Speeding up the IRWLS convergence to the SVM solution

Fernando Pérez-Cruz* and Antonio Artés-Rodríguez

Gatsby Computational Neuroscience unit
University College London
fernando@gatsby.ucl.ac.uk

Signal Theory and Communications Dept.
Universidad Carlos III de Madrid
antonio@ieee.org

Special Session on Least Square Support Vector Machines at IJCNN 2004

27th July 2004

*presenting author

Unconstrained Support Vector Machine

We have a set of i.i.d. examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$, and a transformation to a feature space $\phi(\cdot) \in \mathcal{H}$. The SVC minimizes:

$$L_P(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} L(u_i)$$

with respect to $\mathbf{w}$ and b, where $u_i = 1 - y_i(\phi^T(\mathbf{x}_i)\mathbf{w} + b)$ and

$$L(u) = \begin{cases} 0, & u < 0 \\ Ku^2/2, & 0 \le u < 1/K \\ u - 1/(2K), & u \ge 1/K \end{cases}$$

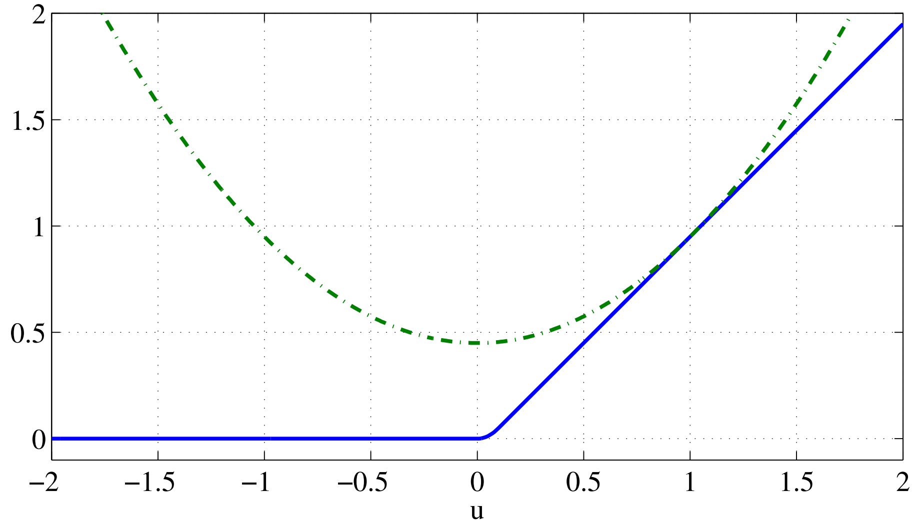
which tends to $\max(u, 0)$ as K approaches infinity.

IRWLS Objective

1.- Place a quadratic upper bound on $L(u)$ for each sample over the previous solution.

2.- Solve the quadratic problem. (linear system of equations)

3.- Repeat until convergence.



Quadratic Approximation

To construct the IRWLS procedure, we modify L_P using a first order Taylor expansion of $L(u)$ over the previous solution, leading to:

$$L_P^{'}(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\left(L(u_i^k) + \left.\frac{dL(u)}{du}\right|_{u_i^k}[u_i - u_i^k]\right)$$

where $u_i^k = 1 - y_i(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{w}^k + b^k)$,

Now, we construct a quadratic approximation:

$$L_P^{''}(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\left(L(u_i^k) + \left.\frac{dL(u)}{du}\right|_{u_i^k}\frac{(u_i)^2 - (u_i^k)^2}{2u_i^k}\right)$$

in which $L_P^{''}(\mathbf{w}^k, b^k) = L_P(\mathbf{w}^k, b^k)$ and $\nabla L_P^{''}(\mathbf{w}^k, b^k) = \nabla L_P(\mathbf{w}^k, b^k)$.

