Part 5. Dual-based Methods

Math 126 Winter 18

Date of current version: February 25, 2018

Abstract This note studies dual-based methods such as dual subgradient method, dual proximal gradient method, augmented Lagrangian method and alternating direction method of multipliers (ADMM). Many parts of this note are based on the chapters [1, Chapter 8,12,15].

Please email me if you find any typos or errors.

1 Dual Projected Subgradient Methods (see [1, Chapter 8])

Consider the problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to } \boldsymbol{g}(\boldsymbol{x}) \preceq \boldsymbol{0}, \tag{1.1}$$

where we assume that

- $f : \mathbb{R}^d \to \mathbb{R}$ is convex.
- $\boldsymbol{g}(\cdot) = (g_1(\cdot), \ldots, g_m(\cdot))^\top$, where $g_1, \ldots, g_m : \mathbb{R}^d \to \mathbb{R}$ are convex.
- The problem has finite optimal value, and the optimal set, denoted by $\mathcal{X}_*$, is nonempty.
- There exists $\hat{\boldsymbol{x}}$ for which $\boldsymbol{g}(\hat{\boldsymbol{x}}) \prec \boldsymbol{0}$.
- For any $\boldsymbol{\lambda} \in \mathbb{R}^m_+$, the problem $\min_{\boldsymbol{x}}\{f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\}$ has an optimal solution.

The Lagrange dual function is

$$q(\boldsymbol{\lambda}) = \inf_{\boldsymbol{x}} \left\{\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \equiv f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\right\}, \tag{1.2}$$

and the dual problem is

$$\max_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} q(\boldsymbol{\lambda}). \tag{1.3}$$

For a given $\boldsymbol{\lambda} \in \mathbb{R}^m_+$, suppose that the minimum in the minimization problem defining $q(\boldsymbol{\lambda})$ in (1.2) is attained at $\boldsymbol{x}_{\boldsymbol{\lambda}}$, *i.e.*,

$$\mathcal{L}(\boldsymbol{x}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) = f(\boldsymbol{x}_{\boldsymbol{\lambda}}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_{\boldsymbol{\lambda}}) = q(\boldsymbol{\lambda}). \tag{1.4}$$

We seek to find a subgradient of the convex function $-q$ at $\boldsymbol{\lambda}$. For any $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^m_+$, we have

$$\begin{aligned} q(\bar{\boldsymbol{\lambda}}) &= \min_{\boldsymbol{x}} \left\{f(\boldsymbol{x}) + \bar{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\boldsymbol{x})\right\} \\ &\le f(\boldsymbol{x}_{\boldsymbol{\lambda}}) + \bar{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\boldsymbol{x}_{\boldsymbol{\lambda}}) \\ &= f(\boldsymbol{x}_{\boldsymbol{\lambda}}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_{\boldsymbol{\lambda}}) + (\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda})^\top \boldsymbol{g}(\boldsymbol{x}_{\boldsymbol{\lambda}}) \\ &= q(\boldsymbol{\lambda}) + \boldsymbol{g}(\boldsymbol{x}_{\boldsymbol{\lambda}})^\top (\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}), \end{aligned} \tag{1.5}$$

Donghwan Kim
Dartmouth College
E-mail: donghwan.kim@dartmouth.edu

which leads to

$$-q(\bar{\boldsymbol{\lambda}}) \geq -q(\boldsymbol{\lambda}) + (-\boldsymbol{g}(\boldsymbol{x}))^\top (\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}). \tag{1.6}$$

This show that

$$-\boldsymbol{g}(\boldsymbol{x}_{\boldsymbol{\lambda}}) \in \partial(-q)(\boldsymbol{\lambda}). \tag{1.7}$$

Algorithm 1 Dual Projected Subgradient (Ascent) Method

1: **Input:** $\boldsymbol{\lambda}_0 \in \mathbb{R}^m_+$.
2: **for** $k \geq 0$ **do**
3: Compute $\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}} \{\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_k) \equiv f(\boldsymbol{x}) + \boldsymbol{\lambda}_k^\top \boldsymbol{g}(\boldsymbol{x})\}$.
4: Choose a step size $s_k > 0$.
5: $\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + s_k \boldsymbol{g}(\boldsymbol{x}_k)]_+$.
6: If a stopping criteria is satisfied, then stop.

Example 1.1 Dual decomposition. Consider the problem

$$\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) \equiv \sum_{i=1}^{B} f_i(\boldsymbol{x}_i) \right\}$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b},$$

where f is (block-)separable, $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_B^\top)^\top$, and $\boldsymbol{A} = (\boldsymbol{A}_1^\top, \dots, \boldsymbol{A}_B^\top)^\top$. Then the Lagrangian $\mathcal{L}(\cdot)$ is also (block-)separable in $\boldsymbol{x}$:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{B} \mathcal{L}_i(\boldsymbol{x}_i, \boldsymbol{\lambda}),$$

where $\mathcal{L}_i(\boldsymbol{x}_i, \boldsymbol{\lambda}) = f_i(\boldsymbol{x}_i) + \boldsymbol{\lambda}^\top (\boldsymbol{A}_i \boldsymbol{x}_i - \boldsymbol{b})$. Thus, at each kth iteration of the dual subgradient method, the $\boldsymbol{x}$-minimization splits into B separate minimizations as

$$[\boldsymbol{x}_k]_i \in \arg\min_{\boldsymbol{x}_i} \mathcal{L}_i(\boldsymbol{x}_i, \boldsymbol{\lambda}_k), \quad i = 1, \dots, B.$$

Convergence analysis of the dual objective function sequence $\{q(\boldsymbol{\lambda}_k)\}_{k\geq 0}$ under various choice of $\{s_k\}_{k\geq 0}$ is already known since the convergence analysis of the primal objective function sequence directly applies here. How about the convergence analysis of a primal sequence? For the primal case, we have to consider a primal sequence other than the sequence $\{\boldsymbol{x}_k\}_{k\geq 0}$:

– full averaging sequence:

$$\bar{\boldsymbol{x}}_k = \sum_{i=0}^{k} \eta_{k,i} \boldsymbol{x}_i, \text{ where } \eta_{k,i} = \frac{s_i}{\sum_{l=0}^{k} s_l}, \; i = 0, \dots, k. \tag{1.8}$$

– partial averaging sequence:

$$\tilde{\boldsymbol{x}}_k = \sum_{i=\lceil k/2 \rceil}^{k} \eta_{k,i} \boldsymbol{x}_i, \text{ where } \eta_{k,i} = \frac{s_i}{\sum_{l=\lceil k/2 \rceil}^{k} s_l}, \; i = 0, \dots, k. \tag{1.9}$$

Lemma 1.1 *Assume that there exists $M > 0$ such that $||\boldsymbol{g}(\boldsymbol{x})||_2 \leq M$ for any $\boldsymbol{x}$. Let $\rho > 0$ be some positive number, and let $\{\boldsymbol{x}_k\}_{k\geq 0}$ and $\{\boldsymbol{\lambda}_k\}_{k\geq 0}$ be the sequences generated by the dual projected subgradient methods with step size $s_k = \frac{\gamma_k}{||g(\boldsymbol{x}_k)||_2}$. Then for any $k \geq 2$,*

$$f(\bar{\boldsymbol{x}}_k) - f(\boldsymbol{x}_*) + \rho ||[\boldsymbol{g}(\bar{\boldsymbol{x}}_k)]_+||_2 \leq \frac{M}{2} \frac{(||\boldsymbol{\lambda}_0||_2 + \rho)^2 + \sum_{i=0}^{k} \gamma_i^2}{\sum_{i=0}^{k} \gamma_i} \tag{1.10}$$

and

$$f(\tilde{\boldsymbol{x}}_k) - f(\boldsymbol{x}_*) + \rho ||[\boldsymbol{g}(\tilde{\boldsymbol{x}}_k)]_+||_2 \leq \frac{M}{2} \frac{(||\boldsymbol{\lambda}_{\lceil k/2 \rceil}||_2 + \rho)^2 + \sum_{i=\lceil k/2 \rceil}^{k} \gamma_i^2}{\sum_{i=\lceil k/2 \rceil}^{k} \gamma_i}. \tag{1.11}$$

2 Dual Proximal Gradient Methods (see [1, Chapter 12])

Consider the problem

$$\min_{\boldsymbol{x}}\{f(\boldsymbol{x})+\phi(\boldsymbol{A}\boldsymbol{x})\}, \tag{2.1}$$

where we assume that

- $f : \mathbb{R}^d \to (-\infty,\infty]$ is proper closed and σ-strongly convex ($\sigma > 0$).
- $\phi : \mathbb{R}^p \to (-\infty,\infty]$ is proper closed and convex.
- $\boldsymbol{A} \in \mathbb{R}^{p\times d}$ is a matrix.
- There exists $\hat{\boldsymbol{x}} \in \operatorname{relint}(\operatorname{dom} f)$ and $\hat{\boldsymbol{z}} \in \operatorname{relint}(\operatorname{dom}\phi)$ such that $\boldsymbol{A}\hat{\boldsymbol{x}} = \hat{\boldsymbol{z}}$.

We first reformulate the problem as

$$\min_{\boldsymbol{x},\boldsymbol{z}}\{f(\boldsymbol{x})+\phi(\boldsymbol{z})\} \tag{2.2}$$

$$\text{subject to } \boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}=\boldsymbol{0}. \tag{2.3}$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\mu}) = f(\boldsymbol{x})+\phi(\boldsymbol{z})-\langle\boldsymbol{\mu},\,\boldsymbol{A}\boldsymbol{x}-\boldsymbol{z}\rangle = f(\boldsymbol{x})+\phi(\boldsymbol{z})-\langle\boldsymbol{A}^\top\boldsymbol{\mu},\,\boldsymbol{x}\rangle+\langle\boldsymbol{\mu},\,\boldsymbol{z}\rangle, \tag{2.4}$$

and the dual function is

$$\begin{aligned} q(\boldsymbol{\mu}) &= \inf_{\boldsymbol{x},\boldsymbol{z}}\left\{f(\boldsymbol{x})+\phi(\boldsymbol{z})-\langle\boldsymbol{A}^\top\boldsymbol{\mu},\,\boldsymbol{x}\rangle+\langle\boldsymbol{\mu},\,\boldsymbol{z}\rangle\right\} \\ &= -\sup_{\boldsymbol{x},\boldsymbol{z}}\left\{-f(\boldsymbol{x})-\phi(\boldsymbol{z})+\langle\boldsymbol{A}^\top\boldsymbol{\mu},\,\boldsymbol{x}\rangle+\langle-\boldsymbol{\mu},\,\boldsymbol{z}\rangle\right\}. \end{aligned} \tag{2.5}$$

Then, the dual problem is

$$\max_{\boldsymbol{\mu}}\left\{q(\boldsymbol{\mu}) \equiv -f^*(\boldsymbol{A}^\top\boldsymbol{\mu})-\phi^*(-\boldsymbol{\mu})\right\}. \tag{2.6}$$

We reformulate the dual problem in its minimization form:

$$\min_{\boldsymbol{\mu}}\{F(\boldsymbol{\mu})+\Phi(\boldsymbol{\mu})\} \tag{2.7}$$

where

$$F(\boldsymbol{\mu}) = f^*(\boldsymbol{A}^\top\boldsymbol{\mu}), \quad \Phi(\boldsymbol{\mu}) = \phi^*(-\boldsymbol{\mu}). \tag{2.8}$$

Theorem 2.1 *Let $\sigma > 0$. Then*

- *If $f : \mathbb{R}^d \to \mathbb{R}$ is a $\frac{1}{\sigma}$-smooth convex function, then f^* is σ-strongly convex.*
- *If $f : \mathbb{R}^d \to (-\infty,\infty]$ is a proper closed σ-strongly convex function, then f^* is $\frac{1}{\sigma}$-smooth.*

Lemma 2.1 *(properties of F and Φ)*

(a) F is convex and L_F-smooth with $L_F = \frac{||\boldsymbol{A}||_2^2}{\sigma}$.
(b) Φ is proper closed and convex.

