Alternating direction method of multipliers

Ye Shi

ShanghaiTech University

Outline

- Augmented Lagrangian method

  $\Downarrow$

- Alternating direction method of multipliers

Two-block problem

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

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

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x},\boldsymbol{z}} & f_1(\boldsymbol{x}) + f_2(\boldsymbol{z}) \\ \text{subject to} & \lambda \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{b} \end{array}$$

$$\Updownarrow$$

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

Augmented Lagrangian method

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

The proximal point method for solving this dual problem:

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

$$\Updownarrow$$

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

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

(x^*, z^*)

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \rho\left(\boldsymbol{A}\boldsymbol{x}^{t+1} + \boldsymbol{B}\boldsymbol{z}^{t+1} - \boldsymbol{b}\right) \qquad (10.1)$$

λ^*

$$\begin{cases} \min & f(x) \\ \text{s.t.} & h(x)=0 \end{cases}$$

optimal

$$\left\{ \min \quad L(x,\lambda) = f(x) + \lambda h(x) \right.$$

λ

loss: Cross-Entropy $+\lambda$ Regularization

Justification of (10.1)

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

$\Updownarrow$ optimality condition

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

$\Updownarrow$

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

where (check: use the conjugate subgradient theorem)

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

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

$$0 \in A^\top\lambda^{t+1} + \partial f_1(x^{t+1}) \Rightarrow -A^\top\lambda^{t+1} \in \partial f_1(x^{t+1}) \Rightarrow x^{t+1} \in \partial f^*(-A^\top\lambda^{t+1})$$

Justification of (10.1)

$$\Updownarrow$$

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

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

$$\Updownarrow$$

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

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

$$\Updownarrow$$

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

Augmented Lagrangian method (ALM)

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

(primal step)

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

(dual step)

where $\rho > 0$ is penalty parameter

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

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

Issues of augmented Lagrangian method

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

- the primal update step is often expensive — as expensive as solving the original problem
- minimization of $\boldsymbol{x}$ and $\boldsymbol{z}$ 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 $\boldsymbol{x}$ and $\boldsymbol{z}$ sequentially via alternating minimization

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

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

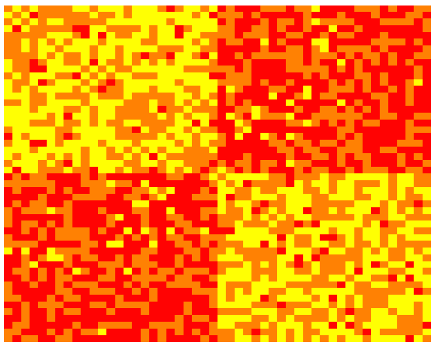
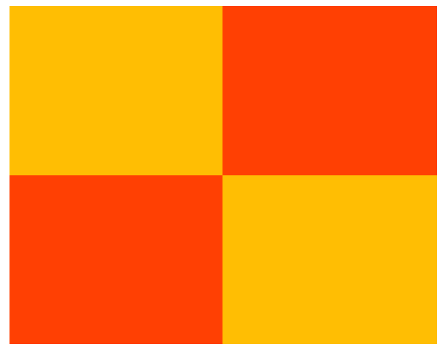
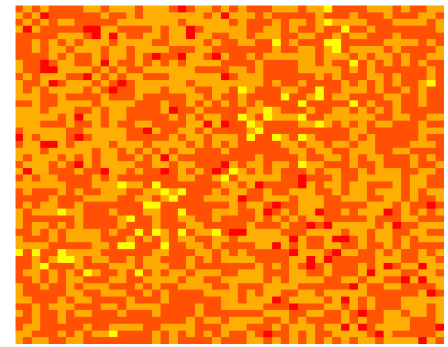
$$\| x^{t+1} - x^* \| \le \ldots \quad \| f(x^{t+1}) - f(x^*) \|$$

Alternating direction method of multipliers

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

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

Example: robust PCA

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

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

Can we hope to disentangle $\boldsymbol{L}$ and $\boldsymbol{S}$?

