JMC
John's Modeling Cult

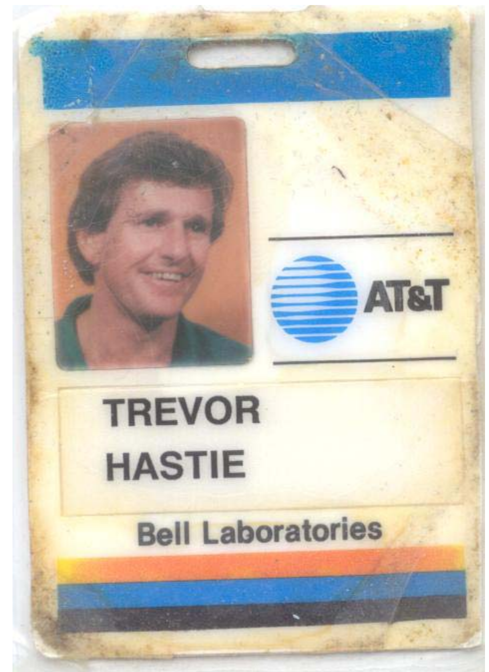
Trevor Hastie
Stanford University

http://www-stat.stanford.edu/∼hastie

Where and When?

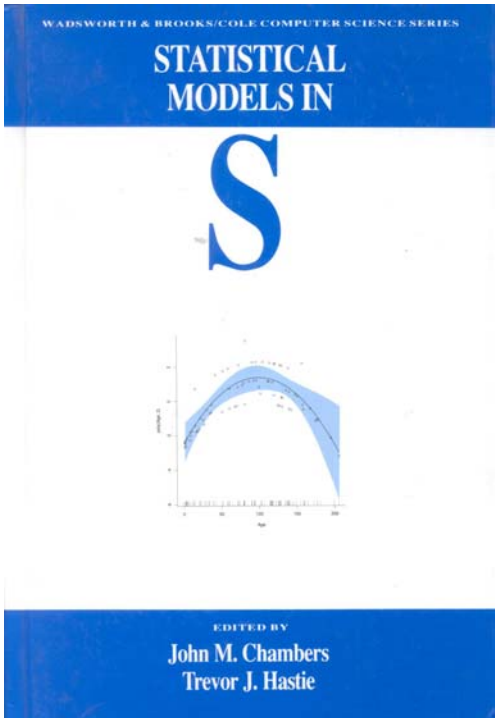
Bell Labs — 1986

- Mid 1985, John gets a phone call from a pushy South African, saying he would rather work at Bell Labs than UBC.
- Eager beaver Hastie joins dept in 1986 - within two weeks visited by Yehuda and Vijay, and told to slow down.
- Very exciting times for a young researcher, with JMC an ideal mentor and role model.



The White Book

- Just finished the GAM book — exhausted — but could not resist being a co-editor of this exciting project with JMC.
- That was before I'd heard of "code review"!
- Just when Daryl and I thought we'd nailed the formula parser, JMC produced his "theorem" (Sec 2.4.1) and cleaned up.
- I am very proud of this collaboration and project.



Rest of the Talk: Regularization Paths

Some of the things I have been doing since leaving the fold.

