DIMACS Technical Report 2007-16
October 2007

SADDLE POINT FEATURE SELECTION IN SVM CLASSIFICATION

by

Yuri Goncharov
Artezio LLC
Moscow, Russia
Goncharov.Yuri@gmail.com

Ilya Muchnik
DIMACS
Rutgers University
New Brunswick, New Jersey 08903
muchnik@dimacs.rutgers.edu

Leonid Shvartser
Ness TSG,
Atidim 10
Tel-Aviv
leonid.shvartser@ness.com

ABSTRACT

SVM wrapper feature selection method for the classification problem, introduced in our previous work [1], is analyzed. The method based on modification of the standard SVM criterion by adding to the basic objective function a third term, which directly penalizes a chosen set of variables. The criterion divides the set of all variables into three subsets: deleted, selected and weighted features. We give more formal derivation of the saddle point problem to which SVM wrapper method reduces. Saddle point algorithm described, proof of its convergence and estimation for the step size of the algorithm done. Effective calculations of projections used in the saddle point algorithm are described. The algorithm is examined on a classification Benchmark and its ability to improve the SVM recognition results is shown.

1 Introduction

This work is an continuation of our previous work [1]. In [1] a new type of SVMs - saddle point SVM (SP-SVM) was introduced and investigated for the feature selection in SVM learning classification problem.

The feature selection problem reduces to the problem of saddle point search of convex-concave function. We presented a survey about saddle points algorithms in [1]. Proposed saddle point search algorithm in [1] for practical computations was based on heuristics proposed in [14]. In current work we propose another saddle point algorithm and prove it convergence. Estimate of step size of the algorithm given. Saddle point search algorithm requires to find projections on intersection of cube and plane. We describe how these projection could be calculated effectively. We show results of computations on benchmark data. The paper is organized into 7 sections. In section 2 we remind setting of SVM problem and prove technical lemma that we need later. In section 3 we repeat the basics of exact wrapped methods of feature selection for classification problem. In sections 4 and 5 we describe our saddle-point algorithm. Section 6 describes experimental results with feature selection algorithm.

2 SVM in Learning Classification

In this section we remind standard SVM problem in learning classification. We denote by $\langle x_1, x_2 \rangle$ inner product of vectors x_1 and x_2. Following [12] suppose that we have a learning sample:

$$\{x_i, y_i\}, x_i \in R^n, y_i \in \{1; -1\}, i = 1, \ldots, l.$$

Below is standard formulation of SVM problem:

$$\begin{gathered} \min_{w,b,\delta_i} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \delta_i \right) \\ y_i(\langle w, x \rangle + b) \geq 1 - \delta_i, \\ \delta_i \geq 0, C > 0, i = 1, ..., l \end{gathered} \tag{1}$$

Solution w^*, b^*, δ^* of the (1) gives optimal hyperplane $\langle w^*, x \rangle + b^* = 0$.

We need the following lemma later:

Lemma 1. *Let for all vectors $x_i, i = 1, ..., l$ in problem* (1) *coordinate j_0 is equal to 0. Let (w^*, b^*, δ^*) - solution of the problem* (1)

Then j_0's coordinate of vector w^* equal to 0.

Proof. Let's suppose contrary: $w^*_{j_0} \neq 0$. Define the following vector:

$$w^1_j = \begin{cases} w^*_j & , j = 1, .., n, \ j \neq j_0, \\ 0 & , j = j_0. \end{cases}$$

Then (w^1, b^*, δ^*) - satisfy constraints of the problem and $\|w^1\|^2 < \|w^*\|^2$. We have contradiction with the fact that (w^*, b^*, δ^*) gives minimum of functional of the problem (1).

Remark 1. *Lemma 1 can be generalized to the case when for all vectors $x_i, i = 1, ..., l$ in problem (1) coordinate j_0 is equal to some constant. In this case j_0's coordinate of vector w^* also equal to 0.*

This property of solution of SVM problem could be seen as simple entire feature selection ability of SVM.

Wolfe's dual for (1) is the following problem:

$$\max_{\lambda}\left(\sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{k=1}^{l}\left(y_i y_k x_i^T x_k \lambda_i \lambda_k\right)\right)$$
$$\sum_{i=1}^{l}\lambda_i y_i = 0, \tag{2}$$
$$0 \le \lambda_i \le C,\ i = 1, \dots, l$$

Let λ^* is a solution of (2), then optimal value of w from problem (1) calculated in the following way:

$$w^* = \sum_{i=1}^{l} y_i \lambda_i^* x_i \tag{3}$$

Vectors x_i, for which $\lambda_i^* \neq 0$, are support vectors.

3 Exact Wrapped Method for Feature Selection in Learning Classification

In this section we introduce new criteria to receive a classifier with maximum margin by searching subspaces of a given space. The methods to find saddle points related to the mentioned optimal classifiers we call as exact wrapped methods. The feature selection problem considered under SVM methods is investigated in [2] - [6]. In [4],[5] was proposed an efficient greedy-like procedure, which worked as a standard wrapped algorithm [7]. We introduce a three terms criterion as a modification of SVM setting and define a problem as a problem of searching for a set of variables that gives optimum to the described criterion. This problem becomes discrete-continuous. We load our problem into a continuous one, which is searching for a transformation of space of variables in such a way that feature selection and margin maximization will be done together. The new problem is also not good for its effective solution because it is not convex. We change it to a problem of successive minimization, which has the same global optimal solutions. The last problem is a problem of non-smoothed convex optimization, which we show using the dual form of this problem. Also the problem could be formulated as a saddle point search problem for a convex-concave function on a close convex compact. In subsection 3.1 we analyze properties of the solution given by the saddle point.

It is convenient to represent vectors x_i from learning sample $\{x_i, y_i\}$ in the form of data matrix D. D has l rows and n columns. Rows of the data matrix represent vectors x_i. Columns of the data matrix represent features.

Let $I = \{1; 2; 3; \ldots; n-1; n\}$- set of coordinates of vector v, $Q \subset I$ - subset of set of coordinates. Let's denote by v^Q - vector with set of coordinates Q, $v_i^Q = v_i$, $i \in Q$. We denote by v_Q the same vector to avoid complex notations. For example for $I = \{1; 2; 3; 4; 5\}, v = (10; 8; 4; 7; 6)^T, Q = \{2; 3; 5\}$, vector v^Q is equal to $(8; 4; 6)^T$.

