Machine Learning 2

6. Sparse Linear Models

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Syllabus

		A. Advanced Supervised Learning
Tue. 5.4.	(1)	A.1 Generalized Linear Models
Tue. 12.4.	(2)	A.2 Gaussian Processes
Tue. 19.4.	(3)	A.2b Gaussian Processes (ctd.)
Tue. 26.4.	(4)	A.3 Advanced Support Vector Machines
Tue. 3.5.	(5)	A.4 Neural Networks
Tue. 10.5.	(6)	A.5 Ensembles (Stacking)
Tue. 17.5.	—	— Pentecoste Break —
Tue. 24.5.	(7)	A.5b Ensembles (Boosting, ctd.)
Tue. 31.5.	(8)	A.5c Ensembles (Mixtures of Experts, ctd.)
Tue. 7.6.	(9)	A.6 Sparse Linear Models — L1 regularization
Tue. 14.6.	(10)	A.6b Sparse Linear Models — L1 regularization (ctd.)
Tue. 21.6.	(11)	A.7. Sparse Linear Models — Further Methods
		B. Complex Predictors
Tue. 28.6.	(12)	B.1 Latent Dirichlet Allocation (LDA)
Tue. 5.7.	(13)	B.2 Deep Learning

Outline

1. Homotopy Methods: Least Angle Regression
2. Proximal Gradient Methods
3. Laplace Priors

Outline

Sparse Models so far

- Variable subset selection
  - forward search, backward search
- L1 regularization / Lasso
  - Coordinate descent (shooting algorithm)

L1 Regularization

$$\begin{aligned} \text{min.} \quad & f(\hat{\theta}) := \ell(y, \hat{y}(\hat{\theta}, X)) + \lambda ||\hat{\theta}||_1 \\ & \hat{\theta} \in \mathbb{R}^P \end{aligned}$$

is equivalent to

$$\begin{aligned} \text{min.} \quad & f(\hat{\theta}) := \ell(y, \hat{y}(\hat{\theta}, X)) \\ & ||\hat{\theta}||_1 \leq B \\ & \hat{\theta} \in \mathbb{R}^P \end{aligned}$$

with

$$B := ||\hat{\theta}^*||_1$$

Homotopy Methods

$$\text{min. } f(\hat{\theta}) := \ell(y, \hat{y}(\hat{\theta}, X)) \ + \lambda ||\hat{\theta}||_1$$

or equivalently

$$\text{min. } f(\hat{\theta}) := \ell(y, \hat{y}(\hat{\theta}, X))$$
$$||\hat{\theta}||_1 \leq B$$

- start with a solution for large $\lambda^{(0)}$ (or equiv. $B := 0$)
  - then $\hat{\theta}^{(0)} = 0$.
- stepwise decrease $\lambda^{(t)}$ (or equiv. increase B)
  - learn $\hat{\theta}^{(t)}$ starting from $\hat{\theta}^{(t-1)}$ (**warmstart**).

Least Angle Regression (LAR)

in step t:

1. choose the predictors with largest correlation with the residuum (**active predictors**):

$$C^{(t-1)} := X^T(y - \hat{y}^{(t-1)})$$

$$A^{(t)} := \arg\max_m |C_m^{(t-1)}|$$

2. regress these predictors on the residuum:

$$X^{(t)} := X_{\cdot, A^{(t)}}$$

$$\hat{\gamma}^{(t)} := \arg\min_\gamma ||y - \hat{y}^{(t-1)} - X^{(t)}\gamma||_2$$

$$= (X^{(t)T} X^{(t)})^{-1} X^{(t)T}(y - \hat{y}^{(t-1)})$$

3. update parameters in this direction:

$$\hat{\beta}^{(t)} := \hat{\beta}^{(t-1)} + \alpha \Delta^{(t)} \hat{\gamma}^{(t)}$$

Note: $\Delta^{(t)}_{m_k,k} := 1$ for $A^{(t)} := \{m_1, m_2, \ldots, m_K\}$, $\Delta^{(t)}_{m,k} := 0$ otherwise.

Least Angle Regression (LAR): step length

Residuum correlations after the update

$$\begin{aligned} C^{(t)} =& X^T(y - \hat{y}^{(t)}) = X^T(y - X\beta^{(t)}) = X^T(y - X(\beta^{(t-1)} + \alpha\Delta^{(t)}\hat{\gamma}^{(t)}) \\ =& C^{(t-1)} - \alpha X^T X \Delta^{(t)} \hat{\gamma}^{(t)} \\ =& C^{(t-1)} - \alpha X^T X^{(t)} \hat{\gamma}^{(t)} \end{aligned}$$

are uniformly reduced for active predictors:

$$C^{(t)}|_{A^{(t)}} = C^{(t-1)}|_{A^{(t)}} - \alpha X^{(t)T} X^{(t)} \hat{\gamma}^{(t)} = (1 - \alpha) C^{(t-1)}|_{A^{(t)}}$$

and may also change for non-active predictors:

$$C_m^{(t)} = C_m^{(t-1)} - \alpha X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)}$$

Note: Maybe a mistake somewhere here. Final formula for α differs from the one in the paper.

