

# ADMM-Type Methods

Ernest K. Ryu and Wotao Yin

Large-Scale Convex Optimization via Monotone Operators

# Overview

This lecture covers

- ▶ FLiP-ADMM, a general and highly versatile variant of ADMM
- ▶ Convergence analysis of FLiP-ADMM
- ▶ A wide range of ADMM-type methods derived from FLiP-ADMM
- ▶ Further generalization to Trip-ADMM and discussions

# Outline

FLiP-ADMM method, main theorem, discussions

Discussions of parameter choices, special cases, and differences

Proof of main theorem

Derived ADMM-type methods

# Function-Linearized Proximal ADMM (FLiP-ADMM)

Consider primal problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^p, y \in \mathbb{R}^q}{\text{minimize}} && \underbrace{f_1(x) + f_2(x)}_{f(x)} + \underbrace{g_1(y) + g_2(y)}_{g(y)} \\ & \text{subject to} && Ax + By = c, \end{aligned} \tag{1}$$

generated by

$$\mathbf{L}(x, y, u) = f(x) + g(y) + \langle u, Ax + By - c \rangle.$$

Assume  $f_1, f_2, g_1, g_2$  are CCP and  $f_2, g_2$  are also differentiable.

Choose

- ▶ scalars  $\rho, \varphi > 0$
- ▶ matrices  $P \in \mathbb{R}^{p \times p}$  and  $Q \in \mathbb{R}^{q \times q}$  obeying  $P, Q \succeq 0$ .

Choose

- ▶ scalars  $\rho, \varphi > 0$
- ▶ matrices  $P \in \mathbb{R}^{p \times p}$  and  $Q \in \mathbb{R}^{q \times q}$  obeying  $P, Q \succeq 0$ .

FLiP-ADMM is the method

$$\begin{aligned}x^{k+1} &\in \operatorname{argmin}_{x \in \mathbb{R}^p} \left\{ f_1(x) + \langle \nabla f_2(x^k) + A^\top u^k, x \rangle \right. \\ &\quad \left. + \frac{\rho}{2} \|Ax + By^k - c\|^2 + \frac{1}{2} \|x - x^k\|_P^2 \right\} \\ y^{k+1} &\in \operatorname{argmin}_{y \in \mathbb{R}^q} \left\{ g_1(y) + \langle \nabla g_2(y^k) + B^\top u^k, y \rangle \right. \\ &\quad \left. + \frac{\rho}{2} \|Ax^{k+1} + By - c\|^2 + \frac{1}{2} \|y - y^k\|_Q^2 \right\} \\ u^{k+1} &= u^k + \varphi \rho (Ax^{k+1} + By^{k+1} - c).\end{aligned}$$

Recover classic ADMM with  $f_2 = 0$ ,  $g_2 = 0$ ,  $P = 0$ ,  $Q = 0$ , and  $\varphi = 1$ .

## Convergence theorem

### Theorem 6.

Assume total duality, that  $x$ - and  $y$ -subproblems always have solutions, that  $f_2$  is  $L_f$ -smooth and  $g_2$  is  $L_g$ -smooth, and there is an  $\varepsilon \in (0, 2 - \varphi)$  such that

$$P \succeq L_f I, \quad Q \succeq 0, \quad \rho \left( 1 - \frac{(1 - \varphi)^2}{2 - \varphi - \varepsilon} \right) B^\top B + Q \succeq 3L_g I.$$

Then FLiP-ADMM iterates  $x^k, y^k$  satisfy

$$f(x^k) + g(y^k) \rightarrow f(x^*) + g(y^*), \quad Ax^k + By^k - c \rightarrow 0,$$

where  $(x^*, y^*)$  is a solution of the primal problem.

The condition

$$\rho \left( 1 - \frac{(1-\varphi)^2}{2-\varphi-\varepsilon} \right) B^\top B + Q \succeq 3L_g I \quad (2)$$

imposes restrictions on  $\varphi, \rho$ : since  $\varphi = \frac{\sqrt{5}+1}{2}$  leads to  $1 - \frac{(1-\varphi)^2}{2-\varphi} = 0$ ,

- ▶ if  $\varphi \in (0, \frac{\sqrt{5}+1}{2})$ , then  $\exists$  small  $\varepsilon$  such that  $1 - \frac{(1-\varphi)^2}{2-\varphi-\varepsilon} > 0$ , so large  $\rho$  helps to meet (2);
- ▶ if  $\varphi \in (\frac{\sqrt{5}+1}{2}, 2)$  and  $\varepsilon \in (0, 2-\varphi)$ , then  $1 - \frac{(1-\varphi)^2}{2-\varphi-\varepsilon} < 0$ , so small  $\rho$  helps to meet (2).



