. Note that $\Lambda^T\bar{w}$ is feasible for the original problem. Hence, the optimality of w for the original problem implies that

$$\begin{aligned} L(\Phi w, y) + c\,\|w\|_2^2 \;&\leq\; L(\Phi\Lambda^T\bar{w}, y) + c\,\|\Lambda^T\bar{w}\|_2^2 \\ &=\; L(\bar{\Phi}\bar{w}, y) + c\,\|\Lambda^T\bar{w}\|_2^2, \end{aligned}$$

where the last equality follows from $\bar{\Phi} = \Phi\Lambda^T$. □

We need one more lemma before we can state and prove our reconstruction result.

Lemma 4. *Under our assumptions it holds for any $\bar{w} \in \mathbb{R}^\ell$ that*

$$\|\Lambda^T \bar{w}\|_2^2 \leq (1+\delta)\|\bar{w}\|_2^2$$

with probability at least $1 - \ell^2 \exp\left(-\frac{m\delta^2}{40\ell^2}\right)$ for large enough ℓ.

Proof. We need to bound $\|\Lambda^T \bar{w}\|_2^2 = \langle \Lambda^T \bar{w}, \Lambda^T \bar{w} \rangle$. The following holds,

$$\begin{aligned}
\langle \Lambda^T \bar{w}, \Lambda^T \bar{w} \rangle &= \sum_{i=1}^{m} \left(\sum_{j=1}^{\ell} \lambda_{ij} \bar{w}_j \right)^2 \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{\ell} \lambda_{ij}^2 \bar{w}_j^2 + 2 \sum_{j=1}^{\ell-1} \sum_{k=j+1}^{\ell} \lambda_{ij} \lambda_{ik} \bar{w}_j \bar{w}_k \right) \\
&= \sum_{j=1}^{\ell} \bar{w}_j^2 \left(\sum_{i=1}^{m} \lambda_{ij}^2 \right) + 2 \sum_{j=1}^{\ell-1} \sum_{k=j+1}^{\ell} \bar{w}_j \bar{w}_k \left(\sum_{i=1}^{m} \lambda_{ij} \lambda_{ik} \right).
\end{aligned}$$

The sums $\sum_{i=1}^{m} (m\lambda_{ij}^2), j = 1, \ldots, \ell$ are χ^2 distributed with m degrees of freedom. From well known concentration bounds for χ^2 distributions, see for example [22], it follows that

$$\begin{aligned}
P\left[\sum_{i=1}^{m} \lambda_{ij}^2 \geq 1 + \frac{\delta}{2} \right] &= P\left[\sum_{i=1}^{m} m\lambda_{ij}^2 \geq \left(1 + \frac{\delta}{2}\right) m \right] \\
&\leq \exp\left(-\frac{m\,\delta^2}{16} \left(1 - \frac{\delta}{2}\right) \right).
\end{aligned}$$

Next we bound the sums $\sum_{i=1}^{m} \lambda_{ij}\lambda_{ik}$, i.e., sums over products of independent identically normally distributed random variables. Its moment generating function is

$$M(t) = (1 - t^2/m^2)^{-m/2}.$$

Applying Chernoff's bounding method for a random variable X and $s > 0$, i.e.,

$$\begin{aligned}
P[X \geq s] &= P\left[\exp(tX) \geq \exp(st)\right] \leq \frac{E\left[\exp(tX)\right]}{\exp(st)} \\
&= \frac{M(t)}{\exp(st)}
\end{aligned}$$

for all $t > 0$, to $\sum_{i=1}^{m} \lambda_{ij}\lambda_{ik}$ gives

$$P\left[\sum_{i=1}^{m} \lambda_{ij}\lambda_{ik} \geq \frac{\delta}{4\ell} \right] \leq (1 - t^2/m^2)^{-m/2} \exp\left(-\frac{t\delta}{4\ell} \right).$$

Setting $t = m\sqrt{1-\tau}$ gives

$$\begin{aligned}
P\left[\sum_{i=1}^{m} \lambda_{ij}\lambda_{ik} \geq \frac{\delta}{4\ell} \right] &\leq \tau^{-m/2} \exp\left(-\frac{\delta m\sqrt{1-\tau}}{4\ell} \right) \\
&= \exp\left(-\frac{m}{2} \log\tau \right) \exp\left(-\frac{\delta m\sqrt{1-\tau}}{4\ell} \right) \\
&= \exp\left(-\frac{m}{4\ell} \left(2\ell \log\tau + \delta\sqrt{1-\tau} \right) \right).
\end{aligned}$$

