A Common Framework for the Convergence of the GSK, MDM and SMO Algorithms

Jorge López and José R. Dorronsoro*

Dpto. de Ingeniería Informática and Instituto de Ingeniería del Conocimiento
Universidad Autónoma de Madrid, 28049 Madrid, Spain

Abstract. Building upon Gilbert's convergence proof of his algorithtm to solve the Minimum Norm Problem, we establish a framework where a much simplified version of his proof allows us to prove the convergence of two algorithms for solving the Nearest Point Problem for disjoint convex hulls, namely the GSK and the MDM algorithms, as well as the convergence of the SMO algorithm for SVMs over linearly separable two–class samples.

1 Introduction

Given a sample $\mathcal{S} = \{(X_i, y_i) : i = 1, \dots, N\}$ with $y_i = \pm 1$, let $\mathcal{I}_\pm$ be the set of indices of the patterns X_i belonging to each class. Writing $W = \sum_i \alpha_i y_i X_i = \sum_{i \in \mathcal{I}_+} \alpha_i X_i - \sum_{i \in \mathcal{I}_-} \alpha_i X_i = W_+ - W_-$, we can solve the Nearest Point Problem (NPP) of finding the two closest points in the convex hulls of each class, by:

$$\min \mathcal{D}(\alpha) = \min \tfrac{1}{2} \|W_+ - W_-\|^2 = \min_\alpha \tfrac{1}{2} \textstyle\sum_i \sum_j \alpha_i \alpha_j y_i y_j X_i \cdot X_j$$

s.t. $\sum_{i \in \mathcal{I}_+} \alpha_i = \sum_{i \in \mathcal{I}_-} \alpha_i = 1, \;\; \alpha_i \geq 0 \;\; \forall i$. To do so, most of the methods proposed in the literature are adaptations of methods for the Minimum Norm Problem (MNP) of finding the point in a convex hull closest to the origin. Classical procedures for MNP are the Gilbert [1] and Mitchell [2] algorithms. These two algorithms were adapted to solve NPP as well as SVM for classification in [3] and [4] respectively. We shall call these NPP adaptations GSK and MDM. We recall that the dual problem solved by an SVM is

$$\min \tilde{\mathcal{D}}(\alpha) = \tfrac{1}{2} \textstyle\sum_i \sum_j \alpha_i \alpha_j y_i y_i X_i \cdot X_j - \sum_i \alpha_i$$

s.t. $\sum_i \alpha_i y_i = 0, \;\; \alpha_i \geq 0 \;\; \forall i$. Convergence proofs for GSK and MDM were given in [1] and Mitchell [2] (we are not aware of such proofs for their extensions to NPP) and a quite general convergence proof for SMO has been given in [5]; for the linearly separable case a much simpler SMO proof was given in [6]. In this work we propose a unified approach for GSK, MDM and SMO that results in much simpler proofs, again for linearly separable samples. More precisely, we will use a common framework with three basic steps, namely: 1) To bound the distance $\|W^t - W^*\|$ between the iterates

* With partial support of Spain's TIN 2007–66862 project and Cátedra UAM–IIC en Modelado y Predicción. The first author is kindly supported by FPU-MICINN grant AP2007–00142.

K. Diamantaras, W. Duch, L.S. Iliadis (Eds.): ICANN 2010, Part II, LNCS 6353, pp. 82–87, 2010.

W^t and the optimal W^* by a quantity Δ^t that appears naturally in the algorithms, 2) To show that $\Delta^{t_j} \to 0$ for some subsequence t_j and, therefore, that $W^{t_j} \to W^*$, 3) To conclude that $W^t \to W^*$ for the full sequence W^t.

The paper is organized as follows. In section 2 we give an overview of the GSK, MDM and SMO methods. Convergence proofs are given in section 3. Finally, section 4 offers some further discussion and pointers to future research.

2 Algorithms for Solving NPP and SVM

GSK uses at each iteration t a single updating pattern to build the new weight vector W^{t+1} by updating one of the components $W^t_\pm$ of the current $W^t = W^t_+ - W^t_-$ through an appropriate convex combination with a pattern $X_{L^\pm}$ of the corresponding class. For instance, assuming we use an $X_{L^t_+}$ in the positive class, we have $W^{t+1} = (1-\lambda^t)W^t_+ + \lambda^t X_{L^t_+} - W^t_- = W^t + \lambda^t(X_{L^t_+} - W^t_+)$ and it is shown in [3] that the optimal λ^t is

$$\lambda^t = \min\left\{1, \Delta^t_+/\|W^t_+ - X_{L^t_+}\|^2\right\} \ , \tag{1}$$