The condition

$$\rho \left( 1 - \frac{(1-\varphi)^2}{2-\varphi-\varepsilon} \right) B^\top B + Q \succeq 3L_g I \quad (2)$$

imposes restrictions on  $\varphi, \rho$ : since  $\varphi = \frac{\sqrt{5}+1}{2}$  leads to  $1 - \frac{(1-\varphi)^2}{2-\varphi} = 0$ ,

- ▶ if  $\varphi \in (0, \frac{\sqrt{5}+1}{2})$ , then  $\exists$  small  $\varepsilon$  such that  $1 - \frac{(1-\varphi)^2}{2-\varphi-\varepsilon} > 0$ , so large  $\rho$  helps to meet (2);
- ▶ if  $\varphi \in (\frac{\sqrt{5}+1}{2}, 2)$  and  $\varepsilon \in (0, 2-\varphi)$ , then  $1 - \frac{(1-\varphi)^2}{2-\varphi-\varepsilon} < 0$ , so small  $\rho$  helps to meet (2).

Besides appearing in the conditions, the choices of LiP-ADMM parameters affect convergence speed and computational cost per iteration. The optimal choice for a given problem balances the speed and the cost.

# Outline

FLiP-ADMM method, main theorem, discussions

Discussions of parameter choices, special cases, and differences

Proof of main theorem

Derived ADMM-type methods

## Golden-ratio ADMM, Dual extrapolation parameter $\varphi$

While  $\varphi = 1$  is common, a larger  $\varphi$  may provide a speedup.

With  $f_2 = 0$ ,  $g_2 = 0$ ,  $P = 0$ , and  $Q = 0$ , FLiP-ADMM reduces to “Golden-ratio ADMM”:

$$\begin{aligned}x^{k+1} &\in \operatorname{argmin}_{x \in \mathbb{R}^p} \mathbf{L}_\rho(x, y^k, u^k) \\y^{k+1} &\in \operatorname{argmin}_{y \in \mathbb{R}^q} \mathbf{L}_\rho(x^{k+1}, y, u^k) \\u^{k+1} &= u^k + \varphi \rho (Ax^{k+1} + By^{k+1} - c),\end{aligned}$$

where

$$\mathbf{L}_\rho(x, y, u) = f(x) + g(y) + \langle u, Ax + By - c \rangle + \frac{\rho}{2} \|Ax + By - c\|^2.$$

Condition (2) reduces to  $0 < \varphi < (1 + \sqrt{5})/2 \approx 1.618$ .

## Penalty parameter $\rho$

Parameter  $\rho$  controls the relative priority between primal and dual convergence.

The Lyapunov function in the proof (below) contains the terms

- ▶ primal error:  $\rho \|B(y^k - y^*)\|^2$ ,
- ▶ dual error:  $\frac{1}{\varphi\rho} \|u^k - u^*\|^2$ .

Large  $\rho$  prioritizes primal accuracy while small  $\rho$  prioritizes dual accuracy.

## Proximal terms via $P$ and $Q$

The letter “P” in FLi**P**-ADMM describes the presence of the *proximal terms*

$$\frac{1}{2}\|x - x^k\|_P^2, \quad \frac{1}{2}\|y - y^k\|_Q^2.$$

Empirically, smaller  $P$  and  $Q$  leads to fewer required iterations.

When  $f_2 = 0$  and  $g_2 = 0$ , the choice  $P = 0$  and  $Q = 0$  is often optimal in the number of required iterations.

However, proper choices of  $P$  and  $Q$  can cancel out unwieldy quadratic terms and thus reduce the costs of subproblems.

## Linearization of $f_2, g_2$

The  $x$ -subproblem of FLiP-ADMM

$$x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^p} \left\{ f_1(x) + f_2(x^k) + \langle \nabla f_2(x^k), x - x^k \rangle + g(y^k) \right. \\ \left. + \langle u^k, Ax + By^k - c \rangle + \frac{\rho}{2} \|Ax + By^k - c\|^2 + \frac{1}{2} \|x - x^k\|_P^2 \right\}$$

uses  $f_2$ 's first-order approximation  $f_2(x^k) + \langle \nabla f_2(x^k), x - x^k \rangle$ , described by “FLi (Function-Linearized)” in **FLiP-ADMM**.

FLiP-ADMM gives us the choice to use  $f_2$  or not. Choosing  $f_2 = 0$  leads to fewer iterations. In some cases, however, nonzero  $f_2$  reduces the cost of subproblem.

The same discussion holds for the  $y$ -subproblem.

## Relation to Method of Multipliers

'MM" in FLiP-ADMM stands for *method of multipliers*, which has only one primal subproblem.

