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
Fri.	24.4.	(1)	A.1 Generalized Linear Models
Fri.	1.5.	—	— *Labour Day* —
Fri.	8.5.	(2)	A.2 Gaussian Processes
Fri.	15.5.	(3)	A.3 Advanced Support Vector Machines
			B. Ensembles
Fri.	22.5.	(4)	B.1 Stacking
			& B.2 Boosting
Fri.	29.5.	(5)	B.3 Mixtures of Experts
Fri.	5.6.	—	— *Pentecoste Break* —
			C. Sparse Models
Fri.	12.6.	(6)	C.1 Homotopy and Least Angle Regression
Fri.	19.6.	(7)	C.2 Proximal Gradients
Fri.	26.6.	(8)	C.3 Laplace Priors
Fri.	3.7.	(9)	C.4 Automatic Relevance Determination
			D. Complex Predictors
Fri.	10.7.	(10)	D.1 Latent Dirichlet Allocation (LDA)
Fri.	17.7.	(11)	Q & A

Outline

Outline

Sparse Models so far

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

L1 Regularization

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

is equivalent to

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

with

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

L1 Regularization / Equivalence

More generally, given

$$x^* := \arg\min_x f(x) + \lambda g(x), \quad \lambda \geq 0 \tag{1}$$

$$\tilde{x} := \arg\min_{x: g(x) \leq g(x^*)} f(x) \tag{2}$$

then

$$x^* = \tilde{x}$$

because

$$f(\tilde{x}) \underset{(2)}{\leq} f(x^*) \underset{(1)}{\leq} f(\tilde{x}) + \lambda \underbrace{(g(\tilde{x}) - g(x^*))}_{\leq 0} \leq f(\tilde{x})$$

$$\rightsquigarrow \quad f(\tilde{x}) = f(x^*) \quad \rightsquigarrow \quad \tilde{x} = x^*$$

assuming x^* is unique.

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

Homotopy Methods / Prerequisites

For homotopy to work,

1. the parameters as function of λ

$$\hat{\theta}(\lambda) := \arg\min_{\hat{\theta}} \ell(y, \hat{y}(\hat{\theta}, X)) + \lambda ||\hat{\theta}||_1$$

   must be continuous, i.e.,

   - $\hat{y}$ must be continuous in $\hat{\theta}$ and
   - ℓ be continuous in $\hat{y}$.

2. the steps in $\lambda^{(t)}$ must be small enough.

Homotopy for the L1 Weight of Linear Regression

Most simple model: linear regression

- model $\hat{y}(\hat{\theta}, X) := X\hat{\theta}$
- loss $\ell(y, \hat{y}) := ||y - \hat{y}||_2^2$

Advantage: can find optimal $\lambda^{(t)}$ sequence analytically! (actually $B^{(t)}$)

Homotopy for the L1 Weight of Linear Regression

Most simple model: linear regression

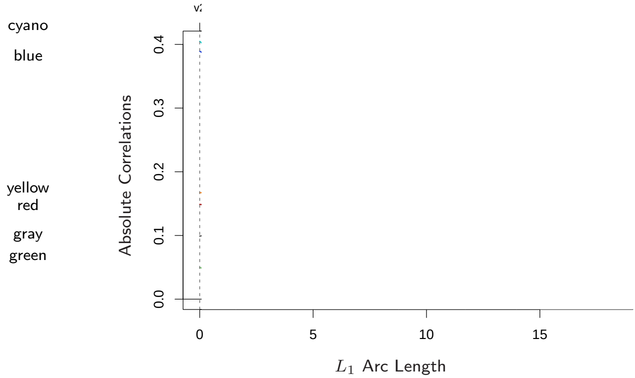
- model $\hat{y}(\hat{\theta}, X) := X\hat{\theta}$
- loss $\ell(y, \hat{y}) := ||y - \hat{y}||_2^2$

Advantage: can find optimal $\lambda^{(t)}$ sequence analytically! (actually $B^{(t)}$)

Imagine the following situation:

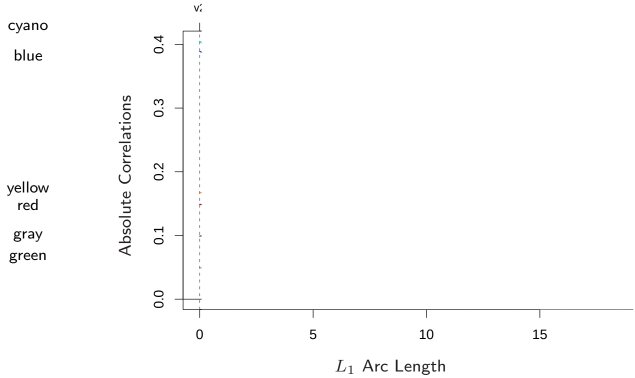
- initially all parameters $\hat{\theta}_m = 0$.
- you can add one variable x_m to the model
  - by setting its parameter $\hat{\theta}_m$ to a small positive or negative value ϵ.
- the goal is to reduce the error as much as possible.
- Q: which parameter $\hat{\theta}_m$ would you choose?

Example



Q: which parameter will you pick initially to reduce the loss maximally?

