Nondifferentiable Convex Functions

DS-GA 1013 / MATH-GA 2824 Optimization-based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_fall17/index.html

Carlos Fernandez-Granda

Applications

Subgradients

Optimization methods

Regression

The aim is to learn a function h that relates

- a response or dependent variable y
- to several observed variables $x_1, x_2, \ldots, x_p$, known as covariates, features or independent variables

The response is assumed to be of the form

$$y = h(\vec{x}) + z$$

where $\vec{x} \in \mathbb{R}^p$ contains the features and z is noise

Linear regression

The regression function h is assumed to be linear

$$y^{(i)} = \vec{x}^{(i)\,T} \vec{\beta}^{*} + z^{(i)}, \quad 1 \leq i \leq n$$

Our aim is to estimate $\vec{\beta}^{*} \in \mathbb{R}^{p}$ from the data

Linear regression

In matrix form

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^{(1)} & \vec{x}_2^{(1)} & \cdots & \vec{x}_p^{(1)} \\ \vec{x}_1^{(2)} & \vec{x}_2^{(2)} & \cdots & \vec{x}_p^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ \vec{x}_1^{(n)} & \vec{x}_2^{(n)} & \cdots & \vec{x}_p^{(n)} \end{bmatrix} \begin{bmatrix} \vec{\beta}_1^* \\ \vec{\beta}_2^* \\ \cdots \\ \vec{\beta}_p^* \end{bmatrix} + \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \cdots \\ z^{(n)} \end{bmatrix}$$

Equivalently,

$$\vec{y} = X\vec{\beta}^* + \vec{z}$$

Sparse linear regression

Only a subset of the features are relevant

Model selection problem

Two objectives:

- Good fit to the data; $\left|\left|X\vec{\beta} - \vec{y}\right|\right|_2^2$ should be as small as possible
- Using a small number of features; $\vec{\beta}$ should be as sparse as possible

Sparse linear regression

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \vec{x}_j^{(1)} & \vec{x}_l^{(1)} \\ \vec{x}_j^{(2)} & \vec{x}_l^{(2)} \\ \cdots & \cdots \\ \vec{x}_l^{(n)} & \vec{x}_l^{(n)} \end{bmatrix} \begin{bmatrix} \vec{\beta}_j^* \\ \vec{\beta}_l^* \end{bmatrix} + \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \cdots \\ z^{(n)} \end{bmatrix}$$

$$= \begin{bmatrix} \vec{x}_1^{(1)} & \cdots & \vec{x}_j^{(1)} & \cdots & \vec{x}_l^{(1)} & \cdots & \vec{x}_p^{(1)} \\ \vec{x}_1^{(2)} & \cdots & \vec{x}_j^{(2)} & \cdots & \vec{x}_l^{(2)} & \cdots & \vec{x}_p^{(2)} \\ & & & \cdots & & & \\ \vec{x}_1^{(n)} & \cdots & \vec{x}_j^{(n)} & \cdots & \vec{x}_l^{(n)} & \cdots & \vec{x}_p^{(n)} \end{bmatrix} \begin{bmatrix} 0 \\ \cdots \\ \vec{\beta}_j^* \\ \cdots \\ \vec{\beta}_l^* \\ \cdots \\ 0 \end{bmatrix} + \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \cdots \\ z^{(n)} \end{bmatrix}$$

$$= X\vec{\beta}^* + \vec{z}$$

Sparse linear regression with 2 features

$$\vec{y} := \alpha\,\vec{x}_1 + \vec{z}$$

$$X := \begin{bmatrix}\vec{x}_1 & \vec{x}_2\end{bmatrix}$$

$$||\vec{x}_1||_2 = 1$$

$$||\vec{x}_2||_2 = 1$$

$$\langle \vec{x}_1, \vec{x}_2 \rangle = \rho$$

Least squares: not sparse

$$\begin{aligned}
\vec{\beta}_{\text{LS}} &= \left(X^T X\right)^{-1} X^T \vec{y} \\
&= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \vec{x}_1^T \vec{y} \\ \vec{x}_2^T \vec{y} \end{bmatrix} \\
&= \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} \alpha + \vec{x}_1^T \vec{z} \\ \alpha\rho + \vec{x}_2^T \vec{z} \end{bmatrix} \\
&= \begin{bmatrix} \alpha \\ 0 \end{bmatrix} + \frac{1}{1-\rho^2} \begin{bmatrix} \langle \vec{x}_1 - \rho \vec{x}_2, \vec{z} \rangle \\ \langle \vec{x}_2 - \rho \vec{x}_1, \vec{z} \rangle \end{bmatrix}
\end{aligned}$$

The lasso

Idea: Use ℓ_1-norm regularization to promote sparse coefficients

$$\vec{\beta}_{\text{lasso}} := \arg\min_{\vec{\beta}} \frac{1}{2} \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2 + \lambda \left|\left|\vec{\beta}\right|\right|_1$$

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$f(\theta \vec{x} + (1-\theta)\vec{y})$$

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$f(\theta \vec{x} + (1-\theta)\vec{y}) = \sum_{i=1}^{m} \alpha_i f_i(\theta \vec{x} + (1-\theta)\vec{y})$$

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$\begin{aligned} f\left(\theta \vec{x} + (1-\theta)\,\vec{y}\right) &= \sum_{i=1}^{m} \alpha_i f_i\left(\theta \vec{x} + (1-\theta)\,\vec{y}\right) \\ &\leq \sum_{i=1}^{m} \alpha_i \left(\theta f_i\left(\vec{x}\right) + (1-\theta)\, f_i\left(\vec{y}\right)\right) \end{aligned}$$

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i \, f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$\begin{aligned} f\left(\theta \vec{x} + (1-\theta)\, \vec{y}\right) &= \sum_{i=1}^{m} \alpha_i \, f_i \left(\theta \vec{x} + (1-\theta)\, \vec{y}\right) \\ &\leq \sum_{i=1}^{m} \alpha_i \left(\theta f_i\left(\vec{x}\right) + (1-\theta)\, f_i\left(\vec{y}\right)\right) \\ &= \theta\, f\left(\vec{x}\right) + (1-\theta)\, f\left(\vec{y}\right) \end{aligned}$$

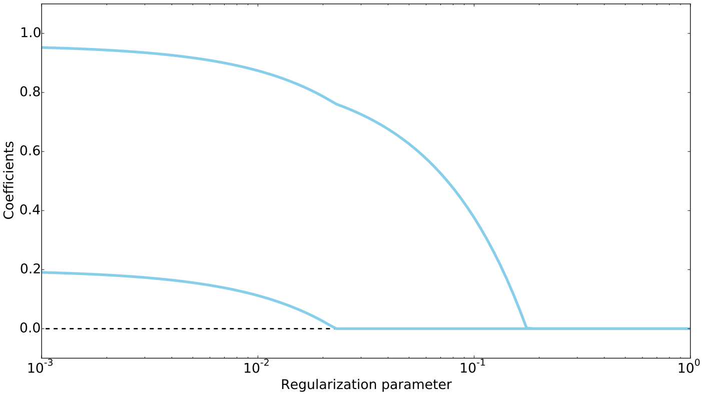
Regularized least-squares

Regularized least-squares cost functions

$$||A\vec{x} - \vec{y}||_2^2 + ||\vec{x}||$$

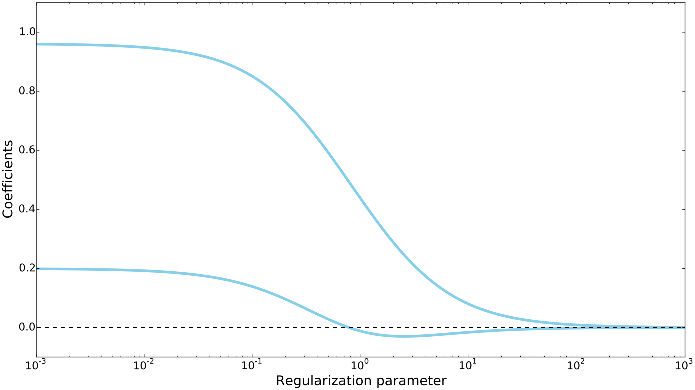
are convex

It works



Coefficients
1.0
0.8
0.6
0.4
0.2
0.0
10^-3
10^-2
10^-1
10^0
Regularization parameter

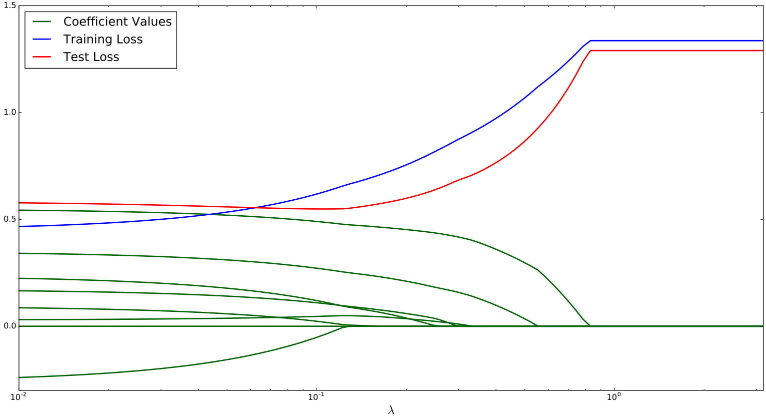
Ridge regression doesn't work



Prostate cancer data set

- 8 features (age, weight, analysis results)
- Response: Prostate-specific antigen (PSA), associated to cancer
- Training set: 60 patients
- Test set: 37 patients

Prostate cancer data set



Principal component analysis

Given n data vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d$,

1. Center the data,

$$\vec{c}_i = \vec{x}_i - \operatorname{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right), \qquad 1 \leq i \leq n$$

2. Group the centered data as columns of a matrix

$$C = \begin{bmatrix} \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \end{bmatrix}.$$

3. Compute the SVD of C

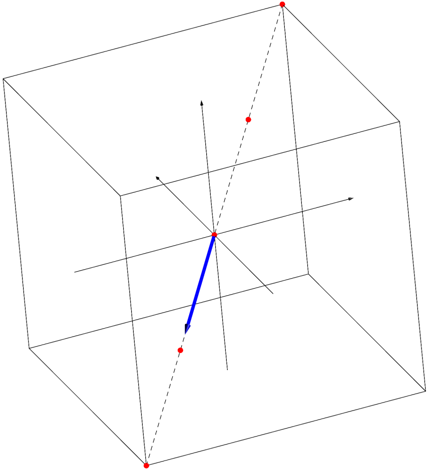
The left singular vectors are the principal directions

The principal values are the coefficients of the centered vectors in the basis of principal directions.