where we write $\Delta^t_+ = y_{L^t_+} W^t \cdot (W^t_+ - X_{L^t_+})$. Moreover, from the expression above for W^{t+1} we have $\|W^t\|^2 - \|W^{t+1}\|^2 = 2\lambda^t W^t \cdot (W^t_+ - X_{L^t_+}) - (\lambda^t)^2\|W^t_+ - X_{L^t_+}\|^2$. Notice that $\|W^t\|^2 - \|W^{t+1}\|^2 = 2(\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}))$ and if we take an unclipped λ^t in (1), we have

$$\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}) = (\Delta^t_+)^2/(2\|W^t_+ - X_{L^t_+}\|^2) \geq (\Delta^t_+)^2/(2D^2) \ , \tag{2}$$

where $D = \max_{i,j} \|X_i - X_j\|$. In this case, the norm decrease is approximately optimal if Δ^t_+ is largest, for which we just choose L^t_+ as $L^t_+ = \arg\min_{i \in \mathcal{I}^+} \{y_i W^t \cdot X_i\}$. If, however, clipping takes place we have $\Delta^t_+ \geq \|W^t_+ - X_{L^t_+}\|^2$, which yields

$$\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}) = W^t \cdot (W^t_+ - X_{L^t_+}) - \|W^t_+ - X_{L^t_+}\|^2/2 \geq \Delta^t_+/2 \ . \tag{3}$$

Similar formulae hold when we choose a X_{L^-} in the negative class and the class chosen in GSK is the one for which Δ^t_+ or Δ^t_- is largest. Once the choice is made we just write Δ^t instead of $\Delta^t_\pm$; see [3] for more details.

Turning our attention to MDM, it updates at each step one of the $W^t_\pm$ components of W^t using now two pattern vectors $X_{L^\pm}$ and $X_{U^\pm}$. For instance, if we update W^t_+ using $X_{L^t_+}$ and $X_{U^t_+}$, we will have $W^t_+ = W^t_+ + \lambda^t(X_{L^t_+} - X_{U^t_+})$ and, therefore, $W^{t+1} = W^t_+ + \lambda^t(X_{L^t_+} - X_{U^t_+}) - W^t_- = W^t + \lambda^t(X_{L^t_+} - X_{U^t_+})$. Now the optimal λ^t is chosen as [4]:

$$\lambda^t = \min\left\{\alpha^t_{U^t_+}, \overline{\Delta}^t_+/\|X_{U^t_+} - X_{L^t_+}\|^2\right\} \ , \tag{4}$$

where this time we write $\overline{\Delta}^t_+ = y_{L^t_+} W^t \cdot (X_{U^t_+} - X_{L^t_+})$. If clipping is not needed, taking $\lambda^t = \overline{\Delta}^t_+/\|X_{U^t_+} - X_{L^t_+}\|^2$ in (4) and arguing as done in the GSK case, we obtain

$$\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}) = (\overline{\Delta}^t_+)^2/(2\|X_{U^t_+} - X_{L^t_+}\|^2) \geq (\overline{\Delta}^t)^2/(2D^2) \ , \tag{5}$$

where we used $\|X_{U^t_+} - X_{L^t_+}\|^2 \le D^2$. The $\mathcal{D}$ decrease is largest when $\overline{\Delta}^t_+$ is largest, which we achieve by choosing $U^t_+ = \arg\max_{i\in\mathcal{I}^+|\alpha^t_i>0}\{y_i W^t \cdot X_i\}$ and L^t_+ as done for GSK. We may have to clip λ^t at $\alpha^t_{U^t_+}$ to ensure that the new coefficient of $X_{U^t_+}$ does not become negative. In this case, we obtain

$$\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}) = \alpha^t_{U^t_+}\overline{\Delta}^t_+ - (\alpha^t_{U^t_+})^2\|X_{U^t_+} - X_{L^t_+}\|^2/2 \ge \alpha^t_{U^t_+}\overline{\Delta}^t/2 \ , \quad (6)$$

since clipping occurs when $\overline{\Delta}^t_+ \ge \alpha^t_{U^t_+}\|X_{U^t_+} - X_{L^t_+}\|^2$. Similar formulae hold when we choose $X_{U^t_-}, X_{L^t_-}$ in the negative class and the class finally selected is the one for which the quantity $\overline{\Delta}^t_\pm$ is largest [4]. Once chosen, we shall just write $\overline{\Delta}^t$.