Example: robust PCA

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

$$\begin{aligned} \text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \quad & \overbrace{\|\boldsymbol{L}\|_*}^{f_1(x)} + \lambda \overbrace{\|\boldsymbol{S}\|_1}^{f_2(z)} \\ \text{s.t.} \quad & \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{M} \end{aligned} \qquad (10.2)$$

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

$$\min_{L} \quad \|L\|_* + \frac{\rho}{2} \left\| L + S^t - M + \frac{1}{\rho} \Lambda^t \right\|_F^2$$

Example: robust PCA

ADMM for solving (10.2):

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

Example: robust PCA

This is equivalent to

$$\boldsymbol{L}^{t+1} = \mathsf{SVT}_{\rho^{-1}}\Big(\boldsymbol{M} - \boldsymbol{S}^t - \frac{1}{\rho}\boldsymbol{\Lambda}^t\Big) \qquad \text{(singular value thresholding)}$$

$$\boldsymbol{S}^{t+1} = \mathsf{ST}_{\lambda\rho^{-1}}\Big(\boldsymbol{M} - \boldsymbol{L}^{t+1} - \frac{1}{\rho}\boldsymbol{\Lambda}^t\Big) \qquad \text{(soft thresholding)}$$

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

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

$$\mathsf{SVT}_\tau(\boldsymbol{X}) = \boldsymbol{U}\mathrm{diag}\big(\{(\sigma_i - \tau)_+\}\big)\boldsymbol{V}^\top$$

and

$$\big(\mathsf{ST}_\tau(\boldsymbol{X})\big)_{i,j} = \begin{cases} X_{i,j} - \tau, & \text{if } X_{i,j} > \tau \\ 0, & \text{if } |X_{i,j}| \le \tau \\ X_{i,j} + \tau, & \text{if } X_{i,j} < -\tau \end{cases}$$

$$\mathrm{prox}_{\|\cdot\|_*}(X) = \arg\min_{Z\in R^{m\times n}} \frac{1}{2}\|X - Z\|_F^2 + \lambda \|Z\|_* \quad \checkmark$$

$$X = U \Sigma V^T \quad (\text{SVD decomposition})$$

$$\Sigma = \mathrm{diag}(\sigma_1, \cdots \sigma_{\min(m,n)}) \quad \checkmark$$

$$\arg\min_{Z\in R^{m\times n}} \frac{1}{2}\| U\Sigma V^T - Z\|_F^2 + \lambda \|Z\|_*$$

$$= \arg\min_{Z\in R^{m\times n}} \frac{1}{2}\| \Sigma - U^T Z V\|_F^2 + \lambda \|Z\|_*$$

$$\tilde{Z} := U^T Z V \Rightarrow Z = U \cdot \tilde{Z} V^T$$

$$\mathrm{prox}_{\|\cdot\|_*}(X) = U \left(\arg\min_{\tilde{Z}} \frac{1}{2}\| \Sigma - \tilde{Z}\|_F^2 + \lambda \| U \tilde{Z} V^T \|_* \right) V^T$$

$$\|\tilde{Z}\|_* = \|Z\|_*, \quad \tilde{Z} = U_{\tilde{Z}} \cdot \Sigma_{\tilde{Z}} \cdot V_{\tilde{Z}}^T,$$

$$Z = U \cdot \tilde{Z} V^T = (U \cdot U_{\tilde{Z}}) \Sigma_{\tilde{Z}} (V_{\tilde{Z}}^T \cdot V^T)$$

$$\operatorname{prox}_{\|\cdot\|_*}(X) = U\Big(\underbrace{\arg\min_{\tilde{Z}} \tfrac{1}{2}\|\Sigma - \tilde{Z}\|_F^2 + \lambda \|\tilde{Z}\|_*}_{\operatorname{prox}_{\|\cdot\|_*}(\Sigma)}\Big) V^T$$

