

9. Proximal Gradient Methods

(近端梯度法)

Proximal Gradient Methods

Proximal Gradient Methods

Proximal gradient methods are a class of iterative techniques used to solve non-differentiable convex optimization problems given by:

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \lambda \cdot g(\boldsymbol{\theta})$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex differentiable function and $g: \mathbb{R}^n \rightarrow \mathbb{R}$ is a continuous convex possibly non-smooth function.

Proximal gradient methods produce a sequence of parameter iterates given by:

$$\boldsymbol{\theta}^{(t+1)} = \text{Prox}_{\frac{\lambda}{L} \cdot g} \left(\boldsymbol{\theta}^{(t)} - \frac{1}{L} \cdot \nabla f(\boldsymbol{\theta}^{(t)}) \right)$$

where $L > 0$ is an upper bound on the Lipschitz constant of ∇f and $\text{Prox}_{\mu \cdot g}$ is the proximal operator given by

$$\text{Prox}_{\mu \cdot g}(x) = \underset{z}{\operatorname{argmin}} \{ 1/2 \cdot \|z - x\|^2 + \mu \cdot g(z) \}$$

NB: The proximal operator is a generalization of the projection operator onto a convex set.

Proximal Gradient Methods: Example

LASSO

The LASSO problem can be written as follows:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \cdot \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \cdot \|\boldsymbol{\theta}\|_1$$

where $\mathbf{X} \in \mathbb{R}^{n \times m}$ is the feature matrix, $\mathbf{y} \in \mathbb{R}^n$ is the predictions vector, and $\boldsymbol{\theta} \in \mathbb{R}^m$ is the unknown parameters vector

Note that the LASSO problem can also be written as follows:

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \lambda \cdot g(\boldsymbol{\theta})$$

where

$$f(\boldsymbol{\theta}) = 1/2 \cdot \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

is a convex differentiable function and

$$g(\boldsymbol{\theta}) = \lambda \cdot \|\boldsymbol{\theta}\|_1$$

is a continuous convex non-smooth function.

Proximal Gradient Methods: Example

LASSO Proximal Operator

The LASSO proximal operator is given by:

$$\text{Prox}_{\mu \cdot g}(x) = \underset{z}{\operatorname{argmin}} \{ 1/2 \cdot \|z - x\|^2 + \mu \cdot \|z\|_1 \}$$

where $\mu = \lambda/L$.

The calculation of the proximal operator involves finding the gradient or subgradient f and g resp.

LASSO Problem

Recall

Subderivative of $|x|$

Let

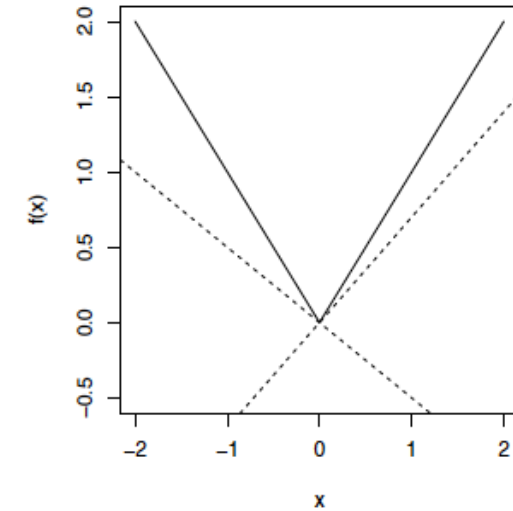
$$f(x) = |x| = \begin{cases} +x & x > 0 \\ -x & x < 0 \\ 0 & x = 0 \end{cases}$$

For $x \neq 0$, the derivative of $f(x) = |x|$ is given by:

$$\frac{df(x)}{dx} = \text{sign}(x)$$

For $x = 0$, the subderivative of $f(x) = |x|$ is given by:

$$[-1, +1]$$



$|x|$

Recall

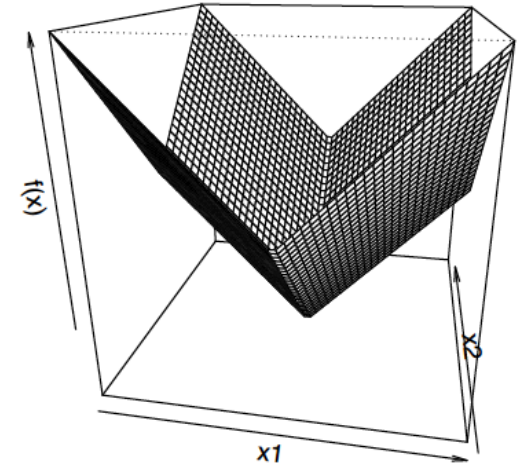
Subgradient of $\|x\|_1$

Let

$$f(x) = \|x\|_1 = |x|_1 + |x|_2 + \cdots + |x|_n$$

For $x_i \neq 0$, the i -th component of the gradient equals $\text{sign}(x_i)$

For $x_i = 0$, the i -th component of the subgradient is any element of $[-1, +1]$



$\|x\|_1$

Proximal Gradient Methods: Example

LASSO Proximal Operator

The LASSO proximal operator is given by:

$$\text{Prox}_{\mu \cdot g}(x) = \underset{z}{\operatorname{argmin}} \{ 1/2 \cdot \|z - x\|^2 + \mu \cdot \|z\|_1 \}$$

where $\mu = \lambda/L$.

The proximal operator reduces to

$$\text{Prox}_{\mu \cdot g}(x_i) = \begin{cases} x_i + \mu & x_i < -\mu \\ 0 & -\mu \leq x_i \leq +\mu, \\ x_i - \mu & x_i > +\mu \end{cases}, \quad i = 1, \dots, n$$

or equivalently to

$$\text{Prox}_{\mu \cdot g}(x_i) = \max(0, 1 - \mu/|x_i|) \cdot x_i, \quad i = 1, \dots, n$$

Iterative Soft Thresholding Algorithm (ISTA)

LASSO Problem

Recall the LASSO problem can also be written as follows:

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \lambda \cdot g(\boldsymbol{\theta})$$

where

$$f(\boldsymbol{\theta}) = 1/2 \cdot \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

is a convex differentiable function and

$$g(\boldsymbol{\theta}) = \lambda \cdot \|\boldsymbol{\theta}\|_1$$

is a continuous convex non-smooth function.

ISTA Algorithm

Proximal gradient method produce a sequence of parameter iterates given by:

$$\boldsymbol{\theta}^{(t+1)} = \text{Prox}_{\frac{\lambda}{L}g} \left(\boldsymbol{\theta}^{(t)} - \frac{1}{L} \cdot \nabla f(\boldsymbol{\theta}^{(t)}) \right)$$

Since $\nabla f(\boldsymbol{\theta}) = -\mathbf{X}^t \cdot (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$, the proximal gradient method also produces the sequence of parameter iterates given by:

$$\boldsymbol{\theta}^{(t+1)} = \text{Prox}_{\frac{\lambda}{L}g} \left(\boldsymbol{\theta}^{(t)} + \frac{1}{L} \cdot \mathbf{X}^t \cdot (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}^{(t)}) \right)$$