We formulate the feature selection problem as modification of the problem (1):

$$\begin{gathered}\min_{Q \subseteq I, w^Q, b, \delta} \left(\frac{1}{2} \left\| w^Q \right\|^2 + C \sum_{i=1}^{l} \delta_i + A\left|Q\right| \right) \\ y_i(\langle w^Q, x_i^Q \rangle + b) \geq 1 - \delta_i, \\ \delta_i \geq 0, i = 1, ..., l, A > 0\end{gathered} \tag{4}$$

The term $A\left|Q\right|$ with the positive constant A is introduced in order to reduce the cardinality of the extreme subspace that we look for. It is easy to see that this discrete-continuous problem is very hard to be solved. Thats why let's extend this problem to its continuous analog. Let's consider the following problem:

$$\begin{gathered}\min_{z, w, b, \delta} \left(\frac{1}{2} \left\| w \right\|^2 + C \sum_{i=1}^{l} \delta_i + A \sum_{j=1}^{n} z_j \right) \\ y_i \left(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b \right) \geq 1 - \delta_i, \\ \delta_i \geq 0, i = 1, ..., l, A > 0, z_j \in \{0; 1\}, j = 1, ..., n\end{gathered} \tag{5}$$

x_i^j - is a coordinate j of vector x_i, w_j - is a coordinate j of vector w. This problem has boolean variables z_j. The value $z_j = 1$ means that feature with index j is chosen, the value $z_j = 0$ means that feature is deleted. Operation of taking square root of z_j is introduced for convenience.

Next proposition asserts that problems (4) and (5) are equivalent in the sense that from solution of one problem we easily get solution of other problem.

Proposition 1. *Let Q, w^Q, b, δ - solution of (4), values z, w in problem (5) calculated by the following rule:*

$$\begin{cases} z_j = 1, w_j = w_j^Q & , \text{ if } j \in Q, \\ z_j = 0, w_j = 0 & , \text{ if } j \notin Q. \end{cases} \tag{6}$$

then z, w, b, δ - solution of the problem (5).
Let z, w, b, δ - solution of (5), values of Q, w^Q calculated by the following rule:

$$Q = \{j | z_j = 1; j = 1, \ldots, n\}, w_j^Q = w_j, j \in Q \tag{7}$$

then Q, w^Q, b, δ *- solution of* (4)*.*

Proof of proposition 1 uses lemma 1 and omitted as elementary.

Variables z_j takes on values 0 or 1 in problem (5). Let's allow for variables z_j to take on values from segment $[0,1]$. In that way we come to continuous analog of the problem (5):

$$\begin{gathered}\min_{z,w,b,\delta}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\delta_i + A\sum_{j=1}^{n}z_j\right)\\ y_i(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b) \geq 1 - \delta_i,\\ \delta_i \geq 0, i = 1, ..., l, A > 0, z_j \in [0,1], j = 1, ..., n\end{gathered} \tag{8}$$

Problem (8) has non-convex constraints on variables z, w, b, δ. We can substitute (8) by another easier problem which solution is coincident to this one.

Let us formulate the new problem of sequential minimization:

$$\min_{0\leq z_j\leq 1, j=1,\ldots,n} \psi(z) \tag{9}$$

where the value of $\psi(z)$ is obtained from the following problem:

$$\begin{gathered}\psi(z) = \min_{w,b,\delta}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\delta_i + A\sum_{j=1}^{n}z_j\right)\\ y_i(\sum_{j=1}^{n} w_j x_i^j \sqrt{z_j} + b) \geq 1 - \delta_i,\\ \delta_i \geq 0, i = 1, ..., l, A > 0\end{gathered} \tag{10}$$

Let z^*-solution of (9) and minimum in (10), when we calculate value of $\psi(z^*)$, obtained on w^*, b^*, δ^*. Then z^*, w^*, b^*, δ^* is a solution of (8). Using transition from primal problem (1) to dual problem (2), we can write (10) in dual form:

$$\begin{gathered}\psi(z) = \max_{\lambda}\left(\sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{k=1}^{l}\left(y_i y_k \sum_{j=1}^{n} z_j x_i^j x_k^j\right)\lambda_i\lambda_k + A\sum_{j=1}^{n} z_j\right)\\ \sum_{i=1}^{l}\lambda_i y_i = 0, 0 \leq \lambda_i \leq C,\ i = 1, \ldots, l\end{gathered} \tag{11}$$

Very important property of function $\psi(z)$ described by the following theorem.

Theorem 1. *Function* $\psi(z)$ *from problem* (11) *is convex.*

Proof of this theorem follows from Theorem 5.5 of book [8].

So the problem (9) is a problem of minimization of non-smooth convex function on a convex set $0 \leq z_j \leq 1, j = 1, ..., n$. The algorithms for such optimization was analyzed in the [1].

Let's write function from problem (9)-(10) as function of two variables:

$$L(z, \lambda) = \sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{k=1}^{l} \left(y_i y_k \sum_{j=1}^{n} z_j x_i^j x_k^j \right) \lambda_i \lambda_k + A \sum_{j=1}^{n} z_j \tag{12}$$

, defined on cartezian product of compact sets:

$$Z = \{z | 0 \leq z_j \leq 1, j = 1, ..., n\} \tag{13}$$

$$\Lambda = \{\lambda | \sum_{i=1}^{l} \lambda_i y_i = 0, 0 \leq \lambda_i \leq C, \ i = 1, \ldots, l\} \tag{14}$$

Function $L(z, \lambda)$ is convex-concave function, i.e. convex by z for fixed value of λ and concave by λ for fixed z. Let's consider the problem of search of a saddle-point $(z^*, \lambda^*) \in Z \times \Lambda$:

$$L(z^*, \lambda) \leq L(z^*, \lambda^*) \leq L(z, \lambda^*), \forall z \in Z, \forall \lambda \in \Lambda. \tag{15}$$

For a saddle point the following equality is true:

$$\min_{z \in Z} \max_{\lambda \in \Lambda} L(z, \lambda) = \max_{\lambda \in \Lambda} \min_{z \in Z} L(z, \lambda) = L(z^*, \lambda^*) \tag{16}$$

Guided by this equality we can substitute minimax problem (9), (11) on saddle point search problem (15).

Remark 2. *We should emphasize that if we have (z^*, λ^0) - solution of the following minimax problem*

$$\min_{z \in Z} \max_{\lambda \in \Lambda} L(z, \lambda) \tag{17}$$

, then we can assert that there is λ^* such that (z^*, λ^*) is a saddle point of function $L(z, \lambda)$. In other words, minimax solution allows to find only component z of a the saddle point and (z^*, λ^0) not necessarily is a saddle point.

Solution (z^*, λ^0) of minimax problem (17) is a saddle point if subproblem

$$\max_{\lambda \in \Lambda} L(z^*, \lambda) \tag{18}$$

has unique solution. Set of solutions of the problem (18) contains point λ^*, such, that (z^*, λ^*) is a saddle point.

3.1 Saddle Point Properties.

Theorem below states existence of saddle point (z^*, λ^*) of function (12) and describes conditions - when coordinates of z^* - are equal to 0 and 1.

