Alternating Direction Method of Multipliers II

Department of Statistics

November 23, 2023

University of Seoul

Introduction

ADMM for non-convex problems

- Focusing on cases in which the individual steps (x-update, z-update) can be carried out exactly.
- Even in this case, ADMM need not converge (when it does converge, it need not converge to an optimal point).
- ADMM converges to different points, depending on the initial values x^0, z^0, y^0 and the parameter ρ.

Definition 1 (Bi-convex problem)

$$\begin{aligned} \min \quad & F(x, z) \\ \text{subject to} \quad & G(x, z) = 0 \end{aligned} \tag{1}$$

where

- $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is bi-convex
  (convex in x for each fixed z and convex in z for each fixed x).
- $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ is bi-affine
  (affine in x for each fixed z and affine in z for each fixed x).

Scaled ADMM form

$$
\begin{aligned}
x^{(k+1)} &:= \arg\min_x \left(F(x, z^{(k)}) + (\rho/2)\|G(x, z^{(k)}) + u^{(k)}\|_2^2 \right) \\
z^{(k+1)} &:= \arg\min_z \left(F(x^{(k+1)}, z) + (\rho/2)\|G(x^{(k+1)}, z) + u^{(k)}\|_2^2 \right) \\
u^{(k+1)} &:= u^{(k)} + G(x^{(k+1)}, z^{(k+1)})
\end{aligned}
$$

- Both the x-updates and z-updates involve convex optimization problems and are tractable.

Definition 2 (Nonnegative Matrix Factorization)

$$\begin{aligned} \min \quad & (1/2)\|X - WV\|_F^2 \\ \text{subject to} \quad & W_{ij} \geq 0, \quad V_{ij} \geq 0 \end{aligned} \tag{2}$$

where the variables $W \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{r \times p}$ and data $X \in \mathbb{R}^{n \times p}$. The objective is bi-convex, and the problem is bi-convex.

What is NMF?

- The analysis method of high-dimensional data as it automatically extracts **sparse** and **interpretable features**.

$$\begin{aligned} \min \quad & f(x_1, x_2) \\ \text{subjec to} \quad & x_2 \geq 0 \end{aligned}$$

- The method of matrix factorization **with element-wise nonnegative constraints**.

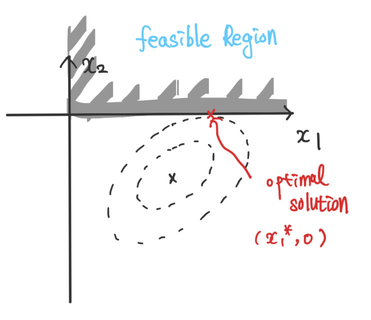

Figure 1: Sparsity obtained from a positivity constraint

Example 3 (Source Appointment Method)

There are n observatories measuring air pollutants. The air pollutants comprise p chemical species, and there are r sources of pollutant emissions."

- $x_i^\top = (x_{i1}, \cdots, x_{ip})$ for $i = 1, \cdots, n$ where x_{ik} is the amount of the kth measured chemical at the ith observatory.
- $v_k^\top = (v_{k1}, \cdots, v_{kp})$ is the (positive valued) chemical profile of the source k.
- $w_i^\top = (w_{i1}, \cdots, w_{ir})$ for $i = 1, \cdots, n$ is the (positive valued) source contribution vector of the ith observatory. w_{ik} denotes the contribution of the source k to the air pollution of the ith observatory.

We assume that

$$x_{ij} = \sum_{k=1}^{r} w_{ik} v_{kj} + \epsilon_{ij},$$

where ϵ_{ij} is an error-variable.