Proof (a) Since f is proper closed and σ-strongly convex, f^* is $\frac{1}{\sigma}$-smooth. Therefore for any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, we have

$$\begin{aligned} ||\nabla F(\boldsymbol{\mu}_1)-\nabla F(\boldsymbol{\mu}_2)||_2 &= ||\boldsymbol{A}(\nabla f^*(\boldsymbol{A}^\top\boldsymbol{\mu}_1))-\boldsymbol{A}(\nabla f^*(\boldsymbol{A}^\top\boldsymbol{\mu}_2))||_2 \\ &\le ||\boldsymbol{A}||_2||\nabla f^*(\boldsymbol{A}^\top\boldsymbol{\mu}_1)-\nabla f^*(\boldsymbol{A}^\top\boldsymbol{\mu}_2)||_2 \\ &\le ||\boldsymbol{A}||_2\frac{1}{\sigma}||\boldsymbol{A}^\top\boldsymbol{\mu}_1-\boldsymbol{A}^\top\boldsymbol{\mu}_2||_2 \\ &\le \frac{1}{\sigma}||\boldsymbol{A}||_2||\boldsymbol{A}^\top||_2||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_2 \\ &= \frac{||\boldsymbol{A}||_2^2}{\sigma}||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_2. \end{aligned}$$

F is convex since F is a composition of a convex function f^* and a linear mapping.

(b) Since ϕ is proper closed and convex, so is ϕ^*. Thus $\Phi(\boldsymbol{\mu}) \equiv \phi^*(-\boldsymbol{\mu})$ is also proper closed and convex.

If f is also L_f-smooth, then we can use the proximal gradient method as below.

Algorithm 2 Primal Proximal Gradient Method

1: **Input:** $\boldsymbol{x}_0$ and $L \geq L_f$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} = \operatorname{prox}_{\frac{1}{L}\phi(\boldsymbol{A}\cdot)}\left(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)\right)$.

This might not be efficient since the proximal operator $\operatorname{prox}_{\frac{1}{L}\phi(\boldsymbol{A}\cdot)}$ may not have a closed-form expression due to $\boldsymbol{A}$ (unless it is a diagonal matrix). Let's see how we can circumvent such issue using the Lagrange dual (even in the case where we do not need the L_f-smoothness assumption on f).

We can easily use the proximal gradient method to solve the dual problem as below.

Algorithm 3 Dual Proximal Gradient Method (dual representation)

1: **Input:** $\boldsymbol{\mu}_0$ and $L \geq L_F = \frac{||\boldsymbol{A}||_2^2}{\sigma}$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{\mu}_{k+1} = \operatorname{prox}_{\frac{1}{L}\Phi}\left(\boldsymbol{\mu}_k - \frac{1}{L}\nabla F(\boldsymbol{\mu}_k)\right)$.

Theorem 2.2 *Let $\{\boldsymbol{\mu}_k\}_{k\geq 0}$ be the sequence generated by the dual proximal gradient method. Then, for any optimal solution $\boldsymbol{\mu}_*$ of the dual problem (2.7), we have*

$$q(\boldsymbol{\mu}_*) - q(\boldsymbol{\mu}_k) \leq \frac{L||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_*||_2^2}{2k}. \tag{2.9}$$

We next find a primal representation of the dual proximal gradient method, which is written in a more explicit way in terms of $f, \phi, \boldsymbol{A}$. (proof omitted)

Algorithm 4 Dual Proximal Gradient Method (primal representation)

1: **Input:** $\boldsymbol{\mu}_0$ and $L \geq L_F = \frac{||\boldsymbol{A}||_2^2}{\sigma}$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_k = \arg\max_{\boldsymbol{x}}\left\{\langle \boldsymbol{x},\, \boldsymbol{A}^\top \boldsymbol{\mu}_k\rangle - f(\boldsymbol{x})\right\}$.
4: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \frac{1}{L}\boldsymbol{A}\boldsymbol{x}_k + \frac{1}{L}\operatorname{prox}_{L\phi}\left(\boldsymbol{A}\boldsymbol{x}_k - L\boldsymbol{\mu}_k\right)$.

Remark 2.1 The sequence $\{\boldsymbol{x}_k\}_{k\geq 0}$ generated by the method will be called the primal sequence. The elements of the sequence are actually not necessarily feasible with respect to the primal problem (2.1) since they are not guaranteed to belong to $\operatorname{dom}\phi(\boldsymbol{A}\cdot)$. Nevertheless, the primal sequence converge to the optimal solution $\boldsymbol{x}_*$.