Example

$$C := \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

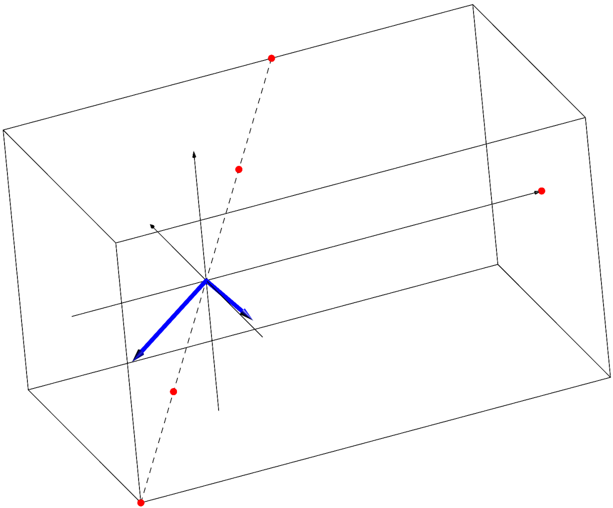
Principal component analysis



Example

$$C := \begin{bmatrix} -2 & -1 & \color{red}{5} & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

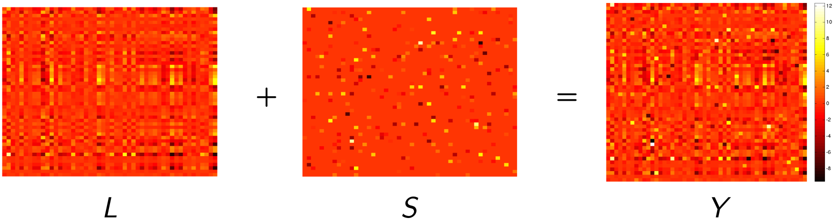
Principal component analysis



Outliers

Problem: Outliers distort principal directions

Model: Data equals low-rank component + sparse component



Idea: Fit model to data, then apply PCA to L

Robust PCA

Data: $Y \in \mathbb{R}^{n \times m}$

Robust PCA estimator of low-rank component:

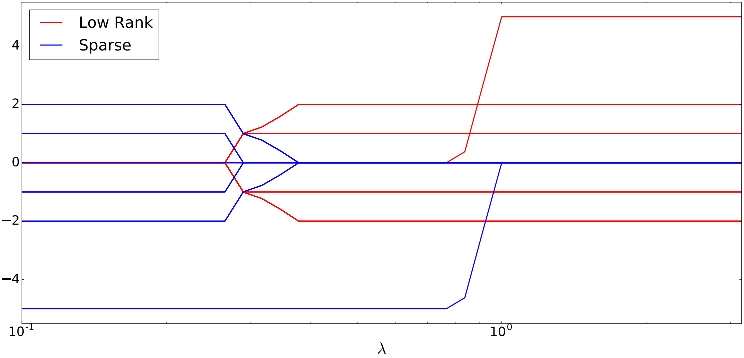
$$L_{\text{RPCA}} := \arg\min_{L} ||L||_* + \lambda \, ||Y - L||_1$$

where $\lambda > 0$ is a regularization parameter

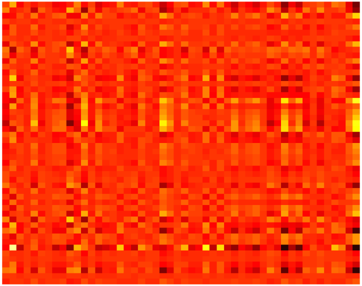
Robust PCA estimator of sparse component: $S_{\text{RPCA}} := Y - L_{\text{RPCA}}$

$||\cdot||_1$ is the ℓ_1 norm of the *vectorized matrix*

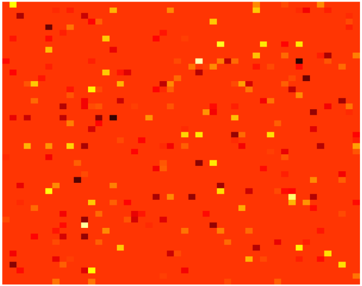
Example



$$\lambda = \frac{1}{\sqrt{n}}$$

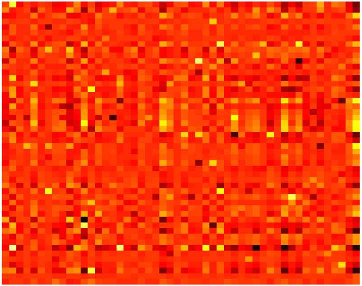


L



S

Large λ

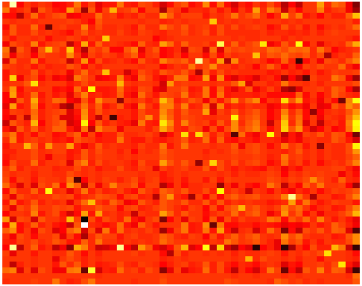


L

S

Small λ

L



S

Background subtraction

Background subtraction

Matrix with vectorized frames as columns

Static image:

$$Y = \begin{bmatrix} \vec{x} & \vec{x} & \cdots & \vec{x} \end{bmatrix} = \vec{x} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$$

Slowly varying background: Low-rank

Rapidly varying foreground: Sparse

Frame 17

Low-rank component

Sparse component

Frame 42

Low-rank component

Sparse component

Frame 75

Low-rank component

Sparse component

Applications

Subgradients

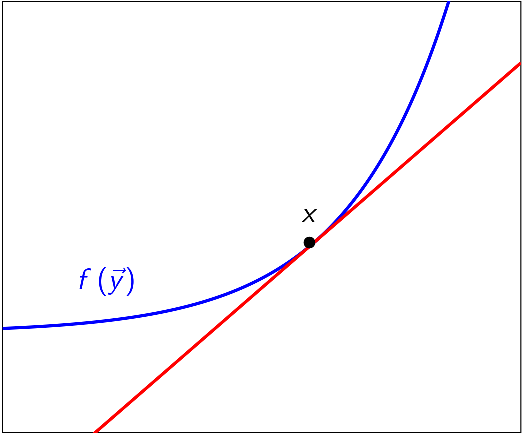
Optimization methods

Gradient

A differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if and only if for every $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$f(\vec{y}) \geq f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

Gradient



Subgradient

The subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\vec{x} \in \mathbb{R}^n$ is a vector $\vec{g} \in \mathbb{R}^n$ such that

$$f(\vec{y}) \geq f(\vec{x}) + \vec{g}^{\,T} (\vec{y} - \vec{x}), \quad \text{for all } \vec{y} \in \mathbb{R}^n$$

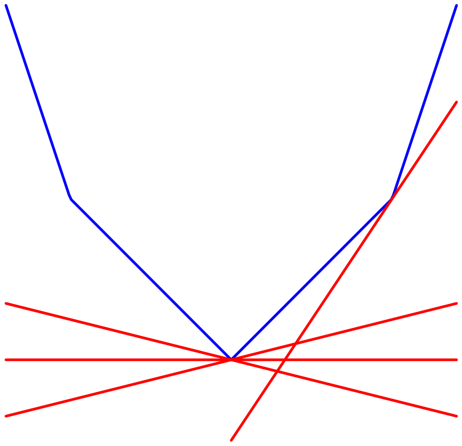
Geometrically, the hyperplane

$$\mathcal{H}_{\vec{g}} := \left\{ \vec{y} \mid \vec{y}[n+1] = \vec{g}^{\,T} \left(\begin{bmatrix} \vec{y}[1] \\ \cdots \\ \vec{y}[n] \end{bmatrix} \right) \right\}$$

is a supporting hyperplane of the epigraph at $\vec{x}$

The set of all subgradients at $\vec{x}$ is called the subdifferential

Subgradients



Subgradient of differentiable function

If a function is differentiable, the only subgradient at each point is the gradient

Proof

Assume $\vec{g}$ is a subgradient at $\vec{x}$, for any $\alpha \geq 0$

$$
\begin{aligned}
f(\vec{x} + \alpha\, \vec{e}_i) &\geq f(\vec{x}) + \vec{g}^{\,T} \alpha\, \vec{e}_i \\
&= f(\vec{x}) + \vec{g}[i]\, \alpha \\
f(\vec{x}) &\geq f(\vec{x} - \alpha\, \vec{e}_i) + \vec{g}^{\,T} \alpha\, \vec{e}_i \\
&= f(\vec{x} - \alpha\, \vec{e}_i) + \vec{g}[i]\, \alpha
\end{aligned}
$$