$\tilde{Z}$ is diagonal to ensure the minimization

$$\operatorname{prox}_{\|\cdot\|_*}(\Sigma) = \arg\min_{\tilde{Z}} \sum_i \Big[\underbrace{\tfrac{1}{2}(\tilde{z}_{ii} - \sigma_i)^2 + \lambda \tilde{z}_{ii}}_{\nabla f = 0}\Big]$$

$$\Rightarrow \quad \tilde{z}_{ii} - \sigma_i + \lambda = 0$$

$$\Rightarrow \quad \tilde{z}_{ii} = \sigma_i - \lambda$$

$$\tilde{z}_{ii} = [\sigma_i - \lambda]_+$$

$$Z = U \cdot \operatorname{diag}([\sigma_i - \lambda]_+)\, V^T$$

Example: graphical lasso

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

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

$$\Updownarrow$$

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

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

$$f(\theta) = \log\det\theta, \qquad \theta \geq 0$$

$$\Downarrow$$

define $g(t) = \log(\det(\theta + tV))$, $\quad \theta + tV > 0$

$$g(t) = \log\det\left(\theta^{\frac{1}{2}}\cdot\theta^{\frac{1}{2}} + t\cdot\theta^{\frac{1}{2}}\theta^{-\frac{1}{2}}\cdot V\cdot\theta^{-\frac{1}{2}}\cdot\theta^{\frac{1}{2}}\right)$$

$$= \log\det\left(\theta^{\frac{1}{2}}\left(I + t\,\theta^{-\frac{1}{2}}V\theta^{-\frac{1}{2}}\right)\theta^{\frac{1}{2}}\right)$$

$$\det(AB) = \det(A)\cdot\det(B)$$

$$g(t) = \log\det(\theta) + \log\det\left(I + t\,\underbrace{\theta^{-\frac{1}{2}}V\cdot\theta^{-\frac{1}{2}}}_{\lambda_1,\cdots\lambda_n}\right)$$

$$= \log\det(\theta) + \sum_{i=1}^{n}\log(1 + t\lambda_i)$$

$$g'(t) = \sum_{i=1}^{n}\frac{\lambda_i}{1 + t\lambda_i}, \quad g''(t) = -\sum_{i=1}^{n}\frac{\lambda_i^2}{(1+t\lambda_i)^2} \leq 0$$

$g(t)$ concave $\therefore$ $f(\theta)$ concave

Example: graphical lasso

ADMM for solving (10.3):

$$
\begin{aligned}
\boldsymbol{\Theta}^{t+1} &= \arg\min_{\boldsymbol{\Theta}\succeq\mathbf{0}} \left\{ -\log\det\boldsymbol{\Theta} + \frac{\rho}{2}\left\|\boldsymbol{\Theta} - \boldsymbol{\Psi}^{t} + \frac{1}{\rho}\boldsymbol{\Lambda}^{t} + \frac{1}{\rho}\boldsymbol{S}\right\|_{\mathrm{F}}^{2} \right\} \\
\boldsymbol{\Psi}^{t+1} &= \arg\min_{\boldsymbol{\Psi}} \left\{ \lambda\|\boldsymbol{\Psi}\|_{1} + \frac{\rho}{2}\left\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Psi} + \frac{1}{\rho}\boldsymbol{\Lambda}^{t}\right\|_{\mathrm{F}}^{2} \right\} \\
\boldsymbol{\Lambda}^{t+1} &= \boldsymbol{\Lambda}^{t} + \rho\left(\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Psi}^{t+1}\right)
\end{aligned}
$$

Example: graphical lasso

This is equivalent to

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

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

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

Example: consensus optimization

Consider solving the following minimization problem

$$\text{minimize}_{\boldsymbol{x}} \quad \sum_{i=1}^{N} f_i(\boldsymbol{x})$$

$$\min_{x \in C_i} \int f_i(x)$$

$$\Updownarrow$$

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(\boldsymbol{x}_i) \quad \text{(block separable)}$$

$$\text{s.t.} \quad \boldsymbol{x}_i = \boldsymbol{z} \qquad 1 \le i \le N$$

$$X_1 \cdots X_n$$

$$\begin{aligned} x_1 &= z \\ x_2 &= z \\ &\vdots \\ x_n &= z \end{aligned} \iff \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} z}_{u}$$