Lemma 2.2 *(primal-dual relation) Let $\bar{\boldsymbol{\mu}} \in \operatorname{dom}\Phi$, and let*

$$\bar{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}}\left\{\langle \boldsymbol{x},\, \boldsymbol{A}^\top \bar{\boldsymbol{\mu}}\rangle - f(\boldsymbol{x})\right\}. \tag{2.10}$$

Then,

$$||\bar{\boldsymbol{x}} - \boldsymbol{x}_*||_2^2 \leq \frac{2}{\sigma}(q(\boldsymbol{\mu}_*) - q(\bar{\boldsymbol{\mu}})). \tag{2.11}$$

Theorem 2.3 *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ and $\{\boldsymbol{\mu}_k\}_{k\geq 0}$ be the primal and dual sequences generated by the dual proximal gradient method. Then, for any optimal solution $\boldsymbol{\mu}_*$ of the dual problem (2.7), we have*

$$||\boldsymbol{x}_k - \boldsymbol{x}_*||_2^2 \leq \frac{L||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_*||_2^2}{\sigma k}. \tag{2.12}$$

We can accelerate dual proximal gradient method by using the fast dual proximal gradient method as below.

Algorithm 5 Fast Dual Proximal Gradient Method (dual representation)

1: **Input:** $\boldsymbol{\mu}_0 = \boldsymbol{\eta}_0$ and $L \geq L_F = \frac{||\boldsymbol{A}||_2^2}{\sigma}$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{\mu}_{k+1} = \text{prox}_{\frac{1}{L}\Phi}\left(\boldsymbol{\eta}_k - \frac{1}{L}\nabla F(\boldsymbol{\eta}_k)\right)$.
4: $\quad t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.
5: $\quad \boldsymbol{\eta}_{k+1} = \boldsymbol{\mu}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k)$.

Theorem 2.4 *Let $\{\boldsymbol{\mu}_k\}_{k\geq 0}$ be the sequence generated by the fast dual proximal gradient method. Then, for any optimal solution $\boldsymbol{\mu}_*$ of the dual problem (2.7), we have*

$$q(\boldsymbol{\mu}_*) - q(\boldsymbol{\mu}_k) \leq \frac{2L||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_*||_2^2}{(k+1)^2} \tag{2.13}$$

Algorithm 6 Fast Dual Proximal Gradient Method (primal representation)

1: **Input:** $\boldsymbol{\mu}_0 = \boldsymbol{\eta}_0$ and $L \geq L_F = \frac{||\boldsymbol{A}||_2^2}{\sigma}$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{u}_k = \arg\max_{\boldsymbol{x}} \left\{\langle \boldsymbol{x}, \boldsymbol{A}^\top \boldsymbol{\eta}_k\rangle - f(\boldsymbol{x})\right\}$.
4: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \frac{1}{L}\boldsymbol{A}\boldsymbol{u}_k + \frac{1}{L}\text{prox}_{L\phi}\left(\boldsymbol{A}\boldsymbol{u}_k - L\boldsymbol{\mu}_k\right)$.
5: $\quad t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.
6: $\quad \boldsymbol{\eta}_{k+1} = \boldsymbol{\mu}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k)$.

Theorem 2.5 *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ and $\{\boldsymbol{\mu}_k\}_{k\geq 0}$ be the primal and dual sequences generated by the fast dual proximal gradient method. Then, for any optimal solution $\boldsymbol{\mu}_*$ of the dual problem (2.7), we have*

$$||\boldsymbol{x}_k - \boldsymbol{x}_*||_2^2 \leq \frac{4L||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_*||_2^2}{\sigma(k+1)^2}. \tag{2.14}$$

Example 2.1 Consider the one-dimensional total variation (TV) problem:

$$\min_{\boldsymbol{x}} \left\{\frac{1}{2}||\boldsymbol{x} - \boldsymbol{b}||_2^2 + \gamma||\boldsymbol{D}\boldsymbol{x}||_1\right\},$$

where

$$\boldsymbol{D} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}.$$

Let $f(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x} - \boldsymbol{b}||_2^2$, $\phi(\boldsymbol{z}) = \gamma||\boldsymbol{z}||_1$ and $\boldsymbol{A} = \boldsymbol{D}$ in (2.1). Then, f is σ-strongly convex with $\sigma = 1$, ϕ is convex, and $||\boldsymbol{D}||_2^2 \leq 4$, so $F(\mu) = f^*(\boldsymbol{D}^\top \mu)$ is L_F-smooth with $L_F = 4$.

Algorithm 7 Dual Proximal Gradient Method for One-dimensional TV Problem (primal representation)

1: **Input:** $\boldsymbol{\mu}_0$ and $L \geq L_F = 4$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_k = \arg\max_{\boldsymbol{x}} \left\{\langle \boldsymbol{x}, \boldsymbol{D}^\top \boldsymbol{\mu}_k\rangle - \frac{1}{2}||\boldsymbol{x} - \boldsymbol{b}||_2^2\right\} = \boldsymbol{D}^\top \boldsymbol{\mu}_k + \boldsymbol{b}$.
4: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \frac{1}{L}\boldsymbol{D}\boldsymbol{x}_k + \frac{1}{L}\text{prox}_{L\phi}\left(\boldsymbol{D}\boldsymbol{x}_k - L\boldsymbol{\mu}_k\right)$.

The dual problem $\min_{\boldsymbol{\mu}} \{F(\boldsymbol{\mu}) + \Phi(\boldsymbol{\mu})\}$, where

$$F(\boldsymbol{\mu}) = f^*(\boldsymbol{D}^\top \boldsymbol{\mu}) = \frac{1}{2}||\boldsymbol{D}^\top \boldsymbol{\mu} + \boldsymbol{b}||_2^2 - \frac{1}{2}||\boldsymbol{b}||_2^2$$
$$\Phi(\boldsymbol{\mu}) = \phi^*(-\boldsymbol{\mu}) = \begin{cases} 0, & ||\boldsymbol{\mu}||_\infty \leq \gamma, \\ \infty, & \text{otherwise}, \end{cases}$$

is a constrained quadratic convex problem. The projection operator onto a norm ball $\mathcal{C} = \{\boldsymbol{\mu} \,:\, ||\boldsymbol{\mu}||_\infty \leq \gamma\}$ is $\boldsymbol{P}_{\mathcal{C}}(\boldsymbol{\mu}) = \min\{|\boldsymbol{\mu}|, \gamma \mathbf{1}\} \odot \text{sign}(\boldsymbol{\mu})$.

