02471 Machine Learning for Signal Processing

Solution

Exercise 7: Sparsity analysis models and time-frequency analysis

7.2 Iterative Shrinkage/thresholding (IST)

Exercise 7.2.1

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{L(\boldsymbol{\theta},\lambda)=\frac{1}{2}\|\boldsymbol{y}-X\boldsymbol{\theta}\|_2^2+\lambda\|\boldsymbol{\theta}\|_1=J(\boldsymbol{\theta})+\lambda\|\boldsymbol{\theta}\|_1\right\}$$

From section 5.2, formula 5.3, we have the gradient decent update defined as

$$\boldsymbol{\theta}^{(i)}=\boldsymbol{\theta}^{(i-1)}-\mu_i J'\big(\boldsymbol{\theta}^{(i-1)}\big)$$

And using a constant step size μ, and taking the derivative of the means squared cost function, we get

$$\begin{aligned}\boldsymbol{\theta}^{(i)}&=\boldsymbol{\theta}^{(i-1)}-\mu J'\big(\boldsymbol{\theta}^{(i-1)}\big)\\&=\boldsymbol{\theta}^{(i-1)}-\mu X^T\big(X\boldsymbol{\theta}^{(i-1)}-\boldsymbol{y}\big)\\&=\boldsymbol{\theta}^{(i-1)}+\mu X^T\boldsymbol{e}^{(i-1)}\end{aligned}$$

where $\boldsymbol{e}^{(i-1)}=\boldsymbol{y}-X\boldsymbol{\theta}^{(i-1)}$.

The gradient decent update that solves the MSE error can identically be written as the solution to the following optimization problem:

$$\boldsymbol{\theta}^{(i)}=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{J\big(\boldsymbol{\theta}^{(i-1)}\big)+\big(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big)^T J'\big(\boldsymbol{\theta}^{(i-1)}\big)+\frac{1}{2\mu}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2\right\}$$

This can easiest be shown by taking the derivative to the expression we are minimizing w.r.t $\boldsymbol{\theta}$, set equal to zero and then solve for $\boldsymbol{\theta}$.

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}\left(J\big(\boldsymbol{\theta}^{(i-1)}\big)+\big(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big)^T J'\big(\boldsymbol{\theta}^{(i-1)}\big)+\frac{1}{2\mu}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2\right)=\\&\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\theta}^T J'\big(\boldsymbol{\theta}^{(i-1)}\big)+\frac{1}{2\mu}\frac{\partial}{\partial\boldsymbol{\theta}}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2=\\&J'\big(\boldsymbol{\theta}^{(i-1)}\big)+\frac{1}{2\mu}\big(2\boldsymbol{\theta}-2\boldsymbol{\theta}^{(i-1)}\big)\end{aligned}$$

Equation to zero, we get

$$\begin{aligned}0&=J'\big(\boldsymbol{\theta}^{(i-1)}\big)+\frac{1}{2\mu}\big(2\boldsymbol{\theta}^{(i)}-2\boldsymbol{\theta}^{(i-1)}\big)\\&=\mu J'\big(\boldsymbol{\theta}^{(i-1)}\big)+\boldsymbol{\theta}^{(i)}-\boldsymbol{\theta}^{(i-1)}\Leftrightarrow\\\boldsymbol{\theta}^{(i)}&=\boldsymbol{\theta}^{(i-1)}-\mu J'\big(\boldsymbol{\theta}^{(i-1)}\big)\end{aligned}$$

Thus we have validated the claim. Hence, by substitution, we rewrite our LASSO problem to

$$\boldsymbol{\theta}^{(i)}=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{J\big(\boldsymbol{\theta}^{(i-1)}\big)+\big(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big)^T\frac{\partial\boldsymbol{J}\big(\boldsymbol{\theta}^{(i-1)}\big)}{\partial\boldsymbol{\theta}}+\frac{1}{2\mu}\big\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\big\|_2^2+\lambda\|\boldsymbol{\theta}\|_1\right\}$$

To shorten the notation burden, define $\boldsymbol{\theta}' := \boldsymbol{\theta}^{(i-1)}$

$$\begin{aligned}
\boldsymbol{\theta}^{(i)} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J\left(\boldsymbol{\theta}^{(i-1)}\right) + \left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\right)^T \frac{\partial \boldsymbol{J}\left(\boldsymbol{\theta}^{(i-1)}\right)}{\partial \boldsymbol{\theta}} + \frac{1}{2\mu}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i-1)}\right\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\} \\
&= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J(\boldsymbol{\theta}') + (\boldsymbol{\theta}-\boldsymbol{\theta}')^T J'(\boldsymbol{\theta}') + \frac{1}{2\mu}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}
\end{aligned}$$