Least Angle Regression (LAR): step length (2/2)

Reduce until another predictor has same (max) residuum correlation:

$$C_m^{(t)} = C_m^{(t-1)} - \alpha X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)} \stackrel{!}{=} (1-\alpha) C_{\max}^{(t-1)}$$

$$\alpha = \frac{C_{\max}^{(t-1)} - C_m^{(t-1)}}{C_{\max}^{(t-1)} - X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)}}$$

or for negative correlations:

$$C_m^{(t)} = C_m^{(t-1)} - \alpha X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)} \stackrel{!}{=} -(1-\alpha) C_{\max}^{(t-1)}$$

$$\alpha = \frac{C_{\max}^{(t-1)} + C_m^{(t-1)}}{C_{\max}^{(t-1)} + X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)}}$$

yielding

$$\alpha := \min\left\{ \left(\frac{C_{\max}^{(t-1)} - C_m^{(t-1)}}{C_{\max}^{(t-1)} - X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)}} \right)_0, \left(\frac{C_{\max}^{(t-1)} + C_m^{(t-1)}}{C_{\max}^{(t-1)} + X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)}} \right)_0 \mid m \in \{1,\ldots,M\} \setminus A^{(t)} \right\}$$

Example

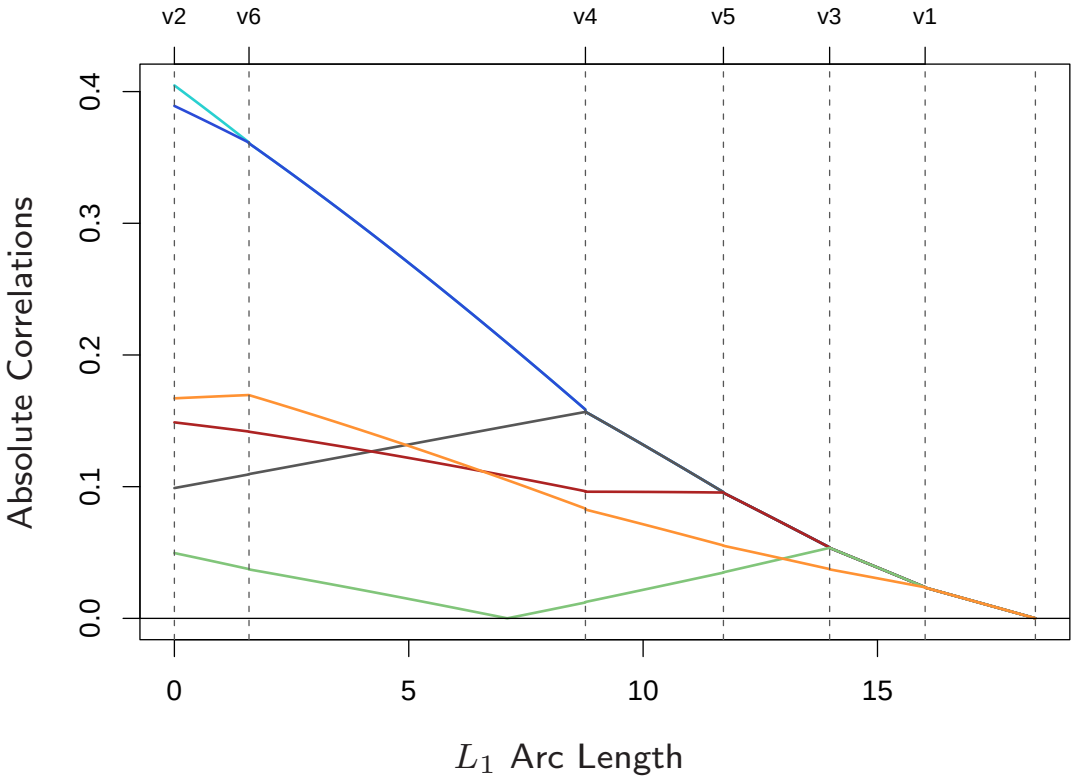
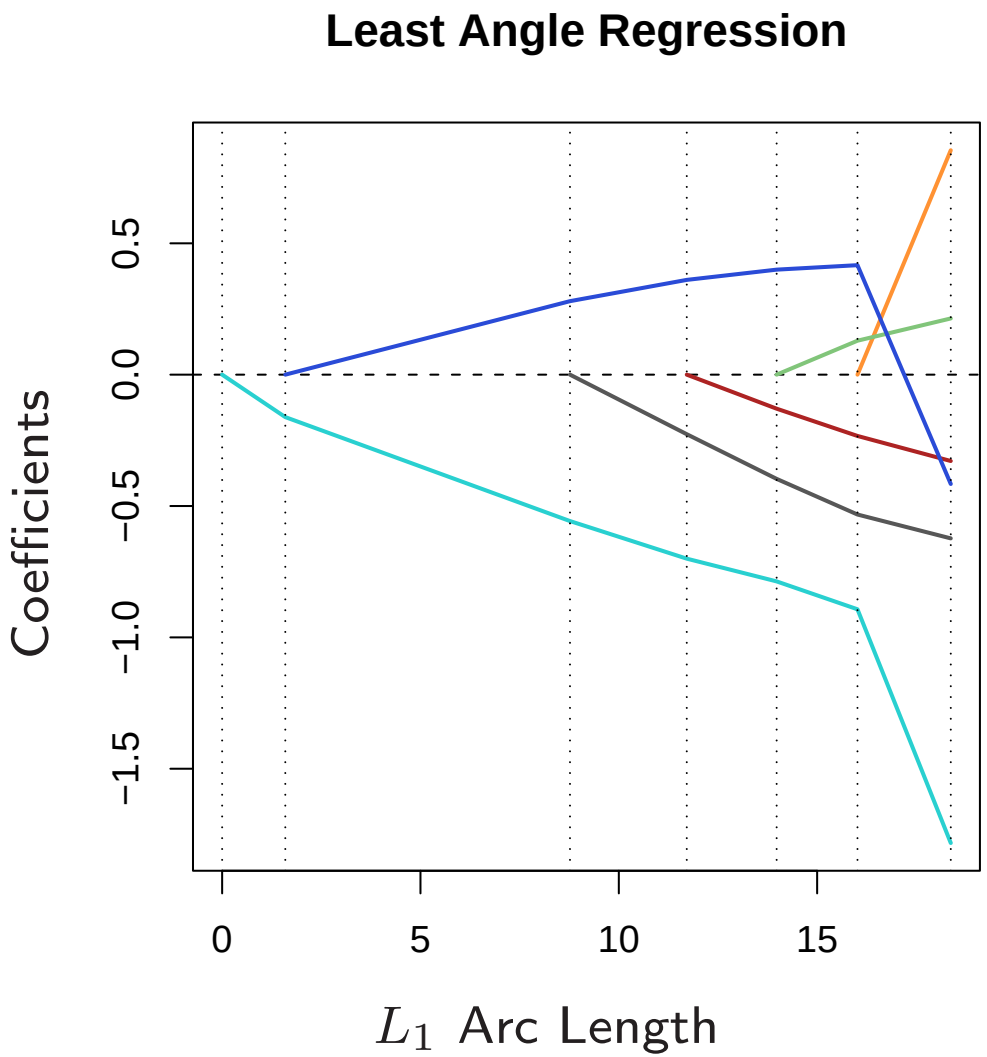