Proof

Assume $\vec{g}$ is a subgradient at $\vec{x}$, for any $\alpha \geq 0$

$$
\begin{aligned}
f(\vec{x} + \alpha \, \vec{e}_i) &\geq f(\vec{x}) + \vec{g}^{\,T} \alpha \, \vec{e}_i \\
&= f(\vec{x}) + \vec{g}[i] \, \alpha \\
f(\vec{x}) &\geq f(\vec{x} - \alpha \, \vec{e}_i) + \vec{g}^{\,T} \alpha \, \vec{e}_i \\
&= f(\vec{x} - \alpha \, \vec{e}_i) + \vec{g}[i] \, \alpha
\end{aligned}
$$

Combining both inequalities

$$
\frac{f(\vec{x}) - f(\vec{x} - \alpha \, \vec{e}_i)}{\alpha} \leq \vec{g}[i] \leq \frac{f(\vec{x} + \alpha \, \vec{e}_i) - f(\vec{x})}{\alpha}
$$

Proof

Assume $\vec{g}$ is a subgradient at $\vec{x}$, for any $\alpha \geq 0$

$$\begin{aligned} f(\vec{x} + \alpha\, \vec{e}_i) &\geq f(\vec{x}) + \vec{g}^{\,T} \alpha\, \vec{e}_i \\ &= f(\vec{x}) + \vec{g}[i]\, \alpha \\ f(\vec{x}) &\geq f(\vec{x} - \alpha\, \vec{e}_i) + \vec{g}^{\,T} \alpha\, \vec{e}_i \\ &= f(\vec{x} - \alpha\, \vec{e}_i) + \vec{g}[i]\, \alpha \end{aligned}$$

Combining both inequalities

$$\frac{f(\vec{x}) - f(\vec{x} - \alpha\, \vec{e}_i)}{\alpha} \leq \vec{g}[i] \leq \frac{f(\vec{x} + \alpha\, \vec{e}_i) - f(\vec{x})}{\alpha}$$

Letting $\alpha \to 0$, implies $\vec{g}[i] = \frac{\partial f(\vec{x})}{\partial \vec{x}[i]}$

Subgradient

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if it has a subgradient at every point

It is strictly convex if and only for all $\vec{x} \in \mathbb{R}^n$ there exists $\vec{g} \in \mathbb{R}^n$ such that

$$f(\vec{y}) > f(\vec{x}) + \vec{g}^{\,T}(\vec{y} - \vec{x}), \quad \text{for all } \vec{y} \neq \vec{x}.$$

Optimality condition for nondifferentiable functions

If $\vec{0}$ is a subgradient of f at $\vec{x}$, then

$$f(\vec{y}) \geq f(\vec{x}) + \vec{0}^T (\vec{y} - \vec{x})$$

Optimality condition for nondifferentiable functions

If $\vec{0}$ is a subgradient of f at $\vec{x}$, then

$$\begin{aligned} f(\vec{y}) &\geq f(\vec{x}) + \vec{0}^T (\vec{y} - \vec{x}) \\ &= f(\vec{x}) \end{aligned}$$

for all $\vec{y} \in \mathbb{R}^n$

Optimality condition for nondifferentiable functions

If $\vec{0}$ is a subgradient of f at $\vec{x}$, then

$$\begin{aligned} f(\vec{y}) &\geq f(\vec{x}) + \vec{0}^T (\vec{y} - \vec{x}) \\ &= f(\vec{x}) \end{aligned}$$

for all $\vec{y} \in \mathbb{R}^n$

Under strict convexity the minimum is unique

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$f(\vec{y}) = f_1(\vec{y}) + f_2(\vec{y})$$

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} f(\vec{y}) &= f_1(\vec{y}) + f_2(\vec{y}) \\ &\geq f_1(\vec{x}) + \vec{g}_1^{\,T}(\vec{y} - \vec{x}) + f_2(\vec{y}) + \vec{g}_2^{\,T}(\vec{y} - \vec{x}) \end{aligned}$$

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} f(\vec{y}) &= f_1(\vec{y}) + f_2(\vec{y}) \\ &\geq f_1(\vec{x}) + \vec{g}_1^{\,T}(\vec{y} - \vec{x}) + f_2(\vec{y}) + \vec{g}_2^{\,T}(\vec{y} - \vec{x}) \\ &\geq f(\vec{x}) + \vec{g}^{\,T}(\vec{y} - \vec{x}) \end{aligned}$$

Subgradient of scaled function

Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \rightarrow \mathbb{R}$

For any $\eta \geq 0$ $\vec{g}_2 := \eta \vec{g}_1$ is a subgradient of $f_2 := \eta f_1$ at $\vec{x}$

Subgradient of scaled function

Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\eta \geq 0$ $\vec{g}_2 := \eta \vec{g}_1$ is a subgradient of $f_2 := \eta f_1$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$f_2(\vec{y}) = \eta f_1(\vec{y})$$

Subgradient of scaled function

Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\eta \geq 0$ $\vec{g}_2 := \eta \vec{g}_1$ is a subgradient of $f_2 := \eta f_1$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} f_2(\vec{y}) &= \eta f_1(\vec{y}) \\ &\geq \eta \left(f_1(\vec{x}) + \vec{g}_1^{\,T} (\vec{y} - \vec{x}) \right) \end{aligned}$$

Subgradient of scaled function

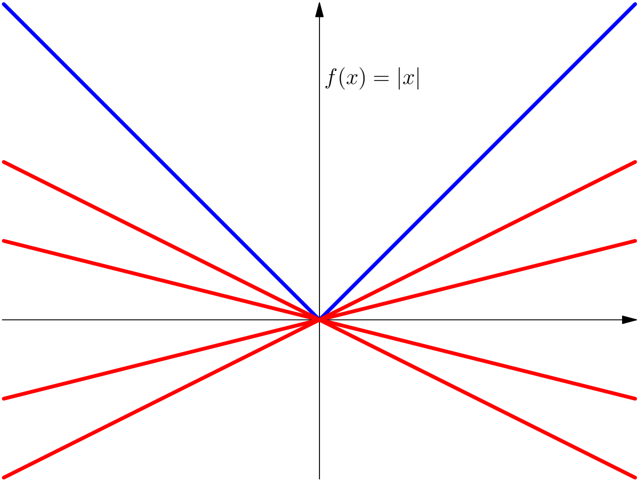
Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\eta \geq 0$ $\vec{g}_2 := \eta \vec{g}_1$ is a subgradient of $f_2 := \eta f_1$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$
\begin{aligned}
f_2(\vec{y}) &= \eta f_1(\vec{y}) \\
&\geq \eta \left(f_1(\vec{x}) + \vec{g}_1^{\,T} (\vec{y} - \vec{x}) \right) \\
&\geq f_2(\vec{x}) + \vec{g}_2^{\,T} (\vec{y} - \vec{x})
\end{aligned}
$$

Subdifferential of absolute value



Subdifferential of absolute value

At $x \neq 0$, $f(x) = |x|$ is differentiable, so $g = \text{sign}(x)$

At $x = 0$, we need

$$f(0 + y) \geq f(0) + g(y - 0)$$

Subdifferential of absolute value

At $x \neq 0$, $f(x) = |x|$ is differentiable, so $g = \text{sign}(x)$

At $x = 0$, we need

$$f(0 + y) \geq f(0) + g(y - 0)$$

$$|y| \geq gy$$

Subdifferential of absolute value

At $x \neq 0$, $f(x) = |x|$ is differentiable, so $g = \text{sign}(x)$

At $x = 0$, we need

$$f(0 + y) \geq f(0) + g(y - 0)$$

$$|y| \geq gy$$

Holds if and only if $|g| \leq 1$

Subdifferential of ℓ_1 norm

$\vec{g}$ is a subgradient of the ℓ_1 norm at $\vec{x} \in \mathbb{R}^n$ if and only if

$$\vec{g}[i] = \operatorname{sign}(x[i]) \qquad \text{if } x[i] \neq 0$$

$$|\vec{g}[i]| \leq 1 \qquad \text{if } \vec{x}[i] = 0$$

Proof

$\vec{g}$ is a subgradient of $||\cdot||_1$ at $\vec{x}$ if and only if $\vec{g}[i]$ is a subgradient of $|\cdot|$ at $\vec{x}[i]$ for all $1 \leq i \leq n$

Proof

If $\vec{g}$ is a subgradient of $||\cdot||_1$ at $\vec{x}$ then for any $y \in \mathbb{R}$

$$|y| = |\vec{x}[i]| + ||\vec{x} + (y - \vec{x}[i])\, \vec{e}_i||_1 - ||\vec{x}||_1$$

Proof

If $\vec{g}$ is a subgradient of $||\cdot||_1$ at $\vec{x}$ then for any $y \in \mathbb{R}$

$$
\begin{aligned}
|y| &= |\vec{x}[i]| + ||\vec{x} + (y - \vec{x}[i])\,\vec{e}_i||_1 - ||\vec{x}||_1 \\
&\geq |\vec{x}[i]| + ||\vec{x}||_1 + \vec{g}^{\,T}\,(y - \vec{x}[i])\,\vec{e}_i - ||\vec{x}||_1
\end{aligned}
$$

Proof

If $\vec{g}$ is a subgradient of $||\cdot||_1$ at $\vec{x}$ then for any $y \in \mathbb{R}$

$$\begin{aligned} |y| &= |\vec{x}[i]| + ||\vec{x} + (y - \vec{x}[i])\, \vec{e}_i||_1 - ||\vec{x}||_1 \\ &\geq |\vec{x}[i]| + ||\vec{x}||_1 + \vec{g}^{\,T} (y - \vec{x}[i])\, \vec{e}_i - ||\vec{x}||_1 \\ &= |\vec{x}[i]| + \vec{g}[i]\, (y - \vec{x}[i]) \end{aligned}$$

so $\vec{g}[i]$ is a subgradient of $|\cdot|$ at $|\vec{x}[i]|$ for all $1 \leq i \leq n$