The SMO updates are of the form $W^{t+1} = W^t + \lambda^t y_{L^t}(X_{L^t} - X_{U^t})$, or, in terms of the α coefficients, $\alpha^{t+1}_{L^t} = \alpha^t_{L^t} + \lambda^t$, $\alpha^{t+1}_{U^t} = \alpha^t_{U^t} - \lambda^t y_{U^t} y_{L^t}$, and the other α_j do not change. An optimal λ^t is now chosen so that the decrease in the SVM dual function is largest, which means [6]

$$\lambda^t = y_{L^t}\tilde{\Delta}^t/\|X_{U^t} - X_{L^t}\|^2 = y_{L^t}\mu \ , \quad (7)$$

where we write now $\tilde{\Delta}^t = W^t\cdot(X_{U^t} - X_{L^t}) - (y_{U^t} - y_{L^t})$ and $\mu = \tilde{\Delta}^t/\|X_{U^t} - X_{L^t}\|^2$. Here the $\tilde{\mathcal{D}}$ decrease is largest when $\tilde{\Delta}^t$ is largest, which can be achieved if we select $L^t = \arg\min_{i\in\mathcal{I}_L}\{W^t\cdot X_i - y_i\}$ and $U^t = \arg\max_{i\in\mathcal{I}_U}\{W^t\cdot X_i - y_i\}$, where $\mathcal{I}_L = \{i : y_i = 1 \text{ or } y_i = -1, \alpha^t_i > 0\}$ and $\mathcal{I}_U = \{i : y_i = 1, \alpha^t_i > 0 \text{ or } y_i = -1\}$. As done before, we may have to clip μ as $\mu^t = \min\{\mu, \alpha^t_{L^t}\}$ if $y_{L^t} = -1$ and as $\mu^t = \min\{\mu, \alpha^t_{U^t}\}$ if $y_{U^t} = 1$. If no clipping is required, making use of (7) yields

$$\tilde{\mathcal{D}}(\alpha^t) - \tilde{\mathcal{D}}(\alpha^{t+1}) = (\tilde{\Delta}^t)^2/(2\|X_{U^t} - X_{L^t}\|^2) \ge (\tilde{\Delta}^t)^2/(2D^2) \ . \quad (8)$$

If, however, μ is clipped at $\alpha^t_{U^t}$, then $\tilde{\Delta}^t \ge \alpha^t_{U^t}\|X_{U^t} - X_{L^t}\|^2$ must hold, yielding

$$\tilde{\mathcal{D}}(\alpha^t) - \tilde{\mathcal{D}}(\alpha^{t+1}) = \alpha^t_{U^t}\tilde{\Delta}^t - (\alpha^t_{U^t})^2\|X_{U^t} - X_{L^t}\|^2/2 \ge \alpha^t_{U^t}\tilde{\Delta}^t/2 \ , \quad (9)$$

while we similarly obtain $\tilde{\mathcal{D}}(\alpha^t) - \tilde{\mathcal{D}}(\alpha^{t+1}) \ge \alpha^t_{L^t}\tilde{\Delta}^t/2$, if μ is clipped at $\alpha^t_{L^t}$. We refer to [6] for more details.

3 Convergence

If the algorithms described previously stop in a finite number t of iterations, the Δ^t, $\overline{\Delta}^t$ and $\tilde{\Delta}^t$ values must be zero, and the KKT conditions imply we are at an optimum. Hence, in the sequel we will consider the case of an infinite number of iterations.

3.1 Convergence of GSK and MDM

We give a unified convergence proof for GSK and MDM. As a first step we show the following.

Proposition 1. *If W^* is the closest vector between the positive and negative class hulls, then the following hold*

$$\|W^t - W^*\|^2 \le 2\Delta^t \le 2\overline{\Delta}^t \ , \tag{10}$$

$$\|W^t - W^*\|^2 \le \|W^t\|^2 - \|W^*\|^2 = 2(\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^*)) \ . \tag{11}$$

Proof. A simple geometric reasoning in disjoint convex hulls shows that $W^t \cdot W^* \ge \|W^*\|^2$ and, therefore,

$$\begin{aligned} \|W^t - W^*\|^2 &= \|W^t\|^2 - W^t \cdot W^* - W^t \cdot W^* + \|W^*\|^2 \le \|W^t\|^2 - W^t \cdot W^* \\ &= \|W^t\|^2 - \sum_{i \in \mathcal{I}^+} \alpha_i^* y_i W^t \cdot X_i - \sum_{i \in \mathcal{I}^-} \alpha_i^* y_i W^t \cdot X_i \\ &\le \|W^t\|^2 - \min_{i \in \mathcal{I}^+} \{y_i W^t \cdot X_i\} - \min_{i \in \mathcal{I}^-} \{y_i W^t \cdot X_i\} \\ &= W^t \cdot W_+^t - \min_{i \in \mathcal{I}^+} \{y_i W^t \cdot X_i\} - W^t \cdot W_-^t - \min_{i \in \mathcal{I}^-} \{y_i W^t \cdot X_i\} \\ &= \Delta_+^t + \Delta_-^t \le 2\Delta^t \ . \end{aligned}$$