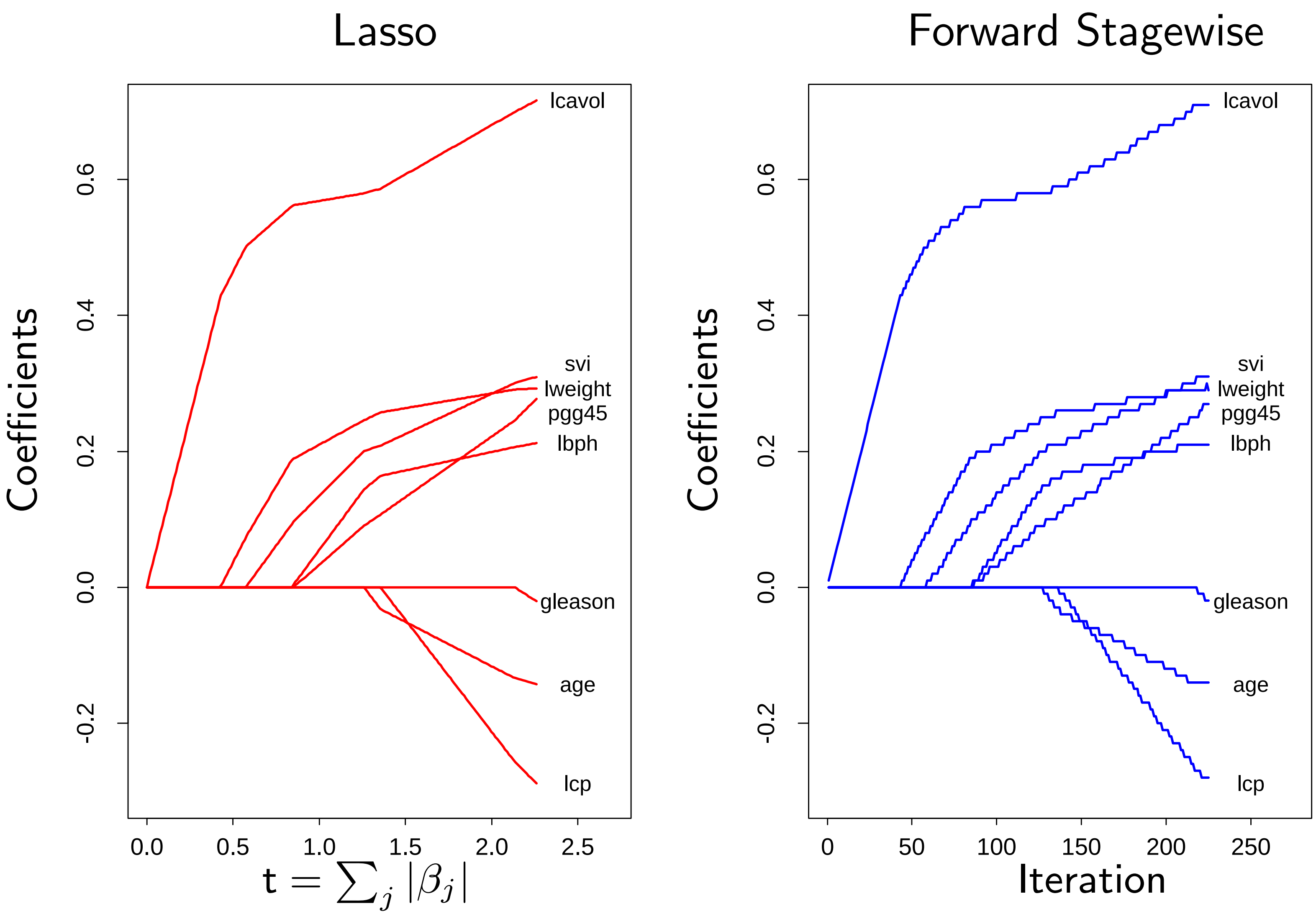
- Least Angle Regression and the Lasso (with Brad Efron, Iain Johnstone and Rob Tibshirani).



- The SVM path (with Saharon Rosset, Ji Zhu and Rob Tibshirani).



Lasso and Boosting



Boosting Linear Regression

Here is a version of least squares boosting for multiple linear regression: (assume predictors are standardized)

(Incremental) Forward Stagewise

1. Start with $r = y$, $\beta_1, \beta_2, \ldots \beta_p = 0$.
2. Find the predictor x_j most correlated with r
3. Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \epsilon \cdot \text{sign}\langle r, x_j \rangle$
4. Set $r \leftarrow r - \delta_j \cdot x_j$ and repeat steps 2 and 3 many times

$\delta_j = \langle r, x_j \rangle$ gives usual forward stagewise; different from forward stepwise

Analogous to regression boosting, with *trees=predictors*

Lasso (Tibshirani, 1995)

- Assume $\bar{y} = 0$, $\bar{x}_j = 0$, $\mathrm{Var}(x_j) = 1$ for all j.
- Minimize $\sum_i (y_i - \sum_j x_{ij}\beta_j)^2$ subject to $\sum_j |\beta_j| \leq s$
- With orthogonal predictors, solutions are soft thresholded version of least squares coefficients:

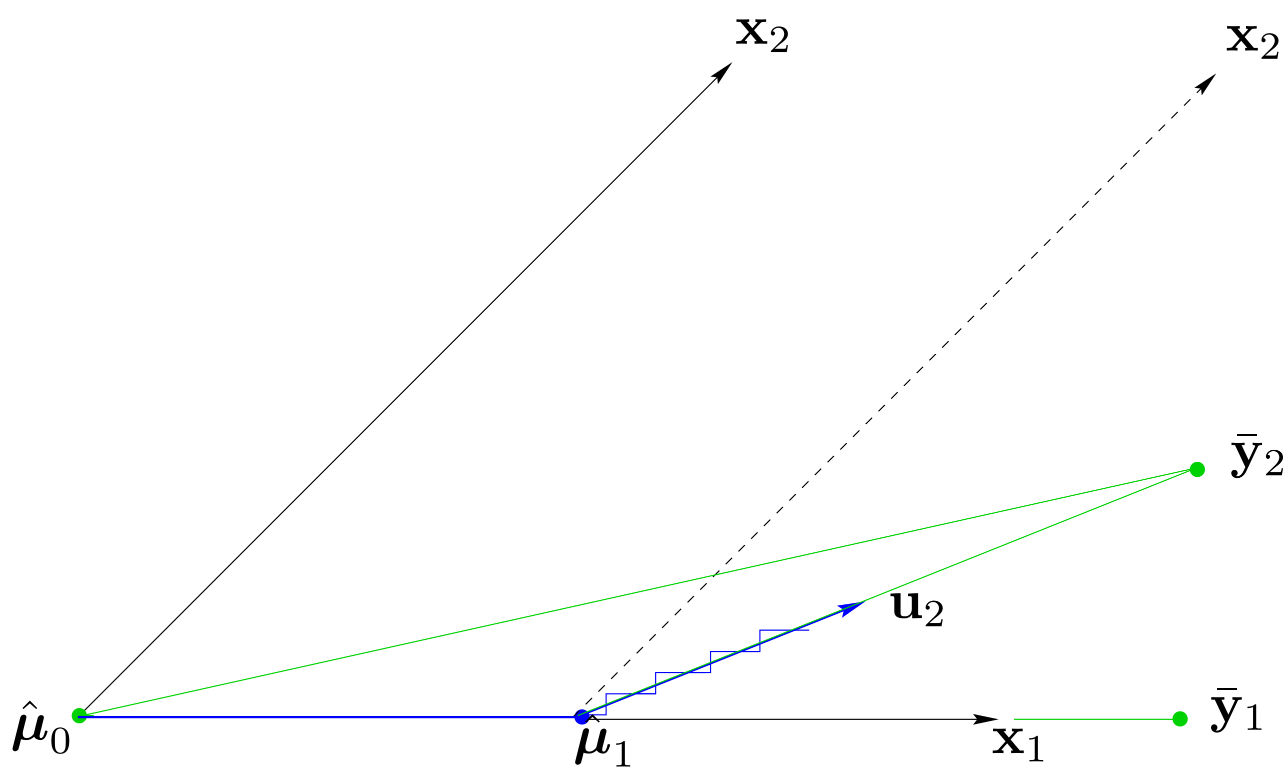
$$\mathrm{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \gamma)_+$$

  (γ is a function of s)
- For small values of the bound s, Lasso does variable selection. See pictures