It is written by

$$X = WV + E$$

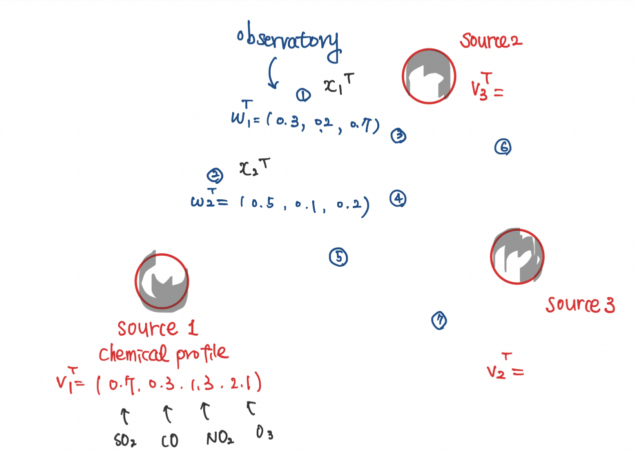

Figure 2: Source appointment methods

Example 4 (Representation learning for image data)

- $x_i^\top = (x_{i1}, \cdots, x_{ip})$ is the ith image consisting of p pixels and $X \in \mathbb{R}_+^{n \times p}$ is the dataset of n images.
- $v_k^\top = (v_{k1}, \cdots, v_{kp})$ is the feature vector representing the kth specific pattern and $V \in \mathbb{R}_+^{r \times p}$ is a feature matrix. V is called a filter bank consisting of r filters.
- $w_i^\top = (w_{i1}, \cdots, w_{ir})$ is the encoding vector of the ith image and $W \in \mathbb{R}^{n \times r}$ is a encoding matrix.
- NMF learns how to combine parts to form a whole (**a parts-based sparse representation**).

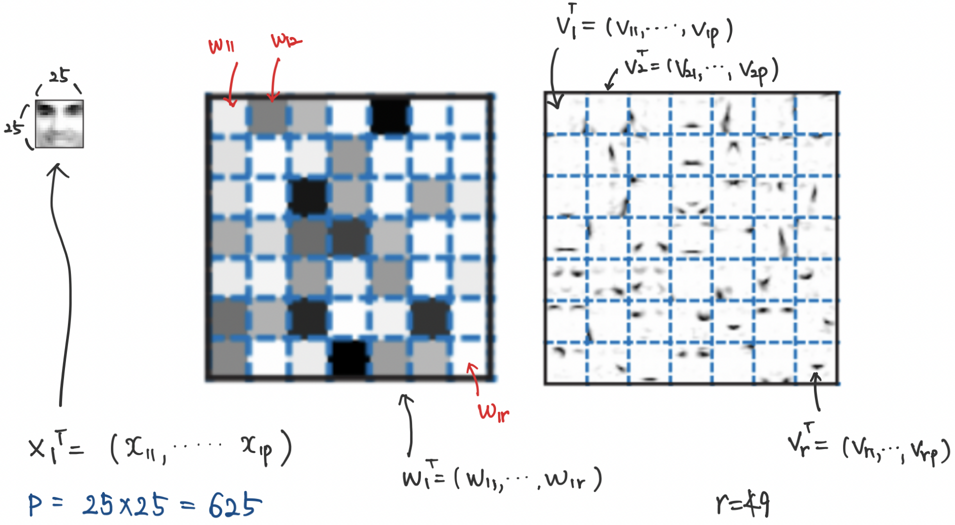

Figure 3: NMF learns a parts-based representation of faces

Example 5 (Application in NLP)

- $X \in \mathbb{R}_+^{n\times p}$ is a document matrix whose each row vector denotes the document represented by p-word frequency.
- $V \in \mathbb{R}_+^{r\times p}$ is a topic matrix whose each row vector denotes the topic (semantic feature) represented by p-word frequency.
- $W \in \mathbb{R}_+ n \times r$ is considered as 'topics' proportion matrix.

Solving NMF by Scaled ADMM

$$\begin{array}{ll} \min\limits_{B,W,V} & (1/2)\|X - B\|_F^2 + I_+(V) + I_+(W) \\ \text{subject to} & B - WV = 0 \end{array}$$

We introduced a new variable X and the indicator function I_+ for element-wise nonnegative matrices.

$$I_+(V) = \begin{cases} 0 & \text{all elements of } V \text{ is non-negative} \\ \infty & \text{otherwise} \end{cases}$$

$$
\begin{aligned}
(B^{k+1}, V^{k+1}) &:= \arg\min_{B,V\geq 0} \left(\|X - B\|_F^2 + (\rho/2)\|B - W^k V + U^k\|_F^2 \right) \\
W^{k+1} &:= \arg\min_{W\geq 0} \|B^{k+1} - W V^{k+1} + U^k\|_F^2 \\
U^{k+1} &:= U^k + B^{k+1} - W^{k+1} V^{k+1}
\end{aligned}
$$

Note that we use the Frobenius norm instead of the L_2-norm.

