Asymptotic Convergence of an SMO Algorithm Without Any Assumptions

Chih-Jen Lin

Department of Computer Science and Information Engineering

National Taiwan University, Taipei 106, Taiwan

cjlin@csie.ntu.edu.tw

Abstract

The asymptotic convergence in Lin [6] can be applied to a modified SMO algorithm by Keerthi et al. [5] with some assumptions. Here we show that for this algorithm those assumptions are not necessary.

I. Introduction

Given training vectors $x_i \in R^n, i = 1, \ldots, l$, in two classes, and a vector $y \in R^l$ such that $y_i \in \{1, -1\}$, the support vector machines (SVM) [9] require the solution of the following optimization problem:

$$\begin{aligned} \min_{\alpha} \quad f(\alpha) = \quad & \frac{1}{2}\alpha^T Q \alpha - e^T \alpha \\ & 0 \le \alpha_i \le C, i = 1, \ldots, l, \\ & y^T \alpha = 0, \end{aligned} \tag{1}$$

where $C > 0$ and e is the vector of all ones. Training vectors x_i are mapped into a higher dimensional space by ϕ and $Q_{ij} \equiv y_i y_j K(x_i, x_j)$ where $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel.

Due to the density of the matrix Q, currently the decomposition method is one of the major methods to solve (1) (e.g. [7], [3], [8]). It is an iterative process where in each iteration the index set of variables are separated to two sets B and N, where B is the working set. Then in that iteration variables corresponding to N are fixed while a sub-problem on variables corresponding to B is minimized.

Among these methods, Platt's Sequential Minimal Optimization (SMO) [8] is a simple algorithm where in each iteration only two variables are selected in the working set so the sub-problem can be analytically solved without using an optimization software. Keerthi et al. [5] pointed out a problem in the original SMO and proposed two modified versions. The

one using the two indices which have the maximal violation of the Karush-Kuhn-Tucker (KKT) condition may be now the most popular implementation among SVM software (e.g. LIBSVM [1], SVMTorch [2]). It is also a special case of another popular software SVM^{light} [3]. For convergence Keerthi and Gilbert [4] has proved that under a stopping criterion and any stopping tolerance, it terminates in finite iterations. However, this result does not imply the asymptotic convergence. On the other hand, the asymptotic convergence of Lin [6] for the software SVM^{light} can be applied to this algorithm when the size of the working set is restricted to two. However, in [6, Assumption IV.1] it requires the assumption that any two by two principal sub-matrix of the Hessian matrix Q is positive definite. This assumption may not be true if, for example, some data points are the same. In this paper we show that without this assumption results in [6] still follow. Hence existing implementations are asymptotically convergent without any problem.

The method by Keerthi et al. is as follows: Using $y_i = \pm 1$, the KKT condition of (1) can be rewritten to

$$\begin{aligned} &\max(\max_{\alpha_i<C, y_i=1} -\nabla f(\alpha)_i, \max_{\alpha_i>0, y_i=-1} \nabla f(\alpha)_i) \\ &\le \min(\min_{\alpha_i<C, y_i=-1} \nabla f(\alpha)_i, \min_{\alpha_i>0, y_i=1} -\nabla f(\alpha)_i), \end{aligned} \tag{2}$$

where $\nabla f(\alpha) = Q\alpha - e$ is the gradient of $f(\alpha)$ defined in (1). Then they consider

$$i \equiv \text{argmax}(\{-\nabla f(\alpha)_t \mid y_t = 1, \alpha_t < C\}, \{\nabla f(\alpha)_t \mid y_t = -1, \alpha_t > 0\}), \tag{3}$$

$$j \equiv \text{argmin}(\{\nabla f(\alpha)_t \mid y_t = -1, \alpha_t < C\}, \{-\nabla f(\alpha)_t \mid y_t = 1, \alpha_t > 0\}), \tag{4}$$

and use $B \equiv \{i, j\}$ as the working set. That is, i and j are the two elements which violate the KKT condition the most.

If $\{\alpha^k\}$ is the sequence generated by the decomposition method, the asymptotic convergence means that any convergent subsequence goes to an optimum of (1). The result of finite termination by Keerthi and Gilbert cannot be extended here because both sides of the inequality (2) are not continuous functions of α. In [6], the asymptotic convergence has been proved but the author has to assume that the matrix Q satisfies

$$\min_I(\min(\text{eig}(Q_{II}))) > 0, \tag{5}$$

where I is any subset of $\{1,\ldots,l\}$ with $|I| \leq 2$ and $\min(\text{eig}(\cdot))$ is the smallest eigenvalue of a matrix ([6, Assumption IV.1]). The main purpose of this paper is to show that (5) is not necessary.

II. Main Results

The only reason why we need (5) is for Lemma IV.2 in [6]. It proves that there exists $\sigma > 0$ such that

$$f(\alpha^{k+1}) \leq f(\alpha^k) - \frac{\sigma}{2}\|\alpha^{k+1} - \alpha^k\|^2, \text{ for all } k. \tag{6}$$

In the following we will show that without (5), (6) is still valid. First we note that if α^k is the current solution and $B = \{i, j\}$ is selected using (3) and (4), the required minimization on the sub-problem takes place in the rectangle $S = [0, C] \times [0, C]$ along a path where $y_i\alpha_i + y_j\alpha_j = -y_N^T\alpha_N^k$ is constant. Let the parametric change in α on this path be given by $\alpha(t)$:

$$\alpha_i(t) \equiv \alpha_i^k + t/y_i, \qquad \alpha_j(t) \equiv \alpha_j^k - t/y_j, \qquad \alpha_s(t) \equiv \alpha_s^k, \ \forall s \neq i, j.$$