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

[HTFF05, p. 75]

Example

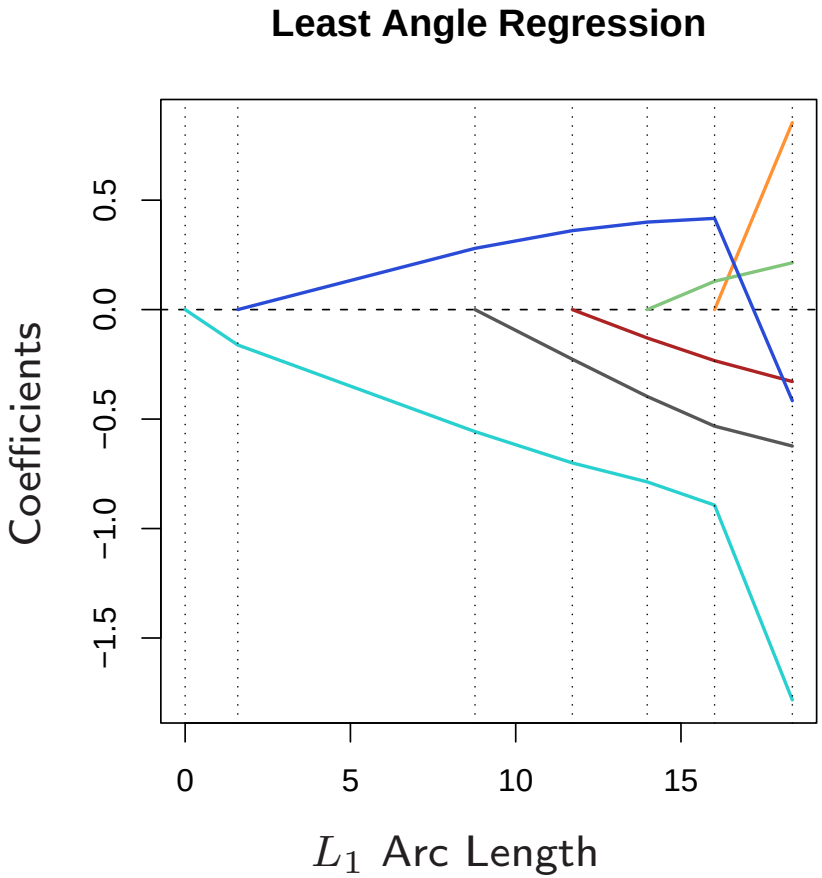


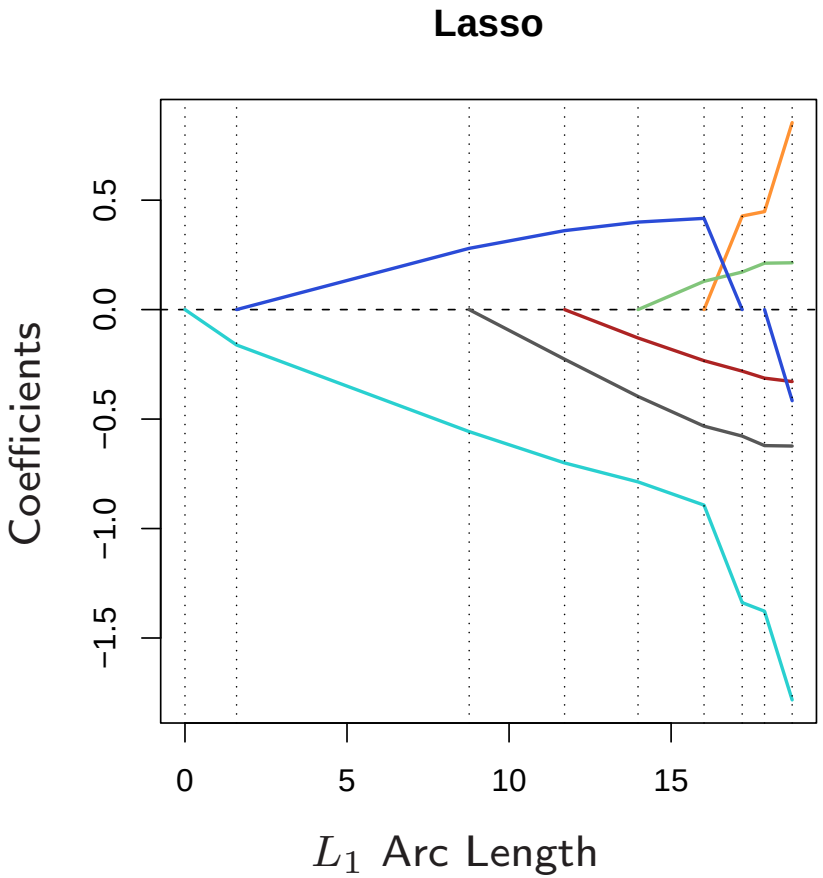
[HTFF05, p. 75]

