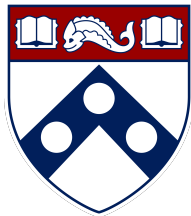


Alternating direction method of multipliers



Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

Outline

- Augmented Lagrangian method
- Alternating direction method of multipliers

Two-block problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}, \mathbf{z}} && F(\mathbf{x}, \mathbf{z}) := f_1(\mathbf{x}) + f_2(\mathbf{z}) \\ & \text{subject to} && \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{b} \end{aligned}$$

where f_1 and f_2 are both convex

- this can also be solved via Douglas-Rachford splitting
- we will introduce another paradigm for solving this problem

Augmented Lagrangian method

Dual problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}, \mathbf{z}} && f_1(\mathbf{x}) + f_2(\mathbf{z}) \\ & \text{subject to} && \mathbf{Ax} + \mathbf{Bz} = \mathbf{b} \end{aligned}$$

\Leftrightarrow

$$\text{maximize}_{\boldsymbol{\lambda}} \min_{\mathbf{x}, \mathbf{z}} \underbrace{f_1(\mathbf{x}) + f_2(\mathbf{z}) + \langle \boldsymbol{\lambda}, \mathbf{Ax} + \mathbf{Bz} - \mathbf{b} \rangle}_{=:\mathcal{L}(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda}) \text{ (Lagrangian)}}$$

\Leftrightarrow

$$\text{maximize}_{\boldsymbol{\lambda}} -f_1^*(-\mathbf{A}^\top \boldsymbol{\lambda}) - f_2^*(-\mathbf{B}^\top \boldsymbol{\lambda}) - \langle \boldsymbol{\lambda}, \mathbf{b} \rangle$$

\Leftrightarrow

$$\text{minimize}_{\boldsymbol{\lambda}} f_1^*(-\mathbf{A}^\top \boldsymbol{\lambda}) + f_2^*(-\mathbf{B}^\top \boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \mathbf{b} \rangle$$

Augmented Lagrangian method

$$\text{minimize}_{\lambda} \quad f_1^*(-\mathbf{A}^\top \lambda) + f_2^*(-\mathbf{B}^\top \lambda) + \langle \lambda, \mathbf{b} \rangle$$

The proximal point method for solving this dual problem:

$$\lambda^{t+1} = \arg \min_{\lambda} \left\{ f_1^*(-\mathbf{A}^\top \lambda) + f_2^*(-\mathbf{B}^\top \lambda) + \langle \lambda, \mathbf{b} \rangle + \frac{1}{2\rho} \|\lambda - \lambda^t\|_2^2 \right\}$$

As it turns out, this is equivalent to the **augmented Lagrangian method** (or the method of multipliers)

$$\begin{aligned} (\mathbf{x}^{t+1}, \mathbf{z}^{t+1}) &= \arg \min_{\mathbf{x}, \mathbf{z}} \left\{ f_1(\mathbf{x}) + f_2(\mathbf{z}) + \frac{\rho}{2} \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{b} + \frac{1}{\rho} \lambda^t\|_2^2 \right\} \\ \lambda^{t+1} &= \lambda^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b}) \end{aligned} \quad (10.1)$$

Justification of (10.1)

$$\boldsymbol{\lambda}^{t+1} = \arg \min_{\boldsymbol{\lambda}} \left\{ f_1^*(-\mathbf{A}^\top \boldsymbol{\lambda}) + f_2^*(-\mathbf{B}^\top \boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \mathbf{b} \rangle + \frac{1}{2\rho} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|_2^2 \right\}$$

\Downarrow optimality condition

$$\mathbf{0} \in -\mathbf{A} \partial f_1^*(-\mathbf{A}^\top \boldsymbol{\lambda}^{t+1}) - \mathbf{B} \partial f_2^*(-\mathbf{B}^\top \boldsymbol{\lambda}^{t+1}) + \mathbf{b} + \frac{1}{\rho} (\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}^t)$$

\Downarrow

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \rho (\mathbf{A} \mathbf{x}^{t+1} + \mathbf{B} \mathbf{z}^{t+1} - \mathbf{b})$$

where (check: use the conjugate subgradient theorem)

$$\mathbf{x}^{t+1} := \arg \min_{\mathbf{x}} \{ \langle \mathbf{A}^\top \boldsymbol{\lambda}^{t+1}, \mathbf{x} \rangle + f_1(\mathbf{x}) \}$$

$$\mathbf{z}^{t+1} := \arg \min_{\mathbf{z}} \{ \langle \mathbf{B}^\top \boldsymbol{\lambda}^{t+1}, \mathbf{z} \rangle + f_2(\mathbf{z}) \}$$