- We know that $\|B\|_F^2 = \sum_{i=1}^{p} \|b_i\|_2^2$ where $X = [b_1, \cdots, b_p]$.
- Using this, we can split the first update step **across the rows** of B and V, and it can be performed by solving a set of quadratic programs in parallel.

$$(b_i^{k+1}, v_i^{k+1}) = \mathrm{argmin}_{b_i, v_i \geq 0} \left(\|x_i - b_i\|_2^2 + (\rho/2)\|b_i - W^{k\top} v_i + u_i^k\|_2^2 \right)$$

for $i = 1, \cdots, p$.

- In the same way, we can split the second update into the columns of W (quadratic programs):

$$w_j^{k+1} := \mathsf{argmin}_{w_j \geq 0} \|b_j^{k+1} - w_j V^{k+1} + u_j^k\|_2^2$$

for $j = 1, \cdots, r$.

Supplementary Note

1. Standard ADMM
2. Augmented ADMM
3. Example(Sparse Fused Lasso)

When we use ADMM algorithm?

We aim to solve the optimization problem of the following form

$$\min_{\theta \in \mathbb{R}^p} f(\theta) + g(A\theta), \tag{3}$$

where f and g are convex functions and $A \in \mathbb{R}^{m \times p}$.

ADMM algorithm can solve convex problems with constraints such as (3) stably but slowly.

Using auxiliary variable γ, ADMM form of problem (3)

$$\begin{aligned} \min_{\theta\in\mathbb{R}^p,\gamma\in\mathbb{R}^m} \quad & f(\theta)+g(\gamma), \\ \text{subject to} \quad & A\theta-\gamma=0 \end{aligned} \tag{4}$$

Updating rules of problem (4)

$$\begin{aligned} \theta^{k+1} &:= \operatorname*{argmin}_{\theta}\left(f(\theta)+\frac{\rho}{2}\|A\theta-\gamma^k+\rho^{-1}\alpha^k\|_2^2\right), \\ \gamma^{k+1} &:= \operatorname*{argmin}_{\gamma}\left(g(\gamma)+\frac{\rho}{2}\|A\theta^{k+1}-\gamma^k+\rho^{-1}\alpha^k\|_2^2\right), \\ \alpha^{k+1} &:= \alpha^k+\rho(A\theta^{k+1}-\gamma^{k+1}), \end{aligned} \tag{5}$$

where α is a dual variable.

In chapter *general patterns* of ADMM, we investigated the quadratic objective function f

$$f(x) = (1/2)x^\top Px + q^\top x + r,$$

and the efficient methods of computing inverse matrix in x-update.

For instance, f is quadratic term of θ and P and A are diagonal matrix, computing cost of $\left(P + \rho A^\top A\right)^{-1}$ is $O(p)$ by comparison with $O(p^3)$ which is general cost of inverse matrix in x-update.

In general case (4), matrix A **has a lot of influence on convergence time**.

Issue

- Many well-known problems like *generalized lasso* can be written in the same form of (4).
- Unless A is not sparse, computing cost is too expensive in θ-update of a high-dimensional problem$(p \gg n)$.

How can we get around this difficulty?

Augmented ADMM

We consider "augmented" variable $(\gamma, \tilde{\gamma})$ and rewrite problem (4)

$$\begin{aligned} \min_{\theta, \gamma \in \mathbb{R}^m, \tilde{\gamma} \in \mathbb{R}^p} \quad & f(\theta) + g(\gamma) \\ \text{subject to} \quad & \begin{pmatrix} A \\ (D - A^\top A)^{1/2} \end{pmatrix} \theta - \begin{pmatrix} \gamma \\ \tilde{\gamma} \end{pmatrix} = 0, \end{aligned} \tag{6}$$

where $D \in \mathbb{R}^{p \times p}$ satisfies $D \succeq A^\top A$.

Note that the augmented variable $\tilde{\gamma}$ and associated constraintally redundant.

