HDDA Final Exam: Question 1

Ian Dover

April 2023

Solve the following optimization problem:

$$\underset{\Theta}{argmin} \frac{1}{2}\sum_{i=1}^{N}(y_i - Tr(X_i\Theta))^2 + \lambda||\Theta||_*$$

The nuclear norm is defined as:

$$||\Theta||_* = \sum_{i=1}^{N}\sigma_i$$

, where σ_i denotes the i-th singular value of the Θ matrix.

1 Part 1

Show that the minimization problem can be solved using the proximal gradient descent algorithm, and write down the procedure.

We know that the first term in the minimzation problem is differentiable (traces are differentiable on all real-numbers in the domain).

However, the nuclear-norm is not necessarily differentiable if matrix Θ does not have full-rank: meaning that the rank is strictly less than the number of rows or columns of the matrix (which is possible).

Because the first term is differentiable and the second term is not, our minimization problem is in the appropriate form for proximal gradient descent.

Let us define the differentiable term:

$$g(\Theta) = \frac{1}{2}\sum_{i=1}^{N}(y_i - Tr(X_i\Theta))^2$$

Let us define the non-differentiable term:

$$h(\Theta) = \lambda||\Theta||_*$$

The minimization problem can be rewritten as:

$$\underset{\Theta}{argmin} \quad g(\Theta) + h(\Theta)$$

In order to find the form of the proximal function, we must determine the gradient (∇) of $g(\Theta)$:

$$\nabla g(\Theta) = \frac{\partial \left[\frac{1}{2} \sum_{i=1}^{N} (y_i - Tr(X_i\Theta))^2 \right]}{\partial \Theta}$$

$$= \sum_{i=1}^{N} \frac{\partial(-Tr(X_i\Theta))}{\partial \Theta} (y_i - Tr(X_i\Theta))$$

We know the identity for the derivative of a product within a trace:

$$\nabla_B Tr(AB) = A^T$$

Therefore, we can solve for the gradient:

$$\nabla g(\Theta) = \sum_{i=1}^{N} -X_i^T (y_i - Tr(X_i\Theta))$$

Now we can define the proximal function as:

$$prox_t(\Theta - t\nabla g(\Theta)) = S_{\lambda t}((\Theta - t\nabla g(\Theta))$$

Where the soft-thresholding function ($S_{\lambda t}$) is defined as:

$$S_{\lambda t}(\Theta) = \begin{cases} \Theta - \lambda t & \Theta \geq \lambda \\ 0 & -\lambda \leq \Theta < \lambda \\ \Theta + \lambda t & \Theta < -\lambda \end{cases}$$

We can now write the proximal gradient descent procedure:

Proximal Gradient Descent Algorithm

```
Initialize $\Theta_0 \in \mathbf{R}^{n \times n}$
Select $\lambda \in \mathbf{R}^1$
for k = 1 : K do
    Compute $\Theta_k = S_{\lambda t}(\Theta_{k-1} - t\nabla g(\Theta_{k-1}))$
end for
```

2 Part 2

Write down the detailed procedure for solving the minimization problem by using the accelerated proximal gradient descent algorithm.

The accelerated proximal gradient descent algorithm uses a similar procedure with some minor alterations. The procedure is written below:

Accelerated Proximal Gradient Descent Algorithm

```
Initialize $\Theta_{-1}, \Theta_0 \in \mathbf{R}^{n \times n}$
Select $\lambda \in \mathbf{R}^1$
for k = 1 : K do
    Compute $v = \Theta_{k-1} + \frac{k-2}{k+1}(\Theta_{k-1} - \Theta_{k-2})$
    Compute $\Theta_k = S_{\lambda t}(v - t\nabla g(v))$
end for
```

3 Part 3

This problem is equivalent to the following problem:

$$\underset{\theta}{argmin} \frac{1}{2}\sum_{i=1}^{N}(y_i - Tr(X_i\theta_x))^2 + \lambda||\theta_y||_* \qquad s.t. \quad \theta_x = \theta_y$$

where $\theta_x, \theta_y \in \mathbf{R}^{n \times n}$.

Solve the optimization problem using the ADMM method. Write down the detailed procedure.