When  $q = 0$ , the entire  $y$ -subproblem and  $B$ -matrix vanish. FLiP-ADMM reduces to the method of multipliers:

$$x^{k+1} \in \operatorname{argmin}_x \left\{ f(x) + \langle u^k, Ax \rangle + \frac{\rho}{2} \|Ax - c\|^2 \right\}$$
$$u^{k+1} = u^k + \varphi \rho (Ax^{k+1} - c),$$

which converges for  $\varphi \in (0, 2)$  by Theorem 6.

# Outline

FLiP-ADMM method, main theorem, discussions

Discussions of parameter choices, special cases, and differences

**Proof of main theorem**

Derived ADMM-type methods



## Difference from previous lectures

Theorem 6 establishes:

- ▶ the convergence of objective values,
- ▶ the convergence of constraint violations,

but not the convergence of iterates.

The convergence proof (below) does not rely on the machinery of monotone operators.

## About the proof

The key challenge is the construction of the *Lyapunov function* (a name borrowed from nonlinear system, used to prove the system's stability).

The proof is not long (only 4 pages in the textbook), easy to follow, but hardly intuitive.

ADMM-type methods are modular. Hence, the proof comes from the insights we accumulated over years of reading (and writing) papers on ADMM-type methods.

## Constants and Lyapunov function

The assumption of total duality means  $\mathbf{L}$  has a saddle point  $(x^*, y^*, u^*)$ .  
Define

$$w^* = \begin{bmatrix} x^* \\ y^* \\ u^* \end{bmatrix}, \quad w^k = \begin{bmatrix} x^k \\ y^k \\ u^k \end{bmatrix} \quad \text{for } k = 0, 1, \dots$$

Define  $\eta = 2 - \varphi - \varepsilon$ . Define the symmetric positive semidefinite matrices

$$M_0 = \frac{1}{2} \begin{bmatrix} P & 0 & 0 \\ 0 & \rho B^\top B + Q & 0 \\ 0 & 0 & \frac{1}{\varphi\rho} I \end{bmatrix}, \quad M_1 = \frac{1}{2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & Q + L_g I & 0 \\ 0 & 0 & \frac{\eta}{\varphi^2\rho} I \end{bmatrix},$$
$$M_2 = \frac{1}{2} \begin{bmatrix} P - L_f I & 0 & 0 \\ 0 & \rho \left(1 - \frac{(1-\varphi)^2}{\eta}\right) B^\top B + Q - 3L_g I & 0 \\ 0 & 0 & \frac{2-\varphi-\eta}{\varphi^2\rho} I \end{bmatrix}.$$

Define the Lyapunov function

$$V^k = \|w^k - w^*\|_{M_0}^2 + \|w^k - w^{k-1}\|_{M_1}^2.$$

## Proof sketch

The proof has 4 stages. We present only the key terms. You should focus on the proof flow rather than the each single term.

**Stage 1:** Use the facts that  $x^{k+1}$  and  $y^{k+1}$  are subproblem minimizers to obtain inequalities that relate  $x^{k+1}$  with  $x^*$  and  $y^{k+1}$  with  $y^*$ . Add those inequalities and combine terms to arrive at:

$$\begin{aligned} & \mathbf{L}(x^{k+1}, y^{k+1}, u^*) - \mathbf{L}(x^*, y^*, u^*) & (3) \\ & \leq \frac{L_f}{2} \|x^{k+1} - x^k\|^2 + \frac{L_g}{2} \|y^{k+1} - y^k\|^2 + \left(1 - \frac{1}{\varphi}\right) \frac{1}{\varphi\rho} \|u^{k+1} - u^k\|^2 \\ & \quad - 2\langle w^{k+1} - w^k, w^{k+1} - w^* \rangle_{M_0} + \frac{1}{\varphi} \langle u^{k+1} - u^k, B(y^{k+1} - y^k) \rangle. \end{aligned}$$

Since we cannot determine the signs of the two inner-product terms, we must transform them.

**Stage 2:** Bound  $\frac{1}{\varphi} \langle u^{k+1} - u^k, B(y^{k+1} - y^k) \rangle$ .

Use the fact that  $y^k, y^{k+1}$  are minimizers to their respective subproblems to obtain inequalities that relate them. Add those inequalities to get

$$\begin{aligned} & \frac{1}{\varphi} \langle u^{k+1} - u^k, B(y^{k+1} - y^k) \rangle \\ & \leq \frac{Lg}{2} \|y^{k+1} - y^k\|^2 + \frac{Lg}{2} \|y^k - y^{k-1}\|^2 - \|y^{k+1} - y^k\|_Q^2 \\ & \quad + \langle y^{k+1} - y^k, y^k - y^{k-1} \rangle_Q - \left(1 - \frac{1}{\varphi}\right) \langle u^k - u^{k-1}, B(y^{k+1} - y^k) \rangle. \end{aligned}$$

Apply Young's inequality  $\langle a, b \rangle \leq \frac{\zeta}{2} \|a\|^2 + \frac{1}{2\zeta} \|b\|^2$  to last 2 terms ...