Proof

If $\vec{g}[i]$ is a subgradient of $|\cdot|$ at $|\vec{x}[i]|$ for $1 \leq i \leq n$ then for any $\vec{y} \in \mathbb{R}^n$

$$||\vec{y}||_1 = \sum_{i=1}^{n} |\vec{y}\,[i]|$$

Proof

If $\vec{g}[i]$ is a subgradient of $|\cdot|$ at $|\vec{x}[i]|$ for $1 \leq i \leq n$ then for any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} ||\vec{y}||_1 &= \sum_{i=1}^{n} |\vec{y}\,[i]| \\ &\geq \sum_{i=1}^{n} |\vec{x}[i]| + \vec{g}[i]\,(\vec{y}\,[i] - \vec{x}[i]) \end{aligned}$$

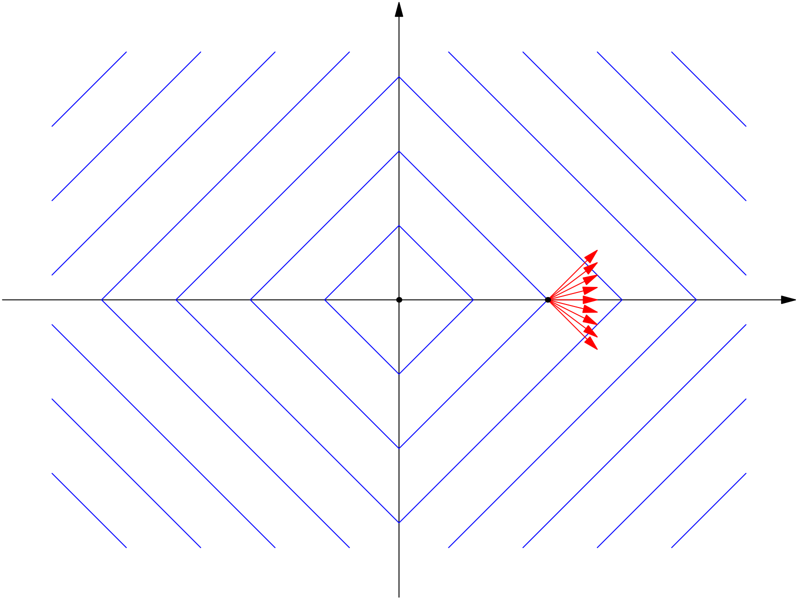
Proof

If $\vec{g}[i]$ is a subgradient of $|\cdot|$ at $|\vec{x}[i]|$ for $1 \leq i \leq n$ then for any $\vec{y} \in \mathbb{R}^n$

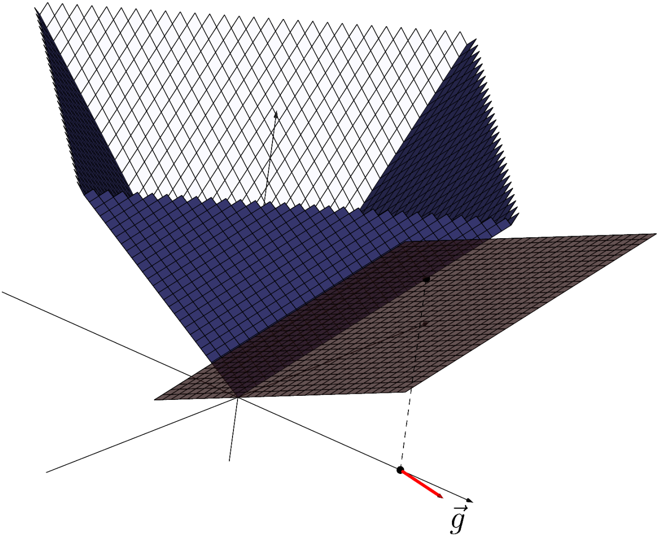
$$
\begin{aligned}
||\vec{y}||_1 &= \sum_{i=1}^{n} |\vec{y}\,[i]| \\
&\geq \sum_{i=1}^{n} |\vec{x}[i]| + \vec{g}[i]\,(\vec{y}\,[i] - \vec{x}[i]) \\
&= ||\vec{x}||_1 + \vec{g}^{\,T}\,(\vec{y} - \vec{x})
\end{aligned}
$$

so $\vec{g}$ is a subgradient of $||\cdot||_1$ at $\vec{x}$

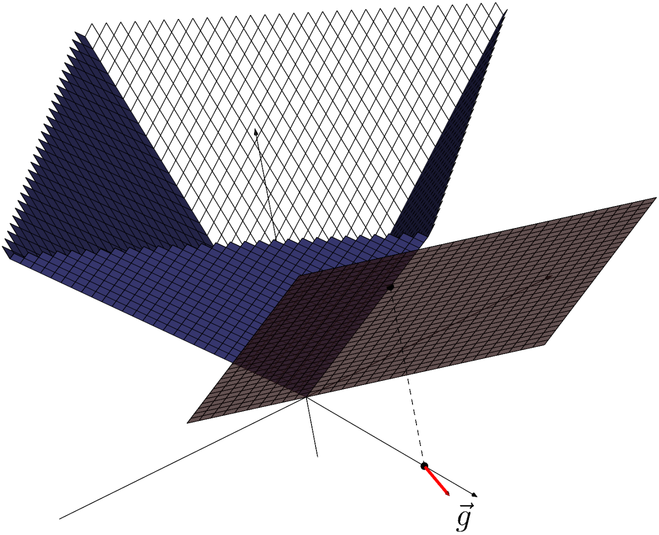
Subdifferential of ℓ_1 norm



Subdifferential of ℓ_1 norm



Subdifferential of ℓ_1 norm



Subdifferential of the nuclear norm

Let $X \in \mathbb{R}^{m \times n}$ be a rank-r matrix with SVD USV^T, where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ and $S \in \mathbb{R}^{r \times r}$

A matrix G is a subgradient of the nuclear norm at X if and only if

$$G := UV^T + W$$

where W satisfies

$$||W|| \leq 1$$
$$U^T W = 0$$
$$W\,V = 0$$

Proof

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = ||\vec{x}||_2^2 = 1$$

Proof

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = \left|\left|\vec{x}\right|\right|_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\left|\left|G\right|\right|^2 := \max_{\left\{\left|\left|\vec{x}\right|\right|_2 = 1 \;|\; \vec{x} \in \mathbb{R}^n\right\}} \left|\left|G\,\vec{x}\right|\right|_2^2$$

Proof

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = \left|\left|\vec{x}\right|\right|_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\begin{aligned} \left|\left|G\right|\right|^2 &:= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|G\,\vec{x}\right|\right|_2^2 \\ &= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\vec{x}\right|\right|_2^2 \end{aligned}$$

Proof

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = \left|\left|\vec{x}\right|\right|_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\begin{aligned}
\left|\left|G\right|\right|^2 &:= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|G\,\vec{x}\right|\right|_2^2 \\
&= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\vec{x}\right|\right|_2^2 \\
&= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|UV^T\,\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2
\end{aligned}$$

Proof

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = \left|\left|\vec{x}\right|\right|_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\begin{aligned}
\left|\left|G\right|\right|^2 &:= \max_{\{||\vec{x}||_2=1 \;|\; \vec{x}\in\mathbb{R}^n\}} \left|\left|G\,\vec{x}\right|\right|_2^2 \\
&= \max_{\{||\vec{x}||_2=1 \;|\; \vec{x}\in\mathbb{R}^n\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\vec{x}\right|\right|_2^2 \\
&= \max_{\{||\vec{x}||_2=1 \;|\; \vec{x}\in\mathbb{R}^n\}} \left|\left|UV^T\,\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 \\
&\leq \left|\left|UV^T\right|\right|^2 \left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\right|\right|^2 \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2
\end{aligned}$$

Proof

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\, \vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\, \vec{x}\right|\right|_2^2 = \left|\left|\vec{x}\right|\right|_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\begin{aligned}
\left|\left|G\right|\right|^2 &:= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|G\,\vec{x}\right|\right|_2^2 \\
&= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\vec{x}\right|\right|_2^2 \\
&= \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} \left|\left|UV^T\,\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 \\
&\leq \left|\left|UV^T\right|\right|^2 \left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\right|\right|^2 \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 \\
&\leq 1
\end{aligned}$$

Hölder's inequality for matrices

For any matrix $A \in \mathbb{R}^{m \times n}$,

$$||A||_* = \sup_{\{||B|| \leq 1 \ | \ B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Proof

For any matrix $Y \in \mathbb{R}^{m \times n}$

$$
\begin{aligned}
||Y||_* &\geq \langle G, Y \rangle \\
&= \langle G, X \rangle + \langle G, Y - X \rangle \\
&= \left\langle UV^T, X \right\rangle + \langle W, X \rangle + \langle G, Y - X \rangle
\end{aligned}
$$

Proof

$U^T W = 0$ implies $\langle W, X \rangle = \left\langle W, USV^T \right\rangle = \left\langle U^T W, SV^T \right\rangle = 0$

$$\left\langle UV^T, X \right\rangle$$

Proof

$U^T W = 0$ implies $\langle W, X \rangle = \left\langle W, USV^T \right\rangle = \left\langle U^T W, SV^T \right\rangle = 0$

$$\left\langle UV^T, X \right\rangle = \operatorname{tr}\left(VU^T X \right)$$

Proof

$U^T W = 0$ implies $\langle W, X \rangle = \left\langle W, USV^T \right\rangle = \left\langle U^T W, SV^T \right\rangle = 0$

$$\begin{aligned}\left\langle UV^T, X \right\rangle &= \operatorname{tr}\left(VU^T X\right) \\ &= \operatorname{tr}\left(VU^T USV^T\right)\end{aligned}$$