A simple calculation shows that $f(\tau) = 2\ell \log \tau + \delta\sqrt{1-\tau}$ is maximized at

$$\tau \;=\; \frac{8\ell^2}{\delta^2}\left(\sqrt{1+\frac{\delta^2}{4\ell^2}}-1\right) \;=\; 1-\frac{\delta^2}{16\ell^2}+\Theta\left(\frac{1}{\ell^4}\right).$$

Setting $\tau = 1-(\delta/4\ell)^2$ gives by using $\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots$ for $x < 1$,

$$\begin{aligned} f(\tau) \;&=\; 2\ell \log\left(1-\left(\frac{\delta}{4\ell}\right)^2\right)+\frac{\delta^2}{4\ell} \\ &=\; -\frac{\delta^2}{8\ell}-\Theta\left(\frac{1}{\ell^3}\right)+\frac{\delta^2}{4\ell} \;=\; \frac{\delta^2}{8\ell}-\Theta\left(\frac{1}{\ell^3}\right), \end{aligned}$$

which is lower bounded by $\frac{\delta^2}{10\ell} > 0$ for large enough ℓ. Hence,

$$P\left[\sum_{i=1}^{m}\lambda_{ij}\lambda_{ik} \geq \frac{\delta}{4\ell}\right] \;\leq\; \exp\left(-\frac{m\delta^2}{40\ell^2}\right).$$

Finally, observing that $\sum_{j=1}^{\ell}\bar{w}_j^2 = \|\bar{w}\|_2^2$, and

$$\sum_{j=1}^{\ell-1}\sum_{k=j+1}^{\ell}\bar{w}_j\bar{w}_k \;\leq\; \|\bar{w}\|_1^2 \;\leq\; \ell\|\bar{w}\|_2^2$$

gives

$$\begin{aligned} \|\Lambda^T\bar{w}\|_2^2 \;&=\; \langle\Lambda^T\bar{w},\Lambda^T\bar{w}\rangle \;\leq\; \left(1+\frac{\delta}{2}\right)\|\bar{w}\|_2^2+2\ell\frac{\delta}{4\ell}\|\bar{w}\|_2^2 \\ &=\; (1+\delta)\|\bar{w}\|_2^2 \end{aligned}$$

with probability at least $1-\ell^2\exp\left(-\frac{m\delta^2}{40\ell^2}\right)$ using a union bound. □

Observe that Lemma 4 together with Lemma 3 implies that with high probability

$$L(\Phi w, y) + c\,\|w\|_2^2 \;\leq\; L(\bar{\Phi}\bar{w}, y) + c(1+\delta)\,\|\bar{w}\|_2^2$$

for the optimal solutions $w \in \mathbb{R}^m$ and $\bar{w} \in \mathbb{R}^\ell$, respectively, of the original and the compressed problem. Thus, as we have shown before by using the optimality of $\bar{w}$ and the feasibility of Λw for the compressed problem, the sparsity of w, and Lemmas 1 and 2,

$$\begin{aligned} &(1-\delta)\left(L(\bar{\Phi}\bar{w},y)+c\,\|\bar{w}\|_2^2\right)-\delta \\ &\quad\leq\; (1-\delta)\left(L(\bar{\Phi}\Lambda w,y)+c\,\|\Lambda w\|_2^2\right)-\delta \\ &\quad\leq\; L(\Phi w,y)+c\,\|w\|_2^2 \\ &\quad\leq\; L(\bar{\Phi}\bar{w},y)+c(1+\delta)\,\|\bar{w}\|_2^2. \end{aligned}$$

Here we are aiming for more, namely an m-dimensional reconstruction of $w \in \mathbb{R}^m$ from $\bar{w} \in \mathbb{R}^\ell$. The following result that follows from Lemmas 1 to 4 shows that $\Lambda^T\bar{w} \in \mathbb{R}^m$ provides such an approximation with strong guarantees.

Theorem 5. *For a given $\delta \in (0, 1/2)$, the optimal l-sparse solution w of the original support vector machine and the reconstruction $\Lambda^T\bar{w}$ from the optimal solution $\bar{w}$ of the compressed machine can be related as follows,*

$$\begin{aligned} &(1+\delta)^{-2}\left(L(\Phi\Lambda^T\bar{w},y)+c\,\|\Lambda^T\bar{w}\|_2^2\right)-\delta \\ &\leq\; L(\Phi w,y)+c\,\|w\|_2^2 \;\leq\; L(\Phi\Lambda^T\bar{w},y)+c\,\|\Lambda^T\bar{w}\|_2^2 \end{aligned}$$

with high probability for sufficiently large m and $\ell \in \Omega((l\log m)/\delta^2)$.