Iterative Re-Weighted Least Square

1. Minimize: $L''_P(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} L_0(u_i) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} a_i u_i^2/2 + d_i$

$$\begin{bmatrix} \mathbf{\Phi}^T \mathbf{D_a} \mathbf{\Phi} + \mathbf{I} & \mathbf{\Phi}^T \mathbf{a} \\ \mathbf{a}^T \mathbf{\Phi} & \mathbf{a}^T \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}^T \mathbf{D}_a \mathbf{y} \\ \mathbf{a}^T \mathbf{y} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \mathbf{H} + \mathbf{D_a}^{-1} & \mathbf{y} \\ \mathbf{y}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{1} \\ 0 \end{bmatrix}$$

$\mathbf{\Phi} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)]^T$ and $(\mathbf{H})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

2. Recompute a_i.

$$a_i = \frac{C}{u_i^k} \frac{dL(u)}{du}\bigg|_{u_i^k} = \begin{cases} 0, & u_i^k < 0 \\ KC, & 0 \leq u_i^k < \frac{1}{K} \\ \frac{C}{u_i^k}, & u_i^k \geq \frac{1}{K} \end{cases}.$$

3. Repeat until convergence.

Proof of Convergence

To prove that the IRWLS actually delivers the SVC solution when it stops, we will need demonstrate the following items:

- the sequence $(\mathbf{w}^0, b^0), \ldots, (\mathbf{w}^k, b^k), \ldots$ converges to $(\mathbf{w}^{op}, b^{op})$; and
- $\mathbf{w}^{op} = \mathbf{w}^*$ and $b^{op} = b^*$.

Wolfe Conditions:

$$L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) \leq L_P(\mathbf{z}^k) + c_1 \nabla L_P(\mathbf{z}^k)^T \mathbf{p}^k$$

$$\nabla L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k)^T \mathbf{p}^k \geq c_2 \nabla L_P(\mathbf{z}^k)^T \mathbf{p}^k$$

for $0 < c_1 < c_2 < 1$. $\mathbf{z}^k = [(\mathbf{w}^k)^T \quad b^k]^T$, $\mathbf{p}^k = [(\mathbf{w}^s - \mathbf{w}^k)^T \quad (b^s - b^k)]^T$

Strictly Decreasing Property

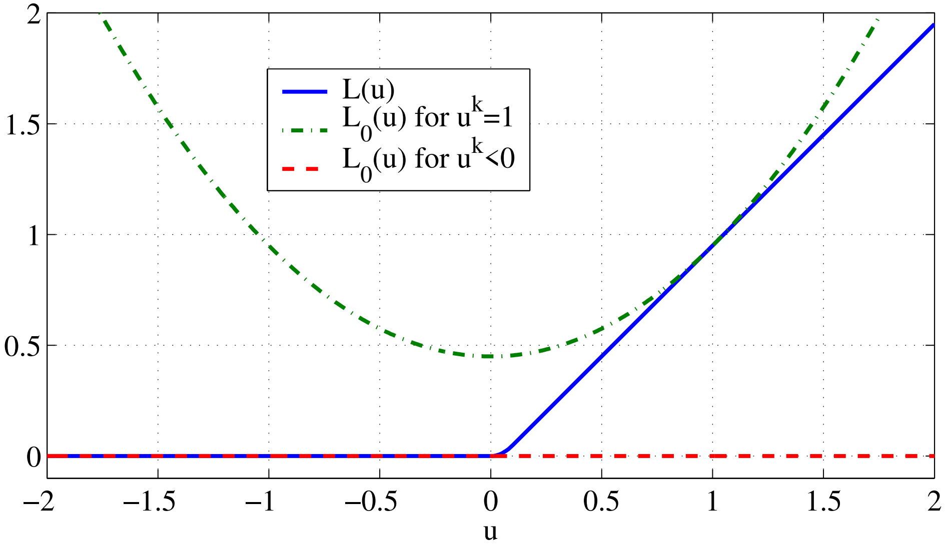
$$\left(L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) < L_P(\mathbf{z}^k)\right)$$

$$L_P(\mathbf{z}^k) = L''_P(\mathbf{z}^k) \geq L''_P(\mathbf{z}^{k+1}) \geq L_P(\mathbf{z}^{k+1})$$

1.- Construction.

2.- Optimization.