... to get

$$\begin{aligned} \frac{1}{\varphi} \langle u^{k+1} - u^k, B(y^{k+1} - y^k) \rangle &\leq \frac{1}{2} \|y^{k+1} - y^k\|_{L_g I - Q + \frac{(1-\varphi)^2}{\eta} \rho B \Upsilon B}^2 \\ &+ \frac{1}{2} \|y^k - y^{k-1}\|_{L_g I + Q}^2 + \frac{\eta}{2\varphi^2 \rho} \|u^k - u^{k-1}\|^2 \end{aligned} \quad (4)$$

If we had applied Young's inequality to  $\frac{1}{\varphi} \langle u^{k+1} - u^k, B(y^{k+1} - y^k) \rangle$  directly, then we couldn't get  $\|y^k - y^{k-1}\|^2$  and  $\|u^k - u^{k-1}\|^2$  terms and thus not  $V^k$  (which is easy to try and verify).

**Stage 3:** Substitute (4) and the generalized cosine identity

$$\|w^{k+1} - w^*\|_{M_0}^2 = \|w^k - w^*\|_{M_0}^2 - \|w^{k+1} - w^k\|_{M_0}^2 + 2\langle w^{k+1} - w^k, w^{k+1} - w^* \rangle_{M_0}$$

into (3); after combine terms, we arrive at the master inequality

$$V^{k+1} \leq V^k - \|w^{k+1} - w^k\|_{M_2}^2 - (\mathbf{L}(x^{k+1}, y^{k+1}, u^*) - \mathbf{L}(x^*, y^*, u^*)).$$

Since  $(x^*, y^*, u^*)$  is a saddle point of  $\mathbf{L}$ ,

$$\mathbf{L}(x^{k+1}, y^{k+1}, u^*) - \mathbf{L}(x^*, y^*, u^*) \geq 0.$$

**Stage 4:** Applying the summability argument on the master inequality tells us

- ▶  $\|w^{k+1} - w^k\|_{M_2}^2 \rightarrow 0$ , from which we conclude  $u^{k+1} - u^k \rightarrow 0$  and thus

$$Ax^k + Bx^k - c \rightarrow 0;$$

- ▶  $\mathbf{L}(x^{k+1}, y^{k+1}, u^*) - \mathbf{L}(x^*, y^*, u^*) \rightarrow 0$ , from which and

$$\mathbf{L}(x^{k+1}, y^{k+1}, u^*) = f(x^{k+1}) + g(y^{k+1}) + \underbrace{\langle u^*, Ax^{k+1} + By^{k+1} - c \rangle}_{\rightarrow 0},$$

we also conclude

$$f(x^k) + g(y^k) \rightarrow f(x^*) + g(y^*).$$



# Outline

FLiP-ADMM method, main theorem, discussions

Discussions of parameter choices, special cases, and differences

Proof of main theorem

Derived ADMM-type methods

## Linearized methods

“Linearization” refers to more than one technique. Most often, it refers to canceling out inconvenient quadratic terms, leaving with linear terms.

Consider

$$\begin{array}{ll} \text{minimize} & f_1(x) + g_1(y) \\ x \in \mathbb{R}^p, y \in \mathbb{R}^q & \\ \text{subject to} & Ax + By = c, \end{array}$$

where  $f_2 = 0$  and  $g_2 = 0$ .

With  $P = (1/\alpha)I - \rho A^\top A$  and  $Q = (1/\beta)I - \rho B^\top B$ , we **recover linearized ADMM** (we saw this method in CH3 with  $\varphi = 1$ ):

$$\begin{aligned} x^{k+1} &= \text{Prox}_{\alpha f} (x^k - \alpha A^\top (u^k + \rho(Ax^k + By^k - c))) \\ y^{k+1} &= \text{Prox}_{\beta g} (y^k - \beta B^\top (u^k + \rho(Ax^{k+1} + By^k - c))) \\ u^{k+1} &= u^k + \varphi \rho (Ax^{k+1} + By^{k+1} - c). \end{aligned}$$

Converge if  $1 \geq \alpha \rho \lambda_{\max}(A^\top A)$ ,  $1 \geq \beta \rho \lambda_{\max}(B^\top B)$ ,  $\varphi < (1 + \sqrt{5})/2$ .