Justification of (10.1)

 \Updownarrow

$$\mathbf{x}^{t+1} := \arg \min_{\mathbf{x}} \{ \langle \mathbf{A}^\top [\boldsymbol{\lambda}^t + \rho (\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b})], \mathbf{x} \rangle + f_1(\mathbf{x}) \}$$

$$\mathbf{z}^{t+1} := \arg \min_{\mathbf{z}} \{ \langle \mathbf{B}^\top [\boldsymbol{\lambda}^t + \rho (\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b})], \mathbf{z} \rangle + f_2(\mathbf{z}) \}$$

 \Updownarrow

$$\mathbf{0} \in \mathbf{A}^\top [\boldsymbol{\lambda}^t + \rho (\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b})] + \partial f_1(\mathbf{x}^{t+1})$$

$$\mathbf{0} \in \mathbf{B}^\top [\boldsymbol{\lambda}^t + \rho (\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b})] + \partial f_2(\mathbf{z}^{t+1})$$

 \Updownarrow

$$(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}) = \arg \min_{\mathbf{x}, \mathbf{z}} \left\{ f_1(\mathbf{x}) + f_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right\|_2^2 \right\}$$

Augmented Lagrangian method (ALM)

$$(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}) = \arg \min_{\mathbf{x}, \mathbf{z}} \left\{ f_1(\mathbf{x}) + f_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{Ax} + \mathbf{Bz} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right\|_2^2 \right\}$$

(primal step)

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \rho(\mathbf{Ax}^{t+1} + \mathbf{Bz}^{t+1} - \mathbf{b})$$

(dual step)

where $\rho > 0$ is penalty parameter

ALM aims to solve the following problem by alternating between primal and dual updates

$$\text{maximize}_{\boldsymbol{\lambda}} \max_{\mathbf{x}, \mathbf{z}} \underbrace{f_1(\mathbf{x}) + f_2(\mathbf{z}) + \rho \langle \mathbf{Ax} + \mathbf{Bz} - \mathbf{b}, \boldsymbol{\lambda} \rangle + \frac{\rho}{2} \left\| \mathbf{Ax} + \mathbf{Bz} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda} \right\|_2^2}_{\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda})}$$

$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda})$: augmented Lagrangian

Issues of augmented Lagrangian method

$$(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}) = \arg \min_{\mathbf{x}, \mathbf{z}} \left\{ f_1(\mathbf{x}) + f_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right\|_2^2 \right\}$$

- the primal update step is often expensive — as expensive as solving the original problem
- minimization of \mathbf{x} and \mathbf{z} cannot be carried out separately

Alternating direction method of multipliers

Alternating direction method of multipliers

Rather than computing exact primal estimate for ALM, we might minimize \mathbf{x} and \mathbf{z} sequentially via alternating minimization

$$\begin{aligned}\mathbf{x}^{t+1} &= \arg \min_{\mathbf{x}} \left\{ f_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}^t - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right\|_2^2 \right\} \\ \mathbf{z}^{t+1} &= \arg \min_{\mathbf{z}} \left\{ f_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right\|_2^2 \right\} \\ \boldsymbol{\lambda}^{t+1} &= \boldsymbol{\lambda}^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b})\end{aligned}$$

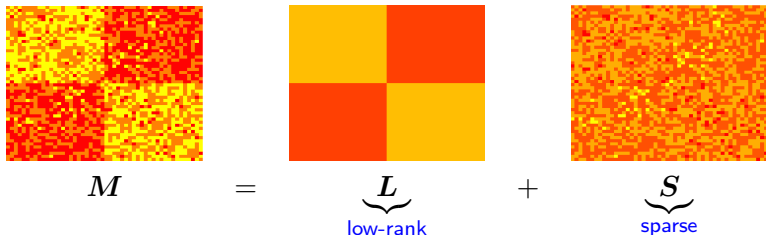
— called the *alternating direction method of multipliers (ADMM)*

Alternating direction method of multipliers

$$\begin{aligned} \mathbf{x}^{t+1} &= \arg \min_{\mathbf{x}} \left\{ f_1(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{Ax} + \mathbf{Bz}^t - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right\|_2^2 \right\} \\ \mathbf{z}^{t+1} &= \arg \min_{\mathbf{z}} \left\{ f_2(\mathbf{z}) + \frac{\rho}{2} \left\| \mathbf{Ax}^{t+1} + \mathbf{Bz} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right\|_2^2 \right\} \\ \boldsymbol{\lambda}^{t+1} &= \boldsymbol{\lambda}^t + \rho(\mathbf{Ax}^{t+1} + \mathbf{Bz}^{t+1} - \mathbf{b}) \end{aligned}$$