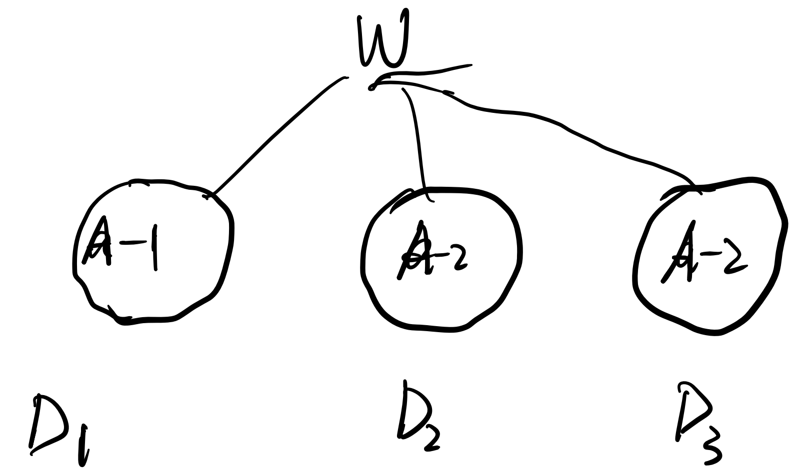
$$\Updownarrow$$

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(\boldsymbol{x}_i)$$

$$\text{s.t.} \quad \boldsymbol{u} := \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} \\ \vdots \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{z}$$

$$Ax + Bz = b$$

$$I \cdot u + \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} z = 0$$



$$W_1 = W_2 = W_3$$

$$W_1 = z$$

Example: consensus optimization

ADMM for solving this problem:

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

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

$$\boldsymbol{\lambda}_i^{t+1} = \boldsymbol{\lambda}_i^t + \rho\left(\boldsymbol{x}_i^{t+1} - \boldsymbol{z}^{t+1}\right), \qquad 1 \le i \le N$$

Example: consensus optimization

This is equivalent to

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

(can be computed in parallel)

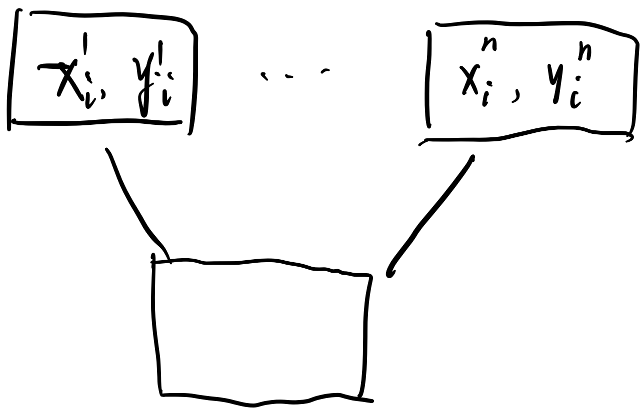
$$\boldsymbol{z}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{x}_i^{t+1} + \frac{1}{\rho} \boldsymbol{\lambda}_i^t \right)$$

(gather all local iterates)

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

("broadcast" $\boldsymbol{z}^{t+1}$ to update all local multipliers)

ADMM is well suited for distributed optimization!