Consider

$$\begin{aligned} & \underset{x \in \mathbb{R}^p, y \in \mathbb{R}^q}{\text{minimize}} && f_1(x) + g_1(y) \\ & \text{subject to} && -Ix + By = 0. \end{aligned}$$

We **recover primal-dual hybrid gradient (PDHG)** with  $\varphi = 1$ ,  $P = 0$ ,  $Q = (1/\beta)I - \rho B^\top B$  in an FLiP-ADMM:

$$\begin{aligned} \mu^{k+1} &= \text{Prox}_{\rho f_1^*}(\mu^k + \rho B(2y^k - y^{k-1})) \\ y^{k+1} &= \text{Prox}_{\beta g_1}(y^k - \beta B^\top \mu^{k+1}). \end{aligned}$$

Converge if  $1 \geq \beta \rho \lambda_{\max}(B^\top B)$ .

## Function-linearized methods

FLiP-ADMM linearizes accesses  $f_2$  and  $g_2$  through their gradient evaluations. This feature provides great flexibility.

Consider

$$\begin{aligned} & \underset{x \in \mathbb{R}^p, y \in \mathbb{R}^q}{\text{minimize}} && f_1(x) + g_1(y) + g_2(y) \\ & \text{subject to} && -Ix + By = 0. \end{aligned}$$

FLiP-ADMM with  $\varphi = 1$ ,  $P = 0$ , and  $Q = (1/\beta)I - \rho B^\top B$  is

$$\begin{aligned} x^{k+1} &= \text{Prox}_{(1/\rho)f_1} \left( (1/\rho)u^k + By^k \right) \\ y^{k+1} &= \text{Prox}_{\beta g_1} \left( y^k - \beta \nabla g_2(y^k) - \beta B^\top (u^k - \rho(x^{k+1} - By^k)) \right) \\ u^{k+1} &= u^k - \rho(x^{k+1} - By^{k+1}). \end{aligned}$$

Apply the Moreau identity to **recover Condat–Vũ**

$$\begin{aligned} \mu^{k+1} &= \text{Prox}_{\rho f_1^*} \left( \mu^k + \rho B(2y^k - y^{k-1}) \right) \\ y^{k+1} &= \text{Prox}_{\beta g_1} \left( y^k - \beta \nabla g_2(y^k) - \beta B^\top \mu^{k+1} \right). \end{aligned}$$

However, FLiP-ADMM condition  $1 \geq \beta \rho \lambda_{\max}(B^\top B) + 3\beta L_g$  is worse than what we have in Ch3.

Consider

$$\begin{aligned} & \underset{x \in \mathbb{R}^p, y \in \mathbb{R}^q}{\text{minimize}} && f_1(x) + f_2(x) + g_1(y) + g_2(y) \\ & \text{subject to} && Ax + By = c. \end{aligned}$$

FLiP-ADMM with  $P = (1/\alpha)I - \rho A^\top A$  and  $Q = (1/\beta)I - \rho B^\top B$  is

$$\begin{aligned} x^{k+1} &= \text{Prox}_{\alpha f_1} \left( x^k - \alpha \left( \nabla f_2(x^k) + A^\top u^k + \rho A^\top (Ax^k + By^k - c) \right) \right) \\ y^{k+1} &= \text{Prox}_{\beta g_1} \left( y^k - \beta \left( \nabla g_2(y^k) + B^\top u^k + \rho B^\top (Ax^{k+1} + By^k - c) \right) \right) \\ u^{k+1} &= u^k + \varphi \rho (Ax^{k+1} + By^{k+1} - c). \end{aligned}$$

We call it **doubly-linearized ADMM**, which generalizes PDHG and Condat-Vũ.

Converge if  $1 \geq \alpha \rho \lambda_{\max}(A^\top A) + \alpha L_f$ ,  $1 \geq \beta \rho \lambda_{\max}(B^\top B) + 3\beta L_g$ , and  $0 < \varphi < (1 + \sqrt{5})/2$ .

## Partial linearization

Consider

$$\begin{aligned} & \underset{x \in \mathbb{R}^p, y \in \mathbb{R}^q}{\text{minimize}} && f_2(x) + g_1(y) + g_2(y) \\ & \text{subject to} && Ax + By = c. \end{aligned}$$

Assume

- ▶  $\gamma I + \rho A^\top A$  is *not* easily invertible
- ▶  $\gamma I + C$  is easily invertible for some  $C \approx \rho A^\top A$

Choose  $P = \gamma I + C - \rho A^\top A$  where  $\gamma > \lambda_{\max}(\rho A^\top A - C)$  is small.

Then, the  $x$ -update of FLiP-ADMM

$$x^{k+1} = x^k - (\gamma I + C)^{-1} (\nabla f_2(x^k) + A^\top u^k + \rho A^\top (Ax^k + By^k - c)),$$

is easy to compute. Call it **partial linearization**. It reduces iterations compared to (full) linearization (with  $P = \gamma I - \rho A^\top A$ ).