Proof

$U^T W = 0$ implies $\langle W, X \rangle = \left\langle W, USV^T \right\rangle = \left\langle U^T W, SV^T \right\rangle = 0$

$$\begin{aligned} \left\langle UV^T, X \right\rangle &= \operatorname{tr}\left(VU^T X \right) \\ &= \operatorname{tr}\left(VU^T USV^T \right) \\ &= \operatorname{tr}\left(V^T V S \right) \end{aligned}$$

Proof

$U^T W = 0$ implies $\langle W, X \rangle = \left\langle W, USV^T \right\rangle = \left\langle U^T W, SV^T \right\rangle = 0$

$$\begin{aligned}
\left\langle UV^T, X \right\rangle &= \operatorname{tr}\left(VU^T X\right) \\
&= \operatorname{tr}\left(VU^T USV^T\right) \\
&= \operatorname{tr}\left(V^T V\, S\right) \\
&= \operatorname{tr}(S)
\end{aligned}$$

Proof

$U^T W = 0$ implies $\langle W, X \rangle = \left\langle W, USV^T \right\rangle = \left\langle U^T W, SV^T \right\rangle = 0$

$$\begin{aligned}
\left\langle UV^T, X \right\rangle &= \text{tr}\left(VU^T X\right) \\
&= \text{tr}\left(VU^T USV^T\right) \\
&= \text{tr}\left(V^T V\, S\right) \\
&= \text{tr}\,(S) \\
&= ||X||_*
\end{aligned}$$

Proof

For any matrix $Y \in \mathbb{R}^{m \times n}$

$$\begin{aligned}
||Y||_* &\geq \langle G, Y \rangle \\
&= \langle G, X \rangle + \langle G, Y - X \rangle \\
&= \left\langle UV^T, X \right\rangle + \langle G, Y - X \rangle \\
&= \left\langle UV^T, X \right\rangle + \langle W, X \rangle + \langle G, Y - X \rangle \\
&= ||X||_* + \langle G, Y - X \rangle
\end{aligned}$$

Sparse linear regression with 2 features

$$\vec{y} := \alpha \, \vec{x}_1 + \vec{z}$$

$$X := \begin{bmatrix} \vec{x}_1 & \vec{x}_2 \end{bmatrix}$$

$$||\vec{x}_1||_2 = 1$$

$$||\vec{x}_2||_2 = 1$$

$$\langle \vec{x}_1, \vec{x}_2 \rangle = \rho$$

Analysis of lasso estimator

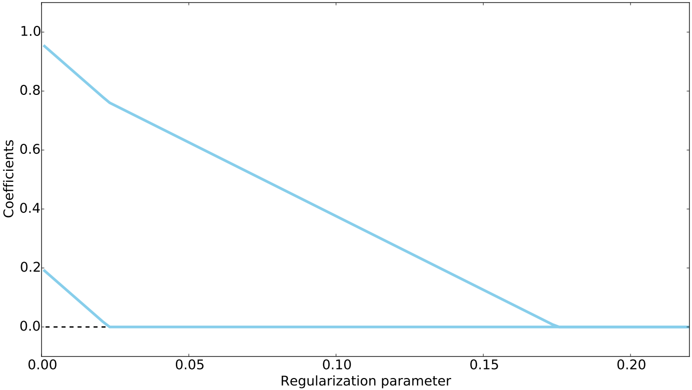
Let $\alpha \geq 0$

$$\vec{\beta}_{\text{lasso}} = \begin{bmatrix} \alpha + \vec{x}_1^{\,T} \vec{z} - \lambda \\ 0 \end{bmatrix}$$

as long as

$$\frac{\left| \vec{x}_2^{\,T} \vec{z} - \rho \vec{x}_1^{\,T} \vec{z} \right|}{1 - |\rho|} \leq \lambda \leq \alpha + \vec{x}_1^{\,T} \vec{z}$$

Lasso estimator



Optimality condition for nondifferentiable functions

If $\vec{0}$ is a subgradient of f at $\vec{x}$, then

$$\begin{aligned} f(\vec{y}) &\geq f(\vec{x}) + \vec{0}^{\,T}(\vec{y} - \vec{x}) \\ &= f(\vec{x}) \end{aligned}$$

for all $\vec{y} \in \mathbb{R}^n$

Under strict convexity the minimum is unique

Proof

The cost function is strictly convex if $n \geq 2$ and $\rho \neq 1$

Aim: Show that there is a subgradient equal to $\vec{0}$ at a 1-sparse solution

Proof

The gradient of the quadratic term

$$q\left(\vec{\beta}\right) := \frac{1}{2}\left|\left|X\vec{\beta} - \vec{y}\right|\right|_2^2$$

at $\vec{\beta}_{\text{lasso}}$ equals

$$\nabla q\left(\vec{\beta}_{\text{lasso}}\right) = X^T\left(X\vec{\beta}_{\text{lasso}} - \vec{y}\right)$$

Proof

If only the first entry is nonzero and nonnegative

$$\vec{g}_{\ell_1} := \begin{bmatrix} 1 \\ \gamma \end{bmatrix}$$

is a subgradient of the ℓ_1 norm at $\vec{\beta}_{\text{lasso}}$ for any $\gamma \in \mathbb{R}$ such that $|\gamma| \leq 1$

Proof

If only the first entry is nonzero and nonnegative

$$\vec{g}_{\ell_1} := \begin{bmatrix} 1 \\ \gamma \end{bmatrix}$$

is a subgradient of the ℓ_1 norm at $\vec{\beta}_{\text{lasso}}$ for any $\gamma \in \mathbb{R}$ such that $|\gamma| \leq 1$

In that case $\vec{g}_{\text{lasso}} := \nabla q\left(\vec{\beta}_{\text{lasso}}\right) + \lambda \vec{g}_{\ell_1}$ is a subgradient of the cost function at $\vec{\beta}_{\text{lasso}}$

Proof

If only the first entry is nonzero and nonnegative

$$\vec{g}_{\ell_1} := \begin{bmatrix} 1 \\ \gamma \end{bmatrix}$$

is a subgradient of the ℓ_1 norm at $\vec{\beta}_{\text{lasso}}$ for any $\gamma \in \mathbb{R}$ such that $|\gamma| \leq 1$

In that case $\vec{g}_{\text{lasso}} := \nabla q \left(\vec{\beta}_{\text{lasso}} \right) + \lambda \vec{g}_{\ell_1}$ is a subgradient of the cost function at $\vec{\beta}_{\text{lasso}}$

If $\vec{g}_{\text{lasso}} = \vec{0}$ then $\vec{\beta}_{\text{lasso}}$ is the unique solution

Proof

$$\vec{g}_{\text{lasso}} := X^T \left(X \vec{\beta}_{\text{lasso}} - \vec{y} \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix}$$

Proof

$$\begin{aligned}\vec{g}_{\text{lasso}} &:= X^T \left(X\vec{\beta}_{\text{lasso}} - \vec{y} \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\ &= X^T \left(\vec{\beta}_{\text{lasso}}[1]\vec{x}_1 - \alpha\vec{x}_1 - \vec{z} \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix}\end{aligned}$$

Proof

$$\begin{aligned}
\vec{g}_{\text{lasso}} &:= X^T \left(X \vec{\beta}_{\text{lasso}} - \vec{y} \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= X^T \left(\vec{\beta}_{\text{lasso}}[1] \vec{x}_1 - \alpha \vec{x}_1 - \vec{z} \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= \begin{bmatrix} \vec{x}_1^{\,T} \left(\vec{\beta}_{\text{lasso}}[1] \vec{x}_1 - \alpha \vec{x}_1 - \vec{z} \right) + \lambda \\ \vec{x}_2^{\,T} \left(\vec{\beta}_{\text{lasso}}[1] \vec{x}_1 - \alpha \vec{x}_1 - \vec{z} \right) + \lambda \gamma \end{bmatrix}
\end{aligned}$$

Proof

$$
\begin{aligned}
\vec{g}_{\text{lasso}} &:= X^T \left(X \vec{\beta}_{\text{lasso}} - \vec{y} \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= X^T \left(\vec{\beta}_{\text{lasso}}[1] \vec{x}_1 - \alpha \vec{x}_1 - \vec{z} \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= \begin{bmatrix} \vec{x}_1^{\,T} \left(\vec{\beta}_{\text{lasso}}[1] \vec{x}_1 - \alpha \vec{x}_1 - \vec{z} \right) + \lambda \\ \vec{x}_2^{\,T} \left(\vec{\beta}_{\text{lasso}}[1] \vec{x}_1 - \alpha \vec{x}_1 - \vec{z} \right) + \lambda \gamma \end{bmatrix} \\
&= \begin{bmatrix} \vec{\beta}_{\text{lasso}}[1] - \alpha - \vec{x}_1^T \vec{z} + \lambda \\ \rho \vec{\beta}_{\text{lasso}}[1] - \rho \alpha - \vec{x}_2^T \vec{z} + \lambda \gamma \end{bmatrix}
\end{aligned}
$$

Proof

$$\vec{g}_{\text{lasso}} = \begin{bmatrix} \vec{\beta}_{\text{lasso}}[1] - \alpha - \vec{x}_1^T \vec{z} + \lambda \\ \rho\vec{\beta}_{\text{lasso}}[1] - \rho\alpha - \vec{x}_2^T \vec{z} + \lambda\gamma \end{bmatrix}$$