Remarks

- algorithm can be used two ways:
  1. Estimate parameters for **all** λ (**regularization path**)
  2. Estimate parameters for **a specific** λ (Homotopy method)
     - start with large $\lambda^{(0)}$, stop once $\lambda^{(t)} < \lambda$ reached.
- not straightforward to extend from regression to GLMs
- LAR can be modified to solve the LASSO:
  - if the parameter $\beta_m^{(t)}$ for an active predictor m becomes 0 or changes sign, drop it from the active set.

Example





[HTFF05, p. 75]

Outline

Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

Can be solved analytically for some typical (possibly non-differentiable) regularization functions:

- $f := \lambda||x||_2^2$: $\text{prox}_f(x^0) = \frac{1}{2\lambda + 1} x^0$

Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

Can be solved analytically for some typical (possibly non-differentiable) regularization functions:

- $f := \lambda||x||_2^2$: $$\text{prox}_f(x^0) = \frac{1}{2\lambda + 1} x^0$$
- $f := \lambda||x||_1$:

$$\text{prox}_f(x^0) = \text{soft}(x^0, \lambda) := (\text{soft}(x_n^0, \lambda))_{n=1,\ldots,N}$$

$$\text{soft}(z, \lambda) := \text{sign}(z)(|z| - \lambda)_0$$

Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

Can be solved analytically for some typical (possibly non-differentiable) regularization functions:

- $f := \lambda||x||_2^2$: $\quad \text{prox}_f(x^0) = \frac{1}{2\lambda + 1} x^0$
- $f := \lambda||x||_1$:

$$\text{prox}_f(x^0) = \text{soft}(x^0, \lambda) := (\text{soft}(x_n^0, \lambda))_{n=1,\ldots,N}$$

$$\text{soft}(z, \lambda) := \text{sign}(z)(|z| - \lambda)_0$$

- $f := \lambda||x||_0$:

$$\text{prox}_f(x^0) = \text{hard}(x^0, \lambda) := (\text{hard}(x_n^0, \lambda))_{n=1,\ldots,N},$$

$$\text{hard}(z, \lambda) := \delta(|z| \geq \lambda)\, z$$

More Analytical Solutions for the Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

$f := I_C$ for a **convex set** C and $I_C(x) := \begin{cases} 0, & \text{if } x \in C \\ \infty, & \text{else} \end{cases}$

$$\text{prox}_f(x^0) = \arg\min_{x \in C} ||x - x^0||_2^2 =: \text{proj}_C(x^0)$$

More Analytical Solutions for the Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

$f := I_C$ for a **convex set** C and $I_C(x) := \begin{cases} 0, & \text{if } x \in C \\ \infty, & \text{else} \end{cases}$

$$\text{prox}_f(x^0) = \arg\min_{x \in C} ||x - x^0||_2^2 =: \text{proj}_C(x^0)$$

- **rectangles** / **box constraints** $C := [l_1, u_1] \times [l_2, u_2] \times \cdots \times [l_N, u_N]$:

$$\text{prox}_f(x^0) = \text{clip}(x^0, C) \quad \text{with } \text{clip}(x^0, C)_n := \min\{\max\{x_n^0, l_n\}, u_n\}$$

More Analytical Solutions for the Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

$f := I_C$ for a **convex set** C and $I_C(x) := \begin{cases} 0, & \text{if } x \in C \\ \infty, & \text{else} \end{cases}$

$$\text{prox}_f(x^0) = \arg\min_{x \in C} ||x - x^0||_2^2 =: \text{proj}_C(x^0)$$

- **rectangles** / **box constraints** $C := [l_1, u_1] \times [l_2, u_2] \times \cdots \times [l_N, u_N]$:

$$\text{prox}_f(x^0) = \text{clip}(x^0, C) \quad \text{with } \text{clip}(x^0, C)_n := \min\{\max\{x_n^0, l_n\}, u_n\}$$

- **euclidean balls** $C := \{x \mid ||x||_2 \leq 1\}$:

$$\text{prox}_f(x^0) = \begin{cases} \frac{x^0}{||x^0||_2}, & \text{if } ||x^0||_2 > 1 \\ x^0, & \text{else} \end{cases}$$

More Analytical Solutions for the Proximal Problem

- find x with minimal f **in a vicinity of a given** x^0:

$$\text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||_2^2$$

$f := I_C$ for

- **L1 balls** $C := \{x \mid ||x||_1 \leq 1\}$:

$$\text{prox}_f(x^0) = \begin{cases} \text{soft}(x^0, \lambda), & \text{if } ||x^0||_1 > 1 \\ x^0, & \text{else} \end{cases}$$

$$\text{for } \lambda \text{ with } \sum_{n=1}^{N} (|x_n^0| - \lambda)_0 \stackrel{!}{=} 1$$

Generalized Gradient Descent

$$\min_x g(x) + h(x), \quad g, h \text{ convex}, g \text{ differentiable}$$

Generalized Gradient Descent:

$$x^{(t+1)} := \text{prox}_{\alpha^{(t)} h}(x^{(t)} - \alpha^{(t)} \nabla g(x^{(t)}))$$

$$\text{with } \text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||^2$$

- two-step approach:
  1. minimize component g via gradient descent
  2. minimize component h via prox operator
- requires control of step size $\alpha^{(t)}$
- generalizes gradient descent to objective functions with non-differentiable additive components
- convergence rate $O(1/t)$.