Algorithm 8 Dual Proximal Gradient Method for One-dimensional TV Problem (dual representation)

1: **Input:** $\boldsymbol{\mu}_0$ and $L \geq L_F = 4$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{P}_{\mathcal{C}}\left(\boldsymbol{\mu}_k - \frac{1}{L}(\boldsymbol{D}^\top \boldsymbol{\mu}_k + \boldsymbol{b})\right)$

3 Augmented Lagrangian Methods (see [1, Chapter 15])

Consider the problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad (3.1)$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where we assume that f is proper closed and convex functions.

The dual function is

$$q(\boldsymbol{\mu}) = \inf_{\boldsymbol{x}} \{\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}) \equiv f(\boldsymbol{x}) + \langle \boldsymbol{\mu},\, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\rangle\} = -f^*(-\boldsymbol{A}^\top \boldsymbol{\mu}) - \langle \boldsymbol{b},\, \boldsymbol{\mu}\rangle,$$

and the dual problem (in the minimization form) is

$$\min_{\boldsymbol{\mu}} \{f^*(-\boldsymbol{A}^\top \boldsymbol{\mu}) + \langle \boldsymbol{b},\, \boldsymbol{\mu}\rangle\}. \quad (3.2)$$

Consider the proximal point method, which has the following update at each kth iteration for given constant $\rho > 0$:

$$\begin{aligned} \boldsymbol{\mu}_{k+1} &= \operatorname{prox}_{\rho(-q)}(\boldsymbol{\mu}_k) = \arg\min_{\boldsymbol{\mu}} \left\{ \frac{1}{2}||\boldsymbol{\mu} - \boldsymbol{\mu}_k||_2^2 + \rho(-q(\boldsymbol{\mu})) \right\} \quad (3.3) \\ &= \arg\min_{\boldsymbol{\mu}} \left\{ f^*(-\boldsymbol{A}^\top \boldsymbol{\mu}) + \langle \boldsymbol{b},\, \boldsymbol{\mu}\rangle + \frac{1}{2\rho}||\boldsymbol{\mu} - \boldsymbol{\mu}_k||_2^2 \right\}. \end{aligned}$$

The sequence $\{\boldsymbol{\mu}_k\}_{k\geq 0}$ satisfies

$$q(\boldsymbol{\mu}_*) - q(\boldsymbol{\mu}_k) \leq \frac{||\boldsymbol{\mu}_0 - \boldsymbol{\mu}_*||_2^2}{2\rho k}. \quad (3.4)$$

Recall that the proximal point method has its accelerated version with rate $O\left(\frac{1}{k^2}\right)$.

The primal representation of the dual proximal point point method is known as the augmented Lagrangian method.

Algorithm 9 Augmented Lagrangian Method

1: **Input:** $\boldsymbol{\mu}_0$ and $\rho > 0$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} \in \arg\min_{\boldsymbol{x}} \left\{\mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{\mu}_k) \equiv f(\boldsymbol{x}) + \langle \boldsymbol{\mu}_k,\, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\rangle + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2^2\right\}$
4: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{b})$

Remark 3.1 This method has a rate faster than than the rate of the dual projected subgradient methods, but the additional quadratic term destroys the separability for the dual decomposition.

We next show that the augmented Lagrangian method is equivalent to the dual proximal point method. The optimality condition of the $\boldsymbol{x}_{k+1}$-update is

$$\mathbf{0} \in \partial f(\boldsymbol{x}_{k+1}) + \boldsymbol{A}^\top(\boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{b})), \quad (3.5)$$

which can be reformulated as

$$-\boldsymbol{A}^\top \boldsymbol{\mu}_{k+1} \in \partial f(\boldsymbol{x}_{k+1}). \quad (3.6)$$

Using the conjugate subgradient theorem [1, Theorem 4.20], we have

$$\boldsymbol{x}_{k+1} \in \partial f^*(-\boldsymbol{A}^\top \boldsymbol{\mu}_{k+1}). \quad (3.7)$$

Multiplying it by $-\boldsymbol{A}$ leads to

$$\mathbf{0} \in -\boldsymbol{A}\partial f^*(-\boldsymbol{A}^\top \boldsymbol{\mu}_{k+1}) + \boldsymbol{A}\boldsymbol{x}_{k+1}, \quad (3.8)$$

which is equivalent to the following optimality condition of the proximal point update:

$$\mathbf{0} \in -\boldsymbol{A}\partial f^*(-\boldsymbol{A}^\top \boldsymbol{\mu}_{k+1}) + \boldsymbol{b} + \frac{1}{\rho}(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k). \quad (3.9)$$

4 Alternating Direction Method of Multipliers (see [1, Chapter 15])

Consider the problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d,\boldsymbol{z}\in\mathbb{R}^p} \{H(\boldsymbol{x},\boldsymbol{z}) \equiv f(\boldsymbol{x}) + \phi(\boldsymbol{z})\} \quad \text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}, \tag{4.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, $\boldsymbol{B} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{c} \in \mathbb{R}^n$. We assume that $f : \mathbb{R}^d \to (-\infty, \infty]$ and $\phi : \mathbb{R}^p \to (-\infty, \infty]$ are proper closed convex functions. There exists $\hat{\boldsymbol{x}} \in \operatorname{relint}(\operatorname{dom} f)$ and $\hat{\boldsymbol{z}} \in \operatorname{relint}(\operatorname{dom} \phi)$ for which $\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{B}\hat{\boldsymbol{z}} = \boldsymbol{c}$. The problem (4.1) has a nonempty optimal set. Note that the objective function $H(\boldsymbol{x}, \boldsymbol{z})$ is separable in $\boldsymbol{x}$ and $\boldsymbol{z}$.