- ρ controls relative priority between primal and dual convergence
- useful if updating \mathbf{x}^t and updating \mathbf{z}^t are both inexpensive
- blend the benefits of dual decomposition and augmented Lagrangian method
- the roles of \mathbf{x} and \mathbf{z} are *almost* symmetric, but not quite

Example: robust PCA



The diagram shows three heatmaps representing matrices. The first heatmap, labeled M , is a noisy matrix with a mix of red and yellow pixels. The second heatmap, labeled L and "low-rank", is a smooth matrix with a clear 2x2 block structure of yellow and red. The third heatmap, labeled S and "sparse", is a matrix with sparse, scattered red and yellow pixels. The equation $M = L + S$ is shown below the heatmaps.

$$M = \underbrace{L}_{\text{low-rank}} + \underbrace{S}_{\text{sparse}}$$

Suppose we observe M , which is the superposition of a low-rank component L and sparse outliers S

Can we hope to disentangle L and S ?

Example: robust PCA

One way to solve it is via convex programming (Candes et al. '08)

$$\begin{aligned} \text{minimize}_{\mathbf{L}, \mathbf{S}} \quad & \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \\ \text{s.t.} \quad & \mathbf{L} + \mathbf{S} = \mathbf{M} \end{aligned} \tag{10.2}$$

where $\|\mathbf{L}\|_* := \sum_{i=1}^n \sigma_i(\mathbf{L})$ is the nuclear norm, and $\|\mathbf{S}\|_1 := \sum_{i,j} |S_{i,j}|$ is the entrywise ℓ_1 norm

Example: robust PCA

ADMM for solving (10.2):

$$\begin{aligned}\mathbf{L}^{t+1} &= \arg \min_{\mathbf{L}} \left\{ \|\mathbf{L}\|_* + \frac{\rho}{2} \left\| \mathbf{L} + \mathbf{S}^t - \mathbf{M} + \frac{1}{\rho} \boldsymbol{\Lambda}^t \right\|_{\text{F}}^2 \right\} \\ \mathbf{S}^{t+1} &= \arg \min_{\mathbf{S}} \left\{ \lambda \|\mathbf{S}\|_1 + \frac{\rho}{2} \left\| \mathbf{L}^{t+1} + \mathbf{S} - \mathbf{M} + \frac{1}{\rho} \boldsymbol{\Lambda}^t \right\|_{\text{F}}^2 \right\} \\ \boldsymbol{\Lambda}^{t+1} &= \boldsymbol{\Lambda}^t + \rho (\mathbf{L}^{t+1} + \mathbf{S}^{t+1} - \mathbf{M})\end{aligned}$$

Example: robust PCA

This is equivalent to

$$\mathbf{L}^{t+1} = \text{SVT}_{\rho^{-1}}\left(\mathbf{M} - \mathbf{S}^t - \frac{1}{\rho}\mathbf{\Lambda}^t\right) \quad (\text{singular value thresholding})$$

$$\mathbf{S}^{t+1} = \text{ST}_{\lambda\rho^{-1}}\left(\mathbf{M} - \mathbf{L}^{t+1} - \frac{1}{\rho}\mathbf{\Lambda}^t\right) \quad (\text{soft thresholding})$$

$$\mathbf{\Lambda}^{t+1} = \mathbf{\Lambda}^t + \rho(\mathbf{L}^{t+1} + \mathbf{S}^{t+1} - \mathbf{M})$$

where for any \mathbf{X} with SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ ($\mathbf{\Sigma} = \text{diag}(\{\sigma_i\})$), one has

$$\text{SVT}_{\tau}(\mathbf{X}) = \mathbf{U}\text{diag}(\{(\sigma_i - \tau)_+\})\mathbf{V}^\top$$

$$\text{and} \quad (\text{ST}_{\tau}(\mathbf{X}))_{i,j} = \begin{cases} X_{i,j} - \tau, & \text{if } X_{i,j} > \tau \\ 0, & \text{if } |X_{i,j}| \leq \tau \\ X_{i,j} + \tau, & \text{if } X_{i,j} < -\tau \end{cases}$$