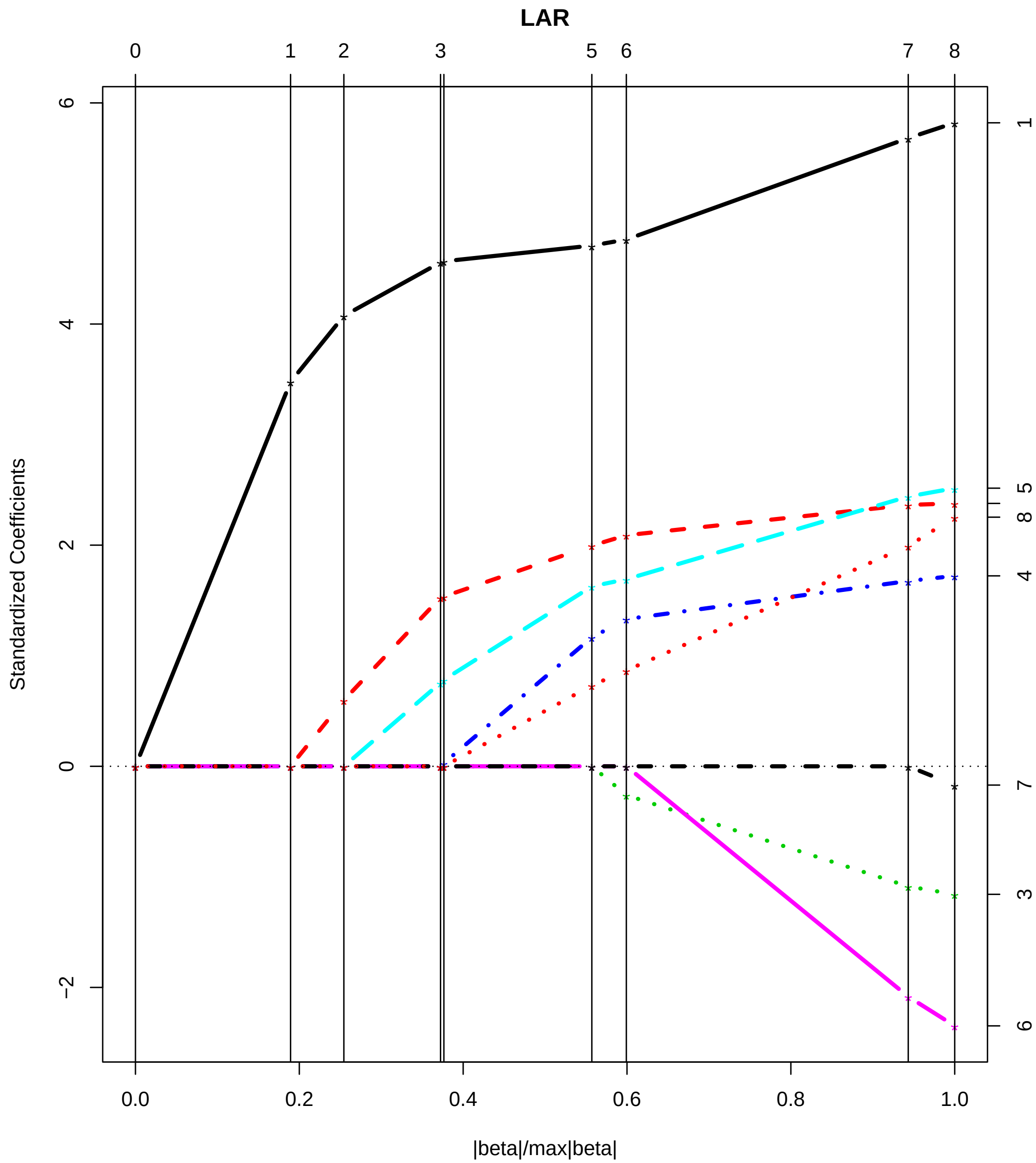
Least Angle Regression — LAR

Like a "more democratic" version of forward stepwise regression.

1. Start with $r = y$, $\hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_p = 0$. Assume x_j standardized.
2. Find predictor x_j most correlated with r.
3. Increase β_j in the direction of $\text{sign}(\text{corr}(r, x_j))$ until some other competitor x_k has as much correlation with current residual as does x_j.
4. Move $(\hat{\beta}_j, \hat{\beta}_k)$ in the joint least squares direction for (x_j, x_k) until some other competitor x_ℓ has as much correlation with the current residual
5. Continue in this way until all predictors have been entered. Stop when $\text{corr}(r, x_j) = 0\ \forall\ j$, i.e. OLS solution.



The LAR direction $\mathbf{u}_2$ at step 2 makes an equal angle with $\mathbf{x}_1$ and $\mathbf{x}_2$.


