

Variable Selection Using Regularized Loss Functions



Adrian Barbu

Logistic Regression

- Binary classification: classes $y=1$ and $y=-1$ (not 0)

$$P(y = 1|\mathbf{x}) = \frac{\exp(w_0 + \sum_i w_i x_i)}{1 + \exp(w_0 + \sum_i w_i x_i)} = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$$

$$P(y = -1|\mathbf{x}) = \frac{1}{1 + \exp(w_0 + \sum_i w_i x_i)} = \frac{1}{1 + \exp(\mathbf{w}^T \mathbf{x})}$$

- Hence

$$P(y|\mathbf{x}) = \frac{1}{1 + \exp(-y\mathbf{w}^T \mathbf{x})}$$

- Conditional negative log likelihood for training examples (\mathbf{x}_j, y_j) :

$$L(\mathbf{w}) = \frac{1}{N} \sum_{j=1}^N \ln(1 + e^{-y_j \mathbf{w}^T \mathbf{x}_j})$$

Variable Selection

■ Why Variable Selection

- Simpler model= better prediction accuracy
- Occam's razor again
- Can have many features e.g. 100000.
- Computationally more efficient.

■ Negative log likelihood + penalty

$$L(\mathbf{w}) = \frac{1}{N} \sum_{j=1}^N \ln(1 + e^{-y_j \mathbf{w}^T \mathbf{x}_j}) + \lambda \sum_{i=1}^M \rho(w_i)$$

Penalty

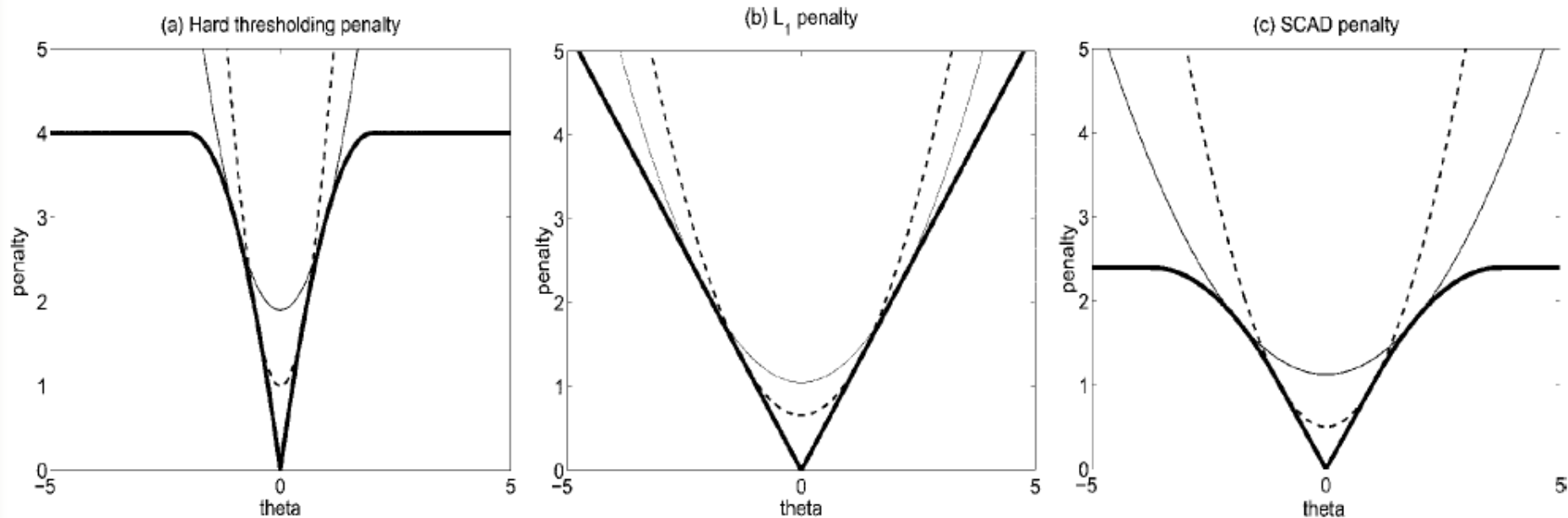
- Penalty = Sparsifying effect

Variable Selection

- Regularized logistic loss:

$$L(\mathbf{w}) = \underbrace{\frac{1}{N} \sum_{j=1}^N \ln(1 + e^{-y_j \mathbf{w}^T \mathbf{x}_j})}_{\text{Logistic Loss}} + \lambda \underbrace{\sum_{i=1}^M \rho(w_i)}_{\text{Penalty}}$$