The sub-problem is to minimize $\psi(t) \equiv f(\alpha(t))$ subject to $(\alpha_i(t), \alpha_j(t)) \in S$. Let $\bar{t}$ denote the solution of this problem and $\alpha^{k+1} = \alpha(\bar{t})$. Clearly,

$$|\bar{t}| = \|\alpha^{k+1} - \alpha^k\|/\sqrt{2}. \tag{7}$$

As $\psi(t)$ is a quadratic function on t,

$$\psi(t) = \psi(0) + \psi'(0)t + \psi''(0)t^2/2. \tag{8}$$

Since

$$\begin{aligned}
\psi'(t) &= \sum_{s=1}^{l} \nabla f(\alpha(t))_s \alpha_s'(t) \\
&= y_i \nabla f(\alpha(t))_i - y_j \nabla f(\alpha(t))_j \\
&= y_i(\sum_{s=1}^{l} Q_{is}\alpha_s(t) - 1) - y_j(\sum_{s=1}^{l} Q_{js}\alpha_s(t) - 1) \text{ and} \qquad (9) \\
\psi''(t) &= Q_{ii} + Q_{jj} - 2y_i y_j Q_{ij}, \qquad (10)
\end{aligned}$$

we have

$$\begin{aligned}\psi'(0) &= y_i\nabla f(\alpha^k)_i - y_j\nabla f(\alpha^k)_j \text{ and} \quad (11)\\ \psi''(0) &= \phi(x_i)^T\phi(x_i) + \phi(x_j)^T\phi(x_j) - 2y_i^2y_j^2\phi(x_i)^T\phi(x_j)\\ &= \|\phi(x_i) - \phi(x_j)\|^2. \quad (12)\end{aligned}$$

Then our new lemma is as follows:

Lemma II.1 *If the working set selection is by using (3) and (4), there exists $\sigma > 0$ such that for any k, (6) holds.*

Proof: Since Q is positive semidefinite, $\psi''(t) \geq 0$ so we can consider the following two cases:

Case 1 $\psi''(0) > 0$. Let t^* denote the unconstrained minimum of ψ, i.e. $t^* = -\psi'(0)/\psi''(0)$. Clearly, $\bar{t} = \gamma t^*$ where $0 < \gamma \leq 1$. Then, by (8),

$$\begin{aligned}\psi(\bar{t}) - \psi(0) &= \gamma\frac{-\psi'(0)^2}{\psi''(0)} + \frac{\gamma^2}{2}\frac{\psi'(0)^2}{\psi''(0)}\\ &\leq -\frac{\gamma^2}{2}\frac{\psi'(0)^2}{\psi''(0)} = -\frac{\psi''(0)}{2}\bar{t}^2 = -\frac{\psi''(0)}{4}\|\alpha^{k+1} - \alpha^k\|^2, \quad (13)\end{aligned}$$

where the last equality is from (7).

Case 2 $\psi''(t) = 0$. By (12), $\phi(x_i) = \phi(x_j)$. Using this, (9), and (11) we get

$$\begin{aligned}\psi'(0) &= y_i\sum_{s=1}^{l} Q_{is}\alpha_s^k - y_j\sum_{s=1}^{l} Q_{js}\alpha_s^k\\ &= y_i(\sum_{s=1}^{l} y_iy_s\phi(x_i)^T\phi(x_s)\alpha_s^k - 1) - y_j(\sum_{s=1}^{l} y_jy_s\phi(x_j)^T\phi(x_s)\alpha_s^k - 1)\\ &= y_j - y_i.\end{aligned}$$

With (11), since descent is assured, $\psi'(0) \neq 0$. Thus $y_i \neq y_j$ and hence $|\psi'(0)| = 2$. Since $\psi''(0) = \psi''(t) = 0$ implies $\psi'(t)$ is a linear function, with $\psi(\bar{t}) \leq \psi(0)$ and $|\bar{t}| \leq C$,

$$\psi(\bar{t}) - \psi(0) = -|\psi'(0)\bar{t}| \leq -\frac{2}{C}\bar{t}^2 = -\frac{\|\alpha^{k+1} - \alpha^k\|^2}{C}. \quad (14)$$

Note that $\psi(0) = f(\alpha^k)$ and $\psi(\bar{t}) = f(\alpha^{k+1})$. Thus, using (10), (7), and (14), if we get

$$\sigma \equiv \min\{\frac{2}{C}, \min_{i,j}\{\frac{Q_{ii} + Q_{jj} - 2y_iy_jQ_{ij}}{2} : Q_{ii} + Q_{jj} - 2y_iy_jQ_{ij} > 0\}\},$$

then the proof is complete. ■

III. Conclusion

Using [6, Theorem IV.1], results here can be extended to the decomposition method for support vector regression which selects the two-component working set in a similar way. The future challenge will be to remove the same assumption when the size of the working set is more than two.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan via the grant NSC 90-2213-E-002-111. The author thanks Sathiya Keerthi for many helpful comments.

References

[1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] R. Collobert and S. Bengio. SVMTorch: A support vector machine for large-scale regression and classification problems. *Journal of Machine Learning Research*, 1:143–160, 2001.

[3] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

[4] S. S. Keerthi and E. G. Gilbert. Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning*, 46:351–360, 2002.

[5] S. S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13:637–649, 2001.

[6] C.-J. Lin. On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 12, 2001. To appear.

[7] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proceedings of CVPR'97*, 1997.

[8] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

[9] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.