Equal to $\vec{0}$ if

Proof

$$\vec{g}_{\text{lasso}} = \begin{bmatrix} \vec{\beta}_{\text{lasso}}[1] - \alpha - \vec{x}_1^T \vec{z} + \lambda \\ \rho \vec{\beta}_{\text{lasso}}[1] - \rho\alpha - \vec{x}_2^T \vec{z} + \lambda\gamma \end{bmatrix}$$

Equal to $\vec{0}$ if

$$\vec{\beta}_{\text{lasso}}[1] = \alpha + \vec{x}_1^T \vec{z} - \lambda$$

Proof

$$\vec{g}_{\text{lasso}} = \begin{bmatrix} \vec{\beta}_{\text{lasso}}[1] - \alpha - \vec{x}_1^T \vec{z} + \lambda \\ \rho\vec{\beta}_{\text{lasso}}[1] - \rho\alpha - \vec{x}_2^T \vec{z} + \lambda\gamma \end{bmatrix}$$

Equal to $\vec{0}$ if

$$\vec{\beta}_{\text{lasso}}[1] = \alpha + \vec{x}_1^T \vec{z} - \lambda$$

$$\gamma = \frac{\rho\alpha + \vec{x}_2^T \vec{z} - \rho\vec{\beta}_{\text{lasso}}[1]}{\lambda}$$

Proof

$$\vec{g}_{\text{lasso}} = \begin{bmatrix} \vec{\beta}_{\text{lasso}}[1] - \alpha - \vec{x}_1^T \vec{z} + \lambda \\ \rho\vec{\beta}_{\text{lasso}}[1] - \rho\alpha - \vec{x}_2^T \vec{z} + \lambda\gamma \end{bmatrix}$$

Equal to $\vec{0}$ if

$$\vec{\beta}_{\text{lasso}}[1] = \alpha + \vec{x}_1^T \vec{z} - \lambda$$

$$\begin{aligned} \gamma &= \frac{\rho\alpha + \vec{x}_2^T \vec{z} - \rho\vec{\beta}_{\text{lasso}}[1]}{\lambda} \\ &= \frac{\vec{x}_2^T \vec{z} - \rho\vec{x}_1^T \vec{z}}{\lambda} + \rho \end{aligned}$$

Proof

We still need to check that it's a valid subgradient at $\vec{\beta}_{\text{lasso}}$, i.e.

- $\vec{\beta}_{\text{lasso}}[1]$ is nonnegative

$$\lambda \leq \alpha + \vec{x}_1^T$$

- $|\gamma| \leq 1$

$$|\gamma| \leq \left| \frac{\vec{x}_2^{\,T} \vec{z} - \rho \vec{x}_1^{\,T} \vec{z}}{\lambda} \right| + |\rho| \leq 1$$

which holds if

$$\lambda \geq \frac{\left| \rho \vec{x}_1^{\,T} \vec{z} + \vec{x}_2^{\,T} \vec{z} \right|}{1 - |\rho|}$$

Robust PCA

Data: $Y \in \mathbb{R}^{n \times m}$

Robust PCA estimator of low-rank component:

$$L_{\text{RPCA}} := \arg\min_{L} ||L||_* + \lambda ||Y - L||_1$$

where $\lambda > 0$ is a regularization parameter

Robust PCA estimator of sparse component: $S_{\text{RPCA}} := Y - L_{\text{RPCA}}$

$||\cdot||_1$ is the ℓ_1 norm of the *vectorized matrix*

Example

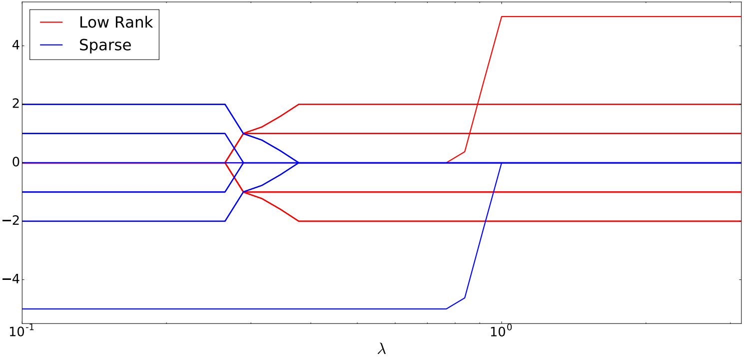
$$Y := \begin{bmatrix} -2 & -1 & \alpha & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

Analysis of robust PCA estimator

The robust PCA estimates of both components are exact for any value of α as long as

$$\frac{2}{\sqrt{30}} < \lambda < \sqrt{\frac{2}{3}}$$

Example



Optimality + uniqueness condition

Let $Y := L^* + S^*$ where $L^*, S^* \in \mathbb{R}^{m\times n}$

$L^* = U_{L^*} S_{L^*} V_{L^*}^T$ has rank r, $U_{L^*} \in \mathbb{R}^{m\times r}$, $V_{L^*} \in \mathbb{R}^{n\times r}$, $S_{L^*} \in \mathbb{R}^{r\times r}$

Assume there exists $G_* := U_{L^*} V_{L^*}^T + W$ where W satisfies

$$||W|| < 1, \qquad U^T W = 0, \qquad W\, V = 0,$$

and there also exists a matrix G_{ℓ_1} satisfying

$$G_{\ell_1}[i,j] = \text{sign}\left(S^*[i,j]\right) \qquad \text{if } S^*[i,j] \neq 0, \tag{1}$$

$$|G_{\ell_1}[i,j]| < 1 \qquad \text{otherwise}, \tag{2}$$

such that $G_* + \lambda G_{\ell_1} = 0$

Then the solution to the robust PCA problem is unique and equal to L^*

Optimality + uniqueness condition

$G_* := U_{L^*} V_{L^*}^T + W$ is a subgradient of the nuclear norm at L^*

G_{ℓ_1} is a subgradient of $||\cdot - Y||_1$ at L^*

$G_* + \lambda G_{\ell_1}$ is a subgradient of the cost function at L^*

$G_* + \lambda G_{\ell_1} = 0$ implies that L^* is a solution (uniqueness is more difficult to prove)

Example

$$Y := \begin{bmatrix} -2 & -1 & \alpha & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

We want to show that the solution is

$$L^* := \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

$$S^* := \begin{bmatrix} 0 & 0 & \alpha & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Example

$$L^* := \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

Example

$$
\begin{aligned}
L^* &:= \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \\
&= \left(\frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) \sqrt{30} \left(\frac{1}{\sqrt{10}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \end{bmatrix} \right)
\end{aligned}
$$

Example

$$L^* := \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

$$= \left(\frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) \sqrt{30} \left(\frac{1}{\sqrt{10}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \end{bmatrix} \right)$$

$$U_{L^*} V_{L^*}^T = \frac{1}{\sqrt{30}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

Example

$$G_* = U_{L^*} V_{L^*}^T + W = \frac{1}{\sqrt{30}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \end{bmatrix} + W$$

Example

$$G_* = U_{L^*} V_{L^*}^T + W = \frac{1}{\sqrt{30}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \end{bmatrix} + W$$

$$G_{\ell_1} = \begin{bmatrix} g_1 & g_2 & -\operatorname{sign}(\alpha) & g_3 & g_4 \\ g_5 & g_6 & g_7 & g_8 & g_9 \\ g_{10} & g_{11} & g_{12} & g_{13} & g_{14} \end{bmatrix}$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} & -\operatorname{sign}(\alpha) & \\ & & \\ & & \end{bmatrix}$$

$$+ \begin{bmatrix} \\ \\ \end{bmatrix}$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} & & -\operatorname{sign}(\alpha) & & \\ & & & & \\ & & & & \end{bmatrix}$$

$$+ \begin{bmatrix} & \lambda \operatorname{sign}(\alpha) & \\ & & \\ & & \end{bmatrix}$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} & & \lambda \operatorname{sign}(\alpha) & & \\ & & & & \\ & & & & \end{bmatrix}$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} & & \lambda \operatorname{sign}(\alpha) & & \\ & & & & \\ & & & & \end{bmatrix}$$

$$WV = 0$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}}\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & & 0 & 0 \end{bmatrix}$$

$$WV = 0$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & & 0 & 0 \end{bmatrix}$$

$$WV = 0 \qquad U^T W = 0$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \end{bmatrix}$$

$$WV = 0 \qquad U^T W = 0$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \end{bmatrix}$$

$$WV = 0 \qquad U^T W = 0$$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \end{bmatrix}$$

$$WV = 0 \qquad U^T W = 0$$

$|G_{\ell_1}[i,j]| < 1$ for $S^*[i,j] = 0$?

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \end{bmatrix}$$

$$WV = 0 \qquad U^T W = 0$$

$|G_{\ell_1}[i,j]| < 1$ for $S^*[i,j] = 0$? $\lambda > 2/\sqrt{30}$

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \end{bmatrix}$$

$$WV = 0 \qquad U^T W = 0$$

$|G_{\ell_1}[i,j]| < 1$ for $S^*[i,j] = 0$? $\lambda > 2/\sqrt{30}$

$||W|| < 1$?