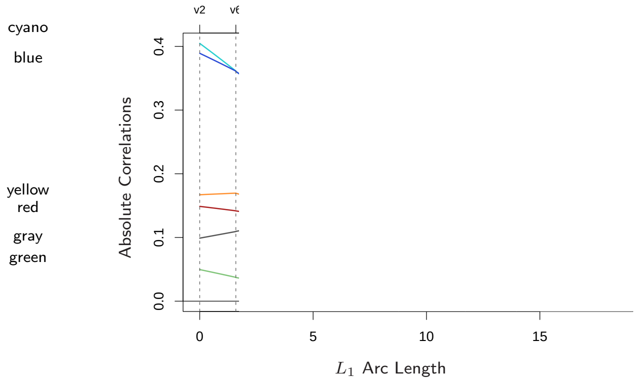
Example



Q: which parameter will you pick initially to reduce the loss maximally?

Q_2: How will the cyano curve go?

Example



Q: which parameter will you pick initially
to reduce the loss maximally?

Q_2: How will the cyano curve go?

Least Angle Regression (LAR)

in step t:

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

$$C^{(t-1)} := X^T(y - \hat{y}^{(t-1)})$$
$$A^{(t)} := \arg\max_m |C_m^{(t-1)}| \qquad \text{(a set!)}$$

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

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

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:

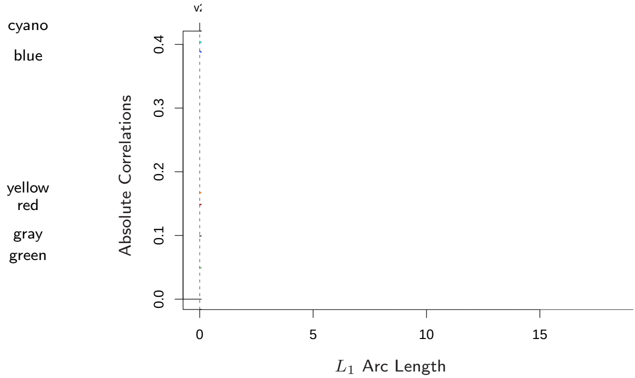
$$C_m^{(t)} = C_m^{(t-1)} - \alpha X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)} \stackrel{!}{=} -(1-\alpha) C_{\max}^{(t-1)}$$

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

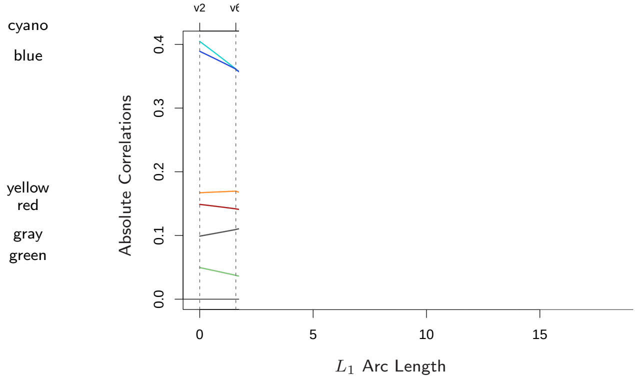
yielding

$$\alpha := \operatorname{minpos}\left\{ \frac{C_{\max}^{(t-1)} - C_m^{(t-1)}}{C_{\max}^{(t-1)} - X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)}}, \frac{C_{\max}^{(t-1)} + C_m^{(t-1)}}{C_{\max}^{(t-1)} + X_{\cdot,m}^T X^{(t)} \hat{\gamma}^{(t)}} \mid m \in \{1, \ldots, M\} \setminus A^{(t)} \right\}, \quad \operatorname{minpos}(X) := \min\{x \in X \mid x > 0\}$$

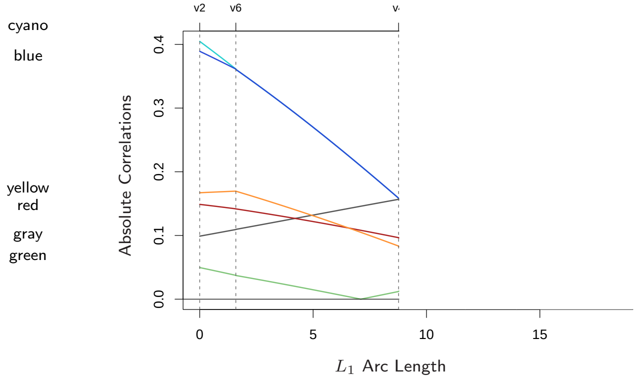
Example


cyano
blue
yellow
red
gray
green
Absolute Correlations
0.0
0.1
0.2
0.3
0.4
0
5
10
15
L_1 Arc Length

Example


cyano
blue
yellow
red
gray
green
v2
v6
Absolute Correlations
0.0
0.1
0.2
0.3
0.4
0
5
10
15
L_1 Arc Length

Example


cyano
blue
yellow
red
gray
green
v2
v6
v4
Absolute Correlations
0.0
0.1
0.2
0.3
0.4
0
5
10
15
L_1 Arc Length