3.- Upper Bound



Strictly Decreasing Property

$\left(L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k) < L_P(\mathbf{z}^k)\right)$

$$L_P(\mathbf{z}^k) = L''_P(\mathbf{z}^k) \geq L''_P(\mathbf{z}^{k+1}) \geq L_P(\mathbf{z}^{k+1})$$

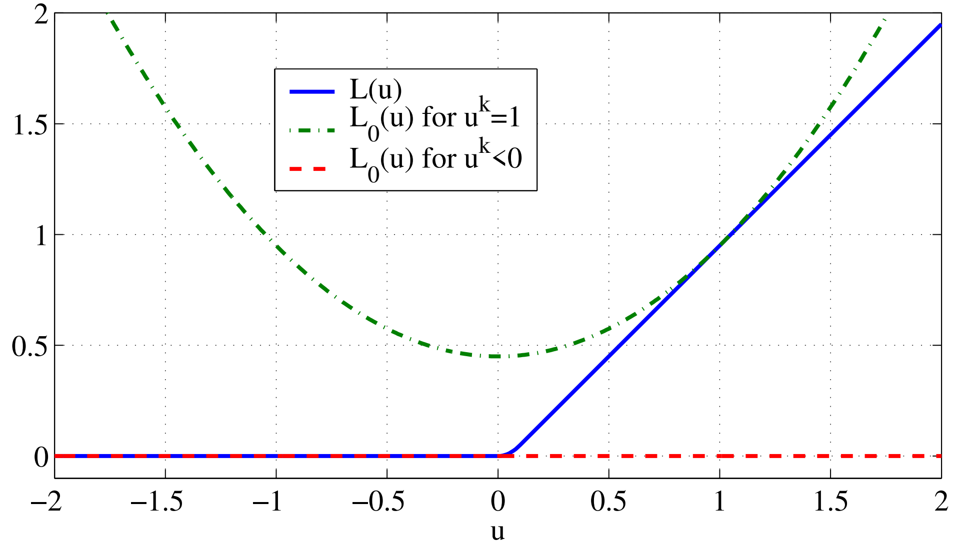
1.- Construction.

2.- Optimization.

3.- Upper Bound (Not Always)

$$\eta^k = \min_{\mathcal{S}} u_i^k / (u_i^k - u_i^s)$$

$\mathcal{S} = \{i|\ u_i^k < 0\ \&\ u_i^s > 0\}$



Sufficient Decrease Property

$$\left(\nabla L_P(\mathbf{z}^k + \eta^k \mathbf{p}^k)^T \mathbf{p}^k > \nabla L_P(\mathbf{z}^k)^T \mathbf{p}^k\right)$$

1.- $\nabla L_P(\mathbf{z}^{k+1})^T \mathbf{p}^k = \dfrac{\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2/2 + L_P(\mathbf{w}^{k+1}, b^{k+1}) - L_P^{'''}(\mathbf{w}^k, b^k)}{\eta^k}$

2.- $\nabla L_P(\mathbf{z}^k)^T \mathbf{p}^k = \dfrac{-\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2/2 - L_P(\mathbf{w}^k, b^k) + L_P^{'}(\mathbf{w}^{k+1}, b^{k+1})}{\eta^k}$

3.- $\nabla L_P(\mathbf{z}^{k+1})^T \mathbf{p}^k - \nabla L_P(\mathbf{z}^k)^T \mathbf{p}^k \overset{?}{>} 0$