We show next that $\Delta_+^t \le \overline{\Delta}_+^t$ and that $\Delta_-^t \le \overline{\Delta}_-^t$. In fact

$$\begin{aligned} \Delta_+^t &= W^t \cdot W_+^t - \min_{i \in \mathcal{I}^+} \{y_i W \cdot X_i\} = \sum_{i \in \mathcal{I}^+} \alpha_i y_i W^t \cdot X_i - \min_{i \in \mathcal{I}^+} \{y_i W \cdot X_i\} \\ &\le \max_{i \in \mathcal{I}^+ | \alpha_i^t > 0} \{y_i W \cdot X_i\} - \min_{i \in \mathcal{I}^+} \{y_i W \cdot X_i\} = \overline{\Delta}_+^t \ , \end{aligned}$$

and a similar argument works for the other bound. Finally, to prove (11), reasoning as just done at the beginning of the previous argument, we have $\|W^t - W^*\|^2 \le \|W^t\|^2 - W^t \cdot W^* \le \|W^t\|^2 - \|W^*\|^2$. □

We show next the following.

Proposition 2. *For GSK we have $\Delta^t \to 0$ as $t \to \infty$. Moreover, for MDM there is a subsequence t_j such that $\overline{\Delta}^{t_j} \to 0$.*

Proof. For GSK (2) and (3) imply $\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}) \ge \min\{(\Delta^t)^2/(2D^2), \Delta^t/2\}$. Thus, since the $\mathcal{D}(\alpha^t)$ sequence decreases and is always positive, it must converge and, therefore, we must have $\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}) \to 0$, so the whole Δ^t sequence goes to 0.

For MDM, (5) and (6) give $\mathcal{D}(\alpha^t) - \mathcal{D}(\alpha^{t+1}) \ge \min\{(\overline{\Delta}^t)^2/(2D^2), \alpha_{U^t}^t \overline{\Delta}^t/2\}$, and arguing as before, the right hand side must tend to 0. Its first term applies when no clipping is done but, arguing as in [6], it can be proved that clipping cannot occur indefinitely after some t. Thus, the bound on $(\overline{\Delta}^t)^2/(2D^2)$ must apply to some subsequence t_j and, therefore, $\overline{\Delta}^{t_j} \to 0$. □

Now we are ready to show the following.

Theorem 1. *The W^t updates of the GSK and MDM algorithms converge to W^* as t goes to ∞.*

Proof. For GSK, the result is immediate by Proposition 2 and (10).

For MDM, Proposition 2 and (10) imply that $W^{t_j} \to W^*$. By continuity of the dual, $\mathcal{D}(\alpha^{t_j}) \to \mathcal{D}(\alpha^*)$. But the sequence $\mathcal{D}(\alpha^t)$ decreases, so we must then have $\mathcal{D}(\alpha^t) \to \mathcal{D}(\alpha^*)$. Finally, it follows by (11) that $W^t \to W^*$.

3.2 Convergence of SMO

The convergence of SMO is also proved along similar lines.

Proposition 3. *If $W^* = \sum \alpha_i^* y_i X_i$ is the optimal SVM solution, we have*

$$\|W^t - W^*\|^2 \le (\tilde{\Delta}^t/2)\left(\sum_i \alpha_i^t + \sum_i \alpha_i^*\right) \ , \tag{12}$$

$$\|W^t - W^*\|^2 \le 2(\tilde{\mathcal{D}}(\alpha^t) - \tilde{\mathcal{D}}(\alpha^*)) \ . \tag{13}$$

Proof. First, notice that W^* is a primal feasible weight vector (i.e., $y_i(W^* \cdot X_i + b^*) \ge 1$ for all i), so we have $W^t \cdot W^* = \sum_i \alpha_i^t y_i W^* \cdot X_i = \sum_i \alpha_i^t y_i (W^* \cdot X_i + b^*) \ge \sum_i \alpha_i^t$. Besides, the KKT conditions imply that $\|W^*\|^2 = \sum_i \alpha_i^*$ and, hence, $\tilde{\mathcal{D}}(\alpha^*) = \|W^*\|^2/2 - \sum_i \alpha_i^* = -\|W^*\|^2/2$. Then

$$\begin{aligned}\|W^t - W^*\|^2 &= \|W^t\|^2 - 2W^t \cdot W^* + \|W^*\|^2 \le \|W^t\|^2 - 2\sum_i \alpha_i^t + \|W^*\|^2 \\ &= 2(\tilde{\mathcal{D}}(\alpha^t) - \tilde{\mathcal{D}}(\alpha^*)) \ ,\end{aligned}$$

so that (13) holds. Observe that, by the results above, we can also write

$$\|W^t - W^*\|^2 \le \|W^t\|^2 - \sum_i \alpha_i^t - W^t \cdot W^* + \sum_i \alpha_i^* \ . \tag{14}$$