Apply standard ADMM to (6), updating rules are

$$
\begin{aligned}
\theta^{k+1} &:= \operatorname*{argmin}_{\theta} \ f(\theta) + \frac{\rho}{2}\|A\theta - \gamma^k + \rho^{-1}\alpha^k\|_2^2 &(7)\\
&\quad +\|(D - A^\top A)^{1/2}\theta - \tilde{\gamma}^k + \rho^{-1}\tilde{\alpha}^k\|_2^2, \\
\gamma^{k+1} &:= \operatorname*{argmin}_{\gamma} \left(g(\gamma) + \frac{\rho}{2}\|A\theta^{k+1} - \gamma + \rho^{-1}\tilde{\alpha}^k\|_2^2 \right), &(8)\\
\tilde{\gamma}^{k+1} &:= (D - A^\top A)^{1/2}\theta^{k+1} + \rho^{-1}\tilde{\alpha}^k &(9)\\
\alpha^{k+1} &:= \alpha^k + \rho(A\theta^{k+1} - \gamma^{k+1}) &(10)\\
\tilde{\alpha}^{k+1} &:= \tilde{\alpha}^k + \rho\left((D - A^\top A)^{1/2}\theta^{k+1} - \tilde{\gamma}^{k+1}\right), &(11)
\end{aligned}
$$

where $\alpha \in \mathbb{R}^m, \tilde{\alpha} \in \mathbb{R}^p$ are dual variables.

Combining (9) and (11) gives $\tilde{\alpha}^{k+1} = 0$. Then plugging (9) into (7), θ-update will be rewritten as

$$
\begin{aligned}
\theta^{k+1} = \quad & \operatorname*{argmin}_{\theta} \; f(\theta) + \frac{\rho}{2}\|A\theta - \gamma^k + \rho^{-1}\alpha^k\|_2^2 \\
& + \|(D - A^\top A)^{1/2}(\theta - \theta^k)\|_2^2.
\end{aligned}
$$

This result cancels out $\theta^\top A^\top A\theta$ in θ-update.

$$
\begin{aligned}
\theta^{k+1} &:= \underset{\theta}{\operatorname{argmin}} \left(f(\theta) + (2\alpha^k - \alpha^{k-1})^\top A\theta + \frac{\rho}{2}(\theta - \theta^k)^\top D(\theta - \theta^k) \right), \\
\gamma^{k+1} &:= \underset{\gamma}{\operatorname{argmin}} \left(g(\gamma) + \frac{\rho}{2}\|A\theta^{k+1} - \gamma + \rho^{-1}\alpha^k\|_2^2 \right), \\
\alpha^{k+1} &:= \alpha^k + \rho(A\theta^{k+1} - \gamma^{k+1})
\end{aligned}
$$

Note that

- In θ-update, we compute inverse matrix of D instead of $A^\top A$
- Updating rules don't involve the augmented $\tilde{\gamma}$ and $\tilde{\alpha}$ at all!

Theorem 1
Under Standard ADMM assumption, for any matrix $D \in \mathbb{R}^{p\times p}$ satisfying $D \succeq A^\top A$ and any positive scalar $\rho > 0$, the following update

$$\begin{aligned}
\theta^{k+1} &:= \operatorname*{argmin}_{\theta} f(\theta) + (2\alpha^k - \alpha^{k-1})^\top A\theta \\
&\quad + \frac{\rho}{2}(\theta - \theta^k)^\top D(\theta - \theta^k), \\
\alpha^{k+1} &:= \alpha^k + \rho(A\theta^{k+1} - \gamma^{k+1})
\end{aligned}$$

converges in the sense that primal objective functions along the sequence of primal variables and dual variable converge to the optimal value: $f(\theta) + g(A\theta^k) \to \inf_\theta f(\theta) + g(A\theta)$ and $\alpha \to \alpha^\star$.

Which D should we choose?

D satisfies $D \succeq A^\top A$. For a simple choice would be $D = \delta I$ with $\delta \geq \|A\|_{op}^2$ where $\|A\|_{op}^2$ denotes the operator norm of A.