Example: graphical lasso

When learning a sparse Gaussian graphical model, one resorts to:

$$\begin{aligned} \underset{\Theta}{\text{minimize}} \quad & \underbrace{-\log \det \Theta + \langle \Theta, S \rangle}_{\text{negative log-likelihood of Gaussian graphical model}} + \underbrace{\lambda \|\Theta\|_1}_{\text{encourage sparsity}} \\ \text{s.t.} \quad & \Theta \succeq \mathbf{0} \end{aligned}$$



$$\begin{aligned} \underset{\Theta}{\text{minimize}} \quad & -\log \det \Theta + \langle \Theta, S \rangle + \mathbb{I}_{\mathbb{S}_+}(\Theta) + \lambda \|\Psi\|_1 \quad (10.3) \\ \text{s.t.} \quad & \Theta = \Psi \end{aligned}$$

where $\mathbb{S}_+ := \{\mathbf{X} \mid \mathbf{X} \succeq \mathbf{0}\}$

Example: graphical lasso

ADMM for solving (10.3):

$$\Theta^{t+1} = \arg \min_{\Theta \succeq \mathbf{0}} \left\{ -\log \det \Theta + \frac{\rho}{2} \left\| \Theta - \Psi^t + \frac{1}{\rho} \Lambda^t + \frac{1}{\rho} \mathbf{S} \right\|_{\text{F}}^2 \right\}$$

$$\Psi^{t+1} = \arg \min_{\Psi} \left\{ \lambda \|\Psi\|_1 + \frac{\rho}{2} \left\| \Theta^{t+1} - \Psi + \frac{1}{\rho} \Lambda^t \right\|_{\text{F}}^2 \right\}$$

$$\Lambda^{t+1} = \Lambda^t + \rho (\Theta^{t+1} - \Psi^{t+1})$$

Example: graphical lasso

This is equivalent to

$$\begin{aligned}\Theta^{t+1} &= \mathcal{F}_\rho\left(\Psi^t - \frac{1}{\rho}\Lambda^t - \frac{1}{\rho}\mathbf{S}\right) \\ \Psi^{t+1} &= \text{ST}_{\lambda\rho^{-1}}\left(\Theta^{t+1} + \frac{1}{\rho}\Lambda^t\right) \quad (\text{soft thresholding}) \\ \Lambda^{t+1} &= \Lambda^t + \rho(\Theta^{t+1} - \Psi^{t+1})\end{aligned}$$

where for $\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^\top \succeq \mathbf{0}$ with $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$, one has

$$\mathcal{F}_\rho(\mathbf{X}) := \frac{1}{2}\mathbf{U}\text{diag}(\{\lambda_i + \sqrt{\lambda_i^2 + \frac{4}{\rho}}\})\mathbf{U}^\top$$

Example: consensus optimization

Consider solving the following minimization problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad \sum_{i=1}^N f_i(\mathbf{x}) \\ & \quad \quad \quad \Downarrow \\ & \text{minimize} \quad \sum_{i=1}^N f_i(\mathbf{x}_i) \quad (\text{block separable}) \\ & \quad \text{s.t.} \quad \mathbf{x}_i = \mathbf{z} \quad 1 \leq i \leq N \\ & \quad \quad \quad \Downarrow \\ & \text{minimize} \quad \sum_{i=1}^N f_i(\mathbf{x}_i) \\ & \quad \text{s.t.} \quad \mathbf{u} := \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \vdots \\ \mathbf{I} \end{bmatrix} \mathbf{z} \end{aligned}$$

Example: consensus optimization

ADMM for solving this problem:

$$\mathbf{u}^{t+1} = \arg \min_{\mathbf{u}=[\mathbf{x}_i]_{1 \leq i \leq N}} \left\{ \sum_{i=1}^N f_i(\mathbf{x}_i) + \frac{\rho}{2} \sum_{i=1}^N \left\| \mathbf{x}_i - \mathbf{z}^t + \frac{1}{\rho} \boldsymbol{\lambda}_i^t \right\|_2^2 \right\}$$

$$\mathbf{z}^{t+1} = \arg \min_{\mathbf{z}} \left\{ \frac{\rho}{2} \sum_{i=1}^N \left\| \mathbf{x}_i^{t+1} - \mathbf{z} + \frac{1}{\rho} \boldsymbol{\lambda}_i^t \right\|_2^2 \right\}$$

$$\boldsymbol{\lambda}_i^{t+1} = \boldsymbol{\lambda}_i^t + \rho(\mathbf{x}_i^{t+1} - \mathbf{z}^{t+1}), \quad 1 \leq i \leq N$$

Example: consensus optimization

This is equivalent to

$$\mathbf{x}_i^{t+1} = \arg \min_{\mathbf{x}_i} \left\{ f_i(\mathbf{x}_i) + \frac{\rho}{2} \left\| \mathbf{x}_i - \mathbf{z}^t + \frac{1}{\rho} \boldsymbol{\lambda}_i^t \right\|_2^2 \right\} \quad 1 \leq i \leq N$$

(can be computed in parallel)

$$\mathbf{z}^{t+1} = \frac{1}{N} \sum_{i=1}^N \left(\mathbf{x}_i^{t+1} + \frac{1}{\rho} \boldsymbol{\lambda}_i^t \right)$$

(gather all local iterates)

$$\boldsymbol{\lambda}_i^{t+1} = \boldsymbol{\lambda}_i^t + \rho(\mathbf{x}_i^{t+1} - \mathbf{z}^{t+1}), \quad 1 \leq i \leq N$$

(“broadcast” \mathbf{z}^{t+1} to update all local multipliers)

ADMM is well suited for distributed optimization!