Theorem 2. *1.There exists saddle point (z^*, λ^*) in problem (15) and the following equalities are true:*

$$\{z^* | (z^*, \lambda)\textbf{\textit{-saddle point}}\} = \arg\min_{z \in Z} \max_{\lambda \in \Lambda} L(z, \lambda) \tag{19}$$

$$\{\lambda^* | (z, \lambda^*)\textbf{\textit{-saddle point}}\} = \arg\max_{\lambda \in \Lambda} \min_{z \in Z} L(z, \lambda) \tag{20}$$

2. If in saddle point the following inequality holds $\sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* > 2A$, then $z_j^* = 1$. If in saddle point the following inequality holds $\sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* < 2A$, then $z_j^* = 0$. If $0 < z_j^* < 1$, then

$$\sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* = 2A \tag{21}$$

Proof.

1. According to well known theorem (see, for example, [8]), continuous convex-concave function defined on product of compact convex sets has saddle point. All conditions of the theorem are satisfied for function $L(z, \lambda)$ and hence exists saddle point (z^*, λ^*). Equalities (19), (20) are properties of saddle point from remark (2).

2. $L(z^*, \lambda^*) = \max_\lambda \min_z L(z, \lambda) = \min_z L(z, \lambda^*)$. For fixed λ function $L(z, \lambda)$ is linear by z. Coefficient of function $L(z, \lambda^*)$ for z_j is equal to $A - \frac{1}{2} \sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^*$. If the coefficient is positive then minimum by z_j achieved on $z_j = 0$. If the coefficient is negative then minimum by z_j achieved on $z_j = 1$. Coordinate z_j can differ from 1 or 0 only if coefficient for z_j is equal to zero. By the following property of a saddle point: $z^* \in \arg_z \min L(z, \lambda^*)$, we prove the second part of theorem.

□

One could think that in case of the equality

$$\sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^* = 2A \ ,$$

we can get rid of it by small variation of parameter A or data from training set.

Below is an example where equality remains true for small variation of parameter A.

Example 1. *Let training set consist of two one-dimensional vectors. First vector belong to first class (+1 class) and has the value of single feature equal to 1. Second vector belong* *to second class (-1 class) and has the value of feature equal to* 0*. It's easy to show that for fixed z, problem* (11) *has solutions that depend on value of z:*

$$\psi(z) = \begin{cases} (A - 2C^2)z + 2C & 0 \le z \le \frac{1}{2C}, \\ \frac{1}{2z} + Az & \frac{1}{2C} \le z \le 1. \end{cases}$$

If the following equalities are true: $A < 2C^2 \quad A > \frac{1}{2}$ *then minimum point of* $\psi(z)$ *is* $z^* = \frac{1}{\sqrt{2A}}$*. We can see that inequalities* $0 < z^* < 1$ *remain for small variations of parameter A. Hence, by theorem 1, equality* (21) *holds true for small variations of parameter A also.*

It's possible to generalize above example on case of linear separable classes of one-dimensional vectors where equality (21) holds true for small variations of training data.

Next section contains description of algorithm for saddle point search that we propose to use to solve problem (15).

4 Saddle Point Algorithm

In the previous section we formulated a problem of learning classification, which simultaneously maximizes margin and minimizes feature space as a saddle point problem. In this section we describe an algorithm for saddle point search.

Let convex-concave function $L(z, \lambda)$ defined on product of convex closed sets $Z \times \Lambda$ and the following inequalities true for some positive constant $M > 0$:

$$\left| L(z, \lambda + h) - L(z, \lambda) - \left\langle \frac{\partial L}{\partial \lambda}(z, \lambda), h \right\rangle \right| \le \frac{1}{2} M \left\| h \right\|^2 \tag{22}$$

$$\left| L(z + h, \lambda) - L(z, \lambda) - \left\langle \frac{\partial L}{\partial z}(z, \lambda), h \right\rangle \right| \le \frac{1}{2} M \left\| h \right\|^2 \tag{23}$$

$$\left\| \frac{\partial L}{\partial z}(z, \lambda + h) - \frac{\partial L}{\partial z}(z, \lambda) \right\| \le M \left\| h \right\| \tag{24}$$

Let π_Z,π_Λ - are projection operators on sets Z and Λ, i.e. $\pi_Z(z)$ - projection of point z on set Z and $\pi_\Lambda(\lambda)$ - projection of point λ on set Λ.

Let's consider the following algorithm, each iteration of which consists of 3 steps:

$$\begin{aligned} \bar{z}_n &= \pi_Z \left(z_n - \alpha \frac{\partial L}{\partial z} \left(z_n, \lambda_n \right) \right) \\ \lambda_{n+1} &= \pi_\Lambda \left(\lambda_n + \alpha \frac{\partial L}{\partial \lambda} \left(\bar{z}_n, \lambda_n \right) \right) \\ z_{n+1} &= \pi_Z \left(z_n - \alpha \frac{\partial L}{\partial z} \left(z_n, \lambda_{n+1} \right) \right) \end{aligned} \tag{25}$$

The following theorem is true.

Theorem 3. *Let $L(z, \lambda)$-convex-concave function on $Z \times \Lambda$, sets Z and Λ-convex and closed, function $L(z, \lambda)$ satisfy to inequalities* (22)-(24)*, for step size the following inequality true $\alpha < \frac{1}{2M}$. Then for any initial point $z_0 \in Z$, $\lambda_0 \in \Lambda$ sequence $(z_n,\ \lambda_n), n = 1, 2, ..,$ calculated by formulas* (25)*, converges to $(z^*,\ \lambda^*)$-saddle point of function $L(z, \lambda)$.*

Proof of the theorem is given in Appendix.

Lets show that function $L(z, \lambda)$ from (12)-(14) satisfy inequalities (22)-(24). Lets make estimation of Lipschitz constant for quadratic function on bounded set.

Let $\|x\| \le T_1$, $\|y\| \le T_2$.

$$\begin{aligned}\|\langle Bx, x\rangle - \langle By, y\rangle\| &= \|\langle Bx, x\rangle - \langle Bx, y\rangle + \langle Bx, y\rangle - \langle By, y\rangle\| = \\ &\|\langle Bx, x-y\rangle + \langle B(x-y), y\rangle\| \le \|\langle Bx, x-y\rangle\| + \|\langle B(x-y), y\rangle\| \le \\ &\|Bx\|\,\|x-y\| + \|B(x-y)\|\,\|y\| \le \|B\|\,\|x\|\,\|x-y\| + \|B\|\,\|x-y\|\,\|y\| = \\ &\|B\|\,(\|x\| + \|y\|)\,\|x-y\| \le \|B\|\,(T_1+T_2)\,\|x-y\| \quad (26)\end{aligned}$$

Let G is matrix obtained from data matrix by multiplying each i-row on y_i . Then matrix of quadratic form in problem (2) is equal to GG^T. Let for vector z we defined the following diagonal matrix:

$$Z = \begin{pmatrix} \sqrt{z_1} & 0 & \cdots & 0 \\ 0 & \sqrt{z_2} & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{z_M} \end{pmatrix}$$