Example

cyano

blue

yellow
red

gray
green

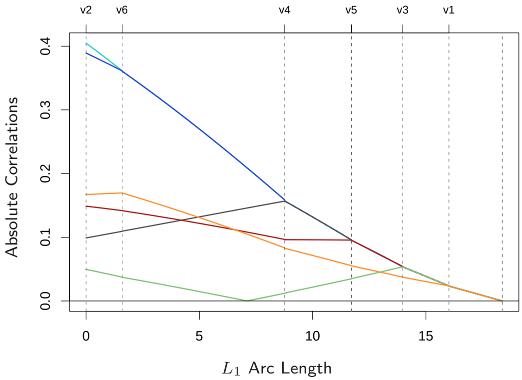
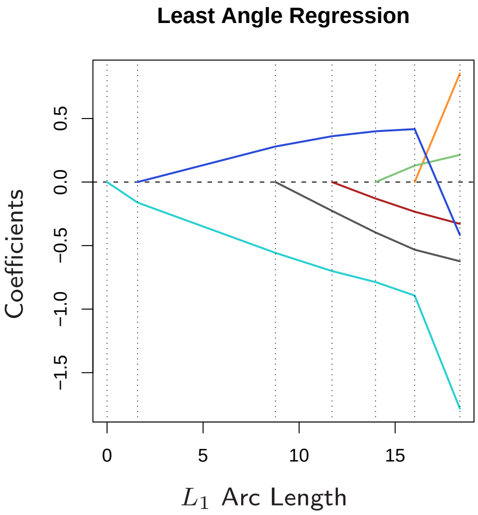


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

Example

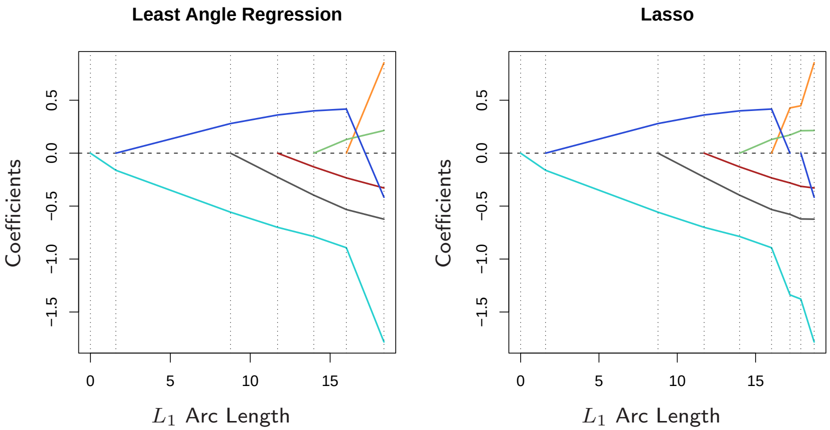

Least Angle Regression
Coefficients
0.5
0.0
−0.5
−1.0
−1.5
0
5
10
15
L_1 Arc Length

[?, 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.
- also called Least Angle Regression and Shrinkage (LARS)

Example


Least Angle Regression
Lasso
Coefficients
0.5
0.0
−0.5
−1.0
−1.5
0
5
10
15
L_1 Arc Length

[?, p. 75]

Outline

1. Homotopy Methods: Least Angle Regression

2. Proximal Gradient Methods

3. Laplace Priors (Bayesian Lasso)

Regularized

We want to compute models

$$\theta^* = \arg\min_\theta \underbrace{L(\theta)}_{\text{Loss}} + \underbrace{R(\theta)}_{\text{Regularization}}$$

Even when R is not differentiable, e.g.

- $R(\theta) = \|\theta\|_1$ (L^1 regularization, LASSO)
- $R(\theta) = I_C(\theta) = \begin{cases} 0 : & \theta \in C \\ \infty : & \theta \notin C \end{cases}$ (hard constraint)

Regularized

We want to compute models

$$\theta^* = \arg\min_{\theta} \underbrace{L(\theta)}_{\text{Loss}} + \underbrace{R(\theta)}_{\text{Regularization}}$$

Even when R is not differentiable, e.g.

- $R(\theta) = \|\theta\|_1$ (L^1 regularization, LASSO)
- $R(\theta) = I_C(\theta) = \begin{cases} 0: & \theta \in C \\ \infty: & \theta \notin C \end{cases}$ (hard constraint)

Observation: For simple loss functions, we can sometimes compute θ^* analytically

$$\arg\min_{\theta} \frac{1}{2}\|\theta - y\|_2^2 + \lambda\|\theta\|_1 = \text{soft}(y, \lambda)$$

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:

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

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial(\lambda x^T x + \frac{1}{2}(x-x^0)^T(x-x^0))}{\partial x} \\ &= 2\lambda x + (x - x^0) = (2\lambda+1)x - x^0 \\ \rightsquigarrow \quad x &= \frac{1}{2\lambda+1}x^0 \end{aligned}$$

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:

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

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

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

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:

- $f := \lambda||x||_2^2$:
$$\text{prox}_f(x^0) = \frac{1}{2\lambda + 1} x^0$$
- $f := \lambda||x||_1$:
$$\text{prox}_f(x^0) = \text{soft}(x^0, \lambda) := (\text{soft}(x_n^0, \lambda))_{n=1,\ldots,N}$$
$$\text{soft}(z, \lambda) := \text{sign}(z)(|z| - \lambda)_0$$
- $f := \lambda||x||_0$:
$$\text{prox}_f(x^0) = \text{hard}(x^0, \lambda) := (\text{hard}(x_n^0, \lambda))_{n=1,\ldots,N},$$
$$\text{hard}(z, \lambda) := \delta(|z| \geq \lambda)\, z$$