Convergence of ADMM

Theorem 10.1 (Convergence of ADMM)

Suppose f_1 and f_2 are closed convex functions, and γ is any constant obeying $\gamma \geq 2\|\lambda^*\|_2$. Then

$$F(\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) - F^{\text{opt}} \leq \frac{\|\mathbf{z}^0 - \mathbf{z}^*\|_{\rho \mathbf{B}^\top \mathbf{B}}^2 + \frac{(\gamma + \|\lambda^0\|_2)^2}{\rho}}{2(t+1)} \quad (10.4a)$$

$$\|\mathbf{A}\mathbf{x}^{(t)} + \mathbf{B}\mathbf{z}^{(t)} - \mathbf{b}\|_2 \leq \frac{\|\mathbf{z}^0 - \mathbf{z}^*\|_{\rho \mathbf{B}^\top \mathbf{B}}^2 + \frac{(\gamma + \|\lambda^0\|_2)^2}{\rho}}{\gamma(t+1)} \quad (10.4b)$$

where $\mathbf{x}^{(t)} := \frac{1}{t+1} \sum_{k=1}^{t+1} \mathbf{x}^k$, $\mathbf{z}^{(t)} := \frac{1}{t+1} \sum_{k=1}^{t+1} \mathbf{z}^k$, and for any \mathbf{C} , $\|\mathbf{z}\|_{\mathbf{C}}^2 := \mathbf{z}^\top \mathbf{C} \mathbf{z}$

- convergence rate: $O(1/t)$
- iteration complexity: $O(1/\varepsilon)$

Fundamental inequality

Define

$$\mathbf{w} := \begin{bmatrix} \mathbf{x} \\ z \\ \lambda \end{bmatrix}, \quad \mathbf{w}^t := \begin{bmatrix} \mathbf{x}^t \\ z^t \\ \lambda^t \end{bmatrix}, \quad \mathbf{G} := \begin{bmatrix} & & \mathbf{A}^\top \\ & & \mathbf{B}^\top \\ -\mathbf{A} & -\mathbf{B} & \end{bmatrix}, \quad \mathbf{d} := \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{b} \end{bmatrix}$$

$$\mathbf{H} := \begin{bmatrix} \mathbf{0} & & \\ & \rho \mathbf{B}^\top \mathbf{B} & \\ & & \rho^{-1} \mathbf{I} \end{bmatrix}, \quad \|\mathbf{w}\|_{\mathbf{H}}^2 := \mathbf{w}^\top \mathbf{H} \mathbf{w}$$

Lemma 10.2

For any \mathbf{x}, z, λ , one has

$$\begin{aligned} F(\mathbf{x}, z) - F(\mathbf{x}^{t+1}, z^{t+1}) + \langle \mathbf{w} - \mathbf{w}^{t+1}, \mathbf{G}\mathbf{w} + \mathbf{d} \rangle \\ \geq \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2 - \frac{1}{2} \|\mathbf{w} - \mathbf{w}^t\|_{\mathbf{H}}^2 \end{aligned}$$

Proof of Theorem 10.1

Set $\mathbf{x} = \mathbf{x}^*$, $\mathbf{z} = \mathbf{z}^*$, and $\mathbf{w} = [\mathbf{x}^{*\top}, \mathbf{z}^{*\top}, \boldsymbol{\lambda}^\top]^\top$ in Lemma 10.2 to reach

$$F(\mathbf{x}^*, \mathbf{z}^*) - F(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) + \langle \mathbf{w} - \mathbf{w}^{k+1}, \mathbf{G}\mathbf{w} + \mathbf{d} \rangle \geq \underbrace{\frac{\|\mathbf{w} - \mathbf{w}^{k+1}\|_H^2}{2} - \frac{\|\mathbf{w} - \mathbf{w}^k\|_H^2}{2}}_{\text{forms telescopic sum}}$$

Summing over all $k = 0, \dots, t$ gives

$$\begin{aligned} & (t+1)F(\mathbf{x}^*, \mathbf{z}^*) - \sum_{k=1}^{t+1} F(\mathbf{x}^k, \mathbf{z}^k) + \left\langle (t+1)\mathbf{w} - \sum_{k=1}^{t+1} \mathbf{w}^k, \mathbf{G}\mathbf{w} + \mathbf{d} \right\rangle \\ & \geq \frac{\|\mathbf{w} - \mathbf{w}^{t+1}\|_H^2 - \|\mathbf{w} - \mathbf{w}^0\|_H^2}{2} \end{aligned}$$