Then for fixed value of z the new data matrix obtained by multiplication of columns on corresponding components of vector z:

$$F = GZ$$

So we can write the following formula for R-matrix of quadratic form of function $L(z, \lambda)$:

$$R = FF^T = GZZ^TG^T = GZ^2G^T$$

and write estimate for norm of this matrix independent from z:

$$\|R\| \le \|G\|\,\|Z^2\|\,\|G^T\| \le \|G\|\,\|G^T\|$$

We used the fact that all components of z are less than 1.

$$\|Z^2\| \le 1$$

For fixed z function $L(z, \lambda)$ is quadratic with matrix of quadratic form equal to R, and hence we get estimate (22):

$$\left| L(z, \lambda + h) - L(z, \lambda) - \left\langle \frac{\partial L}{\partial \lambda}(z, \lambda), h \right\rangle \right| = (Rh, h) \le \|R\|\,\|h\|^2$$

For fixed λ, function $L(z,\lambda)$ is linear function, and we get estimate (23):

$$L(z+h,\lambda)-L(z,\lambda)-\left\langle \frac{\partial L}{\partial z}(z,\lambda),h\right\rangle = 0$$

To get estimate (24) let's write formula for partial derivative of $L(z,\lambda)$ by component of vector z:

$$\frac{\partial L}{\partial z_j}(z,\lambda) = -\frac{1}{2}\sum_{i=1}^{l}\sum_{k=1}^{l}(y_i y_k x_i^j x_k^j)\lambda_i\lambda_k + A \tag{27}$$

Let R^j - matrix of quadratic form in formula for component j of partial derivative, i.e.

$$R_{ik}^j = y_i y_k x_i^j x_k^j$$

Finally we get estimate (24):

$$\begin{aligned}\left\|\frac{\partial L}{\partial z}(z,\lambda+h)-\frac{\partial L}{\partial z}(z,\lambda)\right\| \ &\le\ \sum_{j=1}^{n}\left\|\frac{\partial L}{\partial z_j}(z,\lambda+h)-\frac{\partial L}{\partial z_j}(z,\lambda)\right\| \le \\ &\quad \left(\frac{1}{2}\left(\|\lambda+h\|+\|\lambda\|\right)\sum_{j=1}^{n}\left\|R^j\right\|\right)\|h\| \\ &\le\ \left(Cl\sum_{j=1}^{n}\left\|R^j\right\|\right)\|h\|\end{aligned}$$

We used estimate (26) and boundedness of components of λ and $\lambda+h$:

$$0\le\lambda_i\le C,\quad 0\le\lambda_i+h_i\le C,\quad i=1,\ldots,l.$$

For completeness let's write formula for partial derivative by λ (25):

$$\frac{\partial L}{\partial \lambda_i}(\lambda,z) = 1-\sum_{k=1}^{l}\left(y_i y_k\sum_{j=1}^{n} z_j x_i^j x_k^j\right)\lambda_k \tag{28}$$

5 Calculation of projections

Algorithm (25) requires to calculate projections of points on sets Z and Λ.

Projection of point $\widehat{z}$ on Z: $z^{pr}=\pi_Z(\widehat{z})$, $Z=\{0\le z_j\le 1, j=1,\ldots,n\}$ calculated by formula:

$$z_j^{pr} = \begin{cases} 0 & \widehat{z}_j<0 \\ z_j & 0\le\widehat{z}_j\le 1\ , j=1,\ldots,n \\ 1 & \widehat{z}_j>1\end{cases} \tag{29}$$

Projection of point $\widehat{\lambda}$ on set Λ: $\lambda^{pr} = \pi_\Lambda(\widehat{\lambda}), \Lambda = \{\sum_{j=1}^{l} \lambda_j y_j = 0, 0 \leq \lambda_j \leq C, j = 1, \dots, l\}$, $\widehat{\lambda} \in R^l$ calculated by successive projections on cube $\{0 \leq \lambda_j \leq C, j = 1, \dots, l\}$ and hyperplane $\{\sum_{j=1}^{l} \lambda_j y_j = 0\}$.

Let's remind that projection of point $x_0 \in R^M$ on hyperplane $c^T x = 0, c \in R^M$ calculated by formula:

$$x_0^{pr} = x_0 - \frac{c^T x_0}{c^T c} c \tag{30}$$

So we have simple formulas to find projection on cube and hyperplane. We need to calculate projection on their intersection.

There exists a class of algorithms for calculation projection on intersection of sets using projections on each of the set in intersection. One of the first algorithm from this class is Dykstra's Alternating Projection Algorithm. Results of comparisons of Dykstra's alternating projection algorithm with analogous algorithms reported in [10].

Let's describe Dykstra's algorithm according to [9]. Let A, B convex closed subsets of R^l, point $x \in R^l$. Let for $n \geq 1$ the following consequences defined:

$$\begin{aligned} b_0 &= x & p_0 &= q_0 = 0 \\ a_n &= \pi_A(b_{n-1} + p_{n-1}) & p_n &= b_{n-1} + p_{n-1} - a_n \\ b_n &= \pi_B(a_n + q_{n-1}) & q_n &= a_n + q_{n-1} - b_n \end{aligned} \tag{31}$$

Consequences a_n, b_n converge to $\pi_{A \cap B}(x)$ - projection of x on intersection of sets $A \cap B$. Elements of consequence a_n belong to set A, and elements b_n belong to set B.

Lets notice that consequences a_n, b_n, calculated by the following formulas:

$$\begin{aligned} b_0 &= x \\ a_n &= \pi_A(b_{n-1}) \\ b_n &= \pi_B(a_n) \end{aligned} \tag{32}$$

converge to some point from intersection $A \cap B$. So elements p_n q_n in formulas (31) provide convergence to projection on $A \cap B$. Elements p_n q_n could be thought of as elements of memory about vector x. Indeed according to [9] the following equalities hold:

$$\begin{aligned} x &= a_n + p_n + q_{n-1} \\ x &= b_n + p_n + q_n \end{aligned} \tag{33}$$

Let's notice that in case sets A B are not intersected and the value of distance $d(A, B)$ between sets A and B achieved, i.e.:

$$d(A, B) = inf(\|a - b\| : a \in A, b \in B) \tag{34}$$

$$\exists a^* \in A, \exists b^* \in B, d(A, B) = \|a^* - b^*\| \ , \tag{35}$$

then consequences a_n, b_n convergence to nearest points of sets A and B:

$$a_n \rightarrow a^1 \in A \tag{36}$$

$$b_n \rightarrow b^1 \in B \tag{37}$$

$d(A,B) = \|a^1 - b^1\|$.

If set A is affine subspace(shift of subspace on vector) then we can set $p_n = 0$ for all n in formulas (31), see [11].