More Analytical Solutions for the Proximal Problem

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

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

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

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

More Analytical Solutions for the Proximal Problem

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

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

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

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

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

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

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

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

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

Deriving Generalized Gradient Descent (1/2)

$$\min_x f(x) := g(x) + h(x), \quad g, h \text{ convex}, g \text{ differentiable}, h \text{ possibly not}$$

using a **Taylor expansion of g around previous solution $x^{(t)}$**:

$$g(x) \approx g(x^{(t)}) + \nabla g(x^{(t)})(x - x^{(t)}) + \frac{1}{2}(x - x^{(t)})^T H(x - x^{(t)})$$

and **diagonal approximation of the Hessian** $H \approx \frac{1}{\alpha^{(t)}} I$

$$\begin{aligned}
&\approx g(x^{(t)}) + \nabla g(x^{(t)})(x - x^{(t)}) + \frac{1}{2\alpha^{(t)}}(x - x^{(t)})^T (x - x^{(t)}) \\
&= \frac{1}{2\alpha^{(t)}}(x - x^{(t)} + 2\alpha^{(t)} \nabla g(x^{(t)}))^T (x - x^{(t)}) + \text{const} \\
&= \frac{1}{2\alpha^{(t)}}(x - (x^{(t)} - \alpha^{(t)} \nabla g(x^{(t)}))^T (x - (x^{(t)} - \alpha^{(t)} \nabla g(x^{(t)}))) \\
&\qquad + \text{const} \\
&= \frac{1}{2\alpha^{(t)}} ||x - (x^{(t)} - \alpha^{(t)} \nabla g(x^{(t)}))||^2 + \text{const}
\end{aligned}$$

Deriving Generalized Gradient Descent (2/2)

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

$$g(x) = \frac{1}{2\alpha^{(t)}}||x - (x^{(t)} - \alpha^{(t)}\nabla g(x^{(t)}))||^2 + \text{const}$$

yields a proximal problem

$$\begin{aligned}
\min_x f(x) &= \frac{1}{2\alpha^{(t)}}||x - (x^{(t)} - \alpha^{(t)}\nabla g(x^{(t)}))||^2 + h(x) \\
&\propto \frac{1}{2}||x - (x^{(t)} - \alpha^{(t)}\nabla g(x^{(t)}))||^2 + \alpha^{(t)} h(x) \\
&= \text{prox}_{\alpha^{(t)} h}(x^{(t)} - \alpha^{(t)}\nabla g(x^{(t)}))
\end{aligned}$$

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

Generalized Gradient Descent

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

Generalized Gradient Descent:

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

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

- 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 $I_C(\theta) := \begin{cases} 0, & \theta \in C \\ \infty, & \text{else} \end{cases}$

Special Cases

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

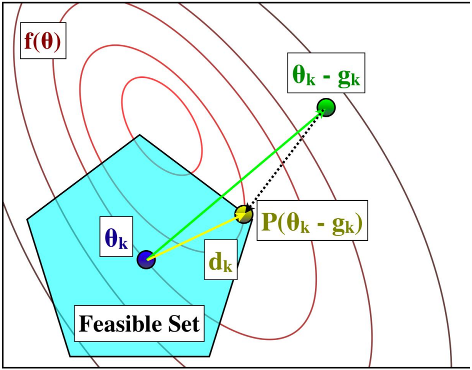
1. $R = 0$ yields **gradient descent**:

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

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

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

Special Cases: Projected Gradient Descent



- Instead of taking a gradient step and then project, we could compute the smallest stepsize that does not leave the feasible area ("guarded gradient descent").
- Q: Which next iterate would "guarded gradient descent" find instead?
- Now assume the current iterate θ_t is on the upper right border of the feasible area.
- Q: Which next iterate would "guarded gradient descent" find now?
- Q: How about projected gradient descent?

[?, fig. 13.11]

Special Cases

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

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

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

Stepsizes $\alpha^{(t)}$

Taylor expansion of the Gradient:

$$\nabla\ell(\theta) \approx \nabla\ell(\theta^{(t)}) + \nabla^2\ell(\theta^{(t)})(\theta - \theta^{(t)}) \approx \nabla\ell(\theta^{(t)}) + \frac{1}{\alpha^{(t)}}(\theta - \theta^{(t)})$$
$$\rightsquigarrow \quad \alpha^{(t)}\nabla\ell(\theta^{(t)}) - \nabla\ell(\theta^{(t-1)}) \approx (\theta^{(t)} - \theta^{(t-1)})$$

Idea:

$$\alpha^{(t)} := \arg\min_{\alpha} ||(\theta^{(t)} - \theta^{(t-1)}) - \alpha(\nabla\ell(\theta^{(t)}) - \nabla\ell(\theta^{(t-1)}))||_2^2$$
$$= \frac{(\theta^{(t)} - \theta^{(t-1)})^T(\theta^{(t)} - \theta^{(t-1)})}{(\theta^{(t)} - \theta^{(t-1)})^T(\nabla\ell(\theta^{(t)}) - \nabla\ell(\theta^{(t-1)}))}$$

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.