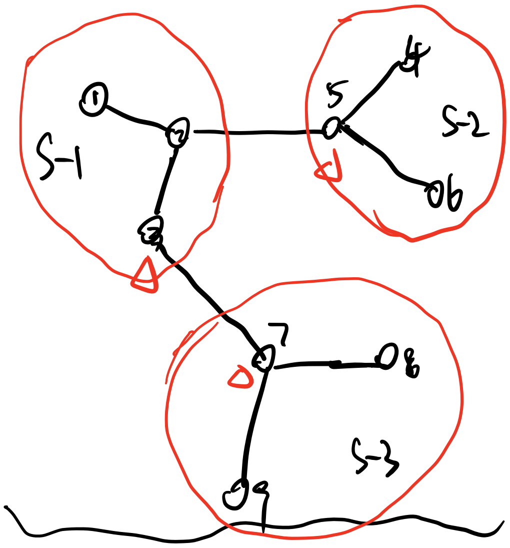
$$\arg\min_{W^1} \|W^1 X^1 - Y^1\|_2^2$$

$$\arg\min_{W^n} \|W^n X^n - Y^n\|_2^2$$

$$\begin{cases} \sum_{k=1}^{n} \|W^k X^k - Y^k\|_2^2 \\ \text{s.t. } W^k = Z, \quad k = 1, \dots, n \end{cases}$$

Centralized

1 Server to control



$$\min_{x_1 \cdots x_9} f(x_1, \cdots x_9)$$

$$\min f_1(x_1, x_2, x_3) + f_2(x_4, x_5, x_6) + f_3(x_7, x_8, x_9)$$

S_1': 1. 2. 3. 5. 7

S_2': 4. 5. 6. 2

S_3': 7. 8. 9. 3

$\min f_1(x_1, x_2, x_3)$ —

$\min f_2(x_4, x_5, x_6)$ —

$\min f_3(x_7, x_8, x_9)$ —

$\min f_1(x_1. x_2, x_3, x_5, x_7)$ —

$\min f_2(x_4, 5. 6. 2)$ —

$\min f_3(x_7. 8. 9. 3)$ —

Z: 1-9

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 $\gamma \geq 2\|\boldsymbol{\lambda}^*\|_2$. Then

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

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

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

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

Fundamental inequality

Define

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

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

Lemma 10.2

For any $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\lambda}$, one has

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

$$Gw + d = \begin{bmatrix} 0 & 0 & A^T \\ 0 & 0 & B^T \\ -A & -B & 0 \end{bmatrix} \begin{bmatrix} x \\ z \\ \lambda \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix}$$

$$= \begin{bmatrix} A^T\lambda \\ B^T\lambda \\ -Ax - Bz + b \end{bmatrix}$$

$$\langle w^t - w,\ Gw + d \rangle = \left\langle \begin{bmatrix} x^{(t)} - x \\ z^{(t)} - z \\ \lambda^{(t)} - \lambda \end{bmatrix},\ \begin{bmatrix} A^T\lambda \\ B^T\lambda \\ -Ax - Bz + b \end{bmatrix} \right\rangle$$

$$= (x^{(t)T} - x^T)(A^T\lambda) + (z^{(t)T} - z^T)(B^T\lambda)$$

$$+ (\lambda^{(t)T} - \lambda^T)(-Ax - Bz + b)$$

$$= \lambda^{(t)T} \underbrace{(-Ax - Bz + b)}_{0} + \langle \lambda,\ Ax^{(t)} + Bz^{(t)} - b \rangle$$

Proof of Theorem 10.1

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

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

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

$$(t+1)F(\boldsymbol{x}^*, \boldsymbol{z}^*) - \sum_{k=1}^{t+1} F(\boldsymbol{x}^k, \boldsymbol{z}^k) + \left\langle (t+1)\boldsymbol{w} - \sum_{k=1}^{t+1} \boldsymbol{w}^k, \boldsymbol{G}\boldsymbol{w} + \boldsymbol{d} \right\rangle$$
$$\geq \frac{\|\boldsymbol{w} - \boldsymbol{w}^{t+1}\|_{\boldsymbol{H}}^2 - \|\boldsymbol{w} - \boldsymbol{w}^{0}\|_{\boldsymbol{H}}^2}{2}$$