If we define

$$\mathbf{w}^{(t)} = \frac{1}{t+1} \sum_{k=1}^{t+1} \mathbf{w}^k, \quad \mathbf{x}^{(t)} = \frac{1}{t+1} \sum_{k=1}^{t+1} \mathbf{x}^k, \quad \mathbf{z}^{(t)} = \frac{1}{t+1} \sum_{k=1}^{t+1} \mathbf{z}^k, \quad \boldsymbol{\lambda}^{(t)} = \frac{1}{t+1} \sum_{k=1}^{t+1} \boldsymbol{\lambda}^k,$$

then from convexity of F we have

$$F(\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) - \underbrace{F(\mathbf{x}^*, \mathbf{z}^*)}_{= F^{\text{opt}}} + \langle \mathbf{w}^{(t)} - \mathbf{w}, \mathbf{G}\mathbf{w} + \mathbf{d} \rangle \leq \frac{1}{2(t+1)} \|\mathbf{w} - \mathbf{w}^0\|_H^2$$

Proof of Theorem 10.1

Further, we claim that

$$\langle \mathbf{w}^{(t)} - \mathbf{w}, \mathbf{G}\mathbf{w} + \mathbf{d} \rangle = \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x}^{(t)} + \mathbf{B}\mathbf{z}^{(t)} - \mathbf{b} \rangle \quad (10.5)$$

which together with preceding bounds yields

$$\begin{aligned} F(\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) - F^{\text{opt}} + \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x}^{(t)} + \mathbf{B}\mathbf{z}^{(t)} - \mathbf{b} \rangle &\leq \frac{1}{2(t+1)} \|\mathbf{w} - \mathbf{w}^0\|_H^2 \\ &= \frac{1}{2(t+1)} \left\{ \|z - z^0\|_{\rho \mathbf{B}^\top \mathbf{B}}^2 + \frac{1}{\rho} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_2^2 \right\} \end{aligned}$$

Notably, this holds for any $\boldsymbol{\lambda}$

Taking maximum of both sides over $\{\boldsymbol{\lambda} \mid \|\boldsymbol{\lambda}\|_2 \leq \gamma\}$ yields

$$\begin{aligned} F(\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) - F^{\text{opt}} + \gamma \|\mathbf{A}\mathbf{x}^{(t)} + \mathbf{B}\mathbf{z}^{(t)} - \mathbf{b}\|_2 \\ \leq \frac{\left\{ \|z - z^0\|_{\rho \mathbf{B}^\top \mathbf{B}}^2 + \frac{(\gamma + \|\boldsymbol{\lambda}^0\|_2)^2}{\rho} \right\}}{2(t+1)} \end{aligned} \quad (10.6)$$

which immediately establishes (10.4a)

Proof of Theorem 10.1 (cont.)

Caution needs to be exercised since, in general, (10.6) does not establish (10.4b), since $F(\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) - F^{\text{opt}}$ may be negative (as $(\mathbf{x}^{(t)}, \mathbf{z}^{(t)})$ is not guaranteed to be feasible)

Fortunately, if $\gamma \geq 2\|\boldsymbol{\lambda}^*\|_2$, then standard results (e.g. Theorem 3.60 in Beck '18) reveal that $F(\mathbf{x}^{(t)}, \mathbf{z}^{(t)}) - F^{\text{opt}}$ will not be “too negative”, thus completing proof

Proof of Theorem 10.1

Finally, we prove (10.5). Observe that

$$\begin{aligned}\langle \mathbf{w}^{(t)} - \mathbf{w}, \mathbf{G}\mathbf{w} + \mathbf{d} \rangle &= \underbrace{\langle \mathbf{w}^{(t)} - \mathbf{w}, \mathbf{G}(\mathbf{w} - \mathbf{w}^{(t)}) \rangle}_{=0 \text{ since } \mathbf{G} \text{ is skew-symmetric}} + \langle \mathbf{w}^{(t)} - \mathbf{w}, \mathbf{G}\mathbf{w}^{(t)} + \mathbf{d} \rangle \\ &= \langle \mathbf{w}^{(t)} - \mathbf{w}, \mathbf{G}\mathbf{w}^{(t)} + \mathbf{d} \rangle\end{aligned}\quad (10.7)$$