Application to Regularized Loss Minimization

$$\min \quad f(\theta) := \ell(\theta) + R(\theta)$$

- ℓ loss, convex and differentiable
  - e.g., RSS.
- R regularization, convex, but possibly not differentiable
  - e.g., $||\theta||_1$ or $I_C(\theta) := \begin{cases} 0, & \theta \in C \\ \infty, & \text{else} \end{cases}$

Application to Regularized Loss Minimization

Minimizing

$$\theta^{(t+1)} := \arg\min_{\theta} R(\theta) + \ell(\theta)$$

using a **Taylor expansion around previous estimate** $\theta^{(t)}$:

$$\ell(\theta^{(t+1)}) \approx \ell(\theta^{(t)}) + \nabla\ell(\theta^{(t)})^T(\theta - \theta^{(t)}) + (\theta - \theta^{(t)})^T H (\theta - \theta^{(t)})$$

and **diagonal approximation of the Hessian** $H \approx \alpha^{(t)} I$

$$\begin{aligned} &\approx \ell(\theta^{(t)}) + \nabla\ell(\theta^{(t)})^T(\theta - \theta^{(t)}) + \alpha^{(t)}||\theta - \theta^{(t)}||_2^2 \\ &\propto \alpha^{(t)}||\theta - (\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))||_2^2 \end{aligned}$$

yields a proximal problem

$$\begin{aligned} &\arg\min_{\theta} \frac{1}{2\alpha^{(t)}} R(\theta) + \frac{1}{2}||\theta - (\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))||_2^2 \\ &= \text{prox}_{\frac{1}{2\alpha^{(t)}}R}(\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)})) \end{aligned}$$

Special Cases

$$\theta^{(t+1)} := \text{prox}_{\frac{1}{2\alpha^{(t)}}R}(\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))$$
$$= \arg\min_{\theta} \frac{1}{2\alpha^{(t)}}R(\theta) + \frac{1}{2}||\theta - (\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))||_2^2$$

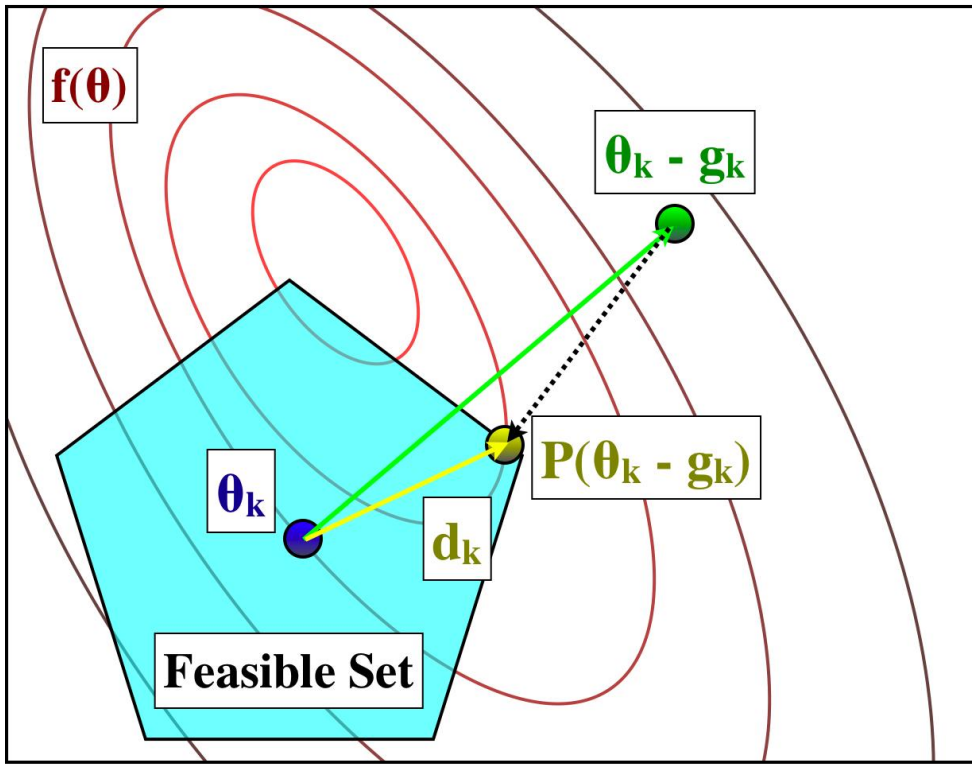
1. $R = 0$ yields **gradient descent**:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)})$$

2. $R = I_C$ yields **projected gradient descent**:

$$\theta^{(t+1)} = \text{proj}_C(\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))$$

Special Cases: Projected Gradient Descent



[Mur12, fig. 13.11]

Special Cases

$$\theta^{(t+1)} := \text{prox}_{\frac{1}{2\alpha^{(t)}}R}(\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))$$
$$= \arg\min_{\theta} \frac{1}{2\alpha^{(t)}}R(\theta) + \frac{1}{2}||\theta - (\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))||_2^2$$

3. $R = \lambda||\theta||_1$ yields **iterative soft thresholding**:

$$\theta^{(t+1)} = \text{soft}(\theta^{(t)} - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}), \frac{\lambda}{2\alpha^{(t)}})$$