ISTA Algorithm

learn-l1reg-ista$(X \in \mathbb{R}^{N\times M}, y \in \mathbb{R}^N, \lambda > 0, s \in (0,1), M)$:

  $\theta := 0, r := y, \tilde{\lambda} := \infty, \alpha := 1$

  for $t := 1, 2, 3, \ldots$ while $\tilde{\lambda} \neq \lambda$:

    $\tilde{\lambda} := \max(\lambda, s||X^T r||_\infty)$

    while $\ell(\theta) + \lambda||\theta||_1$ did not increase too much in the last M steps:

      $\theta^{\text{old}} := \theta$

      $\tilde{\theta} := \theta - \alpha\nabla\ell(\theta)$

      $\theta := \text{soft}(\tilde{\theta}, \tilde{\lambda}\alpha)$

      $\alpha := \frac{(\theta-\theta^{\text{old}})^T(\theta-\theta^{\text{old}})}{(\theta-\theta^{\text{old}})^T(\nabla\ell(\theta)-\nabla\ell(\theta^{\text{old}}))}$

    $r := y - X\theta$

  return θ

Nesterov's Accelerated Generalized Gradient Descent

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

Generalized Gradient Descent:

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

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

- 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)

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

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

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

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

FISTA vs ISTA

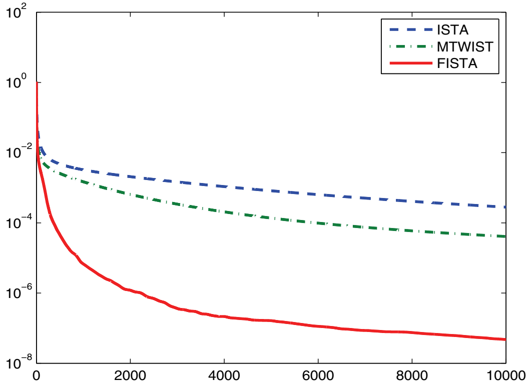


Figure 5. *Comparison of function value errors $F(\mathbf{x}_k) - F(\mathbf{x}^*)$ of ISTA, MTWIST, and FISTA.*

[?, p. 19]

Outline

Bayesian Regression

$$\hat{\beta} = \arg\min_{\beta} \underbrace{L(\beta)}_{\text{Loss}} + \lambda \underbrace{R(\beta)}_{\text{Regularization}}$$

$$\Big\downarrow \text{"Bayesianize"}$$

$$\hat{\beta} = \arg\max_{\beta} p(\beta \mid X, y)$$

Bayesian Regression

$$\hat{\beta} = \arg\min_{\beta} \underbrace{L(\beta)}_{\text{Loss}} + \lambda \underbrace{R(\beta)}_{\text{Regularization}}$$

$$\Big\downarrow \text{"Bayesianize"}$$

$$\hat{\beta} = \arg\max_{\beta} p(\beta \mid X, y)$$

$$\underbrace{p(\beta \mid X, y)}_{\text{posterior}} \propto \underbrace{p(y \mid X, \beta)}_{\text{likelihood}} \cdot \underbrace{p(\beta)}_{\text{prior}}$$

Bayesian Regression

$$\hat{\beta} = \arg\min_{\beta} \underbrace{L(\beta)}_{\text{Loss}} + \lambda \underbrace{R(\beta)}_{\text{Regularization}}$$

$$\Big\downarrow \text{"Bayesianize"}$$

$$\hat{\beta} = \arg\max_{\beta} p(\beta \mid X, y)$$

$$\underbrace{p(\beta \mid X, y)}_{\text{posterior}} \propto \underbrace{p(y \mid X, \beta)}_{\text{likelihood}} \cdot \underbrace{p(\beta)}_{\text{prior}}$$

- $p(y \mid X, \beta) = \mathcal{N}(y \mid X\beta, \sigma^2 I) \leadsto$ Bayesian Linear Regression

Laplace Priors correspond to L1 regularization

L2 regularization:

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

Gaussian priors:

$$\begin{aligned} p(y_n \mid x_n, \beta, \sigma^2) &:= \mathcal{N}(y_n \mid x_n^T \beta, \sigma^2) \\ p(\beta) &:= \mathcal{N}(\beta \mid 0, \tfrac{1}{\lambda} I) \\ &= (2\pi\lambda)^{-M/2}\, e^{-\frac{1}{2}\lambda||\beta||_2^2} \end{aligned}$$

using negative logposterior as objective function:

$$f(\beta; X, y, \sigma^2 \text{ or } \lambda) := -\log p(y \mid X, \beta)\, p(\beta)$$

Laplace Priors correspond to L1 regularization

L2 regularization:

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

Gaussian priors:

$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T\beta, \sigma^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \tfrac{1}{\lambda} I)$$
$$= (2\pi\lambda)^{-M/2}\, e^{-\frac{1}{2}\lambda||\beta||_2^2}$$

L1 regularization:

$$f(\beta) := ||y - X\beta||_2^2 + \lambda||\beta||_1$$

Laplace priors:

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

using negative logposterior as objective function:

$$f(\beta; X, y, \sigma^2 \text{ or } \lambda) := -\log p(y \mid X, \beta)\, p(\beta)$$