Scaling a function (multiplication) or adding a constant to a function does not change the value of the minimizer, i.e. $\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\{f(\boldsymbol{\theta})\} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\{\alpha f(\boldsymbol{\theta}) + k\}, \alpha \in \mathbb{R}, k \in \mathbb{R}$. Using this, we can further rewrite the optimization problem

$$\begin{aligned}
\boldsymbol{\theta}^{(i)} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ J(\boldsymbol{\theta}') + (\boldsymbol{\theta}-\boldsymbol{\theta}')^T J'(\boldsymbol{\theta}') + \frac{1}{2\mu}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\} \\
&= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ \boldsymbol{\theta}^T J'(\boldsymbol{\theta}') + \frac{1}{2\mu}\left(\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{\theta}'\right) + \lambda\|\boldsymbol{\theta}\|_1 \right\} \\
&= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ \mu\boldsymbol{\theta}^T J'(\boldsymbol{\theta}') + \frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{\theta}'\right) + \lambda\mu\|\boldsymbol{\theta}\|_1 \right\} \\
&= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ -\boldsymbol{\theta}^T(\boldsymbol{\theta}' - \mu J'(\boldsymbol{\theta}')) + \frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{\theta}\right) + \lambda\mu\|\boldsymbol{\theta}\|_1 \right\}
\end{aligned}$$

Define $\tilde{\boldsymbol{\theta}} := \boldsymbol{\theta}' - \mu J'(\boldsymbol{\theta}')$ to obtain

$$\boldsymbol{\theta}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^l}\left\{ \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\theta}^T\tilde{\boldsymbol{\theta}} + \lambda\mu\|\boldsymbol{\theta}\|_1 \right\}$$

To find the minimum of this function, we take the derivative if the function to minimize w.r.t. $\boldsymbol{\theta}$ to obtain

$$\frac{\partial}{\partial\boldsymbol{\theta}}\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\theta}^T\tilde{\boldsymbol{\theta}} + \lambda\mu\|\boldsymbol{\theta}\|_1 = \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} + \lambda\mu\partial\|\boldsymbol{\theta}\|_1$$

where ∂ is the subdifferential set. The minimizer must satisfy

$$\mathbf{0} \in \boldsymbol{\theta}^{(i)} - \tilde{\boldsymbol{\theta}} + \lambda\mu\partial\|\boldsymbol{\theta}^{(i)}\|_1$$

What is very useful about this result, as compared to last week, is that no requirements was enforced on X. Last week we assumed that $X^TX = I$. All operations are only applied components wise, and we can therefore write, for the j'th component

$$\begin{cases} \theta_j^{(i)} - \tilde{\theta}_j + \lambda\mu = 0 \Leftrightarrow \theta_j^{(i)} = \tilde{\theta}_j - \lambda\mu & \text{if } \theta_j > 0 \\ \theta_j^{(i)} - \tilde{\theta}_j - \lambda\mu = 0 \Leftrightarrow \theta_j^{(i)} = \tilde{\theta}_j + \lambda\mu & \text{if } \theta_j < 0 \end{cases}$$

These equations are only true for $\tilde{\theta}_j > \lambda\mu$ and $\tilde{\theta}_j < \lambda\mu$ respectively. For $-\lambda\mu \leq \tilde{\theta}_j \leq \lambda\mu$ we get $\tilde{\theta}_j = 0$ (see the solution from last week for further details). These three conditions can be combined into one rule

$$\theta_j^{(i)} = \operatorname{sign}\left(\tilde{\theta}_j^{(i)}\right)\max\left(\left|\tilde{\theta}_j\right| - \lambda\mu, 0\right)$$

where $\operatorname{sign}(x)$ is

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Putting this together, we get the following update for the i'th iteration, where we initialize $\boldsymbol{\theta}^{(0)} = \mathbf{0}$:

$$\begin{aligned}
\boldsymbol{e}^{(i-1)} &= \boldsymbol{y} - X\boldsymbol{\theta}^{(i-1)} \\
\tilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}^{(i-1)} + \mu X^T \boldsymbol{e}^{(i-1)} \\
\boldsymbol{\theta}^{(i)} &= \operatorname{sign}\left(\tilde{\boldsymbol{\theta}}\right) \max\left(\left|\tilde{\boldsymbol{\theta}}\right| - \lambda\mu, 0\right)
\end{aligned}$$

7.3 Signal representation using the Discrete Fourier Transform (DFT)

Exercise 7.3.1

First we define $W_N := e^{-i2\pi/N}$ to rewrite the N-point DFT to

$$\tilde{x}_k = \sum_{n=0}^{N-1} x_n (W_N)^{kn}$$

This sum can be written as an inner product. If define a vector $\boldsymbol{w}^k$ whose elements is $w_i^k = (W_N)^{ki}$, we get

$$\tilde{x}_k = \boldsymbol{x}^T \boldsymbol{w}^k$$

If we want to calculate all the desired $\tilde{x}_k$ components, we can do this using the matrix product

$$\tilde{\boldsymbol{x}} = \Phi^H \boldsymbol{x}$$

Where the matrix Φ^H has the following elements, $\Phi^H_{(i,j)} = (W_N)^{ij}$.