Stepsizes $\alpha^{(t)}$

$$\alpha^{(t)} I\,(\theta^{(t)} - \theta^{(t-1)}) \overset{!}{\approx} H\,(\theta^{(t)} - \theta^{(t-1)}) = \nabla\ell(\theta^{(t)}) - \nabla\ell(\theta^{(t-1)})$$

$$\alpha^{(t)} := \arg\min_{\alpha} ||\alpha^{(t)}(\theta^{(t)} - \theta^{(t-1)}) - (\nabla\ell(\theta^{(t)}) - \nabla\ell(\theta^{(t-1)}))$$

$$= \frac{(\theta^{(t)} - \theta^{(t-1)})^T(\nabla\ell(\theta^{(t)}) - \nabla\ell(\theta^{(t-1)}))}{(\theta^{(t)} - \theta^{(t-1)})^T(\theta^{(t)} - \theta^{(t-1)})}$$

called **Barzilai-Borwein stepsize** or **spectral stepsize**.

- does not guarantee decreasing objective values.
- can be used with any gradient descent method.

Iterative Shrinkage and Thresholding Algorithm(ISTA)

- proximal gradient descent for L1 regularization
  - iterative soft thresholding
- Barzilai-Borwein stepsize
- in outer loop, homotopy on λ
  - i.e., gradually reducing $\lambda^{(t)}$ to λ

Note: This algorithm is called Sparse Reconstruction by Separable Approximation (SpaRSA) in the literature.

Algorithm

Algorithm 13.2: Iterative Shrinkage-Thresholding Algorithm (ISTA)

Input: $\mathbf{X} \in \mathbb{R}^{N \times D}$, $\mathbf{y} \in \mathbb{R}^N$, parameters $\lambda \geq 0$, $M \geq 1$, $0 < s < 1$;
Initialize $\boldsymbol{\theta}_0 = \mathbf{0}$, $\alpha = 1$, $\mathbf{r} = \mathbf{y}$, $\lambda_0 = \infty$;
repeat
  $\lambda_t = \max(s||\mathbf{X}^T\mathbf{r}||_\infty, \lambda)$ // Adapt the regularizer ;
  repeat
    $\mathbf{g} = \nabla L(\boldsymbol{\theta})$;
    $\mathbf{u} = \boldsymbol{\theta} - \frac{1}{\alpha}\mathbf{g}$;
    $\boldsymbol{\theta} = \text{soft}(\mathbf{u}, \frac{\lambda_t}{\alpha})$;
    Update α using BB stepsize in Equation 13.82 ;
  until $f(\boldsymbol{\theta})$ *increased too much within the past* M *steps*;
  $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$ // Update residual ;
until $\lambda_t = \lambda$;

[Mur12, p. 446]

Nesterov's Accelerated Generalized Gradient Descent

$$\min_x g(x) + h(x), \quad g, h \text{ convex}, g \text{ differentiable}$$

Generalized Gradient Descent:

$$x^{(t+1)} := \text{prox}_{\alpha^{(t)}h}(x^{(t)} + \frac{t-2}{t+1}(x^{(t)} - x^{(t-1)}) - \alpha^{(t)}\nabla g(x^{(t)}))$$

$$\text{with } \text{prox}_f(x^0) := \arg\min_x f(x) + \frac{1}{2}||x - x^0||^2$$

- added **momentum term**
- works also for vanilla gradient descent ($h = 0$)
- convergence rate $O(1/t^2)$!
- beware, there are at least 3 versions of **Nesterov's method**.

Fast Iterative Shrinkage and Thresholding Alg. (FISTA)

$$\theta^{(t+1)} := \text{prox}_{\frac{1}{2\alpha^{(t)}}R}(\theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)}) - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}))$$

for $R = \lambda||\theta||_1$ yields iterative soft thresholding:

$$\theta^{(t+1)} = \text{soft}(\theta^{(t)} + \frac{t-1}{t+2}(\theta^{(t)} - \theta^{(t-1)}) - \frac{1}{\alpha^{(t)}}\nabla\ell(\theta^{(t)}), \frac{\lambda}{2\alpha^{(t)}})$$

using **Nesterov's Accelerated Generalized Gradient Descent**.