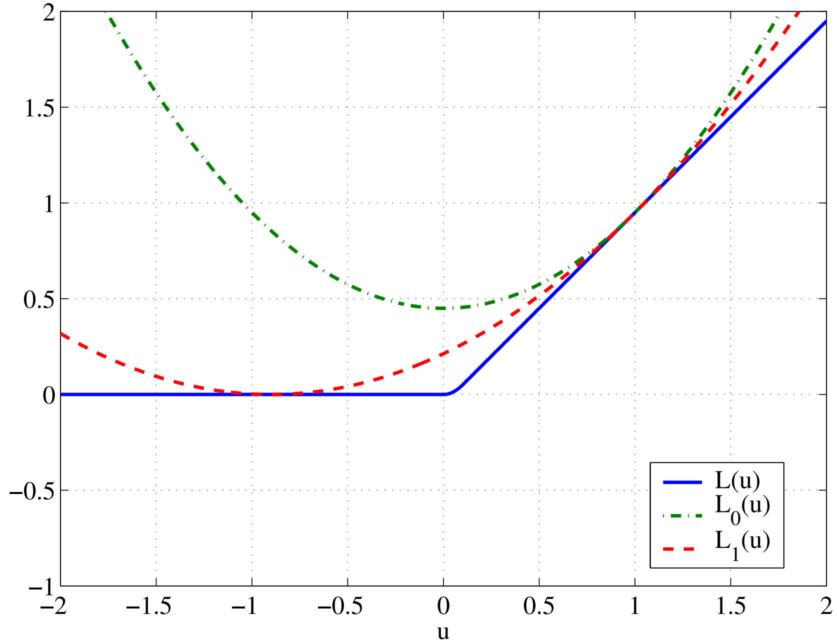
Novel Approximation I

We can propose a closer fit to the SVM loss function

$$L_1(u_i) = a_i u_i^2/2 + t_i u_i + d_i \geq L(u_i)$$

$$a_i = \begin{cases} 0, & u_i^k < 0 \\ CK, & 0 \leq u_i^k < \frac{1}{K} \\ \frac{CK}{2Ku_i^k - 1}, & u_i^k \geq \frac{1}{K} \end{cases}$$

$$t_i = \begin{cases} 0, & u_i^k < \frac{1}{K} \\ C\frac{Ku_i^k - 1}{2Ku_i^k - 1}, & u_i^k \geq \frac{1}{K} \end{cases}$$

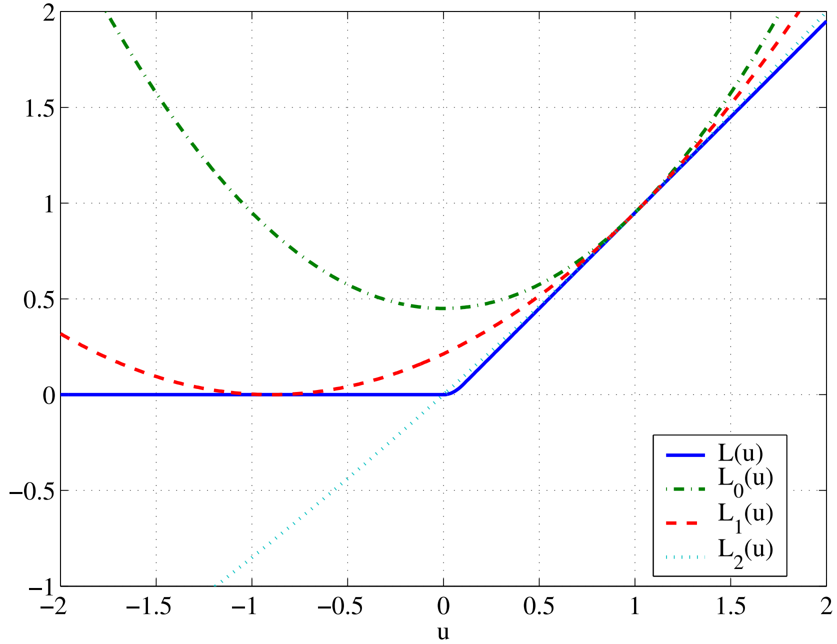


Novel Approximation II

$$L_2(u_i) = a_i u_i^2/2 + t_i u + d_i \geq L(u_i) \text{ for } \forall u_i \geq 0.$$

$$a_i = \begin{cases} 0, & u_i^k < 0 \\ CK, & 0 \leq u_i^k < \frac{1}{K} \\ \frac{C}{K {u_i^k}^2}, & u_i^k \geq \frac{1}{K} \end{cases}$$

$$t_i = \begin{cases} 0, & u_i^k < \frac{1}{K} \\ C\frac{K u_i^k - 1}{K u_i^k}, & u_i^k \geq \frac{1}{K} \end{cases}$$



Algorithm

1. Initialization: set $k = 0$, $\mathbf{w}^0 = \mathbf{0}$, $b^0 = 0$ and $u_i^0 = 1 \quad \forall i = 1, \ldots, n$.

2. Minimize: $L''_P(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} L_j(u_i) \qquad j = 0, 1, 2.$

3. Compute u_i^s and $\mathcal{S}$. If $\mathcal{S} = \emptyset$, $\mathbf{w}^{k+1} = \mathbf{w}^s$ and $b^{k+1} = b^s$ go to 5.