The dual function is

$$\begin{aligned} q(\boldsymbol{\mu}) &= \inf_{\boldsymbol{x}\in\mathbb{R}^d,\boldsymbol{z}\in\mathbb{R}^p} \{\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\mu}) \equiv f(\boldsymbol{x}) + \phi(\boldsymbol{z}) + \langle \boldsymbol{\mu},\, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}\rangle\} \\ &= -f^*(-\boldsymbol{A}^\top\boldsymbol{\mu}) - \phi^*(-\boldsymbol{B}^\top\boldsymbol{\mu}) - \langle \boldsymbol{c},\, \boldsymbol{\mu}\rangle, \end{aligned}$$

and the dual problem (in the minimization form) is

$$\min_{\boldsymbol{\mu}} \{f^*(-\boldsymbol{A}^\top\boldsymbol{\mu}) + \phi^*(-\boldsymbol{B}^\top\boldsymbol{\mu}) + \langle \boldsymbol{c},\, \boldsymbol{\mu}\rangle\} \tag{4.2}$$

The dual subgradient method (in primal representation) for this problem is as follows.

Algorithm 10 Dual Subgradient Method

1: **Input:** $\boldsymbol{\mu}_0 \in \mathbb{R}^n$.
2: **for** $k \geq 0$ **do**
3: $(\boldsymbol{x}_{k+1}, \boldsymbol{z}_{k+1}) \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d,\boldsymbol{z}\in\mathbb{R}^p} \{\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\mu}_k) \equiv f(\boldsymbol{x}) + \phi(\boldsymbol{z}) + \langle \boldsymbol{\mu}_k,\, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}\rangle\}$
4: Choose a step size $s_k > 0$.
5: $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + s_k(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c})$

The $(\boldsymbol{x}_{k+1}, \boldsymbol{z}_{k+1})$-update of dual subgradient method is separable, which preserves the separability of the objective function $H(\boldsymbol{x}, \boldsymbol{z})$, but the overall method suffers from a slow worst-case convergence rate $O\left(\frac{1}{\sqrt{k}}\right)$.

The dual proximal point method (in primal representation), also known as augmented Lagrangian method or the method of multipliers, for this problem is as follows.

Algorithm 11 Augmented Lagrangian Method

1: **Input:** $\boldsymbol{\mu}_0 \in \mathbb{R}^n$ and $\rho > 0$.
2: **for** $k \geq 0$ **do**
3: $(\boldsymbol{x}_{k+1}, \boldsymbol{z}_{k+1}) \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d,\boldsymbol{z}\in\mathbb{R}^p} \{\mathcal{L}_\rho(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\mu}_k) \equiv f(\boldsymbol{x}) + \phi(\boldsymbol{z}) + \langle \boldsymbol{\mu}_k,\, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}\rangle + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}||_2^2\}$
4: $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c})$

Remark 4.1 The augmented Lagrangian method is equivalent to a dual subgradient method (with step size $s_k = \rho > 0$) for solving the following equivalent problem of (4.1):

$$\min_{\boldsymbol{x}\in\mathbb{R}^d,\boldsymbol{z}\in\mathbb{R}^p} \left\{H(\boldsymbol{x},\boldsymbol{z}) + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}||_2^2\right\} \quad \text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}, \tag{4.3}$$

because the function $\mathcal{L}_\rho(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\mu})$ is a Lagrangian function for this equivalent problem.

The augmented Lagrangian method has a worst-case convergence rate $O\left(\frac{1}{k}\right)$ that is faster than that of dual subgradient method, but it looses separability due to the additional quadratic term, unlike the dual subgradient method. Also, the $(\boldsymbol{x}_{k+1}, \boldsymbol{z}_{k+1})$-update step of the augmented Lagrangian method is almost as difficult to solve as the original problem.

To circumvent this non-separability issue, the following alternating direction method of multipliers (ADMM) solves the $(\boldsymbol{x}_{k+1}, \boldsymbol{z}_{k+1})$-update of the augmented Lagrangian method by one iteration of the alternating minimization method.

Algorithm 12 Alternating Direction Method of Multipliers

Input: $\boldsymbol{z}_0 \in \mathbb{R}^p$, $\boldsymbol{\mu}_0 \in \mathbb{R}^n$ and $\rho > 0$.
for $k \geq 0$ **do**
  $\boldsymbol{x}_{k+1} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{\mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{z}_k, \boldsymbol{\mu}_k) \equiv f(\boldsymbol{x}) + \phi(\boldsymbol{z}_k) + \langle \boldsymbol{\mu}_k,\, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z}_k - \boldsymbol{c}\rangle + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z}_k - \boldsymbol{c}||_2^2\right\}$
  $\boldsymbol{z}_{k+1} \in \arg\min_{\boldsymbol{z}\in\mathbb{R}^p} \left\{\mathcal{L}_\rho(\boldsymbol{x}_{k+1}, \boldsymbol{z}, \boldsymbol{\mu}_k) \equiv f(\boldsymbol{x}_{k+1}) + \phi(\boldsymbol{z}) + \langle \boldsymbol{\mu}_k,\, \boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}\rangle + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}||_2^2\right\}$
  $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c})$

ADMM can be generalized as below by adding quadratic proximity terms with given two positive semidefinite matrices $\boldsymbol{G} \in \mathbb{S}_+^d$ and $\boldsymbol{Q} \in \mathbb{S}_+^p$.