To further simplify this inner product, we use $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{b}$ to obtain

$$\begin{aligned}\langle \mathbf{w}^{(t)} - \mathbf{w}, \mathbf{G}\mathbf{w}^{(t)} + \mathbf{d} \rangle &= \langle \mathbf{x}^{(t)} - \mathbf{x}^*, \mathbf{A}^\top \boldsymbol{\lambda}^{(t)} \rangle + \langle \mathbf{z}^{(t)} - \mathbf{z}^*, \mathbf{B}^\top \boldsymbol{\lambda}^{(t)} \rangle \\ &\quad + \langle \boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}, -\mathbf{A}\mathbf{x}^{(t)} - \mathbf{B}\mathbf{z}^{(t)} + \mathbf{b} \rangle \\ &= \langle -\mathbf{A}\mathbf{x}^* - \mathbf{B}\mathbf{z}^* + \mathbf{b}, \boldsymbol{\lambda}^{(t)} \rangle + \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x}^{(t)} + \mathbf{B}\mathbf{z}^{(t)} - \mathbf{b} \rangle \\ &= \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x}^{(t)} + \mathbf{B}\mathbf{z}^{(t)} - \mathbf{b} \rangle\end{aligned}$$

Proof of Lemma 10.2

To begin with, ADMM update rule requires

$$\begin{aligned} -\rho \mathbf{A}^\top \left(\mathbf{A} \mathbf{x}^{t+1} + \mathbf{B} \mathbf{z}^t - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right) &\in \partial f_1(\mathbf{x}^{t+1}) \\ -\rho \mathbf{B}^\top \left(\mathbf{A} \mathbf{x}^{t+1} + \mathbf{B} \mathbf{z}^{t+1} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right) &\in \partial f_2(\mathbf{z}^{t+1}) \end{aligned}$$

Therefore, for any \mathbf{x}, \mathbf{z} ,

$$\begin{aligned} f_1(\mathbf{x}) - f_1(\mathbf{x}^{t+1}) + \left\langle \rho \mathbf{A}^\top \left(\mathbf{A} \mathbf{x}^{t+1} + \mathbf{B} \mathbf{z}^t - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right), \mathbf{x} - \mathbf{x}^{t+1} \right\rangle &\geq 0 \\ f_2(\mathbf{z}) - f_2(\mathbf{z}^{t+1}) + \left\langle \rho \mathbf{B}^\top \left(\mathbf{A} \mathbf{x}^{t+1} + \mathbf{B} \mathbf{z}^{t+1} - \mathbf{b} + \frac{1}{\rho} \boldsymbol{\lambda}^t \right), \mathbf{z} - \mathbf{z}^{t+1} \right\rangle &\geq 0 \end{aligned}$$

Proof of Lemma 10.2 (cont.)

Using $\lambda^{t+1} = \lambda^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b})$, setting $\tilde{\lambda}^t := \lambda^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^t - \mathbf{b})$, and adding above two inequalities give

$$\begin{aligned} & F(\mathbf{x}, \mathbf{z}) - F(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}) \\ & + \left\langle \begin{bmatrix} \mathbf{x} - \mathbf{x}^{t+1} \\ \mathbf{z} - \mathbf{z}^{t+1} \\ \lambda - \tilde{\lambda}^t \end{bmatrix}, \begin{bmatrix} \mathbf{A}^\top \tilde{\lambda}^t \\ \mathbf{B}^\top \tilde{\lambda}^t \\ -\mathbf{A}\mathbf{x}^{t+1} - \mathbf{B}\mathbf{z}^{t+1} + \mathbf{b} \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \rho \mathbf{B}^\top \mathbf{B}(\mathbf{z}^t - \mathbf{z}^{t+1}) \\ \frac{1}{\rho}(\lambda^t - \lambda^{t+1}) \end{bmatrix} \right\rangle \geq 0 \end{aligned} \tag{10.8}$$

Next, we'd like to simplify above inner product. Let $\mathbf{C} := \rho \mathbf{B}^\top \mathbf{B}$, then

$$(\mathbf{z} - \mathbf{z}^{t+1})^\top \mathbf{C}(\mathbf{z}^t - \mathbf{z}^{t+1}) = \frac{1}{2} \|\mathbf{z} - \mathbf{z}^{t+1}\|_{\mathbf{C}}^2 - \frac{1}{2} \|\mathbf{z} - \mathbf{z}^t\|_{\mathbf{C}}^2 + \frac{1}{2} \|\mathbf{z}^t - \mathbf{z}^{t+1}\|_{\mathbf{C}}^2$$

Proof of Lemma 10.2 (cont.)