## CT imaging with total variation regularization

Let  $x$  represent a 2D or 3D image to recover from CT measurements  $b$ :

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \quad \ell(Ax - b) + \lambda \|Dx\|_1,$$

where  $A$  is the discrete Radon transform operator,  $D$  is a finite difference operator, and  $\ell$  is a CCP function.

PDHG has low-cost steps and but requires too many iterations. Classic ADMM requires (much) fewer iterations but an expensive step:

$$x^{k+1} = x^k - (\rho A^T A + \rho D^T D)^{-1} (A^T u^{k+1} + Dv^{k+1}).$$

Since  $A^T A$  and  $D^T D$  are discretizations of shift-invariant continuous operators, they can be approximated by circulant matrices, so we use a *circulant matrix*

$$C \approx \rho A^T A + \rho D^T D.$$

Let  $c^\top$  be the first row of  $C$  and  $\hat{c}$  is its discrete Fourier transform. Let  $F(\cdot)$  be a discrete Fourier transform. By the *convolution theorem*, for any vector  $x$  and its inverse Fourier transform  $\check{x}$ , we have

$$\begin{aligned}(\gamma I + C)x &= F(\text{Diag}(\gamma \mathbf{1} + \hat{c})\check{x}) \\ (\gamma I + C)^{-1}x &= F(\text{Diag}^{-1}(\gamma \mathbf{1} + \hat{c})\check{x}),\end{aligned}$$

so  $(\gamma I + C)^{-1}$  is easy by fast Fourier transform (FFT).

FLiP-ADMM with partial linearization  $P = \gamma I + C - \rho A^\top A - \rho D^\top D$

$$\begin{aligned}u^{k+1} &= \text{Prox}_{\rho \ell^*}(u^k + \rho A(2x^k - x^{k-1}) - \rho b) \\ v^{k+1} &= \Pi_{[-\lambda, \lambda]}(v^k + \rho D(2x^k - x^{k-1})) \\ x^{k+1} &= x^k - (\gamma I + C)^{-1}(A^\top u^{k+1} + Dv^{k+1})\end{aligned}$$

has easy-to-compute steps. A small  $\gamma > \lambda_{\max}(\rho A^\top A + \rho D^\top D - C)$  leads a minimal increase in iterations over classic ADMM.



## Multi-block ADMM problem

Partition  $x \in \mathbb{R}^p$  into  $m$  non-overlapping blocks of sizes  $p_1, \dots, p_m$ .

Partition matrix  $A = [A_{:,1} \quad A_{:,2} \quad \cdots \quad A_{:,m}]$  such that

$$Ax = A_{:,1}x_1 + A_{:,2}x_2 + \cdots + A_{:,m}x_m.$$

Multi-block ADMM problem or extended monotropic program is

$$\begin{aligned} & \underset{(x_1, \dots, x_m) \in \mathbb{R}^p}{\text{minimize}} && f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m) \\ & \text{subject to} && A_{:,1}x_1 + A_{:,2}x_2 + \cdots + A_{:,m}x_m = c. \end{aligned} \tag{5}$$

Unless the column-blocks of  $A$  are orthogonal, i.e.,  $A_{:,i}^\top A_{:,j} = 0$  for all  $i \neq j$ , the blocks  $x_1^{k+1}, \dots, x_m^{k+1}$  cannot be computed independently.

Next, we present two splitting techniques with which  $x_1^{k+1}, \dots, x_m^{k+1}$  can be computed independently in parallel.

## Jacobi ADMM

In numerical linear algebra, the Jacobi method is an iterative method for solving certain linear systems. It updates the blocks of  $x$  independently.

Consider problem (5) and matrix

$$P = \begin{bmatrix} \gamma I & -\rho A_{:,1}^\top A_{:,2} & \cdots & \cdots & -\rho A_{:,1}^\top A_{:,m} \\ -\rho A_{:,2}^\top A_{:,1} & \gamma I & \cdots & \cdots & -\rho A_{:,2}^\top A_{:,m} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ -\rho A_{:,m}^\top A_{:,1} & -\rho A_{:,m}^\top A_{:,2} & \cdots & -\rho A_{:,m}^\top A_{:,(m-1)} & \gamma I \end{bmatrix},$$

which is positive semidefinite for  $\gamma \geq \rho \lambda_{\max}(A^\top A)$ .