- Sparsifying penalty functions ρ must be singular at origin



Variable Selection

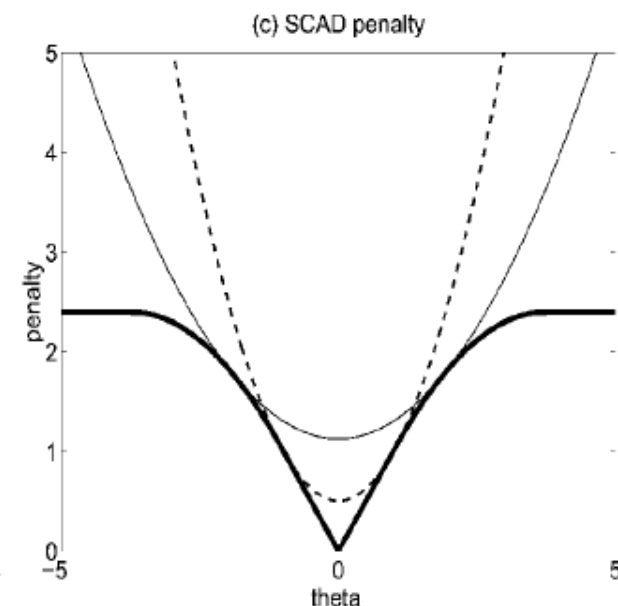
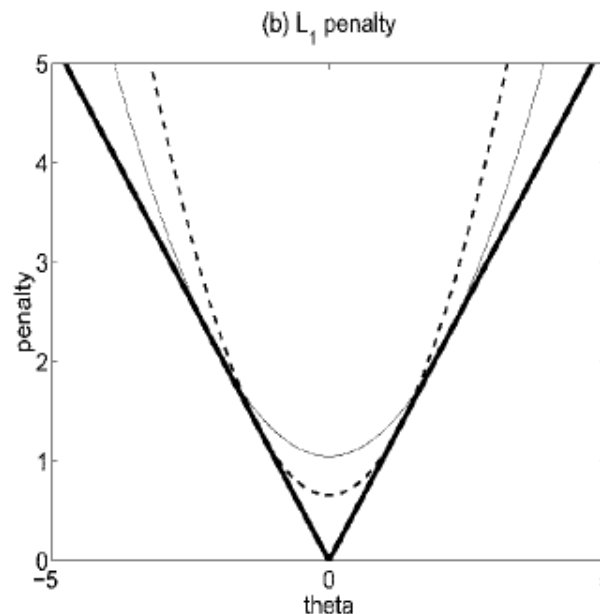
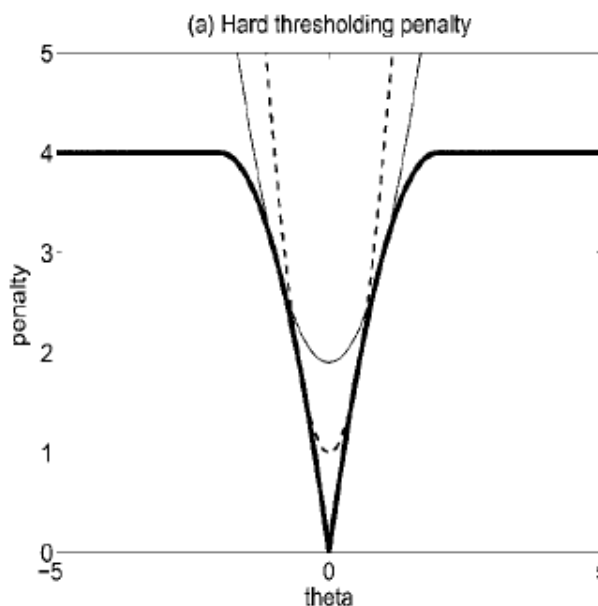
■ Sparsifying penalty functions