Finally algorithm for finding projection of a point x on set $\Lambda = \{\sum_{j=1}^{l} \lambda_j y_j = 0, 0 \leq \lambda_j \leq C, j = 1, \dots, l\}$ is:

$$\begin{aligned} b_0 &= x & q_0 &= 0 \\ a_n &= \pi_A(b_{n-1}) & & \\ b_n &= \pi_B(a_n + q_{n-1}) & q_n &= a_n + q_{n-1} - b_n \end{aligned} \tag{38}$$

where sets A and B are: $A = \{\sum_{j=1}^{l} \lambda_j y_j = 0\}$, $B = \{0 \leq \lambda_j \leq C, j = 1, \dots, l\}$ and projections on sets A and B calculated by formulas (30) and (29).

6 Experimental Results

6.1 Experimental Framework

The proposed algorithm is time consuming. It works rather fast on data arrays with ~ 30 observations (calculation time < 1 sec), but it takes about an hour to process real data arrays with ~ 1000 observations. Therefore we divide initial data set into two parts. The first part is a training data set. The second part is a testing data set. The quality of classification rule that was built on training data set is tested on testing data set. Initial data set is represented by initial data matrix with N rows and $n+1$ columns. Row with number i contains information about class membership and values of coordinates of vector x_i. First column contains class membership sign: $y_i \in \{+1; -1\}$. We compare results of classification by decision rules obtained by feature selection algorithm and standard SVM algorithm.

Below is the scheme of algorithm testing:

Step 1. Let integer value l is randomly uniform distributed on interval $[15, 30]$. Take realization of random variable l. Then l-rows are taken randomly from initial data matrix in a way to preserve approximately proportion of members of classes as in initial data set. These rows form data matrix in problem (9)-(10). The rest $(N - l)$ rows of initial data matrix form the second (testing) data set. Thus each matrix containes l objects and all n variables.

Step 2. Parameter C is chosen according to [13], in the following way. Let's calculate for all vectors of training set the following value:

$$C_0 = \left(\frac{1}{l} \sum_{i=1}^{l} \|x_i\| \right)^{-1} \tag{39}$$

For $C = C_0$ the problem is solved (1). Let S- set of support vectors in problem (1) [1]. Calculate the following value of parameter C:

$$C = \left(\frac{1}{|S|} \sum_{i \in S} \|x_i\| \right)^{-1} \tag{40}$$

Obtained value of parameter C is used in problem (9)-(10).

Step 3. Solve problem (1), where parameter C is calculated by formula (40). Decision rule that we obtain on that step is compared with decision rule obtained in feature selection algorithm.

Step 4. We set variation interval $[A_{min}, A_{max}]$ for parameter A. Feature selection problem is solved for various values of A taken from variation interval $[A_{min}, A_{max}]$. The bounds of interval $[A_{min}, A_{max}]$ are empirical values and are chosen in such way that solutions of feature selection problem are not trivial. For small values of A all features are selected while for large values of A the set of selected features are empty. Based on equality (21), the values A_{min} and A_{max} are calculated in the following way:

a. Take (z^*, λ^*) - solution of the problem (9)-(10) where $A = 0$.

b. Calculate $s_j = \sum_{i=1}^{l} \sum_{k=1}^{l} y_i y_k x_i^j x_k^j \lambda_i^* \lambda_k^*$, $j = 1, \ldots, n$.

c. Calculate $\hat{A}_{min} = 0,5 \min_{j=1,\ldots,n} s_j$.

d. Calculate $\hat{A}_{max} = 0,5 \max_{j=1,\ldots,n} s_j$.

As starting point obtained above $\hat{A}_{min}$ and $\hat{A}_{max}$ are taken as bounds of variation interval for parameter A. In general bounds of variation interval are calculated as function of the values s_j.

Step 5. Take value of A from $[A_{min}, A_{max}]$ and solve saddle point search problem. Saddle point training algorithm found $(\hat{z}, \hat{\lambda})$ -approximation to saddle point (z^*, λ^*).

Step 6. The following 3 groups of features defined:

a. Deleted features - those numbers of coordinates i, for which the equality $\hat{z}_i = 0$ holds.

b. Selected features - those numbers of coordinates i, for which the equality $\hat{z}_i = 1$ holds.

c. Weighted features - those numbers of coordinates i, for which the inequalities $0 < \hat{z}_i < 1$ hold.

Step 7. For the first (training) and second (testing) parts of initial data set the following 3 recognition quality evaluations for SVM were calculated: :

[1] See definition of support vectors in the end of section 2 of current report

a. $Accuracy = 100\% * (T_1 + T_2) / l$

b. $Precision = 100\% * T_1/N_1$

c. $Recall = 100\% * T_1/ (T_1 + T_2)$,

where T_1-number of true classified vectors x_i-objects of first class, T_2-number of true classified vectors x_i-objects of second class, N_1-overall number of objects in first class, l - overall number of objects in both classes.

Step 8. The same set of 3 evaluations is calculated for the solution of standard SVM problem, obtained on the Step 3.

6.2 Benchmark

We applied testing scheme to various experimental data sets. We describe results of testing feature selection algorithm to data set that characterize 11 vowels of English language.

Data set is available from website: http://www-stat-class.Stanford.EDU/ tibs/ElemStatLearn/, in section DATA and has name VOWEL. The features are derived from analysis of sample windowed segments of the speech signal and are real-valued. We extracted only the first two classes from this data set. Initial data set has 180 elements.

There were generated 20 data matrices according to step 1 of the testing scheme. Variation interval $[A_{min}, A_{max}]$ for parameter A, calculated on step 4, is equal to $[0, 17\hat{A}_{min}, 0, 6\hat{A}_{max}]$. Feature selection problem was solved for 31 values of parameter A uniformly distributed on variation interval. There was selected a solution with maximal value of *Accuracy* from 31 solutions.

Testing results presented in table 1. For each data matrix presented quality of results for solution with maximal value of *Accuracy*. Last two columns of the table contain number of deleted and selected features. For comparison of recognition quality the recognition quality of solution of standard SVM problem (1) is given.

Results, presented in table 1, show that in 15 cases of 20 cases quality characteristic *Accuracy* is greater than *Accuracy* of solutions obtained by standard SVM. For data matrices with numbers 5,13,15 and 20 the values of *Accuracy* are equal for feature selection solution and SVM. The only case when SVM solution has greater value of *Accuracy* is the case of matrix number 18. There were deleted features in 14 cases of 20 cases. For example, 7 features from 10 were deleted for matrix number 4 and the value of *Accuracy* is $80, 13$ while *Accuracy* of SVM solution is $68, 59$.

So results of experiments show that feature selection algorithm can delete features and in the same time improve quality of classification. Moreover we can see from the table that in almost all cases when there were deleted features in solution of feature selection problem the value of *Accuracy* is larger than value of *Accuracy* for SVM solution. The only case when it is not true is case of matrix number 20.

Table shows that even in case of such simple choice of parameter A the algorithm does feature selection without big losses in Accuracy.