Proof. The second inequality has been shown in Lemma 3, and the first inequality follows from

$$\begin{aligned} & L(\Phi\Lambda^T\bar{w}, y) + c\,\|\Lambda^T\bar{w}\|_2^2 \\ & \quad = L(\bar{\Phi}\bar{w}, y) + c\,\|\Lambda^T\bar{w}\|_2^2 \\ & \quad \leq L(\bar{\Phi}\bar{w}, y) + c(1+\delta)\,\|\bar{w}\|_2^2 \\ & \quad \leq L(\bar{\Phi}\Lambda w, y) + c(1+\delta)\,\|\Lambda w\|_2^2 \\ & \quad \leq L(\Phi w, y) + \delta + c(1+\delta)\,\|\Lambda w\|_2^2 \\ & \quad \leq L(\Phi w, y) + c(1+\delta)^2\,\|w\|_2^2 + \delta \\ & \quad \leq (1+\delta)^2\left(L(\Phi w, y) + c\,\|w\|_2^2\right) + \delta \end{aligned}$$

where the first inequality has been shown in Lemma 4, the second inequality follows from the optimality of $\bar{w}$ and the feasibility of Λw for the compressed problem, the third and fourth inequality have been shown in Lemma 1 and Lemma 2, respectively, and the fifth inequality follows from $1 + \delta \leq (1-\delta)^{-1}$ for $\delta < 1$. □

4 Feature Selection

Our compression schemes works when we can assume that the optimal solution of the original support vector machine is sparse. Sparse solutions can be encouraged by L_1-regularization, i.e., by adding an L_1-regularization term $c_1\|w\|_1$ with regularization parameter c_1 to the original support vector machine problem. Unfortunately, Theorem 5 is no longer valid for the modified support vector machine since Lemmas 2 and 4 do not hold for the L_1-norm. This problem can be circumvented by the feature selection approach that we describe in the following, see [14] for the idea. The feature selection approach is best motivated by using the adjoint problem formulation of support vector machines which is a consequence of the Representer Theorem, see [13]. The standard L_2-regularization parameter c is referred to as c_2 in the following.

Theorem 6. [Representer Theorem] *For any loss function $L(\cdot,\cdot)$ if*

$$w^* \;=\; \mathit{argmin}_{w\in\mathbb{R}^m}\; L(\Phi w, y) + c_2\|w\|_2^2$$

exists, then $w^* = \Phi^T a^*$ for some $a^* \in \mathbb{R}^n$. □

It follows that we can optimize over $a \in \mathbb{R}^n$ instead of $w \in \mathbb{R}^m$, namely by substituting $w = \Phi^T a$ in the original optimization problem, which results in the equivalent adjoint formulation

$$\min_{a\in\mathbb{R}^n}\; \left(L(\Phi\Phi^T a, y) + c_2\big(a^T\Phi\Phi^T a\big)\right).$$

The matrix $\Phi\Phi^T \in \mathbb{R}^{n\times n}$ can also be written as

$$\Phi\Phi^T = \sum_{j=1}^{m} \Psi_j\Psi_j^T,$$

where Ψ_j is the j-th column of the matrix Φ. That is, $\Phi\Phi^T$ has been written as sum of m rank-1 matrices $\Psi_j\Psi_j^T$ that correspond to the feature functions e_j, $j = 1,\ldots,m$. Weighting the j-th feature function by $0 \leq \mu_j \leq 1$ results in new feature functions $e_j^{(\mu)} = \mu_j e_j$ and

$$\sum_{j=1}^{m} \Psi_j^{(\mu)}\left(\Psi_j^{(\mu)}\right)^T \;:=\; \sum_{j=1}^{m} \mu_j^2\Psi_j\Psi_j^T \;=\; \Phi^{(\mu)}\left(\Phi^{(\mu)}\right)^T,$$

with $\Phi^{(\mu)} := \Phi D$, where $D = D(\mu_1, \ldots, \mu_m) \in \mathbb{R}^{m \times m}$ is the diagonal matrix whose diagonal is the weight vector $\mu = (\mu_1, \ldots, \mu_m) \in [0,1]^m$. Sparse solutions can now be encouraged by adding an L_1-regularization term for the weight vector μ which results in the following feature selective support vector machine

$$\min_{\mu \in [0,1]^m} \min_{w \in \mathbb{R}^m} \left(L(\Phi^{(\mu)} w, y) + c_2 \|w\|_2^2 + c_1 \|\mu\|_1 \right)$$