FISTA vs ISTA

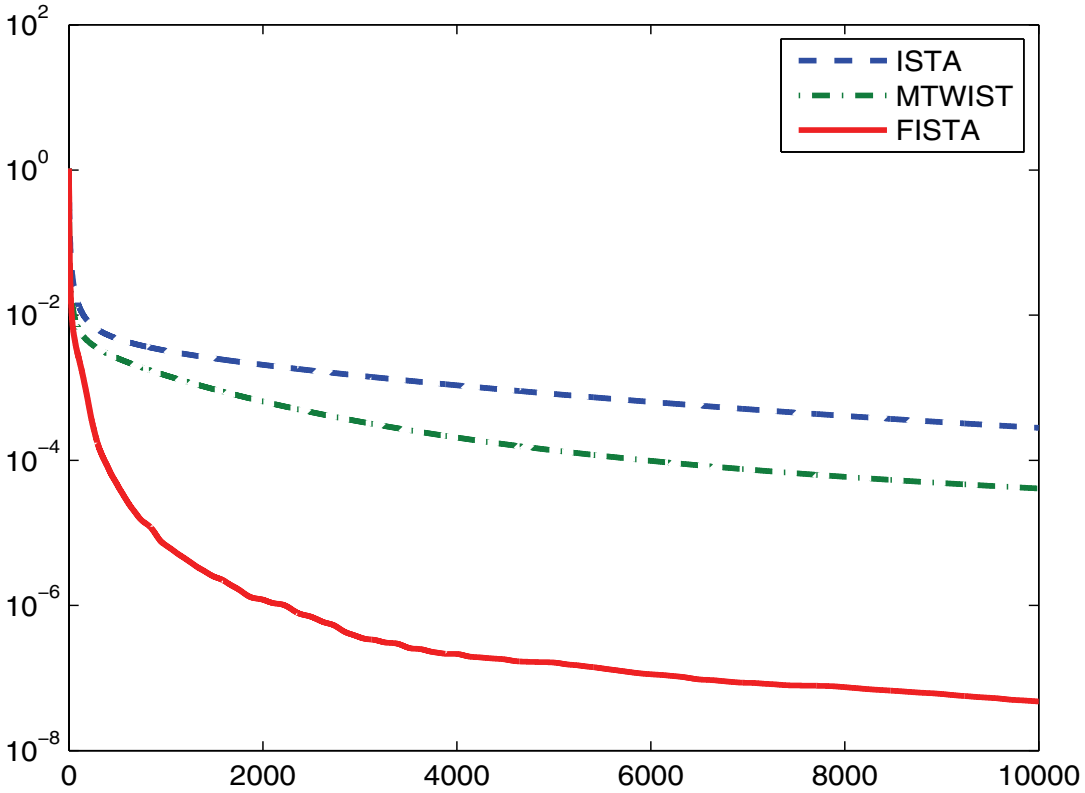


Figure 5. *Comparison of function value errors $F(\mathbf{x}_k) - F(\mathbf{x}^*)$ of ISTA, MTWIST, and FISTA.*

[BT09, p. 19]

Outline

Laplace Priors correspond to L1 regularization

L2 regularization:

$$f(\theta) := ||y - X\beta||_2^2 + \lambda||\beta||_2^2$$

Gaussian priors:

$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T \beta, \sigma^2)$$

$$p(\beta) := \mathcal{N}(\beta \mid 0, \tau^2)$$

$$\propto (\tau^2)^{-M/2} e^{-\frac{1}{2\tau^2}\beta^T\beta}$$

using negative loglikelihood as objective function:

$$f(\beta) := -\log p(y \mid X, \beta)$$

Laplace Priors correspond to L1 regularization

L2 regularization:

$$f(\theta) := ||y - X\beta||_2^2 + \lambda||\beta||_2^2$$

L1 regularization:

$$f(\theta) := ||y - X\beta||_2^2 + \lambda||\beta|$$

Gaussian priors:

$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T\beta, \sigma^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \tau^2)$$
$$\propto (\tau^2)^{-M/2}\, e^{-\frac{1}{2\tau^2}\beta^T\beta}$$

Laplace priors:

$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T\beta, \sigma^2)$$
$$p(\beta_m) := \mathrm{Lap}(\beta_m \mid 0, 1/\gamma)$$
$$= \frac{\gamma}{2}\, e^{-\gamma|\beta_m|}$$

using negative loglikelihood as objective function:

$$f(\beta) := -\log p(y \mid X, \beta)$$

Laplace as Gaussian Scale Mixture

$$\mathrm{Lap}(\beta_m \mid 0, 1/\gamma) = \int \mathcal{N}(\beta_m \mid 0, \tau_m^2)\, \mathrm{Exp}(\tau_m^2 \mid \frac{\gamma^2}{2})\, d\tau_m^2$$

with **exponential distribution**

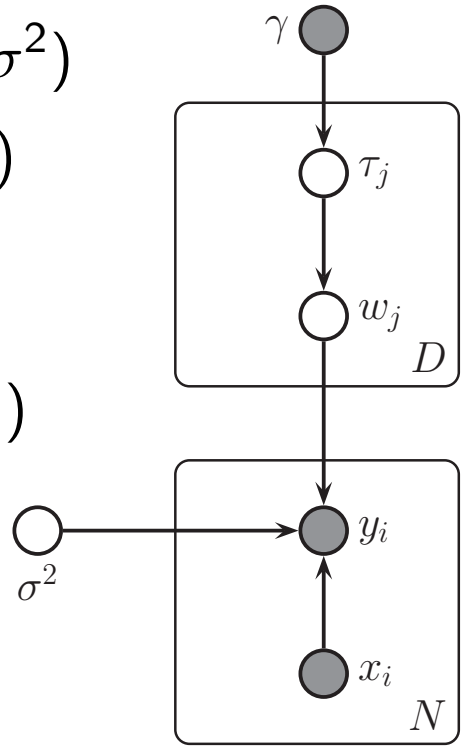
$$\mathrm{Exp}(x \mid \lambda) := \lambda\, e^{-\lambda x}$$

Laplace Prior as Gaussian Scale Mixture

$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T \beta, \sigma^2)$$
$$p(\beta_m \mid \tau_m^2) := \mathcal{N}(\beta_m \mid 0, \tau_m^2)$$
$$p(\tau_m^2) := \mathrm{Exp}(\tau_m^2 \mid \frac{\gamma^2}{2})$$
$$p(\sigma^2) := \mathrm{IG}(\sigma^2 \mid a_\sigma, b_\sigma)$$