If we define

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

then from convexity of F we have

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

$$F\left(\frac{1}{t+1}\sum_{k=1}^{t+1} x^k, \frac{1}{t+1}\sum_{k=1}^{t+1} z^k\right)$$

$$= F(x^{(t)}, z^{(t)})$$

$$\leq \frac{1}{t+1}\sum_{k=1}^{t+1} F(x^k, z^k)$$

Proof of Theorem 10.1

Further, we claim that

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

which together with preceding bounds yields

$$F(\boldsymbol{x}^{(t)}, \boldsymbol{z}^{(t)}) - F^{\mathsf{opt}} + \left\langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x}^{(t)} + \boldsymbol{B}\boldsymbol{z}^{(t)} - \boldsymbol{b} \right\rangle \leq \frac{1}{2(t+1)} \|\boldsymbol{w} - \boldsymbol{w}^0\|_{\boldsymbol{H}}^2$$

$$= \frac{1}{2(t+1)} \left\{ \|\boldsymbol{z} - \boldsymbol{z}^0\|_{\rho \boldsymbol{B}^\top \boldsymbol{B}}^2 + \frac{1}{\rho} \underbrace{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^0\|_2^2}_{\leq (\|\lambda\|_2 + \|\lambda^0\|_2)^2} \right\}$$

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

Taking maximum of both sides over $\{\boldsymbol{\lambda} \mid \|\boldsymbol{\lambda}\|_2 \leq \gamma\}$ yields $\leq (\gamma + \|\lambda^0\|_2)^2$

$$F(\boldsymbol{x}^{(t)}, \boldsymbol{z}^{(t)}) - F^{\mathsf{opt}} + \gamma \|\boldsymbol{A}\boldsymbol{x}^{(t)} + \boldsymbol{B}\boldsymbol{z}^{(t)} - \boldsymbol{b}\|_2$$

$$\leq \frac{\left\{ \|\boldsymbol{z} - \boldsymbol{z}^0\|_{\rho \boldsymbol{B}^\top \boldsymbol{B}}^2 + \frac{\left(\gamma + \|\boldsymbol{\lambda}^0\|_2\right)^2}{\rho} \right\}}{2(t+1)} \tag{10.6}$$

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 $F(\boldsymbol{x}^{(t)}, \boldsymbol{z}^{(t)}) - F^{\mathsf{opt}}$ may be negative (as $(\boldsymbol{x}^{(t)}, \boldsymbol{z}^{(t)})$ 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 $F(\boldsymbol{x}^{(t)}, \boldsymbol{z}^{(t)}) - F^{\mathsf{opt}}$ will not be "too negative", thus completing proof

Proof of Theorem 10.1

Finally, we prove (10.5). Observe that

$$\left\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}, \boldsymbol{G}\boldsymbol{w} + \boldsymbol{d} \right\rangle = \underbrace{\left\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}, \boldsymbol{G}\overset{0}{(\boldsymbol{w} - \boldsymbol{w}^{(t)})} \right\rangle}_{=0 \text{ since } \boldsymbol{G} \text{ is skew-symmetric}} + \left\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}, \boldsymbol{G}\boldsymbol{w}^{(t)} + \boldsymbol{d} \right\rangle$$

$$= \left\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}, \boldsymbol{G}\boldsymbol{w}^{(t)} + \boldsymbol{d} \right\rangle \tag{10.7}$$

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

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

Proof of Lemma 10.2

To begin with, ADMM update rule requires

$$-\rho \boldsymbol{A}^\top \Big(\boldsymbol{A}\boldsymbol{x}^{t+1} + \boldsymbol{B}\boldsymbol{z}^t - \boldsymbol{b} + \frac{1}{\rho}\boldsymbol{\lambda}^t \Big) \in \partial f_1(\boldsymbol{x}^{t+1})$$

$$-\rho \boldsymbol{B}^\top \Big(\boldsymbol{A}\boldsymbol{x}^{t+1} + \boldsymbol{B}\boldsymbol{z}^{t+1} - \boldsymbol{b} + \frac{1}{\rho}\boldsymbol{\lambda}^t \Big) \in \partial f_2(\boldsymbol{z}^{t+1})$$