Also,

$$\begin{aligned} & 2(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1})^\top (\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t+1}) \\ &= \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1}\|_2^2 - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|_2^2 + \|\tilde{\boldsymbol{\lambda}}^t - \boldsymbol{\lambda}^t\|_2^2 - \|\tilde{\boldsymbol{\lambda}}^t - \boldsymbol{\lambda}^{t+1}\|_2^2 \\ &= \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1}\|_2^2 - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|_2^2 + \rho^2 \|\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^t - \mathbf{b}\|_2^2 \\ &\quad - \|\boldsymbol{\lambda}^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^t - \mathbf{b}) - \boldsymbol{\lambda}^t - \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{b})\|_2^2 \\ &= \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1}\|_2^2 - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|_2^2 + \rho^2 \|\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^t - \mathbf{b}\|_2^2 \\ &\quad - \rho^2 \|\mathbf{B}(\mathbf{z}^t - \mathbf{z}^{t+1})\|_2^2 \end{aligned}$$

which implies that

$$\begin{aligned} & 2(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1})^\top (\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t+1}) \\ & \geq \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t+1}\|_2^2 - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|_2^2 - \rho^2 \|\mathbf{B}(\mathbf{z}^t - \mathbf{z}^{t+1})\|_2^2 \end{aligned}$$

Proof of Lemma 10.2 (cont.)

Combining above results gives

$$\begin{aligned} & \left\langle \begin{bmatrix} \mathbf{x} - \mathbf{x}^{t+1} \\ \mathbf{z} - \mathbf{z}^{t+1} \\ \boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}^t \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \rho \mathbf{B}^\top \mathbf{B} (\mathbf{z}^t - \mathbf{z}^{t+1}) \\ \frac{1}{\rho} (\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t+1}) \end{bmatrix} \right\rangle \\ & \geq \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{t+1}\|_H^2 - \frac{1}{2} \|\mathbf{w} - \mathbf{w}^t\|_H^2 + \frac{1}{2} \|\mathbf{z}^t - \mathbf{z}^{t+1}\|_C^2 - \frac{\rho}{2} \|\mathbf{B}(\mathbf{z}^t - \mathbf{z}^{t+1})\|_2^2 \\ & = \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{t+1}\|_H^2 - \frac{1}{2} \|\mathbf{w} - \mathbf{w}^t\|_H^2 \end{aligned}$$

This together with (10.8) yields

$$\begin{aligned} & F(\mathbf{x}, \mathbf{z}) - F(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}) + \langle \mathbf{w} - \mathbf{w}^{t+1}, \mathbf{G}\mathbf{w}^{t+1} + \mathbf{d} \rangle \\ & \geq \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{t+1}\|_H^2 - \frac{1}{2} \|\mathbf{w} - \mathbf{w}^t\|_H^2 \end{aligned}$$

Since \mathbf{G} is skew-symmetric, repeating prior argument in (10.7) gives

$$\langle \mathbf{w} - \mathbf{w}^{t+1}, \mathbf{G}\mathbf{w}^{t+1} + \mathbf{d} \rangle = \langle \mathbf{w} - \mathbf{w}^{t+1}, \mathbf{G}\mathbf{w} + \mathbf{d} \rangle$$

This immediately completes proof

Convergence of ADMM in practice

- ADMM is slow to converge to high accuracy
- ADMM often converges to modest accuracy within a few tens of iterations, which is sufficient for many large-scale applications

Beyond two-block models

Convergence is not guaranteed when there are 3 or more blocks

- e.g. consider solving

$$x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + x_3 \mathbf{a}_3 = \mathbf{0}$$

where

$$[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

3-block ADMM is divergent for solving this problem (Chen et al. '16)

Reference

- "*Distributed optimization and statistical learning via the alternating direction method of multipliers*," S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *Foundations and Trends in Machine learning*, 2011.
- "*A First Course in Convex Optimization Theory*," E. Ryu, W. Yin.
- "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.
- "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.
- "*Optimization methods for large-scale systems, EE236C lecture notes*," L. Vandenberghe, UCLA.
- "*Large scale optimization for machine learning, ISE633 lecture notes*," M. Razaviyayn, USC.

Reference

- "*Modern big data optimization, IE487/587 lecture notes*," M. Hong, ISU.
- "*Convex optimization and algorithms*," D. Bertsekas, 2015.
- "*Proximal algorithms*," N. Parikh and S. Boyd, *Foundations and Trends in Optimization*, 2013.
- "*The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent*," C. Chen, B. He, Y. Ye, X. Yuan, *Mathematical Programming*, 2016.
- "*Robust principal component analysis?*," E. Candes, X. Li, Y. Ma, J. Wright, *Journal of the ACM*, 2011.
- "*Sparse inverse covariance estimation with the graphical lasso*," J. Friedman, T. Hastie, and R. Tibshirani, *Biostatistics*, 2008.