Let

$$\mathbf{L}_\rho(x, u) = \sum_{i=1}^m f_i(x_i) + \langle u, Ax - c \rangle + \frac{\rho}{2} \|Ax - c\|^2.$$

Let  $x_{\neq i}^k$  denote all components of  $x^k$  excluding  $x_i^k$ . FLiP-ADMM with the matrix  $P$  is

$$x_i^{k+1} = \operatorname{argmin}_{x_i \in \mathbb{R}^{p_i}} \left\{ \mathbf{L}_\rho(x_i, x_{\neq i}^k, u^k) + \frac{\gamma}{2} \|x_i - x_i^k\|^2 \right\} \quad \text{for } i = 1, \dots, m$$

$$u^{k+1} = u^k + \varphi \rho (Ax^{k+1} - c).$$

This method is called **Jacobi ADMM** in analogy to the Jacobi method.

See Exercise 8.3 for other choices of  $P$ , where the diagonal  $\gamma I$  is replaced by diagonal blocks.

## Dummy variable technique + FLiP-ADMM

Consider the following generalization to problem (5):

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m f_i(x_i) + g(y) \\ (x_1, \dots, x_m) \in \mathbb{R}^p & \\ y \in \mathbb{R}^n & \\ \text{subject to} & Ax + y = c. \end{array}$$

Introduce dummy variables  $z_1, \dots, z_m$  and eliminate  $y$  to get the equivalent problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m f_i(x_i) + g\left(c - \sum_{i=1}^m z_i\right) \\ (x_1, \dots, x_m) \in \mathbb{R}^p & \\ z_1, \dots, z_m \in \mathbb{R}^n & \\ \text{subject to} & A_{:,i}x_i - z_i = 0 \quad \text{for } i = 1, \dots, m. \end{array}$$

Apply FLiP-ADMM with  $P = 0$ ,  $Q = 0$ , no function linearization, and initial  $u$ -variables satisfying  $u_1^0 = \dots = u_m^0$ . Then we can show  $u_1^k = \dots = u_m^k$  for  $k = 1, \dots, m$ , and the iteration simplifies to

$$x_i^{k+1} \in \operatorname{argmin}_{x_i \in \mathbb{R}^{p_i}} \left\{ f_i(x_i) + \left\langle u^k + \frac{\rho}{m} (Ax^k - z_{\text{sum}}^k), A_{:,i} x_i \right\rangle + \frac{\rho}{2} \|A_{:,i}(x_i - x_i^k)\|^2 \right\}$$

for  $i = 1, \dots, m$

$$z_{\text{sum}}^{k+1} = c - \operatorname{Prox}_{\frac{m}{\rho} g} \left( c - Ax^{k+1} - \frac{m}{\rho} u^k \right)$$

$$u^{k+1} = u^k + \frac{\varphi \rho}{m} (Ax^{k+1} - z_{\text{sum}}^{k+1}).$$

The method converges if  $\varphi \in (0, (1 + \sqrt{5})/2)$ .

## Consensus technique + FLiP-ADMM

Consider

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^n f_i(x).$$

Use the consensus technique to get the equivalent problem

$$\begin{aligned} & \underset{x_1, \dots, x_n, z \in \mathbb{R}^p}{\text{minimize}} && \sum_{i=1}^n f_i(x_i) \\ & \text{subject to} && x_i = z, \quad \text{for } i = 1, \dots, n. \end{aligned}$$

Here,  $x_i \in \mathbb{R}^p$  is a copy of  $x \in \mathbb{R}^p$ . This contrasts with block splitting, where each  $x_i$  represented a single block of  $x$ .

Apply FLiP-ADMM with  $P = 0$ ,  $Q = 0$ , no function linearization, and initial  $u$ -variables satisfying  $u_1^0 + \dots + u_n^0 = 0$  to get

$$x_i^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^p} \left\{ f_i(x_i) + \langle u_i^k, x_i \rangle + \frac{\rho}{2} \|x_i - z^k\|^2 \right\} \quad \text{for } i = 1, \dots, n$$

$$z^{k+1} = \frac{1}{n} \sum_{i=1}^n x_i^{k+1}$$

$$u_i^{k+1} = u_i^k + \varphi \rho (x_i^{k+1} - z^{k+1}) \quad \text{for } i = 1, \dots, n.$$

Converge if  $\varphi \in (0, (1 + \sqrt{5})/2)$ .

The consensus technique is versatile. Instead of constraining  $x_1, \dots, x_n$  to equal a single  $z$  here, we will equal them to multiple  $z$ -variables through a graph structure to obtain a decentralized method in CH11.

## 2-1-2 ADMM

This is a technique that applies an ADMM method to a problem with one or two more blocks (if they are strongly-convex quadratic) than what it is designed for.