- Hard thresholding $\rho_\lambda(x) = \begin{cases} \lambda^2 - (|x| - \lambda)^2 & \text{if } |x| < \lambda \\ \lambda^2 & \text{else} \end{cases}$

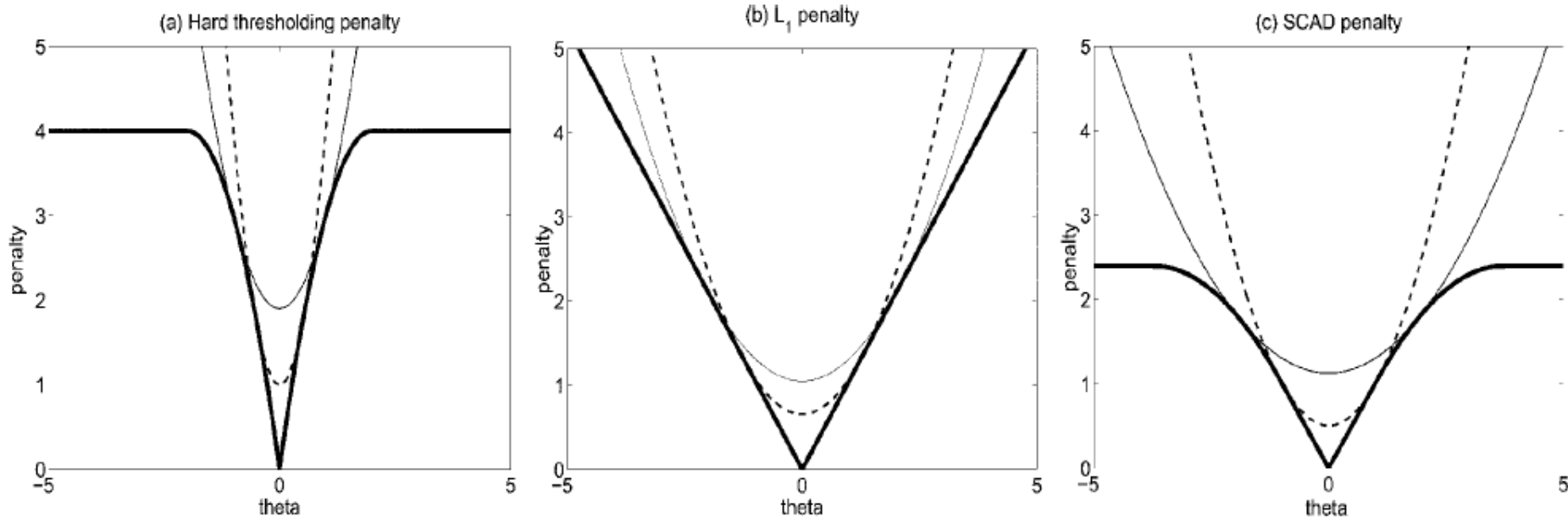
- L1 penalty $\rho(x) = |x|$

- Smoothly clipped absolute deviation (SCAD) penalty

■ L_2 penalty to avoid overfitting $\rho(x) = x^2$



Penalty Functions



- The L_1 penalty biases coefficients (makes them smaller)
- The other penalties result in non-convex optimization
 - Suboptimal result
 - Computationally intensive

Regularized Loss Functions

- More generally:

$$L(\mathbf{w}) = \frac{1}{N} \sum_{j=1}^N f(\mathbf{w}^T \mathbf{x}_j, y_j) + \lambda \sum_{i=1}^M \rho(w_i)$$

- Loss functions f

- Logistic (binomial deviance)