Laplace as Gaussian Scale Mixture

- problem: MAP cannot be found analytically.
- idea: rewrite the Laplace as a **Gaussian-Scale-Mixture** with Exponential priors:

$$\text{Lap}(\beta_i \mid 0, \tfrac{1}{\lambda}) = \int \mathcal{N}(\beta_i \mid 0, \tau_i^2)\, \text{Exp}(\tau_i^2 \mid \tfrac{1}{2}\lambda^2) d\tau^2$$

i.e. each parameter is distributed as $\beta_i \sim \mathcal{N}(0, \tau_i^2)$ with $\tau_i^2 \sim \text{Exp}(\frac{1}{2}\lambda^2)$

Laplace as Gaussian Scale Mixture

- problem: MAP cannot be found analytically.
- idea: rewrite the Laplace as a **Gaussian-Scale-Mixture** with Exponential priors:

$$\mathrm{Lap}(\beta_i \mid 0, \tfrac{1}{\lambda}) = \int \mathcal{N}(\beta_i \mid 0, \tau_i^2)\, \mathrm{Exp}(\tau_i^2 \mid \tfrac{1}{2}\lambda^2) d\tau^2$$

i.e. each parameter is distributed as $\beta_i \sim \mathcal{N}(0, \tau_i^2)$ with $\tau_i^2 \sim \mathrm{Exp}(\frac{1}{2}\lambda^2)$

$\rightsquigarrow$ posterior distribution:

$$p(\beta, \sigma^2 \mid X, y, \tau^2) \propto \underbrace{p(y \mid X, \beta, \sigma^2)}_{=\mathcal{N}(y|X\beta,\sigma^2 I)} \cdot \underbrace{p(\beta \mid \tau^2)}_{=\mathcal{N}(\beta|0,\mathrm{diag}(\tau^2))} \cdot \underbrace{p(\tau^2 \mid \lambda)}_{=\mathrm{Exp}(\tau^2|\frac{1}{2}\lambda^2)} \cdot \underbrace{p(\sigma^2)}$$

with $(\tau_i)_{m=1...M}$ **latent variables**,
λ the regularization strength hyperparameter,

Laplace as Gaussian Scale Mixture

- problem: MAP cannot be found analytically.
- idea: rewrite the Laplace as a **Gaussian-Scale-Mixture** with Exponential priors:

$$\mathrm{Lap}(\beta_i \mid 0, \tfrac{1}{\lambda}) = \int \mathcal{N}(\beta_i \mid 0, \tau_i^2)\, \mathrm{Exp}(\tau_i^2 \mid \tfrac{1}{2}\lambda^2) d\tau^2$$

i.e. each parameter is distributed as $\beta_i \sim \mathcal{N}(0, \tau_i^2)$ with $\tau_i^2 \sim \mathrm{Exp}(\frac{1}{2}\lambda^2)$

$\rightsquigarrow$ posterior distribution:

$$p(\beta, \sigma^2 \mid X, y, \tau^2) \propto \underbrace{p(y \mid X, \beta, \sigma^2)}_{=\mathcal{N}(y|X\beta,\sigma^2 I)} \cdot \underbrace{p(\beta \mid \tau^2)}_{=\mathcal{N}(\beta|0,\mathrm{diag}(\tau^2))} \cdot \underbrace{p(\tau^2 \mid \lambda)}_{=\mathrm{Exp}(\tau^2|\frac{1}{2}\lambda^2)} \cdot \underbrace{p(\sigma^2)}_{=\mathrm{IG}(\sigma^2|a,b)}$$

with $(\tau_i)_{m=1...M}$ **latent variables**,
λ the regularization strength hyperparameter,
$\mathrm{IG}(\sigma^2 \mid a, b)$ an **Inverse-Gamma** prior on the variance.

Laplace as Gaussian Scale Mixture

- problem: MAP cannot be found analytically.
- idea: rewrite the Laplace as a **Gaussian-Scale-Mixture** with Exponential priors:

$$\text{Lap}(\beta_i \mid 0, \tfrac{1}{\lambda}) = \int \mathcal{N}(\beta_i \mid 0, \tau_i^2)\,\text{Exp}(\tau_i^2 \mid \tfrac{1}{2}\lambda^2) d\tau^2$$

i.e. each parameter is distributed as $\beta_i \sim \mathcal{N}(0, \tau_i^2)$ with $\tau_i^2 \sim \text{Exp}(\frac{1}{2}\lambda^2)$

$\rightsquigarrow$ posterior distribution:

$$p(\beta, \sigma^2 \mid X, y, \tau^2) \propto \underbrace{p(y \mid X, \beta, \sigma^2)}_{=\mathcal{N}(y|X\beta,\sigma^2 I)} \cdot \underbrace{p(\beta \mid \tau^2)}_{=\mathcal{N}(\beta|0,\text{diag}(\tau^2))} \cdot \underbrace{p(\tau^2 \mid \lambda)}_{=\text{Exp}(\tau^2|\frac{1}{2}\lambda^2)} \cdot \underbrace{p(\sigma^2)}_{=\text{IG}(\sigma^2|a,b)}$$

with $(\tau_i)_{m=1\ldots M}$ **latent variables**,
λ the regularization strength hyperparameter,
$\text{IG}(\sigma^2 \mid a, b)$ an **Inverse-Gamma** prior on the variance.

- p is now smooth in all parameters! We can apply EM-algorithm!