Example

$$G_* + \lambda G_{\ell_1} =$$

$$\frac{1}{\sqrt{30}} \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} + \lambda \begin{bmatrix} \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & -\operatorname{sign}(\alpha) & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \\ \frac{2}{\lambda\sqrt{30}} & \frac{1}{\lambda\sqrt{30}} & \frac{\lambda \operatorname{sign}(\alpha)}{2\lambda} & -\frac{1}{\lambda\sqrt{30}} & -\frac{2}{\lambda\sqrt{30}} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 & \lambda \operatorname{sign}(\alpha) & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \\ 0 & 0 & -\frac{\lambda \operatorname{sign}(\alpha)}{2} & 0 & 0 \end{bmatrix}$$

$$WV = 0 \qquad U^T W = 0$$

$|G_{\ell_1}[i,j]| < 1$ for $S^*[i,j] = 0$? $\lambda > 2/\sqrt{30}$

$||W|| < 1$? $\lambda < \sqrt{2/3}$

Applications

Subgradients

Optimization methods

Subgradient method

Optimization problem

$$\text{minimize} \quad f(\vec{x})$$

where f is convex but nondifferentiable

Subgradient-method iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$
$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \alpha_k \vec{g}^{(k)}$$

where $\vec{g}^{(k)}$ is a subgradient of f at $\vec{x}^{(k)}$

Least-squares regression with ℓ_1-norm regularization

$$\text{minimize} \quad \frac{1}{2} \left|\left| A\vec{x} - \vec{y} \right|\right|_2^2 + \lambda \left|\left| \vec{x} \right|\right|_1$$

Subgradient at $\vec{x}^{(k)}$

$$\vec{g}^{(k)}$$

Least-squares regression with ℓ_1-norm regularization

$$\text{minimize} \quad \frac{1}{2} \left|\left|A\vec{x} - \vec{y}\right|\right|_2^2 + \lambda \left|\left|\vec{x}\right|\right|_1$$

Subgradient at $\vec{x}^{(k)}$

$$\vec{g}^{(k)} = A^T \left(A\vec{x}^{(k)} - \vec{y}\right) + \lambda \operatorname{sign}\left(\vec{x}^{(k)}\right)$$

Least-squares regression with ℓ_1-norm regularization

$$\text{minimize} \quad \frac{1}{2} \left|\left| A\vec{x} - \vec{y} \right|\right|_2^2 + \lambda \left|\left| \vec{x} \right|\right|_1$$

Subgradient at $\vec{x}^{(k)}$

$$\vec{g}^{(k)} = A^T \left(A\vec{x}^{(k)} - \vec{y} \right) + \lambda \, \text{sign}\left(\vec{x}^{(k)} \right)$$

Subgradient-method iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \alpha_k \left(A^T \left(A\vec{x}^{(k)} - \vec{y} \right) + \lambda \, \text{sign}\left(\vec{x}^{(k)} \right) \right)$$

Convergence of subgradient method

It is not a descent method

Convergence rate can be shown to be $\mathcal{O}\left(1/\epsilon^2\right)$

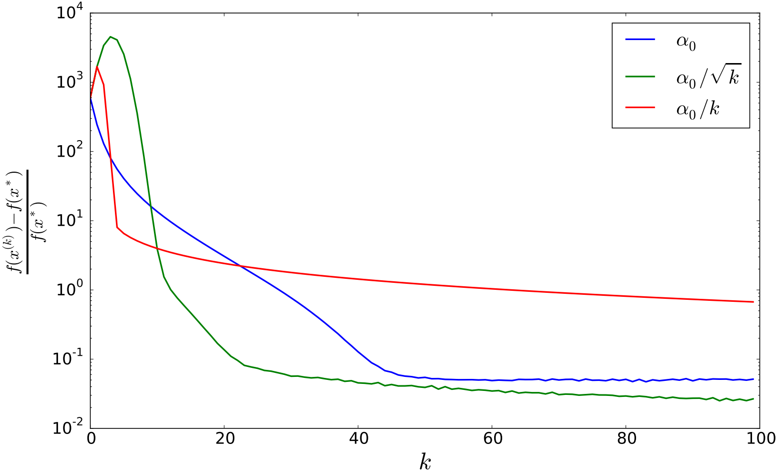
Diminishing step sizes are necessary for convergence

Experiment:

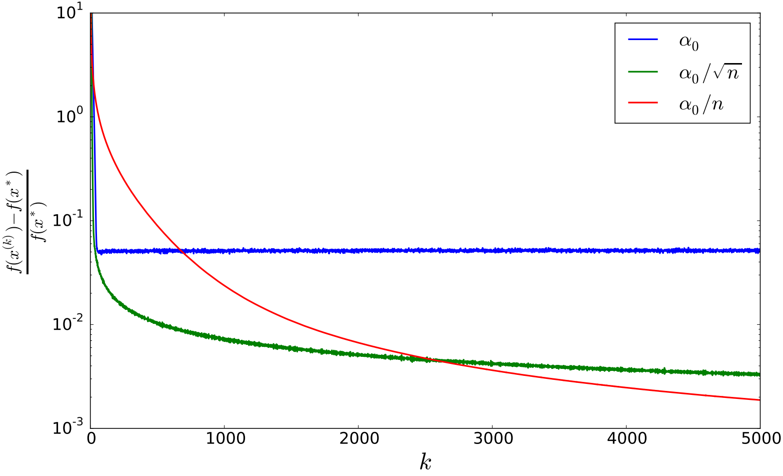
$$\text{minimize} \quad \frac{1}{2}\left|\left|A\vec{x}-\vec{y}\right|\right|_2^2+\lambda\left|\left|\vec{x}\right|\right|_1$$

$A \in \mathbb{R}^{2000\times 1000}$, $y = A\vec{x}^* + \vec{z}$ where $\vec{x}^*$ is 100-sparse and $\vec{z}$ is iid Gaussian

Convergence of subgradient method



Convergence of subgradient method



Composite functions

Interesting class of functions for data analysis

$$f(\vec{x}) + h(\vec{x})$$

f convex and differentiable, h convex but not differentiable

Example:

$$\frac{1}{2} ||A\vec{x} - \vec{y}||_2^2 + \lambda \, ||\vec{x}||_1$$

Motivation

Aim: Minimize convex differentiable function f

Idea: Iteratively minimize first-order approximation, while staying close to current point

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$

$$\vec{x}^{(k+1)} = \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2$$

where α_k is a parameter that determines how close we stay

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$

$$= \nabla f\left(\vec{x}^{(k)}\right) + \frac{\vec{x} - \vec{x}^{(k)}}{\alpha_k}$$

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$

$$= \nabla f\left(\vec{x}^{(k)}\right) + \frac{\vec{x} - \vec{x}^{(k)}}{\alpha_k}$$

Setting the gradient to zero

$$\vec{x}^{(k+1)} = \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2$$

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$
$$= \nabla f\left(\vec{x}^{(k)}\right) + \frac{\vec{x} - \vec{x}^{(k)}}{\alpha_k}$$

Setting the gradient to zero

$$\vec{x}^{(k+1)} = \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2$$
$$= \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

Proximal gradient method

Idea: Minimize local first-order approximation $+ h$

$$
\begin{aligned}
\vec{x}^{(k+1)} &= \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \\
&\qquad + h\left(\vec{x}\right) \\
&= \arg\min_{\vec{x}} \frac{1}{2} \left|\left|x - \left(\vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right)\right|\right|_2^2 + \alpha_k\, h\left(\vec{x}\right) \\
&= \operatorname{prox}_{\alpha_k h}\left(\vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right)
\end{aligned}
$$

Proximal operator:

$$
\operatorname{prox}_h\left(y\right) := \arg\min_{\vec{x}} h\left(\vec{x}\right) + \frac{1}{2} \left|\left|y - x\right|\right|_2^2
$$

Proximal gradient method

Method to solve the optimization problem

$$\text{minimize} \quad f(\vec{x}) + h(\vec{x}),$$

where f is differentiable and prox_h is tractable

Proximal-gradient iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$
$$\vec{x}^{(k+1)} = \text{prox}_{\alpha_k h}\left(\vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right)$$

Interpretation as a fixed-point method

A vector $\vec{x}^*$ is a solution to

$$\text{minimize} \quad f(\vec{x}) + h(\vec{x}),$$

if and only if it is a fixed point of the proximal-gradient iteration for any $\alpha > 0$

$$\vec{x}^* = \text{prox}_{\alpha\, h}\left(\vec{x}^* - \alpha\, \nabla f(\vec{x}^*)\right)$$

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h(\vec{x}) + \frac{1}{2} \left|\left|\vec{x}^* - \alpha \nabla f(\vec{x}^*) - \vec{x}\right|\right|_2^2 \tag{3}$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h\left(\vec{x}\right) + \frac{1}{2}\left|\left|\vec{x}^* - \alpha\, \nabla f\left(\vec{x}^*\right) - \vec{x}\right|\right|_2^2 \tag{3}$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that
$\alpha \nabla f\left(\vec{x}^*\right) + \alpha \vec{g} = \vec{0}$

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h(\vec{x}) + \frac{1}{2}\,||\vec{x}^* - \alpha\,\nabla f(\vec{x}^*) - \vec{x}||_2^2 \tag{3}$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that
$\alpha\nabla f(\vec{x}^*) + \alpha\vec{g} = \vec{0}$

$\vec{x}^*$ minimizes $f + h$ if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h\left(\vec{x}\right) + \frac{1}{2}\left|\left|\vec{x}^* - \alpha\, \nabla f\left(\vec{x}^*\right) - \vec{x}\right|\right|_2^2 \tag{3}$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that $\alpha\nabla f\left(\vec{x}^*\right) + \alpha\vec{g} = \vec{0}$

$\vec{x}^*$ minimizes $f + h$ if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that $\nabla f\left(\vec{x}^*\right) + \vec{g} = \vec{0}$