Algorithm 13 Alternating Direction Proximal Method of Multipliers

Input: $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{z}_0 \in \mathbb{R}^p$, $\boldsymbol{\mu}_0 \in \mathbb{R}^n$, $\rho > 0$, $\boldsymbol{G} \in \mathbb{S}_+^d$ and $\boldsymbol{Q} \in \mathbb{S}_+^p$.
for $k \geq 0$ **do**
  $\boldsymbol{x}_{k+1} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{\mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{z}_k, \boldsymbol{\mu}_k) + \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_k||_{\boldsymbol{G}}^2\right\}$
  $\boldsymbol{z}_{k+1} \in \arg\min_{\boldsymbol{z}\in\mathbb{R}^p} \left\{\mathcal{L}_\rho(\boldsymbol{x}_{k+1}, \boldsymbol{z}, \boldsymbol{\mu}_k) + \frac{1}{2}||\boldsymbol{z} - \boldsymbol{z}_k||_{\boldsymbol{Q}}^2\right\}$
  $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c})$

For the following choice:

$$\boldsymbol{G} = \alpha\boldsymbol{I} - \rho\boldsymbol{A}^\top\boldsymbol{A}, \quad \boldsymbol{Q} = \beta\boldsymbol{I} - \rho\boldsymbol{B}^\top\boldsymbol{B} \tag{4.4}$$

with $\alpha \geq \rho||\boldsymbol{A}^\top\boldsymbol{A}||_2$ and $\beta \geq \rho||\boldsymbol{B}^\top\boldsymbol{B}||_2$, the alternating direction proximal method of multipliers becomes the alternating direction linearized proximal method of multipliers, also known as a linearized ADMM.

Algorithm 14 Alternating Direction Linearized Proximal Method of Multipliers

Input: $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{z}_0 \in \mathbb{R}^p$, $\boldsymbol{\mu}_0 \in \mathbb{R}^m$, $\rho > 0$, $\alpha \geq \rho||\boldsymbol{A}^\top\boldsymbol{A}||_2$ and $\beta \geq \rho||\boldsymbol{B}^\top\boldsymbol{B}||_2$.
for $k \geq 0$ **do**
  $\boldsymbol{x}_{k+1} = \mathrm{prox}_{\frac{1}{\alpha}f}\left[\boldsymbol{x}_k - \frac{\rho}{\alpha}\boldsymbol{A}^\top\left(\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{z}_k - \boldsymbol{c} + \frac{1}{\rho}\boldsymbol{\mu}_k\right)\right]$
  $\boldsymbol{z}_{k+1} = \mathrm{prox}_{\frac{1}{\beta}\phi}\left[\boldsymbol{z}_k - \frac{\rho}{\beta}\boldsymbol{B}^\top\left(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_k - \boldsymbol{c} + \frac{1}{\rho}\boldsymbol{\mu}_k\right)\right]$
  $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{z}_{k+1} - \boldsymbol{c})$

Assumption 4.1 For any $\boldsymbol{a} \in \mathbb{R}^d$, $\boldsymbol{b} \in \mathbb{R}^p$, $\boldsymbol{G} \in \mathbb{S}_+^d$, $\boldsymbol{Q} \in \mathbb{S}_+^p$, and $\rho > 0$, the optimal sets of the problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{f(\boldsymbol{x}) + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x}||_2^2 + \frac{1}{2}||\boldsymbol{x}||_{\boldsymbol{G}}^2 + \langle \boldsymbol{a},\, \boldsymbol{x}\rangle\right\} \tag{4.5}$$

and

$$\min_{\boldsymbol{z}\in\mathbb{R}^p} \left\{\phi(\boldsymbol{z}) + \frac{\rho}{2}||\boldsymbol{B}\boldsymbol{z}||_2^2 + \frac{1}{2}||\boldsymbol{z}||_{\boldsymbol{Q}}^2 + \langle \boldsymbol{b},\, \boldsymbol{z}\rangle\right\} \tag{4.6}$$

are nonempty.

Theorem 4.1 *Suppose Assumption 4.1 (in addition to the assumptions for (4.1)) holds. Let* $\{(\boldsymbol{x}_k, \boldsymbol{z}_k)\}_{k\geq 0}$ *be sequence generated by the alternating direction proximal method of multipliers for solving (4.1). Let* $(\boldsymbol{x}_*, \boldsymbol{z}_*)$ *be an optimal solution of (4.1) and* $\boldsymbol{\mu}_*$ *be an optimal solution of (4.2). Suppose that* $\gamma > 0$ *is any constant satisfying* $\gamma \geq 2||\boldsymbol{y}_*||_2$*. Then for all* $k \geq 0$*,*

$$H(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{z}}_k) - H(\boldsymbol{x}_*, \boldsymbol{z}_*) \leq \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_*||_{\boldsymbol{G}}^2 + ||\boldsymbol{z}_0 - \boldsymbol{z}_*||_{\boldsymbol{D}}^2 + \frac{1}{\rho}(\gamma + ||\boldsymbol{\mu}_0||_2)^2}{2(k+1)}, \tag{4.7}$$

$$||\boldsymbol{A}\bar{\boldsymbol{x}}_k + \boldsymbol{B}\bar{\boldsymbol{z}}_k - \boldsymbol{c}||_2 \leq \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_*||_{\boldsymbol{G}}^2 + ||\boldsymbol{z}_0 - \boldsymbol{z}_*||_{\boldsymbol{D}}^2 + \frac{1}{\rho}(\gamma + ||\boldsymbol{\mu}_0||_2)^2}{\gamma(k+1)}, \tag{4.8}$$

where $\boldsymbol{D} = \rho \boldsymbol{B}^\top \boldsymbol{B} + \boldsymbol{Q}$ *and*

$$\bar{\boldsymbol{x}}_k = \frac{1}{k+1}\sum_{i=0}^{k} \boldsymbol{x}_{i+1}, \quad \bar{\boldsymbol{z}}_k = \frac{1}{k+1}\sum_{i=0}^{k} \boldsymbol{z}_{i+1}. \tag{4.9}$$

Let's now consider the following simpler problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \{f(\boldsymbol{x}) + \phi(\boldsymbol{A}\boldsymbol{x})\}, \tag{4.10}$$

which is equivalent to (2.1) without the σ-strong convexity of f. As for (2.1), we can reformulate (4.10) as

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^d, \boldsymbol{z}\in\mathbb{R}^p} \{f(\boldsymbol{x}) + \phi(\boldsymbol{z})\} \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} - \boldsymbol{z} = \boldsymbol{0}, \end{aligned} \tag{4.11}$$

which is equivalent to (4.1) with $\boldsymbol{B} = -\boldsymbol{I}$ and $\boldsymbol{c} = \boldsymbol{0}$.