   $\mathcal{S} = \{i | \ u_i^k < 0 \ \& \ u_i^s > 0\}$ or $\mathcal{S} = \{i | \ (u_i^k < 0 \ \& \ u_i^s > 0) | (u_i^k > 1/K \ \& \ u_i^s < 0)\}$

4. Compute $\eta^k = \arg\min_{\eta \in \mathcal{S}} L(\eta \mathbf{w}^k + (1-\eta)\mathbf{w}^s, \eta b^k + (1-\eta) b^s)$, and set $\mathbf{w}^{k+1} = \eta^k \mathbf{w}^k + (1-\eta^k)\mathbf{w}^s$ and $b^{k+1} = \eta^k b^k + (1-\eta^k) b^s$.

5. Set $k = k + 1$ and go to 2 until convergence.

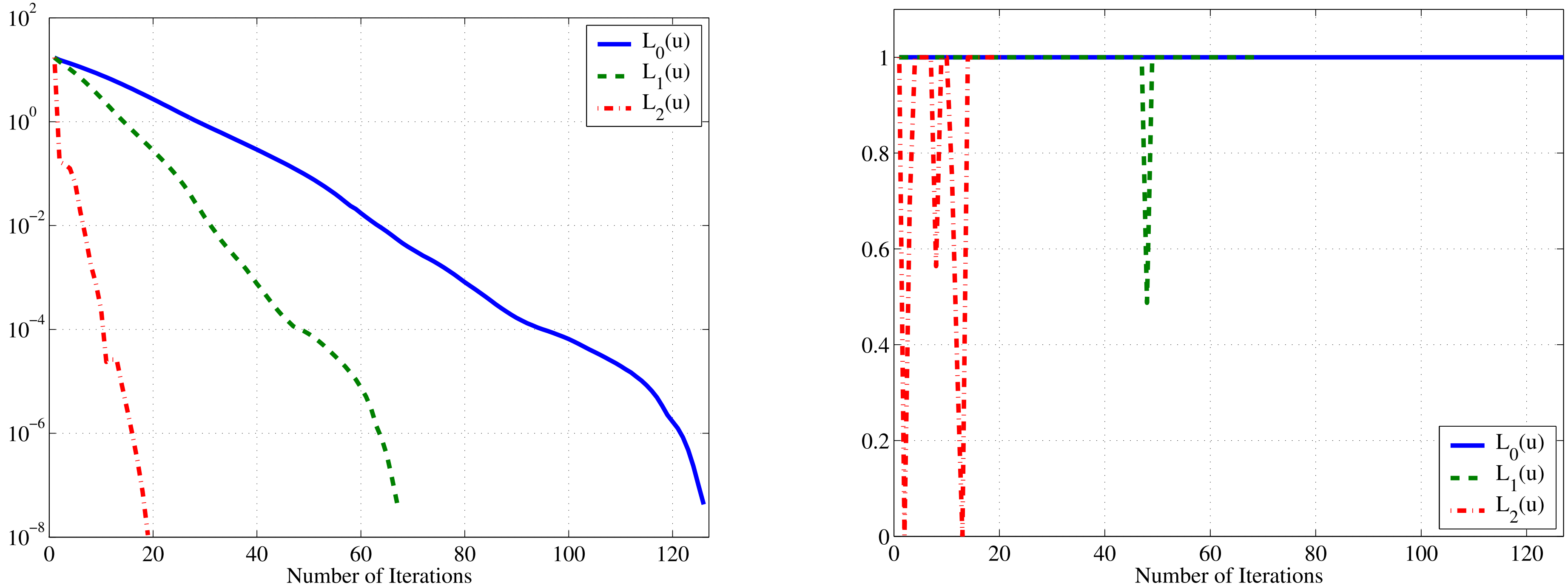
Experiments Results

Data set	L_0	L_1	L_2	N.S.	Dim.
Banana	8.11	4.36	**1.13**	400	2
Thyroid	0.25	0.13	**0.07**	140	5
Breast-Cancer	0.87	**0.53**	0.68	200	9
German	104.95	77.69	**73.08**	700	20
Ringnorm	2.88	**0.79**	1.78	400	20
Twonorm	2.02	**0.84**	1.22	400	20

Computing time in seconds over 20 sets.

Experiments Results II

Results for one trial of Banana data set.



Discussion

1. Faster algorithm using the convergence demonstration.

2. Combination of $L_1(u)$ and $L_2(u)$ for optimal results.

3. Large Scale Problems.

For more information, visit my web site at

http://www.gatsby.ucl.ac.uk/~fernando