Proximal operator of ℓ_1 norm

The proximal operator of the ℓ_1 norm is the soft-thresholding operator

$$\text{prox}_{\alpha\,||\cdot||_1}(y) = \mathcal{S}_\alpha(y)$$

where $\alpha > 0$ and

$$\mathcal{S}_\alpha(y)_i := \begin{cases} y_i - \text{sign}(y_i)\,\alpha & \text{if } |y_i| \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

Proof

$$\alpha \left|\left|\vec{x}\right|\right|_1 + \frac{1}{2} \left|\left|\vec{y} - \vec{x}\right|\right|_2^2 = \alpha \sum_{i=1}^{m} \left|\vec{x}[i]\right| + \frac{1}{2} \left(\vec{y}[i] - \vec{x}[i]\right)^2$$

We can just consider

$$w\left(x\right) := \alpha \left|x\right| + \frac{1}{2} \left(y - x\right)^2 = \frac{y^2 + x^2}{2} + \alpha \left|x\right| - yx$$

Proof

If $x \geq 0$

$$
\begin{aligned}
w(x) &= \frac{y^2 + x^2}{2} - (y - \alpha)\, x \\
w'(x) &=
\end{aligned}
$$

Proof

If $x \geq 0$

$$\begin{aligned} w(x) &= \frac{y^2 + x^2}{2} - (y - \alpha)\, x \\ w'(x) &= x - (y - \alpha) \end{aligned}$$

Proof

If $x \geq 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y - \alpha)\, x$$
$$w'(x) = x - (y - \alpha)$$

If $y \geq \alpha$ minimum at

Proof

If $x \geq 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y - \alpha)\, x$$
$$w'(x) = x - (y - \alpha)$$

If $y \geq \alpha$ minimum at $x := y - \alpha$

Proof

If $x \geq 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y - \alpha)x$$
$$w'(x) = x - (y - \alpha)$$

If $y \geq \alpha$ minimum at $x := y - \alpha$

If $y < \alpha$ minimum at

Proof

If $x \geq 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y - \alpha)x$$
$$w'(x) = x - (y - \alpha)$$

If $y \geq \alpha$ minimum at $x := y - \alpha$

If $y < \alpha$ minimum at 0

Proof

If $x < 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y + \alpha)x$$
$$w'(x) =$$

Proof

If $x < 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y + \alpha)\,x$$
$$w'(x) = x - (y + \alpha)$$

Proof

If $x < 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y + \alpha)x$$
$$w'(x) = x - (y + \alpha)$$

If $y \leq -\alpha$ minimum at

Proof

If $x < 0$

$$\begin{aligned} w(x) &= \frac{y^2 + x^2}{2} - (y + \alpha)\,x \\ w'(x) &= x - (y + \alpha) \end{aligned}$$

If $y \leq -\alpha$ minimum at $x := y + \alpha$

Proof

If $x < 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y + \alpha)x$$
$$w'(x) = x - (y + \alpha)$$

If $y \leq -\alpha$ minimum at $x := y + \alpha$

If $y \geq -\alpha$ minimum at

Proof

If $x < 0$

$$w(x) = \frac{y^2 + x^2}{2} - (y + \alpha)x$$
$$w'(x) = x - (y + \alpha)$$

If $y \leq -\alpha$ minimum at $x := y + \alpha$

If $y \geq -\alpha$ minimum at 0

Proof

If $-\alpha \leq y \leq \alpha$ minimum at $x := 0$

If $y \geq \alpha$ minimum at $x := y - \alpha$ or at $x := 0$, but

$$w\,(y - \alpha)$$

Proof

If $-\alpha \leq y \leq \alpha$ minimum at $x := 0$

If $y \geq \alpha$ minimum at $x := y - \alpha$ or at $x := 0$, but

$$w(y - \alpha) = \alpha(y - \alpha) + \frac{\alpha^2}{2}$$

Proof

If $-\alpha \leq y \leq \alpha$ minimum at $x := 0$

If $y \geq \alpha$ minimum at $x := y - \alpha$ or at $x := 0$, but

$$\begin{aligned} w(y-\alpha) &= \alpha(y-\alpha) + \frac{\alpha^2}{2} \\ &= \alpha y - \frac{\alpha^2}{2} \end{aligned}$$

Proof

If $-\alpha \leq y \leq \alpha$ minimum at $x := 0$

If $y \geq \alpha$ minimum at $x := y - \alpha$ or at $x := 0$, but

$$\begin{aligned} w(y-\alpha) &= \alpha(y-\alpha) + \frac{\alpha^2}{2} \\ &= \alpha y - \frac{\alpha^2}{2} \\ &\leq \frac{y^2}{2} = w(0) \end{aligned}$$

because $(y-\alpha)^2 \geq 0$

Proof

If $-\alpha \leq y \leq \alpha$ minimum at $x := 0$

If $y \geq \alpha$ minimum at $x := y - \alpha$ or at $x := 0$, but

$$\begin{aligned} w(y-\alpha) &= \alpha(y-\alpha) + \frac{\alpha^2}{2} \\ &= \alpha y - \frac{\alpha^2}{2} \\ &\leq \frac{y^2}{2} = w(0) \end{aligned}$$

because $(y-\alpha)^2 \geq 0$

Same argument for $y < \alpha$

Iterative Shrinkage-Thresholding Algorithm (ISTA)

The proximal gradient method for the problem

$$\text{minimize} \quad \frac{1}{2} \left|\left|A\vec{x} - \vec{y}\right|\right|_2^2 + \lambda \left|\left|\vec{x}\right|\right|_1$$

is called ISTA

ISTA iteration:

$$\begin{aligned}
\vec{x}^{(0)} &= \text{arbitrary initialization} \\
\vec{x}^{(k+1)} &= \mathcal{S}_{\alpha_k \lambda}\left(\vec{x}^{(k)} - \alpha_k A^T \left(A\vec{x}^{(k)} - \vec{y}\right)\right)
\end{aligned}$$

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

ISTA can be accelerated using Nesterov's accelerated gradient method

FISTA iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$
$$\vec{z}^{(0)} = \vec{x}^{(0)}$$
$$\vec{x}^{(k+1)} = \mathcal{S}_{\alpha_k \lambda}\left(\vec{z}^{(k)} - \alpha_k A^T\left(A\vec{z}^{(k)} - \vec{y}\right)\right)$$
$$\vec{z}^{(k+1)} = \vec{x}^{(k+1)} + \frac{k}{k+3}\left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right)$$

Convergence of proximal gradient method

Without acceleration:

- Descent method
- Convergence rate can be shown to be $\mathcal{O}(1/\epsilon)$ with constant step or backtracking line search

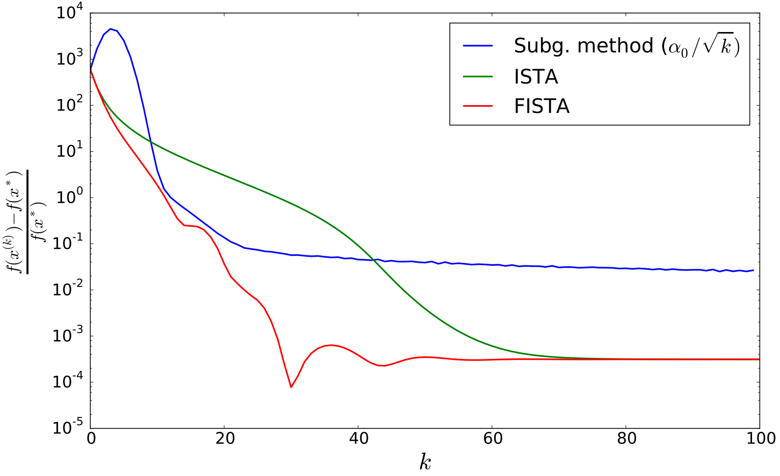
With acceleration:

- Not a descent method
- Convergence rate can be shown to be $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$ with constant step or backtracking line search

Experiment: minimize $\quad \frac{1}{2}\left|\left|A\vec{x}-\vec{y}\right|\right|_2^2+\lambda\left|\left|\vec{x}\right|\right|_1$

$A \in \mathbb{R}^{2000\times 1000}$, $y = A\vec{x}_0 + \vec{z}$, x_0 100-sparse and z iid Gaussian

Convergence of proximal gradient method



Coordinate descent

Idea: Solve the n-dimensional problem

$$\text{minimize} \quad c\left(\vec{x}[1], \vec{x}[2], \ldots, \vec{x}[n]\right)$$

by solving a sequence of 1D problems

Coordinate-descent iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$

$$\vec{x}^{(k+1)}[i] = \arg\min_{\alpha} c\left(\vec{x}^{(k)}[1], \ldots, \alpha, \ldots, \vec{x}^{(k)}[n]\right) \quad \text{for some } 1 \leq i \leq n$$

Coordinate descent

Convergence is guaranteed for functions of the form

$$f(\vec{x}) + \sum_{i=1}^{n} h_i(\vec{x}[i])$$

where f is convex and differentiable and $h_1, \ldots, h_n$ are convex

Least-squares regression with ℓ_1-norm regularization

$$h(\vec{x}) := \frac{1}{2} \left|\left| A\vec{x} - \vec{y} \right|\right|_2^2 + \lambda \left|\left| \vec{x} \right|\right|_1$$

The solution to the subproblem $\min_{\vec{x}[i]} h(\vec{x}[1], \ldots, \vec{x}[i], \ldots, \vec{x}[n])$ is

$$\vec{x}^*[i] = \frac{\mathcal{S}_\lambda(\gamma_i)}{\left|\left| A_i \right|\right|_2^2}$$

where A_i is the ith column of A and

$$\gamma_i := \sum_{l=1}^{m} A_{li} \left(\vec{y}[l] - \sum_{j \neq i} A_{lj} \vec{x}[j] \right)$$