$$f(x, y) = \log_2(1 + e^{-xy})$$

- Exponential loss

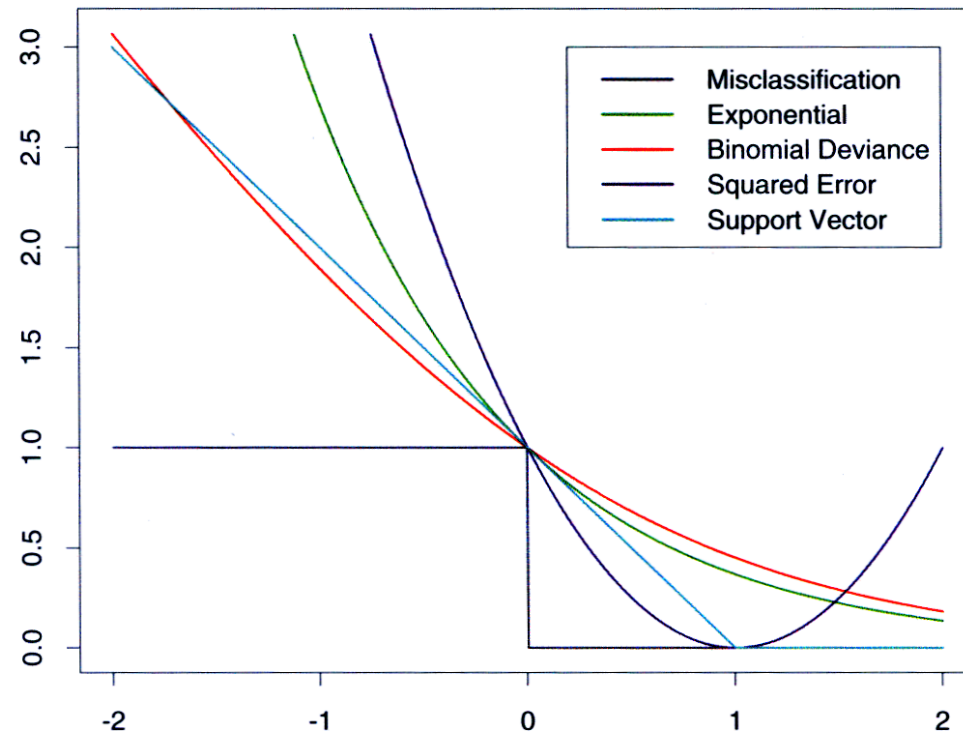
$$f(x, y) = e^{-xy}$$

- Square loss

$$f(x, y) = (x - y)^2$$

- Hinge (SVM) loss

$$f(x, y) = \begin{cases} 1 - xy & \text{if } xy > 1 \\ 0 & \text{else} \end{cases}$$



Learning with Regularized Loss Functions

■ Convex Optimization

- Any convex loss with L_1 or L_2 penalty
- E.g:
 - Logistic loss with L_1 penalty (Boyd 2007)
 - SVM (Hinge loss with L_2 penalty)
- Can be slow

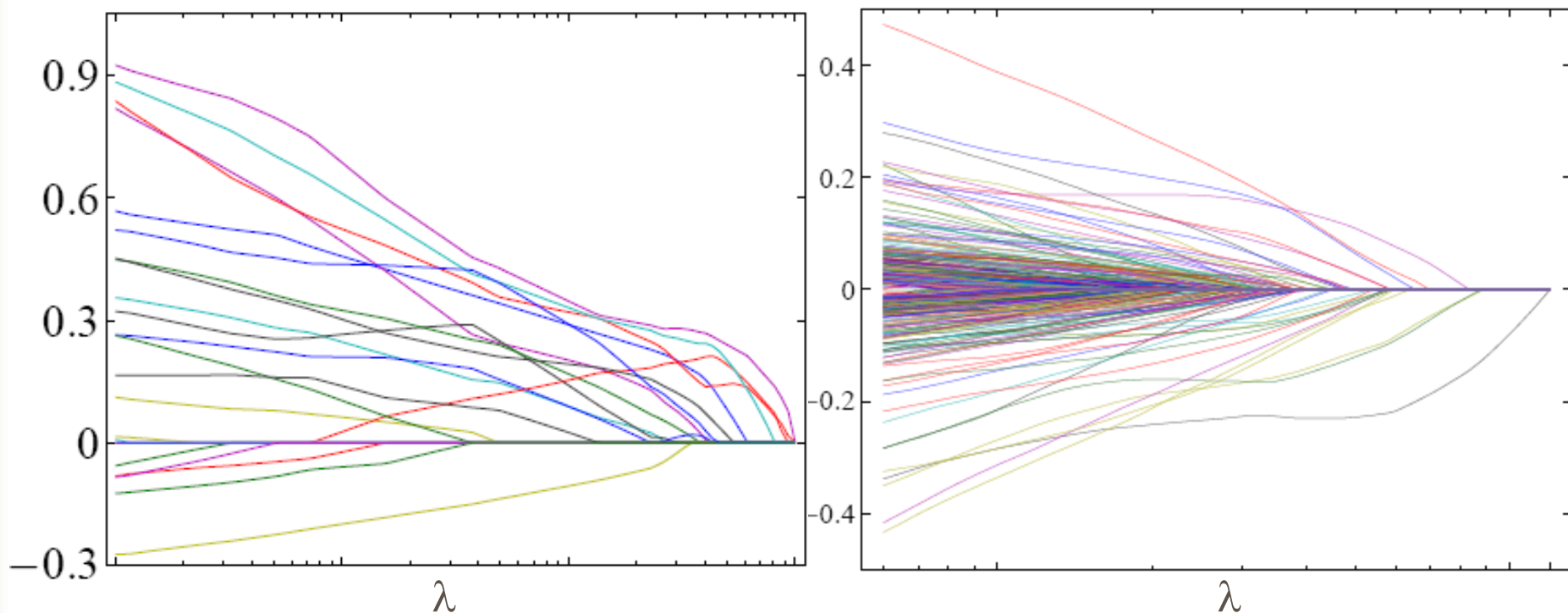
■ Analytical:

- Least Angle Regression (LARS) (Efron et al, 2004)
for Square loss with L_1 penalty
- For other combinations, see Rosset & Zhu, 2007
- Faster but only defined for some combinations Loss/Penalty

■ Greedy:

- Grafting (Perking et al, 2003)
 - Keeps a set F of nonzero weights w_i
 - Adds variables to F based on a gradient criterion
 - Gradient descent using the weights in F only
 - Fast but can be suboptimal

Regularization Path



- Graph of the weights vs. λ
- In some cases it is piecewise linear
 - Can be obtained using the LARS algorithm

TISP – Thresholding-based Iterative Selection Procedure

- Penalized Regression:

$$L(\mathbf{w}) = \frac{1}{N} \sum_{j=1}^N \|\mathbf{w}^T \mathbf{x}_j - y_j\|^2 + \lambda \sum_{i=1}^M \rho(w_i)$$

- Algorithm iterates:

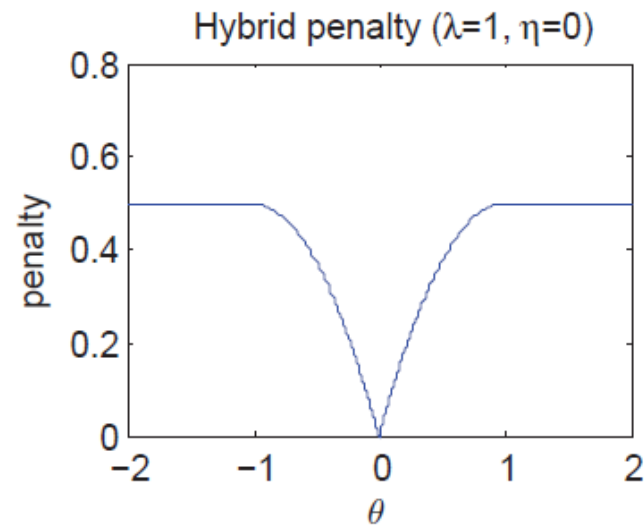
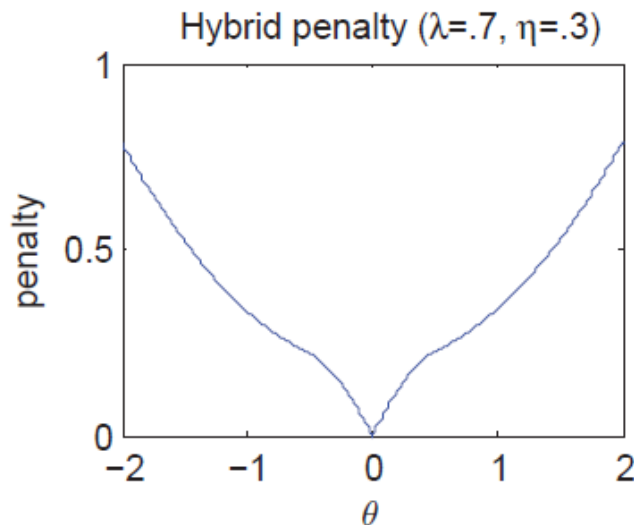
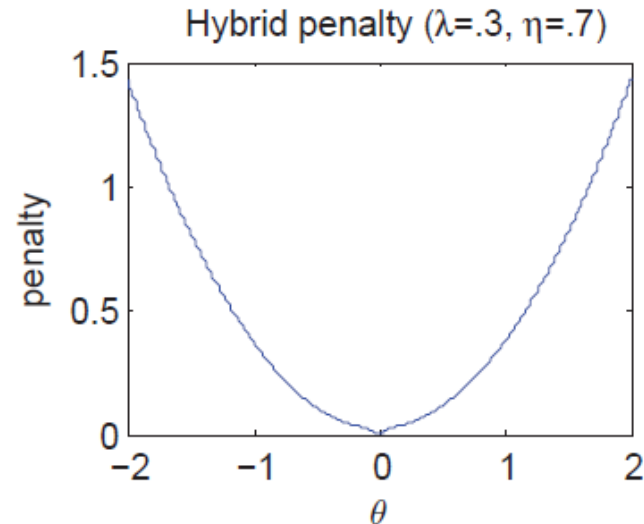
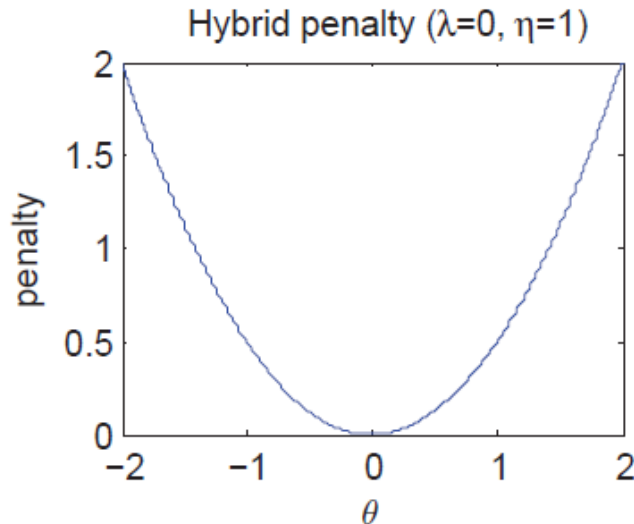
$$\mathbf{w}^{(t+1)} = \Theta(\mathbf{w}^{(t)} + \eta(X^T \mathbf{y} - X^T X \mathbf{w}^{(t)}), \lambda)$$

where η could be $1/N$ and Θ is such that

$$\rho(x) = \int_0^{|x|} [\sup\{t, \Theta(t, \lambda) \leq u\} - u] du$$

TISP