Table 1:

Matrix number	Number of objects (l)	Max. Accuracy	Accuracy SVM	Precision	Precision SVM	Recall	Recall SVM	Number of $z_i = 0$	Number of $z_i = 1$
1	18	79,63	75,31	97,53	95,06	61,24	63,11	4	3
2	30	80,67	75,33	81,33	76,00	50,41	50,44	5	2
3	20	53,13	52,50	22,62	25,00	22,35	25,00	3	3
4	24	80,13	68,59	92,31	70,51	57,60	51,40	7	0
5	15	50,30	50,30	4,65	4,65	4,82	4,82	0	6
6	29	77,48	74,17	84,42	84,42	55,56	58,04	4	1
7	21	84,28	80,50	98,72	93,59	57,46	57,03	5	3
8	24	67,95	67,31	51,28	51,28	37,74	38,10	0	9
9	26	72,08	68,83	100	100,0	65,77	68,87	6	3
10	21	73,58	64,78	65,43	65,43	45,30	51,46	6	1
11	28	81,58	75,66	95,89	84,93	56,45	53,91	6	1
12	24	77,56	76,28	74,07	67,90	49,59	46,22	0	6
13	15	68,94	68,94	100	100,0	71,17	71,17	0	7
14	19	77,50	71,88	75,31	64,20	49,19	45,22	4	3
15	20	76,32	76,32	74,68	74,68	50,86	50,86	0	5
16	28	76,40	72,05	79,27	89,02	52,85	62,93	6	2
17	19	75,15	74,55	75,90	74,70	50,81	50,41	3	4
18	15	82,89	84,21	84,81	83,54	53,17	51,56	0	8
19	24	79,49	76,92	82,05	80,77	51,61	52,50	6	1
20	21	58,49	58,49	84,62	84,62	70,97	70,97	2	6

7 Acknowledgements

Ilya Muchnik thanks to NSF for partial support of this work with grant # 0325398.

References

[1] Goncharov Yu., Muchnik I., Shvartser L. ,*Simultaneous Feature Selection and Margin Maximization Using Saddle Point Approach*, DIMACS Technical Report 2004-08, April 2004

[2] J. Bi and V. Vapnik ,*Learning with Rigorous Support Vector Machines*,COLT 2003: Conference on Learning Theory, Washington D.C. August 24-27 2003.

[3] J. Bi, *Multi-Objective Programming in SVMs*, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC 2003.

[4] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, *Feature Selection for SVMs.*,Advances in Neural Information Proccesing Systems, No. 13, MIT Press 668-674, 2000.

[5] J. Weston et. al., *Feature Selection and Transduction for Prediction of Molecular Bioactivity for Drug Design*,Bioinformatics Vol. 1, No. 1 2002: 1-8

[6] J. Bi et al., *Dimensionality Reduction via Sparse SVMs*,Journal of Machine Learning Research, No. 1:1-48, 2002.

[7] L. Molina, L. Belanche and A. Nebot, *Feature Selection Algorithms: A Survey and Experimental Evaluation*, IEEE Internat. Conf. on Data Mining, Japan.

[8] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970

[9] Bauschke H.H., Borwein J.M.,*Dykstra's Alternating Projection Algorithm for Two Sets*, Journal of Approximation Theory 79,418-443, 1994.

[10] Censor Y.,*Computational Acceleration of Projection Algorithms for the Linear Best Approximation Problem*, Technical report, Department of Mathematics, University of Haifa, Israel, May, 2005.

[11] N. Gaffke, R. Mathar,*A cyclic projection algorithm via duality*, Metrika 36 (1989),pp 29-54

[12] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 1995,New York

[13] Andrei V. Anghelescu and Ilya B. Muchnik *Optimization of SVN in a Space of Two Parameters: Weak Margin and Intercept*, DIMACS Working Group on Monitoring Message Streams, May 2003.

[14] B.T.Polyak *Introduction to optimization*, Moscow (in Russian), 1983.

8 Appendix/Proof of theorem 3.

Let π_Z,π_Λ - projection operator on sets Z and Λ, i.e. $\pi_Z(z)$ - projection of z on set Z and $\pi_\Lambda(\lambda)$ - projection of λ on set Λ. Projection operator π_X on convex set X has the property not to increase distance:

$$\|\pi_X(x) - \pi_X(y)\| \leq \|x - y\| \tag{41}$$

and the following property:

$$\langle \pi_X(x) - x, y - \pi_X(x) \rangle \geq 0, \forall y \in X \tag{42}$$

As $L(z, \lambda)$ is a convex-concave, continuous function, Z Λ are closed, convex sets then there exist saddle point (z^*, λ^*) of function $L(z, \lambda)$.

Necessary and sufficient conditions for minimum of convex by z function $L(z, \lambda^*)$ in point z^* and for maximum of concave by λ function $L(z^*, \lambda)$ in λ^* are:

$$\begin{aligned} z^* &= \pi_Z \left(z^* - \alpha \frac{\partial L}{\partial z}(z^*, \lambda^*) \right) & (43) \\ \lambda^* &= \pi_\Lambda \left(\lambda^* + \alpha \frac{\partial L}{\partial \lambda}(z^*, \lambda^*) \right) & (44) \\ \alpha > 0 & \end{aligned}$$

Let's notice that if (43)-(44) is true for some $\alpha > 0$, then (43)-(44) is true for all $\alpha > 0$.

From equalities (25) that describe an iteration of algorithm and property of project operator (42) we get the following three inequalities:

$$\left\langle \bar{z}_n - z_n + \alpha \frac{\partial L}{\partial z}(z_n, \lambda_n), z - \bar{z}_n \right\rangle \geq 0, \forall z \in Z \tag{45}$$

$$\left\langle \lambda_{n+1} - \lambda_n - \alpha \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda - \lambda_{n+1} \right\rangle \geq 0, \forall \lambda \in \Lambda \tag{46}$$

$$\left\langle z_{n+1} - z_n + \alpha \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z - z_{n+1} \right\rangle \geq 0, \forall z \in Z \tag{47}$$

Let's substitute $z = z_{n+1}$, $\lambda = \lambda^*$ and $z = z^*$ correspondingly in inequalities (45), (46) and (47)

$$\left\langle \bar{z}_n - z_n + \alpha \frac{\partial L}{\partial z}(z_n, \lambda_n), z_{n+1} - \bar{z}_n \right\rangle \geq 0 \tag{48}$$

$$\left\langle \lambda_{n+1} - \lambda_n - \alpha \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda^* - \lambda_{n+1} \right\rangle \geq 0 \tag{49}$$

$$\left\langle z_{n+1} - z_n + \alpha \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z^* - z_{n+1} \right\rangle \geq 0 \tag{50}$$

As function $L(\bar{z}_n, \lambda)$ - concave by λ, we have:

$$\left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_n - \lambda^* \right\rangle \leq L(\bar{z}_n, \lambda_n) - L(\bar{z}_n, \lambda^*) \tag{51}$$