For the first two terms, we have

$$\begin{aligned}\left\|W^t\right\|^2 - \sum_i \alpha_i^t &= \sum_i \alpha_i^t y_i W \cdot X_i - \sum_i \alpha_i^t y_i^2 = \sum_i \alpha_i^t y_i (W \cdot X_i - y_i) \\ &= \sum_{\mathcal{I}_+} \alpha_i^t (W \cdot X_i - y_i) - \sum_{\mathcal{I}_-} \alpha_i^t (W \cdot X_i - y_i) \\ &\le \left(\max_{\mathcal{I}_U}\left\{W^t \cdot X_i - y_i\right\} - \min_{\mathcal{I}_L}\left\{W^t \cdot X_i - y_i\right\}\right) \sum_i \alpha_i^t/2 \\ &= (\tilde{\Delta}^t/2) \sum_i \alpha_i^t \ ,\end{aligned}$$

where we use $\sum_{\mathcal{I}_+} \alpha_i^t = \sum_{\mathcal{I}_-} \alpha_i^t = \sum \alpha_i^t/2$. Analogously to what has just been done, we get $W^t \cdot W^* - \sum_i \alpha_i^* \ge (-\tilde{\Delta}^t/2)\sum_i \alpha_i^*$ for the last two terms. Hence, putting it all together in (14), we arrive at (12). □

We point out that the above argument for inequality (12) can also be applied to complete the partial proof of Lemma 1 in [6] given there.

Proposition 4. *There is a subsequence W^{t_j} that tends to W^* as $t \to \infty$.*

Proof. As an easy consequence of estimates (8) and (9) we get $\tilde{\mathcal{D}}(\alpha^t) - \tilde{\mathcal{D}}(\alpha^{t+1}) \geq \min\{(\tilde{\Delta}^t)^2/(2D^2), \alpha^t_{U^t}\tilde{\Delta}^t/2, \alpha^t_{L^t}\tilde{\Delta}^t/2\}$. The first term at the right hand side applies when there is no clipping and, as just argued for MDM, clipping cannot go on indefinitely. Thus, there must be a subsequence t_j such that $\tilde{\Delta}^{t_j} \to 0$ and, by Proposition 3, $W^{t_j} \to W^*$, since $\sum_i \alpha_i^t$ can be shown to be bounded [6]. □

Now convergence of the full W^t sequence is proved just as in the MDM case, with Proposition 4 and (13).

Theorem 2. *The W^t updates of the SMO algorithm converge to W^* as t goes to ∞.*

4 Conclusions and Further Work

In this work we present, for linearly separable samples, simple proofs of convergence for the GSK and MDM algorithms for NPP, and the SMO algorithm for SVM training, all three under a common framework. This results in much simpler proofs for GSK and MDM than the ones in [1] and [2], and also generalize them to the NPP case. Our proof for SMO is also simpler than the ones given in [6] and [5], but in its present form it is only applicable to linearly–separable tasks. We are currently working on its extension to the non–linearly separable case.

References

1. Gilbert, E.G.: Minimizing the Quadratic Form on a Convex Set. SIAM J. Contr. 4, 61–79 (1966)
2. Mitchell, B.F., Dem'yanov, V.F., Malozemov, V.N.: Finding the Point of a Polyhedron Closest to the Origin. SIAM J. Contr. 12, 19–26 (1974)
3. Franc, V., Hlavăc, V.: An Iterative Algorithm Learning the Maximal Margin Classifier. Pattern Recognition 36, 1985–1996 (2003)
4. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: A Fast Iterative Nearest Point Algorithm for Support Vector Machine Classifier Design. IEEE Transactions on Neural Networks 11(1), 124–136 (2000)
5. Lin, C.-J.: On the Convergence of the Decomposition Method for Support Vector Machines. IEEE Transactions on Neural Networks 12(6), 1288–1298 (2001)
6. López, J., Dorronsoro, J.R.: A Simple Proof of the Convergence of the SMO Algorithm for Linearly Separable Problems. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009, Part I. LNCS, vol. 5768, pp. 904–912. Springer, Heidelberg (2009)