LAR
0
1
2
3
5
6
7
8
6
4
2
0
−2
Standardized Coefficients
1
5
8
4
7
3
6
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|

LAR vs Lasso

- A modification of LAR fits the entire Lasso path.
- Start with LAR. If a coefficient crosses zero, stop. Drop that predictor, recompute the best direction and continue. This gives the Lasso path
- Proof (lengthy): use Karush-Kuhn-Tucker theory of convex optimization. Informally:

$$\frac{\partial}{\partial \beta_j}\left\{||\mathbf{y}-\mathbf{X}\beta||^2 \quad + \quad \lambda \sum_j |\beta_j| \right\} = 0$$

$$\Leftrightarrow$$

$$\langle \mathbf{x}_j, \mathbf{r} \rangle \quad = \quad \frac{\lambda}{2}\text{sign}(\hat{\beta}_j) \quad \text{if } \hat{\beta}_j \neq 0 \text{ (active)}$$

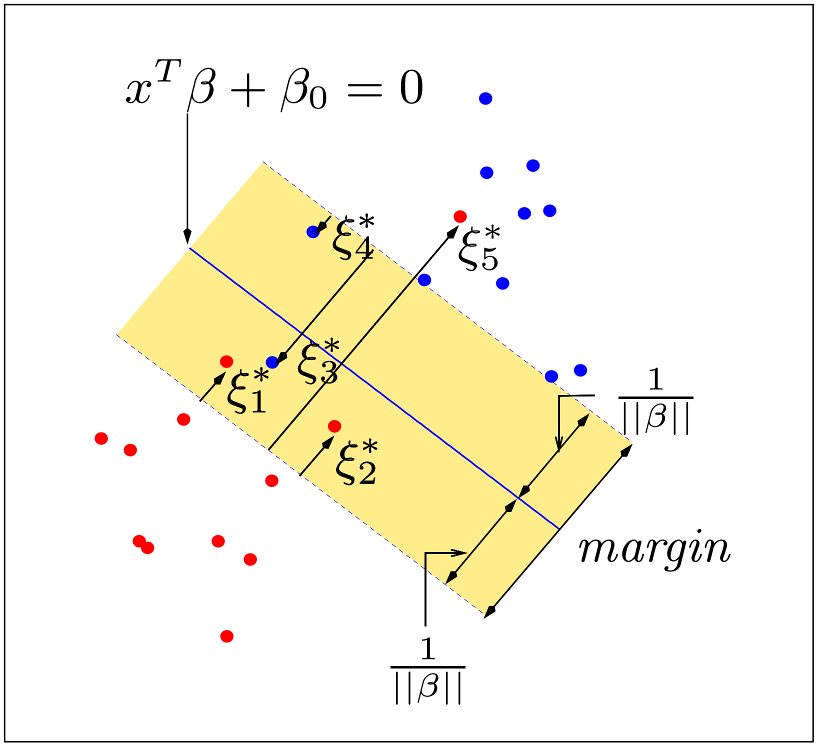
Benefits

- Possible explanation of the benefit of "slow learning" in boosting: it is approximately fitting via an L_1 (lasso) penalty
- new algorithm computes entire Lasso path in same order of computation as one full least squares fit. Splus/R Software on my website or CRAN.
- Degrees of freedom formula for LAR:

  After k steps, degrees of freedom of fit $= k$ (with some regularity conditions)
- For Lasso, the procedure often takes $> p$ steps, since predictors can drop out. Corresponding formula (conjecture):

  Degrees of freedom for last model in sequence with k predictors is equal to k.