Assume  $g$  is a strongly convex quadratic function with affine constraints, i.e.,

$$g(y) = y^T M y + \mu^T y + \delta_{\{y \in \mathbb{R}^q \mid N y = \nu\}}(y)$$

for  $M \in \mathbb{R}^{q \times q}$  with  $M \succ 0$ ,  $N \in \mathbb{R}^{s \times q}$ , and  $\nu \in \mathcal{R}(N)$ . If no affine constraint, we set  $s = 0$ .

Define

$$\mathbf{L}_\rho(x, y, u) = f(x) + g(y) + \langle u, Ax + By - c \rangle + \frac{\rho}{2} \|Ax + By - c\|^2.$$



**2-1-2 ADMM** is the method:

$$y^{k+1/2} = \operatorname{argmin}_{y \in \mathbb{R}^q} \mathbf{L}_\rho(x^k, y, u^k)$$

$$x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^p} \mathbf{L}_\rho(x, y^{k+1/2}, u^k)$$

$$y^{k+1} = \operatorname{argmin}_{y \in \mathbb{R}^q} \mathbf{L}_\rho(x^{k+1}, y, u^k)$$

$$u^{k+1} = u^k + \varphi \rho(Ax^{k+1} + By^{k+1} - c).$$

It is equivalent to (single-block) FLiP-ADMM applied to

$$(x^{k+1}, y^{k+1}) \in \operatorname{argmin}_{x \in \mathbb{R}^p, y \in \mathbb{R}^q} \left\{ \mathbf{L}_\rho(x, y, u^k) + \frac{\rho}{2} \|x - x^k\|_P^2 \right\}$$
$$u^{k+1} = u^k + \varphi \rho(Ax^{k+1} + By^{k+1} - c)$$

with  $P = A^\top B^\top B A$ . Converge if  $\varphi \in (0, 2)$ .

See Exercises 8.11 for 4-block ADMM with 2-1-2-4-3-4 updates and 8.12 for generalization with function linearization and proximal terms.

## Trip-ADMM

Consider the more problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^p, y \in \mathbb{R}^q}{\text{minimize}} && f_1(Cx) + f_2(x) + g_1(Dy) + g_2(y) \\ & \text{subject to} && Ax + By = c. \end{aligned}$$

**Trip-ADMM** (Triple-linearized ADMM) is the method

$$\begin{cases} x^{k+1/2} = x^k - \sigma (C^\top v^k + \nabla f_2(x^k) + A^\top u^k + \rho A^\top (Ax^k + By^k - c)) \\ v^{k+1} = \text{Prox}_{\tau f_1^*} (v^k + \tau Cx^{k+1/2}) \\ x^{k+1} = x^{k+1/2} - \sigma C^\top (v^{k+1} - v^k) \end{cases}$$
$$\begin{cases} y^{k+1/2} = y^k - \sigma (D^\top w^k + \nabla g_2(y^k) + B^\top u^k + \rho B^\top (Ax^{k+1} + By^k - c)) \\ w^{k+1} = \text{Prox}_{\tau g_1^*} (w^k + \tau Dy^{k+1/2}) \\ y^{k+1} = y^{k+1/2} - \sigma D^\top (w^{k+1} - w^k) \\ u^{k+1} = u^k + \rho (Ax^{k+1} + By^{k+1} - c), \end{cases}$$

which generalizes FLiP-ADMM.

If the parameters satisfy  $\rho > 0$ ,  $\sigma > 0$ ,  $\tau > 0$ ,

$$\begin{aligned} 1 &\geq \sigma\rho\lambda_{\max}(A^\top A) + \sigma L_f, & 1 &\geq \sigma\rho\lambda_{\max}(B^\top B) + 3\sigma L_g, \\ 1 &\geq \sigma\tau\lambda_{\max}(CC^\top), & 1 &\geq \sigma\tau\lambda_{\max}(DD^\top) \end{aligned}$$

and assume total duality, we have

$$\begin{aligned} &f_1\left(Cx^k - \sigma\tilde{C}^\top\tilde{C}(v^{k+1} - v^k)\right) + f_2(x^k) \\ &+ g_1\left(Dy^k - \sigma\tilde{D}^\top\tilde{D}(w^{k+1} - w^k)\right) + g_2(y^k) \\ &\rightarrow f_1(Cx^\star) + f_2(x^\star) + g_1(Dy^\star) + g_2(y^\star), \\ &Ax^k + Bx^k - c \rightarrow 0. \end{aligned}$$

## Conclusion

FLiP-ADMM is a combination of four techniques: alternating update, method of multipliers, linearization, function-linearization, and use of proximal terms.

These techniques can be combined like modules to solve problems with complicated structures.

ADMM-type methods are “splitting methods”, intimately related to monotone operator methods, though are not monotone operator methods themselves.

Fully general FLiP-ADMM (with dual extrapolation  $\varphi$ ) cannot be reduced to a monotone operator splitting method and must be analyzed directly.