Therefore, for any $\boldsymbol{x}, \boldsymbol{z}$,

$$f_1(\boldsymbol{x}) - f_1(\boldsymbol{x}^{t+1}) + \Big\langle \rho \boldsymbol{A}^\top \Big(\boldsymbol{A}\boldsymbol{x}^{t+1} + \boldsymbol{B}\boldsymbol{z}^t - \boldsymbol{b} + \frac{1}{\rho}\boldsymbol{\lambda}^t \Big), \boldsymbol{x} - \boldsymbol{x}^{t+1} \Big\rangle \geq 0$$

$$f_2(\boldsymbol{z}) - f_2(\boldsymbol{z}^{t+1}) + \Big\langle \rho \boldsymbol{B}^\top \Big(\boldsymbol{A}\boldsymbol{x}^{t+1} + \boldsymbol{B}\boldsymbol{z}^{t+1} - \boldsymbol{b} + \frac{1}{\rho}\boldsymbol{\lambda}^t \Big), \boldsymbol{z} - \boldsymbol{z}^{t+1} \Big\rangle \geq 0$$

Proof of Lemma 10.2 (cont.)

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

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

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

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

$(a-b)^\top C(e-f) =$

Proof of Lemma 10.2 (cont.)

Also,

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

which implies that

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

Proof of Lemma 10.2 (cont.)

Combining above results gives

$$\left\langle \begin{bmatrix} \boldsymbol{x} - \boldsymbol{x}^{t+1} \\ \boldsymbol{z} - \boldsymbol{z}^{t+1} \\ \boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}^{t} \end{bmatrix}, \begin{bmatrix} \boldsymbol{0} \\ \rho \boldsymbol{B}^{\top} \boldsymbol{B}(\boldsymbol{z}^{t} - \boldsymbol{z}^{t+1}) \\ \frac{1}{\rho}(\boldsymbol{\lambda}^{t} - \boldsymbol{\lambda}^{t+1}) \end{bmatrix} \right\rangle$$

$$\geq \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}^{t+1}\|_{\boldsymbol{H}}^{2} - \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}^{t}\|_{\boldsymbol{H}}^{2} + \frac{1}{2}\|\boldsymbol{z}^{t} - \boldsymbol{z}^{t+1}\|_{\boldsymbol{C}}^{2} - \frac{\rho}{2}\|\boldsymbol{B}(\boldsymbol{z}^{t} - \boldsymbol{z}^{t+1})\|_{2}^{2}$$

$$= \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}^{t+1}\|_{\boldsymbol{H}}^{2} - \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}^{t}\|_{\boldsymbol{H}}^{2}$$

This together with (10.8) yields

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

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

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

This immediately completes proof

Convergence of ADMM in practice

$O(\frac{1}{\varepsilon})$

- 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

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

where

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

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

Reference

[1] "*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.

[2] "*A First Course in Convex Optimization Theory*," E. Ryu, W. Yin.

[3] "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.

[4] "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.

[5] "*Optimization methods for large-scale systems, EE236C lecture notes*," L. Vandenberghe, UCLA.

[6] "*Large scale optimization for machine learning, ISE633 lecture notes*," M. Razaviyayn, USC.

[7] "*Modern big data optimization, IE487/587 lecture notes*," M. Hong, ISU.

Reference

[8] "*Convex optimization and algorithms*," D. Bertsekas, 2015.

[9] "*Proximal algorithms*," N. Parikh and S. Boyd, *Foundations and Trends in Optimization*, 2013.

[10] "*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.

[11] "*Robust principal component analysis?*," E. Candes, X. Li, Y. Ma, J. Wright, *Journal of the ACM*, 2011.

[12] "*Sparse inverse covariance estimation with the graphical lasso*," J. Friedman, T. Hastie, and R. Tibshirani, *Biostatistics*, 2008.