Inverse Gamma Distribution

- Gamma distribution:

$$\Gamma(x \mid a, b) := \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$
$$\mathbb{E}(x) = \frac{a}{b}$$

- **Inverse Gamma** distribution:

$$\mathrm{IG}(x \mid a, b) := \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-\frac{b}{x}}$$

- $X \sim \Gamma(a, b) \iff X^{-1} \sim \mathrm{IG}(a, b)$

$$\rightsquigarrow \quad \mathbb{E}(\frac{1}{x}) = \frac{a}{b}$$

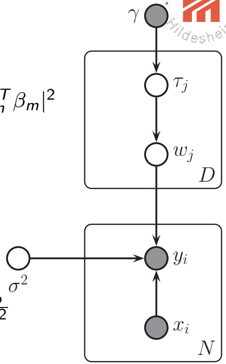
Laplace Prior as Gaussian Scale Mixture

$$p(y_n \mid x_n, \beta, \sigma^2) := \mathcal{N}(y_n \mid x_n^T \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}|y_n - x_n^T \beta_m|^2}$$

$$p(\beta_m \mid \tau_m^2) := \mathcal{N}(\beta_m \mid 0, \tau_m^2) = \frac{1}{\sqrt{2\pi\tau_m^2}} e^{-\frac{1}{2\tau_m^2}|\beta_m|^2}$$

$$p(\tau_m^2) := \mathrm{Exp}(\tau_m^2 \mid \frac{1}{2}\lambda^2) = \frac{1}{2}\lambda^2 e^{-\frac{1}{2}\lambda^2\tau_m^2}$$

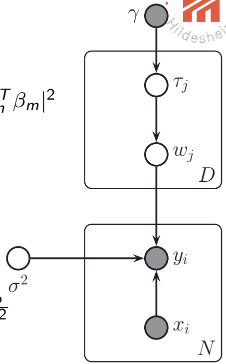
$$p(\sigma^2) := \mathrm{IG}(\sigma^2 \mid a, b) = \frac{b^a}{\Gamma(a)} \sigma^{-2(1+a)} e^{-\frac{b}{\sigma^2}}$$



Note: $T := \mathrm{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_M^2)$

[?, p. 446]

Laplace Prior as Gaussian Scale Mixture



$$
\begin{aligned}
p(y_n \mid x_n, \beta, \sigma^2) &:= \mathcal{N}(y_n \mid x_n^T \beta, \sigma^2) &&= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}|y_n - x_n^T \beta_m|^2} \\
p(\beta_m \mid \tau_m^2) &:= \mathcal{N}(\beta_m \mid 0, \tau_m^2) &&= \frac{1}{\sqrt{2\pi\tau_m^2}} e^{-\frac{1}{2\tau_m^2}|\beta_m|^2} \\
p(\tau_m^2) &:= \mathrm{Exp}(\tau_m^2 \mid \frac{1}{2}\lambda^2) &&= \frac{1}{2}\lambda^2 e^{-\frac{1}{2}\lambda^2 \tau_m^2} \\
p(\sigma^2) &:= \mathrm{IG}(\sigma^2 \mid a, b) &&= \frac{b^a}{\Gamma(a)} \sigma^{-2(1+a)} e^{-\frac{b}{\sigma^2}}
\end{aligned}
$$

negative logposterior:

$$
\begin{aligned}
\ell(\beta, \sigma^2 \mid X, y, \tau^2) = {} & \frac{1}{2} N \log \sigma^2 + \frac{1}{2\sigma^2} ||y - X\beta||_2^2 \\
& + \sum_{m=1}^{M} \log \tau_m^2 + \frac{1}{2}\beta^T T^{-1} \beta + \frac{1}{2}\lambda^2 \sum_{m=1}^{M} \tau_m^2 + (1+a)\log \sigma^2 + \frac{b}{\sigma^2}
\end{aligned}
$$

Note: $T := \mathrm{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_M^2)$

[?, p. 446]

E-step for τ^2

We need to compute the expectation of

$$p(\tau^2 \mid X, y, \beta, \sigma^2) \propto p(\beta \mid \tau^2) p(\tau^2)$$

where $p(\beta_m \mid \tau_m^2) = \mathcal{N}(\beta_m \mid 0, \tau_m^2)$ and $p(\tau_m^2) = \text{Exp}(\tau_m^2 \mid \frac{1}{2}\lambda^2)$

E-step for τ^2

We need to compute the expectation of

$$p(\tau^2 \mid X, y, \beta, \sigma^2) \propto p(\beta \mid \tau^2) p(\tau^2)$$

where $p(\beta_m \mid \tau_m^2) = \mathcal{N}(\beta_m \mid 0, \tau_m^2)$ and $p(\tau_m^2) = \mathrm{Exp}(\tau_m^2 \mid \frac{1}{2}\lambda^2)$

It turns out simpler to estimate $\frac{1}{\tau^2}$: One can show that (tutorial)

$$\frac{1}{\tau^2} \mid \beta \sim \mathrm{InvGauss}(\sqrt{\frac{\lambda^2}{\beta^2}}, \lambda^2)$$

Where the **Inverse Gaussian distribution** is given by

$$\mathrm{InvGauss}(x \mid \mu, \nu) = \sqrt{\frac{\nu}{2\pi x^3}} e^{-\frac{\nu}{2\mu^2 x}(x-\mu)^2}$$

with mean $\mathbb{E}[x] = \mu$ and variance $\mathrm{Var}[x] = \mu^3/\nu \implies \boxed{\mathbb{E}\left[\frac{1}{\tau_m^2}\right] = \frac{\lambda}{|\beta_m|}}$