- Example of penalties: $\Theta(x, \lambda, \eta) = \begin{cases} 0 & \text{if } |x| \leq \lambda \\ \frac{x}{1+\eta} & \text{if } |x| > \lambda \end{cases}$



TISP for Classification

- Penalized logistic regression:

$$L(\mathbf{w}) = \frac{1}{N} \sum_{j=1}^N \ln(1 + e^{-y_j \mathbf{w}^T \mathbf{x}_j}) + \lambda \sum_{i=1}^M \rho(w_i)$$

- Algorithm iterates:

$$\mathbf{w}^{(t+1)} = \Theta(\mathbf{w}^{(t)} + \eta X^T \left[\mathbf{y} - \frac{1}{1 + \exp(-X \mathbf{w}^{(t)})} \right], \lambda)$$

where η could be $1/N$

References

- B. Efron, T. Hastie, I.M. Johnstone and R. Tibshirani, Least angle regression (with discussion). *Annals of Statistics* 32 407–499. (2004).
- J Fan, R Li - Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 2001
- K. Koh, S.-J. Kim, and S. Boyd. An Interior-Point Method for Large-Scale l_1 -Regularized Logistic Regression. *Journal of Machine Learning Research*, 8:1519-1555, July 2007.
- S Perkins, K Lacker, J Theiler. Grafting: Fast, incremental feature selection by gradient descent in function space *Journal of Machine Learning Research*, 2003
- S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics* 35, Number 3 (2007)
- Y. She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, 2009.
- Y She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, 2012