Operator norm

Given two normed vector spaces V and W, linear map $A : V \to W$ and operator norm is

$$\begin{aligned} \|A\|_{op} &:= \inf\{c \geq 0 \,|\, \|Av\| \leq c\|v\| \text{ for all } v \in V\} \\ &:= \sup\left\{\frac{\|Av\|}{\|v\|} : v \in V \text{ and } v \neq 0\right\} \end{aligned}$$

Well-known lemma

$$\left(\sigma_{\text{max}} I - A^\top A\right) \text{ is a positive semi-definite matrix,}$$

where σ_{max} is maximum singular value of $A^\top A$.

Therefore we can choose σ_{max} for δ.

Example 6 (Sparse fused lasso over a graph)

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. Often, the node set $\mathcal{V}$ represents the features in the model, and the edge set $\mathcal{E}$ represents their relationship.

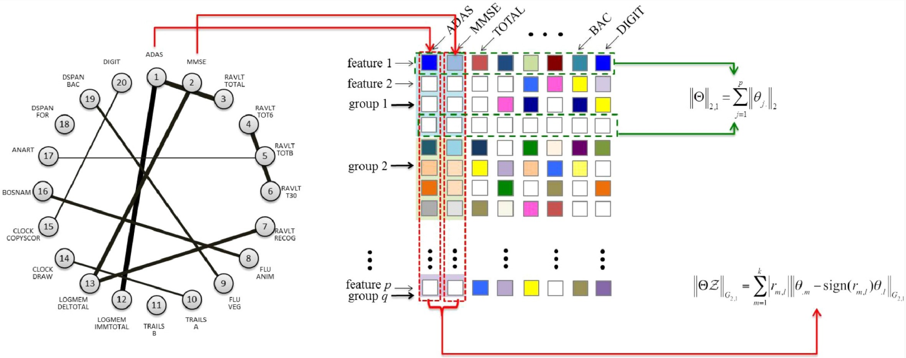

Figure 4: Genetic Graph

Based on such a graph, we consider the following optimization problem

$$\min_{\beta \in \mathbb{R}^p} \underbrace{(1/2)\|y - X\beta\|_2^2}_{=f(\beta)} + \underbrace{\lambda_1 \|\beta\|_1 + \lambda_2 \sum_{(i,j)\in\mathcal{E}} |\beta_i - \beta_j|}_{=g(\beta)}, \tag{12}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is a data matrix.

This regularization term g desires the structure where β_i and β_j have a similar or same value in $(i,j) \in \mathcal{E}$ and makes β sparse.

Write (12) in the form of ADMM with $A = \begin{bmatrix} I \\ C \end{bmatrix}$ and $g(\gamma) = \lambda_1|\gamma_1| + \lambda_2|\gamma_2|$, where C is matrix associated with graph $\mathcal{G}$ and $\gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$.

The constraints are $\beta = \gamma_1$ and $C\beta = \gamma_2$.

For example, assume that $p = 3$ and there is a connection between the first and second features. Then

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}, \gamma_1 = \beta \text{ and } \gamma_2 = \beta_1 - \beta_2$$

The augmented ADMM gives the following updates

$$
\begin{aligned}
\beta^{k+1} &:= (\rho D + X^\top X)^{-1}(\rho D \beta^k + X^\top y - A^\top(2\alpha^k - \alpha^{k-1})) \\
\alpha^{k+1} &:= \alpha^k + \rho(A\beta^{k+1} - \gamma^{k+1})
\end{aligned}
$$

where $\alpha = (\alpha_1^\top \alpha_2^\top)^\top \in \mathbb{R}^{p+m}$ is dual variable.

	$p \leq n$	$p > n$
stanADMM	$O(N_{\text{chol}} p^2 n + N_{\text{admm}} p^2)$	$O(N_{\text{chol}} p^3 + N_{\text{admm}} p^2)$
augADMM	$O(N_{\text{chol}} p^2 n + N_{\text{admm}} p^2)$	$O(N_{\text{chol}} n^2 p + N_{\text{admm}} [pn \vee m])$

Table 1: Computational complexity

When $p > n$, the augmented ADMM gains computation efficiency, which is linear in p(if $m < np$).

Summary

- The matrix A has a lot of influence on convergence time in ADMM algorithm.
- By using augmented ADMM, we can gain huge computational efficiency.