that can be compressed as follows

$$\min_{\bar{\mu} \in [0,1]^m} \min_{\bar{w} \in \mathbb{R}^\ell} \left(L(\bar{\Phi}^{(\bar{\mu})} \bar{w}, y) + c_2 \|\bar{w}\|_2^2 + c_1 \|\bar{\mu}\|_1 \right),$$

where $\bar{\Phi}^{(\bar{\mu})} \in \mathbb{R}^{n \times \ell}$ is the matrix whose rows are the ℓ-dimensional feature vectors

$$\begin{aligned} \bar{\Phi}_i^{(\bar{\mu})} &= \left(\sum_{j=1}^m \lambda_{1j} e_j^{(\bar{\mu})}(x_i), \ldots, \sum_{j=1}^m \lambda_{\ell j} e_j^{(\bar{\mu})}(x_i) \right) \\ &= \left(\sum_{j=1}^m \lambda_{1j} \bar{\mu}_j e_j(x_i), \ldots, \sum_{j=1}^m \lambda_{\ell j} \bar{\mu}_j e_j(x_i) \right) \end{aligned}$$

for $i = 1, \ldots, n$. That is, $\bar{\Phi}^{(\bar{\mu})} = \Phi^{(\bar{\mu})} \Lambda^T$.

In the spirit of Theorem 5 we can prove an analogous theorem for the relationship between the feature selective support vector machine and its compressed counterpart.

Theorem 7. *For a given $\delta \in (0, 1/2)$, the optimal solution (w, μ) of the original feature selective support vector machine, where w is l-sparse, and the reconstruction $(\Lambda^T \bar{w}, \bar{\mu})$ from the optimal solution $(\bar{w}, \bar{\mu})$ of the compressed feature selective support vector machine can be related as follows,*

$$\begin{aligned} (1-\delta)^2 \left(L(\Phi^{(\bar{\mu})} \Lambda^T \bar{w}, y) + c_2 \|\Lambda^T \bar{w}\|_2^2 + c_1 \|\bar{\mu}\|_1 \right) - \delta \\ \leq \quad L(\Phi^{(\mu)} w, y) + c_2 \|w\|_2^2 + c_1 \|\mu\|_1 \\ \leq \quad L(\Phi^{(\bar{\mu})} \Lambda^T \bar{w}, y) + c_2 \|\Lambda^T \bar{w}\|_2^2 + c_1 \|\bar{\mu}\|_1. \end{aligned}$$

with high probability for sufficiently large m and $\ell \in \Omega((l \log m)/\delta^2)$.

Proof. The first inequality follows from

$$\begin{aligned} & L(\Phi^{(\bar{\mu})} \Lambda^T \bar{w}, y) + c_2 \|\Lambda^T \bar{w}\|_2^2 + c_1 \|\bar{\mu}\|_1 \\ &= L(\bar{\Phi}^{(\bar{\mu})} \bar{w}, y) + c_2 \|\Lambda^T \bar{w}\|_2^2 + c_1 \|\bar{\mu}\|_1 \\ &\leq L(\bar{\Phi}^{(\bar{\mu})} \bar{w}, y) + c_2 (1+\delta) \|\bar{w}\|_2^2 + c_1 \|\bar{\mu}\|_1 \\ &\leq L(\bar{\Phi}^{(\mu)} \Lambda w, y) + c_2 (1+\delta) \|\Lambda w\|_2^2 + c_1 \|\mu\|_1 \\ &\leq (1-\delta)^{-1} L(\Phi^{(\mu)} w, y) + \delta + c_2 (1+\delta) \|\Lambda w\|_2^2 \\ &\quad + c_1 \|\mu\|_1 \\ &\leq (1-\delta)^{-1} L(\Phi^{(\mu)} w, y) + c_2 (1+\delta)^2 \|w\|_2^2 \\ &\quad + c_1 \|\mu\|_1 + \delta \\ &\leq (1-\delta)^{-2} \left(L(\Phi^{(\mu)} w, y) + c_2 \|w\|_2^2 + c_1 \|\mu\|_1 \right) + \delta, \end{aligned}$$

where the first inequality has been shown in Lemma 4, the second inequality is implied by the optimality of $(\bar{w}, \bar{\mu})$ and the feasibility of $(\Lambda w, \mu)$ for the compressed problem, the third inequality has been shown in Lemma 1, and the fourth inequality has been shown in Lemma 2.