Let's rewrite (49) as:

$$\langle \lambda_{n+1} - \lambda_n, \lambda^* - \lambda_{n+1} \rangle + \alpha \left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda^* \right\rangle \geq 0 \tag{52}$$

Let's estimate inner product in the second member of inequality (52), taking in account (51):

$$\begin{aligned} \left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda^* \right\rangle &= \left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda_n \right\rangle + \left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_n - \lambda^* \right\rangle \\ &\leq \left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda_n \right\rangle + L(\bar{z}_n, \lambda_n) - L(\bar{z}_n, \lambda^*) \\ &= L(\bar{z}_n, \lambda_{n+1}) - L(\bar{z}_n, \lambda^*) + \left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda_n \right\rangle + L(\bar{z}_n, \lambda_n) - L(\bar{z}_n, \lambda_{n+1}) \end{aligned} \tag{53}$$

From (22) we get:

$$\left| \left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda_n \right\rangle + L(\bar{z}_n, \lambda_n) - L(\bar{z}_n, \lambda_{n+1}) \right| \leq \frac{1}{2} M \left\| \lambda_{n+1} - \lambda_n \right\|^2 \tag{54}$$

Using (54) and estimation (53), we get inequality:

$$\left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda^* \right\rangle \quad \leq \quad L(\bar{z}_n, \lambda_{n+1}) \quad - \quad L(\bar{z}_n, \lambda^*) \quad + \quad \frac{1}{2} M \left\| \lambda_{n+1} - \lambda_n \right\|^2 \tag{55}$$

The next expression follows from the fact that (z^*, λ^*) is a saddle point:

$$L(\bar{z}_n, \lambda^*) \geq L(z^*, \lambda^*) \geq L(z^*, \lambda_{n+1}) \tag{56}$$

From (56) and (55) we get:

$$\left\langle \frac{\partial L}{\partial \lambda}(\bar{z}_n, \lambda_n), \lambda_{n+1} - \lambda^* \right\rangle \quad \leq \quad L(\bar{z}_n, \lambda_{n+1}) \quad - \quad L(z^*, \lambda_{n+1}) \quad + \quad \frac{1}{2} M \left\| \lambda_{n+1} - \lambda_n \right\|^2 \tag{57}$$

From (52) and estimate (57) we have inequality we need further and that gives relation between saddle value of variable λ and value of λ in iteration of algorithm:

$$\langle \lambda_{n+1} - \lambda_n, \lambda^* - \lambda_{n+1} \rangle + \alpha \left[L(\bar{z}_n, \lambda_{n+1}) - L(z^*, \lambda_{n+1}) + \frac{1}{2} M \left\| \lambda_{n+1} - \lambda_n \right\|^2 \right] \geq 0 \qquad (58)$$

Let's add (48) and (50), write the sum as following:

$$\langle \bar{z}_n - z_n, z_{n+1} - \bar{z}_n \rangle + \langle z_{n+1} - z_n, z^* - z_{n+1} \rangle + \\ \alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_n), z_{n+1} - \bar{z}_n \right\rangle + \alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z^* - z_{n+1} \right\rangle \geq 0 \qquad (59)$$

Let's get estimate for third member of inequality (59).

At first estimate the value $\|\bar{z}_n - z_{n+1}\|$. By property of projection operator (41) and according (24) we have:

$$\|\bar{z}_n - z_{n+1}\| \leq \alpha \left\| \frac{\partial L}{\partial z}(z_n, \lambda_n) - \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}) \right\| \leq \alpha M \left\| \lambda_n - \lambda_{n+1} \right\| \qquad (60)$$

From (24) and (60), we have the following estimate:

$$\alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_n), z_{n+1} - \bar{z}_n \right\rangle = \alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_n) - \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z_{n+1} - \bar{z}_n \right\rangle \\ + \alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z_{n+1} - \bar{z}_n \right\rangle \leq \alpha^2 M^2 \left\| \lambda_n - \lambda_{n+1} \right\|^2 + \alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z_{n+1} - \bar{z}_n \right\rangle \qquad (61)$$

Applying estimate (61) in inequality (59), we get the following inequality:

$$\langle \bar{z}_n - z_n, z_{n+1} - \bar{z}_n \rangle + \alpha^2 M^2 \left\| \lambda_n - \lambda_{n+1} \right\|^2 \qquad (62) \\ + \langle z_{n+1} - z_n, z^* - z_{n+1} \rangle + \alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z^* - \bar{z}_n \right\rangle \geq 0$$

As $L(z, \lambda)$ - convex by z, we have:

$$L(z^*, \lambda_{n+1}) \geq L(z_n, \lambda_{n+1}) + \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z^* - z_n \right\rangle \qquad (63)$$

Let's multiply this inequality on α and add to (62):

$$\langle \bar{z}_n - z_n, z_{n+1} - \bar{z}_n \rangle + \alpha^2 M^2 \left\| \lambda_n - \lambda_{n+1} \right\|^2 + \langle z_{n+1} - z_n, z^* - z_{n+1} \rangle \qquad (64) \\ + \alpha \left[L(z^*, \lambda_{n+1}) - L(z_n, \lambda_{n+1}) \right] + \alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z_n - \bar{z}_n \right\rangle \geq 0$$

Let's add obtained inequality to (58), we have:

$$\begin{aligned} \langle \bar{z}_n - z_n, z_{n+1} - \bar{z}_n \rangle + \alpha^2 M^2 \left\| \lambda_n - \lambda_{n+1} \right\|^2 \\ + \langle z_{n+1} - z_n, z^* - z_{n+1} \rangle - \alpha L(z_n, \lambda_{n+1}) \\ +\alpha \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z_n - \bar{z}_n \right\rangle + \langle \lambda_{n+1} - \lambda_n, \lambda^* - \lambda_{n+1} \rangle \\ +\alpha \left[L(\bar{z}_n, \lambda_{n+1}) + \frac{1}{2} M \left\| \lambda_{n+1} - \lambda_n \right\|^2 \right] \geq 0 \end{aligned} \tag{65}$$

Let's estimate the following expression with (23):

$$\begin{aligned} -L(z_n, \lambda_{n+1}) + \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), z_n - \bar{z}_n \right\rangle = \\ -L(\bar{z}_n, \lambda_{n+1}) + L(\bar{z}_n, \lambda_{n+1}) - L(z_n, \lambda_{n+1}) - \\ \left\langle \frac{\partial L}{\partial z}(z_n, \lambda_{n+1}), \bar{z}_n - z_n \right\rangle \leq -L(\bar{z}_n, \lambda_{n+1}) + \frac{1}{2} M \left\| \bar{z}_n - z_n \right\|^2 \end{aligned} \tag{66}$$