We study ADMM and linearized ADMM for the problem (4.10). (Compare these methods with the dual proximal gradient method that requires additional σ-strongly convex property of f.)

Algorithm 15 Alternating Direction Method of Multipliers for (4.10)

1: **Input:** $\boldsymbol{z}_0 \in \mathbb{R}^p$, $\boldsymbol{\mu}_0 \in \mathbb{R}^n$ and $\rho > 0$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{f(\boldsymbol{x}) + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{z}_k + \frac{1}{\rho}\boldsymbol{\mu}_k||_2^2\right\}$
4: $\quad \boldsymbol{z}_{k+1} = \text{prox}_{\frac{1}{\rho}\phi}\left(\boldsymbol{A}\boldsymbol{x}_{k+1} + \frac{1}{\rho}\boldsymbol{\mu}_k\right)$
5: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{z}_{k+1})$

Algorithm 16 Alternating Direction Linearized Proximal Method of Multipliers for (4.10)

1: **Input:** $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{z}_0 \in \mathbb{R}^p$, $\boldsymbol{\mu}_0 \in \mathbb{R}^m$, $\rho > 0$, $\alpha \geq \rho||\boldsymbol{A}^\top\boldsymbol{A}||_2$ and $\beta \geq \rho||\boldsymbol{B}^\top\boldsymbol{B}||_2$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} = \text{prox}_{\frac{1}{\alpha}f}\left[\boldsymbol{x}_k - \frac{\rho}{\alpha}\boldsymbol{A}^\top\left(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{z}_k + \frac{1}{\rho}\boldsymbol{\mu}_k\right)\right]$
4: $\quad \boldsymbol{z}_{k+1} = \text{prox}_{\frac{1}{\beta}\phi}\left[\boldsymbol{z}_k + \frac{\rho}{\beta}\left(\boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{z}_k + \frac{1}{\rho}\boldsymbol{\mu}_k\right)\right]$
5: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{z}_{k+1})$

Example 4.1 Basis pursuit. Consider the problem

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^d} ||\boldsymbol{x}||_1 \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

This can be reformulated in the form of (4.11) as

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{||\boldsymbol{x}||_1 + I_{\{\boldsymbol{v}\,:\,\boldsymbol{v}=\boldsymbol{b}\}}(\boldsymbol{z})\right\}$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}$$

by letting $f(\boldsymbol{x}) = ||\boldsymbol{x}||_1$ and $\phi(\boldsymbol{z}) = I_{\{\boldsymbol{v}\,:\,\boldsymbol{v}=\boldsymbol{b}\}}(\boldsymbol{z})$. For any $\gamma > 0$, $\text{prox}_{\gamma f}(\cdot)$ is a soft-thresholding operator and $\text{prox}_{\gamma\phi}(\cdot) = \boldsymbol{b}$.

Algorithm 17 Alternating Direction Linearized Proximal Method of Multipliers

1: **Input:** $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\boldsymbol{\mu}_0 \in \mathbb{R}^m$, $\rho > 0$ and $L = ||\boldsymbol{A}^\top\boldsymbol{A}||_2$ ($\alpha = \rho L$ and $\beta = \rho$).
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} = \text{prox}_{\frac{1}{\rho L}f}\left[\boldsymbol{x}_k - \frac{1}{L}\boldsymbol{A}^\top\left(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b} + \frac{1}{\rho}\boldsymbol{\mu}_k\right)\right]$
4: $\quad \boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \rho(\boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{b})$

Remark 4.2 The direct extension of ADMM for three or more variables is not necessarily convergent [2], while it works well in practice.

Remark 4.3 ADMM is a primal representation of the dual Douglas-Rachford splitting method [4], analogous to the fact that augmented Lagrangian method is a primal representation of the dual proximal point method.

The following is Douglas-Rachford splitting method [3] that solves $\min_{\boldsymbol{x}}\{\rho(f(\boldsymbol{x}) + \phi(\boldsymbol{x}))\}$, where f and ϕ are proper closed convex functions and $\rho > 0$.

Algorithm 18 Douglas-Rachford splitting method

1: **Input:** $\boldsymbol{y}_0 \in \mathbb{R}^d$ and $\rho > 0$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} = \text{prox}_{\rho f}(\boldsymbol{y}_k)$
4: $\quad \boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \text{prox}_{\rho\phi}(2\boldsymbol{x}_{k+1} - \boldsymbol{y}_k) - \boldsymbol{x}_k$

References

1. Beck, A.: First-Order Methods in Optimization. Soc. Indust. Appl. Math., Philadelphia (2017). URL http://epubs.siam.org/doi/book/10.1137/1.9781611974997
2. Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Mathematical Programming **155**(1), 57–79 (2016). DOI 10.1007/s10107-014-0826-5
3. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Amer. Math. Soc. **82**(2), 421–39 (1956). URL http://www.jstor.org/stable/1993056
4. Gabay, D.: Applications of the method of multipliers to variational inequalities (1983). In: Augmented Lagrangian Methods: Applications to the numerical solution of boundary-value problems