Laplace Prior as Gaussian Scale Mixture

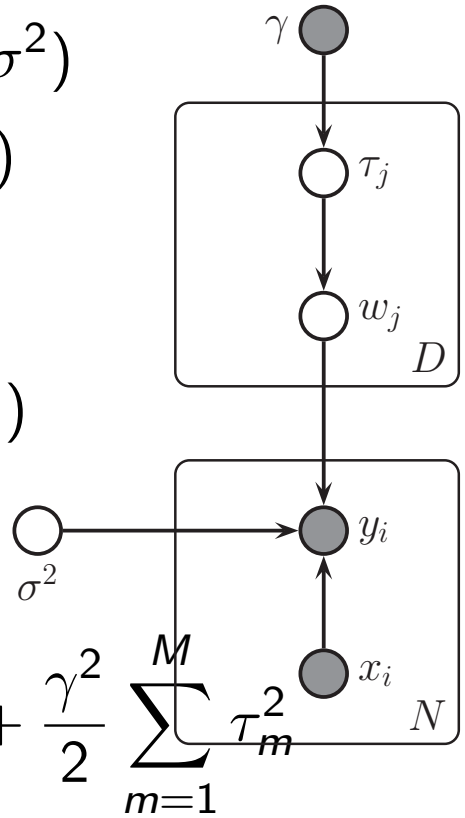
$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T \beta, \sigma^2)$$
$$p(\beta_m \mid \tau_m^2) := \mathcal{N}(\beta_m \mid 0, \tau_m^2)$$
$$p(\tau_m^2) := \mathsf{Exp}(\tau_m^2 \mid \frac{\gamma^2}{2})$$
$$p(\sigma^2) := \mathsf{IG}(\sigma^2 \mid a_\sigma, b_\sigma)$$



[Mur12, p. 446]

Laplace Prior as Gaussian Scale Mixture

$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T \beta, \sigma^2)$$
$$p(\beta_m \mid \tau_m^2) := \mathcal{N}(\beta_m \mid 0, \tau_m^2)$$
$$p(\tau_m^2) := \mathsf{Exp}(\tau_m^2 \mid \frac{\gamma^2}{2})$$
$$p(\sigma^2) := \mathsf{IG}(\sigma^2 \mid a_\sigma, b_\sigma)$$



NLL:

$$f(\beta, \tau^2) = -\frac{1}{2\sigma^2}||y - X\beta||_2^2 - \frac{1}{2}\beta^T \Lambda \beta + \frac{\gamma^2}{2} \sum_{m=1}^{M} \tau_m^2$$

$$\Lambda := \mathsf{diag}(\frac{1}{\tau_1^2}, \frac{1}{\tau_2^2}, \ldots, \frac{1}{\tau_M^2})$$

[Mur12, p. 446]

EM for Laplace Prior

NLL:

$$f(\beta, \tau^2) = -\frac{1}{2\sigma^2}||y - X\beta||_2^2 - \frac{1}{2}\beta^T \Lambda \beta + \frac{\gamma^2}{2}\sum_{m=1}^{M} \tau_m^2$$

$$\Lambda := \text{diag}(\frac{1}{\tau_1^2}, \frac{1}{\tau_2^2}, \ldots, \frac{1}{\tau_M^2})$$

1. Expectaction of τ^2:

$$p(1/\tau_m^2 \mid \beta) = \text{InverseGaussian}(\sqrt{\frac{\gamma^2}{\beta_m^2}}, \gamma^2)$$

$$E(1/\tau_m^2 \mid \beta) = \frac{\gamma}{|\beta_m|}$$

EM for Laplace Prior

NLL:

$$f(\beta, \tau^2) = -\frac{1}{2\sigma^2}||y - X\beta||_2^2 - \frac{1}{2}\beta^T \Lambda \beta + \frac{\gamma^2}{2}\sum_{m=1}^{M} \tau_m^2$$

$$\Lambda := \text{diag}(\frac{1}{\tau_1^2}, \frac{1}{\tau_2^2}, \ldots, \frac{1}{\tau_M^2})$$

2. Expectaction of σ^2:

$$p(\sigma^2 \mid \beta) = \text{IG}(a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(y - X\beta)^T(y - X\beta))$$

$$E(1/\sigma^2 \mid \beta) = \frac{a_\sigma + \frac{N}{2}}{b_\sigma + \frac{1}{2}(y - X\beta)^T(y - X\beta)}$$

EM for Laplace Prior

NLL:

$$f(\beta, \tau^2) = -\frac{1}{2\sigma^2}||y - X\beta||_2^2 - \frac{1}{2}\beta^T \Lambda \beta + \frac{\gamma^2}{2}\sum_{m=1}^{M} \tau_m^2$$

$$\Lambda := \text{diag}(\frac{1}{\tau_1^2}, \frac{1}{\tau_2^2}, \ldots, \frac{1}{\tau_M^2})$$

3. Maximizing β: ridge regression

$$\beta = (\sigma^2 \Lambda + X^T X)^{-1} X^T y$$

Why Laplace Prior?

- Bayesian Lasso
  - provides posterior distribution, not just point estimates
- Can be generalized to other models / losses
- Motivates to experiment with other types of priors, too
- Less scalable than the other methods, though.

Further Readings

- L1 regularization: [Mur12, chapter 13.3–5], [HTFF05, chapter 3.4, 3.8, 4.4.4], [Bis06, chapter 3.1.4].
  - LAR, LARS: [HTFF05, chapter 3.4.4], [Mur12, chapter 13.4.2],
- Non-convex regularizers: [Mur12, chapter 13.6].
- Automatic Relevance Determination (ARD): [Mur12, chapter 13.7], [HTFF05, chapter 11.9.1], [Bis06, chapter 7.2.2].
- Sparse Coding: [Mur12, chapter 13.8].

References

Christopher M. Bishop.
Pattern recognition and machine learning, volume 1.
springer New York, 2006.

Amir Beck and Marc Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin.
The elements of statistical learning: data mining, inference and prediction, volume 27.
2005.

Kevin P. Murphy.
Machine learning: a probabilistic perspective.
The MIT Press, 2012.