From (65) and (66) we get:

$$\begin{aligned} \langle \bar{z}_n - z_n, z_{n+1} - \bar{z}_n \rangle + \left(\frac{1}{2} \alpha M + \alpha^2 M^2 \right) \left\| \lambda_n - \lambda_{n+1} \right\|^2 \\ + \langle z_{n+1} - z_n, z^* - z_{n+1} \rangle + \frac{1}{2} \alpha M \left\| \bar{z}_n - z_n \right\|^2 \\ + \langle \lambda_{n+1} - \lambda_n, \lambda^* - \lambda_{n+1} \rangle \geq 0 \end{aligned} \tag{67}$$

From the next equality:

$$\left\| y_1 - y_2 \right\|^2 = \left\| y_1 - y_3 + y_3 - y_2 \right\|^2 = \left\| y_1 - y_3 \right\|^2 + \left\| y_3 - y_2 \right\|^2 + 2 \langle y_1 - y_3, y_3 - y_2 \rangle$$

we have:

$$\begin{gathered} \langle \lambda_{n+1} - \lambda_n, \lambda^* - \lambda_{n+1} \rangle = \frac{1}{2} \left[\left\| \lambda^* - \lambda_n \right\|^2 - \left\| \lambda_n - \lambda_{n+1} \right\|^2 - \left\| \lambda^* - \lambda_{n+1} \right\|^2 \right] \\ \langle \bar{z}_n - z_n, z_{n+1} - \bar{z}_n \rangle = \frac{1}{2} \left[\left\| z_n - z_{n+1} \right\|^2 - \left\| z_n - \bar{z}_n \right\|^2 - \left\| \bar{z}_n - z_{n+1} \right\|^2 \right] \\ \langle z_{n+1} - z_n, z^* - z_{n+1} \rangle = \frac{1}{2} \left[\left\| z_n - z^* \right\|^2 - \left\| z_n - z_{n+1} \right\|^2 - \left\| z_{n+1} - z^* \right\|^2 \right] \end{gathered} \tag{68}$$

Let's multiply (67) on 2 and use equalities (68):

$$\begin{aligned} &\|z_n - z_{n+1}\|^2 - \|z_n - \bar{z}_n\|^2 - \|\bar{z}_n - z_{n+1}\|^2 + \left(\alpha M + 2\alpha^2 M^2\right) \|\lambda_n - \lambda_{n+1}\|^2 \\ &\quad + \|z_n - z^*\|^2 - \|z_n - z_{n+1}\|^2 - \|z_{n+1} - z^*\|^2 + \alpha M \|\bar{z}_n - z_n\|^2 \\ &\quad + \|\lambda^* - \lambda_n\|^2 - \|\lambda_n - \lambda_{n+1}\|^2 - \|\lambda^* - \lambda_{n+1}\|^2 \geq 0 \end{aligned} \tag{69}$$

Let's reduce homogeneous members:

$$\begin{aligned} &\|z_n - z^*\|^2 + \|\lambda^* - \lambda_n\|^2 \geq \|\lambda^* - \lambda_{n+1}\|^2 + \|z_{n+1} - z^*\|^2 \\ &+ \left(1 - \alpha M - 2\alpha^2 M^2\right) \|\lambda_n - \lambda_{n+1}\|^2 + (1 - \alpha M) \|\bar{z}_n - z_n\|^2 + \|\bar{z}_n - z_{n+1}\|^2 \end{aligned} \tag{70}$$

Let's sum the last inequality by n:

$$\begin{aligned} &\|z_0 - z^*\|^2 + \|\lambda^* - \lambda_0\|^2 \geq \|\lambda^* - \lambda_{N+1}\|^2 + \|z_{N+1} - z^*\|^2 \\ &+ \left(1 - \alpha M - 2\alpha^2 M^2\right) \sum_{k=0}^{N} \|\lambda_k - \lambda_{k+1}\|^2 + (1 - \alpha M) \sum_{k=0}^{N} \|\bar{z}_k - z_k\|^2 + \sum_{k=0}^{N} \|\bar{z}_k - z_{k+1}\|^2 \end{aligned} \tag{71}$$

If we get $\alpha < \frac{1}{2M}$, then we have $(1 - \alpha M - 2\alpha^2 M^2) > 0 \;\; (1 - \alpha M) > 0$.

Let's notice that if equalities $\lambda_n = \lambda_{n+1}$ and $\bar{z}_n = z_n = z_{n+1}$ are true, we have in virtue of (43)-(44) that (z_n, λ_n) - is a saddle point. So according to (70), we could assume that consequence $\|\lambda^* - \lambda_{N+1}\|^2 + \|z_{N+1} - z^*\|^2$ - monotone decrease by N.

From (71) we have:

$$\|z_0 - z^*\|^2 + \|\lambda^* - \lambda_0\|^2 \geq \|\lambda^* - \lambda_{N+1}\|^2 + \|z_{N+1} - z^*\|^2$$

From the last inequality follows that consequences $\lambda_N \;\; z_N$ are bounded.

$\sum_{k=0}^{N} \|\lambda_k - \lambda_{k+1}\|^2$ and $\sum_{k=0}^{N} \|\bar{z}_k - z_k\|^2$ bounded and hence converge. From convergence of series we have:

$$\|\lambda_k - \lambda_{k+1}\|^2 \to 0 \tag{72}$$

$$\|\bar{z}_k - z_k\|^2 \to 0 \tag{73}$$

Let's take convergent subsequences $\lambda_{i_k} \to \lambda' \;\; z_{i_k} \to z'$.

From (72) follows: $\lambda_{i_k+1} \to \lambda'$.

From (73) follows: $\bar{z}_{i_k} \to z'$.

According to (25) we have:

$$\begin{aligned}\bar{z}_{i_k} &= \pi_Z\left(z_{i_k} - \alpha\frac{\partial L}{\partial z}(z_{i_k},\lambda_{i_k})\right) \\ \lambda_{i_k+1} &= \pi_\Lambda\left(\lambda_{i_k} + \alpha\frac{\partial L}{\partial \lambda}(\bar{z}_{i_k},\lambda_{i_k})\right)\end{aligned} \tag{74}$$

Passing to limit in (74) we get:

$$\begin{aligned}z^{'} &= \pi_Z\left(z^{'} - \alpha\frac{\partial L}{\partial z}(z^{'},\lambda^{'})\right) \\ \lambda^{'} &= \pi_\Lambda\left(\lambda^{'} + \alpha\frac{\partial L}{\partial \lambda}(z^{'},\lambda^{'})\right)\end{aligned} \tag{75}$$

According to (43)-(44) $(z^{'},\lambda^{'})$ - saddle point.

As $\left\|\lambda^* - \lambda_{N+1}\right\|^2 + \left\|z_{N+1} - z^*\right\|^2$ - monotone decrease by N for any saddle point (z^*,λ^*) then $\left\|\lambda^{'} - \lambda_{N+1}\right\|^2 + \left\|z_{N+1} - z^{'}\right\|^2$ monotone decrease by N.

Finally we get $\lambda_{N+1} \to \lambda^{'}, z_{N+1} \to z^{'}$. Theorem is proved.