Using the constraint, we can determine Lagrangian and Augmented terms. Below is the Lagrangian term:

$$\mathscr{L}(\theta_x, \theta_y, u) = u^T(\theta_x - \theta_y)$$

where $u \in \mathbf{R}^{n \times n}$

Below is the Augmented term:

$$\mathscr{A}(\theta_x, \theta_y; \rho) = \frac{\rho}{2}||\theta_x - \theta_y||_2^2$$

where $\rho \in \mathbf{R}^1, \rho > 0$

Given that the initial minization problem is defined as $f(\theta_x, \theta_y)$, the new minimization problem can be written as:

$$L(\theta_x, \theta_y, u; \rho) = f(\theta_x, \theta_y) + \mathscr{L}(\theta_x, \theta_y, u) + \mathscr{A}(\theta_x, \theta_y; \rho)$$

The new minimization formulation can be written as:

$$\underset{\theta}{argmin} \quad L(\theta_x, \theta_y, u; \rho) = \underset{\theta}{argmin} \quad f(\theta_x, \theta_y) + \mathscr{L}(\theta_x, \theta_y, u) + \mathscr{A}(\theta_x, \theta_y; \rho)$$

We can now take the gradient (∇_θ) of the new formulation:

$$\nabla_\theta L(\theta_x, \theta_y, u; \rho) = \left[\frac{\partial L(\theta_x, \theta_y, u; \rho)}{\partial \theta_x}, \frac{\partial L(\theta_x, \theta_y, u; \rho)}{\partial \theta_y} \right]$$

Let us solve for θ_x:

$$\frac{\partial L(\theta_x, \theta_y, u; \rho)}{\partial \theta_x} = \frac{\partial \left[f(\theta_x, \theta_y) + \mathscr{L}(\theta_x, \theta_y, u) + \mathscr{A}(\theta_x, \theta_y; \rho) \right]}{\partial \theta_x} = 0$$

$$\frac{\partial L(\theta_x, \theta_y, u; \rho)}{\partial \theta_x} = \frac{\partial \left[\frac{1}{2} \sum_{i=1}^N (y_i - Tr(X_i \theta_x))^2 + \lambda ||\theta_y||_* + u^T(\theta_x - \theta_y) + \frac{\rho}{2} ||\theta_x - \theta_y||_2^2 \right]}{\partial \theta_x} = 0$$

$$\sum_{i=1}^N -X_i^T y_i + \sum_{i=1}^N X_i^T Tr(X_i \theta_x) + u + \rho \theta_x - \rho \theta_y = 0$$

We can see that this cannot be further reduced. Let us solve for θ_y:

$$\frac{\partial L(\theta_x, \theta_y, u; \rho)}{\partial \theta_y} = \frac{\partial \left[f(\theta_x, \theta_y) + \mathscr{L}(\theta_x, \theta_y, u) + \mathscr{A}(\theta_x, \theta_y; \rho) \right]}{\partial \theta_y} = 0$$

$$\frac{\partial L(\theta_x, \theta_y, u; \rho)}{\partial \theta_y} = \frac{\partial \left[\frac{1}{2} \sum_{i=1}^N (y_i - Tr(X_i \theta_x))^2 + \lambda ||\theta_y||_* + u^T(\theta_x - \theta_y) + \frac{\rho}{2} ||\theta_x - \theta_y||_2^2 \right]}{\partial \theta_y} = 0$$

Because the nuclear-norm, $\lambda ||\theta_y||_*$, is non-differentiable, we must perform soft-thresholding:

$$-u - \rho \theta_x + \rho \theta_y + \lambda \left(\frac{\partial ||\theta_y||_*}{\partial \theta_y} \right) = 0$$

This can be rewritten by separating the non-differentiable term from the differentiable terms:

$$-u - \rho \theta_x + \rho \theta_y \pm \lambda = 0$$

$$\alpha = \frac{u}{\rho} + \theta_x$$

$$\gamma = \frac{\lambda}{\rho}$$

$$\text{where } \theta_y = \alpha \pm \gamma$$

Re-defining θ_y to utilize soft-thresholding, we get:

$$\theta_y = S_\gamma(\alpha) = \begin{cases} \alpha - \gamma & \alpha \geq \gamma \\ 0 & -\gamma \leq \alpha < \gamma \\ \alpha + \gamma & \alpha < -\gamma \end{cases}$$

Alternating Direction Method of Multipliers

```
Initialize $\theta_{x0}, \theta_{y0}, u_0 \in \mathbf{R}^{n \times n}$
Initialize $\lambda, \rho \in \mathbf{R}^1$ and $\rho > 0$
for k = 1 : K do
    Solve for $\theta_{x(k+1)}$ given:
        $\left\{ \sum_{i=1}^{N} -X_i^T y_i + \sum_{i=1}^{N} X_i^T Tr(X_i \theta_{x(k+1)}) + u_k + \rho\theta_{x(k+1)} - \rho\theta_{yk} = 0 \right\}$
    Compute $\alpha = \frac{u_k}{\rho} + \theta_{x(k+1)}$
    Compute $\gamma = \frac{\lambda}{\rho}$
    Compute $\theta_{y(k+1)} = S_\gamma(\alpha)$
    Compute $u_{k+1} = u_k + \rho(\theta_x - \theta_y)$
end for
return $\theta_{xk}, \theta_{yk}$
```