Maximal (Soft) Margin Classifier

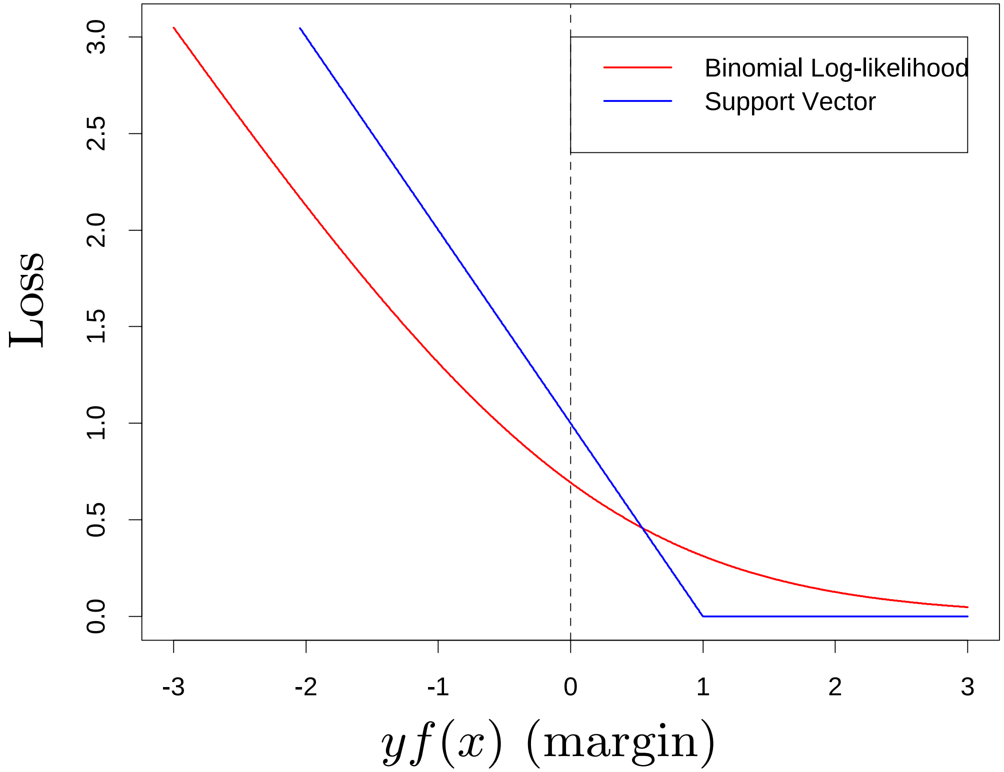


$\xi_i^* = \xi_i/||\beta||$

$$\min_{\beta,\beta_0} ||\beta||^2$$

subject to $y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; \sum_i \xi_i \leq B \; (Budget)$

SVM via Loss + Penalty



With $f(x) = x^T\beta + \beta_0$ and $y_i \in \{-1, 1\}$, consider

$$\min_{\beta_0,\,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2$$

This *hinge loss* criterion is equivalent to the SVM, with $\lambda \sim B$.

Quadratic Programming

$$L_P: \sum_{i=1}^N \xi_i + \frac{\lambda}{2}\beta^T\beta + \sum_{i=1}^N \alpha_i(1 - y_i f(x_i) - \xi_i) - \sum_{i=1}^N \gamma_i \xi_i$$

$$\frac{\partial}{\partial \beta}: \qquad \beta = \frac{1}{\lambda}\sum_{i=1}^N \alpha_i y_i x_i$$

$$\frac{\partial}{\partial \beta_0}: \qquad \sum_{i=1}^N y_i \alpha_i = 0,$$

along with the KKT conditions

$$\begin{aligned} \alpha_i(1 - y_i f(x_i) - \xi_i) &= 0 \\ \gamma_i \xi_i &= 0 \\ 1 - \alpha_i - \gamma_i &= 0 \end{aligned}$$

Implications of the KKT conditions

Observations are in one of three states:

$\mathcal{L} = \{i : y_i f(x_i) < 1,\ \alpha_i = 1\}$, $\mathcal{L}$ for *Left* of the elbow

$\mathcal{E} = \{i : y_i f(x_i) = 1,\ 0 \le \alpha_i \le 1\}$, $\mathcal{E}$ for *Elbow*

$\mathcal{R} = \{i : y_i f(x_i) > 1,\ \alpha_i = 0\}$, $\mathcal{R}$ for *Right* of the elbow

- Start with λ large, and the margin very wide. All $\alpha_i = 1$. As $\lambda \downarrow 0$, the margin gets narrower.
- For the narrowing margin to pass through a point, it's α has to change from 1 to 0 (or from 0 to 1). While this is happening, the point has to *linger* on the margin. Hence the point moves from $\mathcal{L}$ via $\mathcal{E}$ to $\mathcal{R}$.

The Path

- The α_i are piecewise-linear in λ (or $1/C$) [*MOVIES*].
- The points in $\mathcal{E}$ characterize these paths, since points must stay on the margin ($y_i f(x_i) = 1$) while their α_i lie in $(0, 1)$.
- Points can revisit the margin more than once.
- The coefficients β_0 and β are piecewise-linear in $C = 1/\lambda$.
- The margins can stay wedged while their α_i change, if they are "loaded to capacity".
- For non-separable data, the loss $\sum_i \xi_i$ achieves a minimum value, with a positive margin.