E-step for σ^2

We need to compute the expectation of

$$p(\sigma^2 \mid X, y, \beta, \tau^2) \propto p(y \mid X, \beta, \sigma^2)\, p(\sigma^2)$$
$$= \mathcal{N}(y \mid X\beta, \sigma^2 I)\ \mathrm{IG}(\sigma^2 \mid a, b)$$

E-step for σ^2

We need to compute the expectation of

$$\begin{aligned} p(\sigma^2 \mid X, y, \beta, \tau^2) &\propto p(y \mid X, \beta, \sigma^2)\, p(\sigma^2) \\ &= \mathcal{N}(y \mid X\beta, \sigma^2 I)\ \mathrm{IG}(\sigma^2 \mid a, b) \end{aligned}$$

One can show that (tutorial)

$$\begin{aligned} p(\sigma^2 \mid X, y, \beta, \tau^2) &= \mathrm{IG}(\sigma^2, a', b') \\ \text{with} \quad a' &:= a + \tfrac{1}{2}N, \quad b' := b + \tfrac{1}{2}\|y - X\beta\|_2^2 \end{aligned}$$

Remark on Conjugate Prior

Note that the posterior of σ^2 is again an Inverse Gamma distribution!

$$\underbrace{p(\sigma^2 \mid X, y, \beta)}_{=\mathrm{IG}(a',b')} \propto \underbrace{p(y \mid X, \beta, \sigma^2)}_{\mathcal{N}(\mu,\nu)} \underbrace{p(\sigma^2)}_{=\mathrm{IG}(a,b)}$$

This is because the IG is a **conjugate prior** to the normal distribution. Conjugate priors let you interpret how the data changes the believe about the parameters. $\longrightarrow$ Main reason for choosing this prior!

Remark on Conjugate Prior

Note that the posterior of σ^2 is again an Inverse Gamma distribution!

$$\underbrace{p(\sigma^2 \mid X, y, \beta)}_{=\mathrm{IG}(a', b')} \propto \underbrace{p(y \mid X, \beta, \sigma^2)}_{\mathcal{N}(\mu, \nu)} \underbrace{p(\sigma^2)}_{=\mathrm{IG}(a, b)}$$

This is because the IG is a **conjugate prior** to the normal distribution. Conjugate priors let you interpret how the data changes the believe about the parameters. $\longrightarrow$ Main reason for choosing this prior!

Remark: inverse distributions

Note that the Inverse Gamma distribution is called Inverse Gamma because

$$X \sim \Gamma(a, b) \iff X^{-1} \sim \mathrm{IG}(a, b) \tag{1}$$

However, despite the name, the same is **not true** for the Inverse Gaussian!

M-step for β

We need to compute

$$\hat{\beta} = \arg\min_{\beta} \ell(\beta, \sigma^2, \tau^2) = \arg\min_{\beta} \frac{1}{2\sigma^2} \|y - X\beta\|_2^2 + \frac{1}{2}\beta^T T^{-1} \beta$$

where we dropped all terms independent of β. Then

$$\nabla_\beta \ell = 0 \iff (\frac{1}{\sigma^2} X^T X + T^{-1})\hat{\beta} = \frac{1}{\sigma^2} X^T y$$

So $\boxed{\hat{\beta} = (X^T X + (\frac{1}{\sigma^2} T)^{-1})^{-1} X^T y}$ which is a ridge regression objective!

EM summary

1. Expectation of τ^2:

$$p(\frac{1}{\tau_m^2} \mid \beta) = \text{Inv-Gauss}(\sqrt{\frac{\lambda^2}{\beta_m^2}}, \lambda^2)$$

$$\mathbb{E}\left[\frac{1}{\tau_m^2}\right] = \frac{\lambda}{|\beta_m|}$$

2. Expectation of σ^2:

$$p(\sigma^2 \mid X, y, \beta) = \text{IG}(\sigma^2 \mid a', b')$$

$$a' := a + \frac{1}{2}N, \quad b' := b + \frac{1}{2}\|y - X\beta\|_2^2$$

$$\mathbb{E}[\frac{1}{\sigma^2}] = \frac{a'}{b'}$$

3. Maximization w.r.t. β:

$$\ell(\beta) = \frac{1}{2\sigma^2}\|y - X\beta\|_2^2 + \frac{1}{2}\beta^T T^{-1}\beta$$

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

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: [**?**, chapter 13.3–5], [**?**, chapter 3.4, 3.8, 4.4.4], [**?**, chapter 3.1.4].
  - LAR, LARS: [**?**, chapter 3.4.4], [**?**, chapter 13.4.2],
- Non-convex regularizers: [**?**, chapter 13.6].
- Automatic Relevance Determination (ARD): [**?**, chapter 13.7], [**?**, chapter 11.9.1], [**?**, chapter 7.2.2].
- Sparse Coding: [**?**, chapter 13.8].
- Multivariate Laplace Distribution: [**?**]

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.

Torbjørn Eltoft, Taesu Kim, and Te-Won Lee.
On the multivariate laplace distribution.
IEEE Signal Processing Letters, 13(5):300–303, 2006.

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

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