The second inequality follows from the optimality of (w, μ) and the feasibility of $(\Lambda^T \bar{w}, \bar{\mu})$ for the original feature selective support vector machine, i.e.,

$$\begin{aligned} & L(\Phi^{(\mu)} w, y) + c_2 \|w\|_2^2 + c_1 \|\mu\|_1 \\ & \quad \leq L(\Phi^{(\bar{\mu})} \Lambda^T \bar{w}, y) + c_2 \|\Lambda^T \bar{w}\|_2^2 + c_1 \|\bar{\mu}\|_1. \end{aligned}$$ □

The original and the compressed feature selective support vector machine can be reformulated as convex-concave optimization problems that have a unique solution and can be solved efficiently.

5 Conclusions

We have presented a compression technique for high-dimensional support vector machines. Our approach takes feature vectors as input and projects them by a random projection to compressed feature vectors with exponentially fewer entries. The random projection matrix that we use here is a Johnson-Lindenstrauss transform that satisfies the restricted isometry property from compressive sensing. Both properties are needed to show that the solution of the original problem can be reconstructed up to a small error from the solution to the compressed problem. The achievable compression rate depends on the sparsity of the solution to the original problem. Hence, our compression technique works well together with a feature selection approach that encourages sparse solutions. Our approach is not restricted to standard support vector machines. Only Lemma 1 needs to be adapted for handling other loss functions, e.g., logistic loss function or the Crammer Singer loss function [8] for multi-class classification. Unlike previous work we do not assume that the data matrix is sparse or of low rank. Instead, we rely on the weaker assumption that the optimal solution is sparse. We have also provided an algorithm for recovering the solution to the original problem from the solution to the compressed problem.

References

[1] Dimitris Achlioptas, Frank McSherry, and Bernhard Schölkopf. Sampling techniques for kernel methods. In *Advances in Neural Information Processing Systems (NIPS)*, pages 335–342, 2001.

[2] Maria-Florina Balcan and Avrim Blum. A PAC-Style Model for Learning from Labeled and Unlabeled Data. In *Computational Learning Theory (COLT)*, pages 111–126, 2005.

[3] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. On Kernels, Margins, and Low-Dimensional Mappings. In *Algorithmic Learning Theory (ALT)*, pages 194–205, 2004.

[4] Robert Calderbank, Sina Jafarpour, and Robert Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, 2009.

[5] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

[6] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference (STOC)*, pages 81–90, 2013.

[7] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[8] Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. In *Computational Learning Theory (COLT)*, pages 35–46, 2000.

[9] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[10] Massimo Fornasier and Holger Rauhut. *Compressive sensing*, chapter 2. Springer, 2011.

[11] Thorsten Joachims. Training linear SVMs in linear time. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 217–226, 2006.

[12] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[13] George S. Kimeldorf and Grace Wahba. A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. *The Annals of Mathematical Statistics*, 41:595–502, 1970.

[14] Gert R. G. Lanckriet, Nello Cristianini, Peter L. Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the Kernel Matrix with Semi-Definite Programming. In *Proceedings of the Nineteenth International Conference (ICML)*, pages 323–330, 2002.

[15] Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi. Trust region newton method for logistic regression. *Journal of Machine Learning Research*, 9:627–650, 2008.

[16] Odalric-Ambrym Maillard and Rémi Munos. Compressed least-squares regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1213–1221, 2009.

[17] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Symposium on Theory of Computing Conference (STOC)*, pages 91–100, 2013.

[18] Saurabh Paul, Christos Boutsidis, Malik Magdon-Ismail, and Petros Drineas. Random projections for support vector machines. In *CoRR*, volume International Conference on Artificial Intelligence (AISTATS), 2013.

[19] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[20] Qinfeng Shi, Chunhua Shen, Rhys Hill, and Anton van den Hengel. Is margin preserved after random projection? In *International Conference on Machine Learning (ICML)*, 2012.

[21] Andrea Vedaldi and Andrew Zisserman. Sparse Kernel Approximations for Efficient Classification and Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, 2012.

[22] Santosh Vempala. *The Random Projection Method*. DIMACS: Series in Discrete Mathematics and Theoretical Computer Science Series. American Mathematical Society, 2004.

[23] Hsiang-Fu Yu, Cho-Jui Hsieh, Kai-Wei Chang, and Chih-Jen Lin. Large linear classification when data cannot fit in memory. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 833–842, 2010.

[24] Lijun Zhang, Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Recovering optimal solution by dual random projection. In *Conference on Learning Theory (COLT)*, 2013.

[25] Shuheng Zhou, John D. Lafferty, and Larry A. Wasserman. Compressed regression. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.