Distributed nonsmooth composite optimization via the proximal augmented Lagrangian

Neil K. Dhingra
neilkdh.com

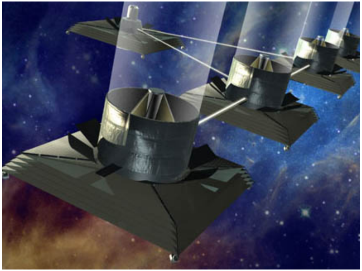
joint work with

Sei Zhen Khong   Mihailo Jovanović

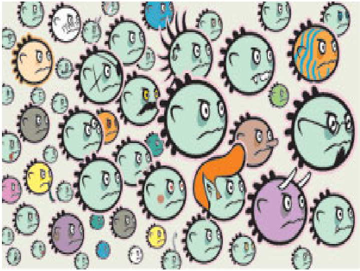
LCCC Focus Period on Large-Scale and Distributed Optimization
June 9, 2017

Applications

SATELLITE FORMATIONS

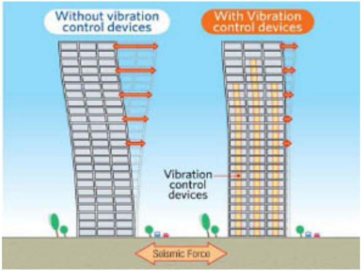


COMBINATION DRUG THERAPY



POWER NETWORKS

CONTROL OF BUILDINGS



Structure via composite optimization

$$\text{minimize} \quad \underbrace{f(x)}_{\text{performance}} \; + \; \underbrace{g(Tx)}_{\text{structure}}$$

minimize $f(x)$ + $g(Tx)$

$\downarrow$ **performance** $\qquad$ $\downarrow$ **structure**

- f – possibly nonconvex; cts-differentiable
- g – convex; often non-differentiable

- Tx – promote structure in alternate coordinates
- $g(x)$ admits easily computable proximal operator, $g(Tx)$ does not

Outline

I Proximal augmented Lagrangian

- centralized approach – method of multipliers

II Primal-dual method

- distributable
- convergence for convex problems
- linear convergence for strongly convex problems

Proximal gradient method

$$\text{minimize} \quad f(x) \; + \; g(x)$$

Generalizes gradient descent

$$x^{k+1} \; = \; \mathbf{prox}_{\alpha_k g}\big(x^k \; - \; \alpha_k \nabla f(x^k)\big)$$

- cannot be used for $g(Tx)$ in general

Nesterov '07
Beck & Teboulle '09

Proximal operator and Moreau envelope

- Proximal operator

$$\mathbf{prox}_{\mu g}(v) \;:=\; \underset{z}{\operatorname{argmin}} \;\; g(z) \;+\; \frac{1}{2\mu}\,\|z \,-\, v\|^2$$

- Moreau envelope

$$M_{\mu g}(v) \;:=\; \inf_{z} \;\; g(z) \;+\; \frac{1}{2\mu}\,\|z \,-\, v\|^2$$

  - **continuously differentiable** even when g is not

$$\nabla M_{\mu g}(v) \;=\; \frac{1}{\mu}\left(v \;-\; \mathbf{prox}_{\mu g}(v)\right)$$

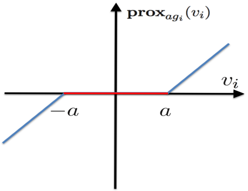
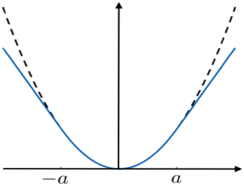
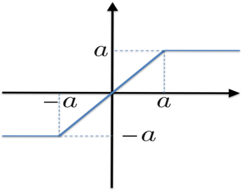
Parikh & Boyd, FnT in Optimization '14

Example

- Soft-thresholding – proximal operator for ℓ_1 norm

$$\underset{z_i}{\text{minimize}} \quad \sum_i \left(\gamma \, |z_i| \; + \; \frac{1}{2\mu} \, (z_i \; - \; v_i)^2 \right)$$

separability $\Rightarrow$ **element-wise analytical solution**

prox operator soft-thresholding	**Moreau envelope** Huber function	∇M saturation
		
	$a \; = \; \mu\gamma$	

Auxiliary variable

$$\underset{x,z}{\text{minimize}} \quad f(x) + g(z)$$
$$\text{subject to} \quad Tx - z = 0$$

- **Decouples** f and g
- Can use methods for constrained optimization

Augmented Lagrangian

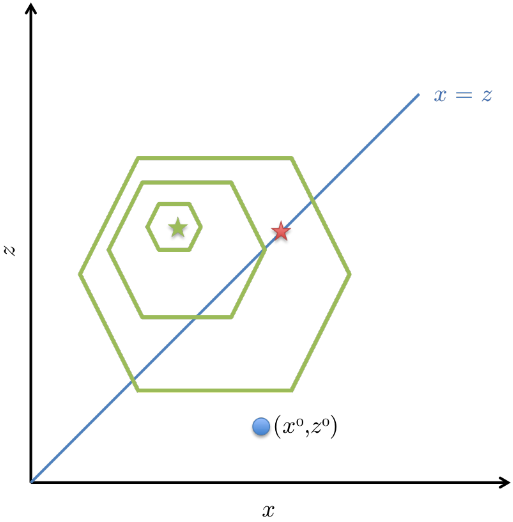
$$\mathcal{L}_\mu(x, z; y) = f(x) + g(z) + \langle y, Tx - z \rangle + \frac{1}{2\mu} \|Tx - z\|^2$$

Method of multipliers

$$
\begin{aligned}
(x^{k+1}, z^{k+1}) &= \operatorname*{argmin}_{x,z} \; \mathcal{L}_\mu(x, z; y^k) \\
y^{k+1} &= y^k + \frac{1}{\mu}(Tx^{k+1} - z^{k+1})
\end{aligned}
$$

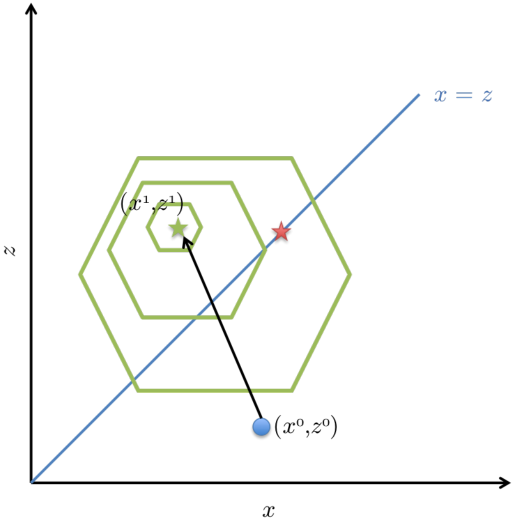
- Gradient ascent on a strengthened dual problem
- Requires *joint* minimization over x and z
- Well-studied: convergence to local minimum, adaptive μ update, inexact subproblems, etc.

MM cartoon



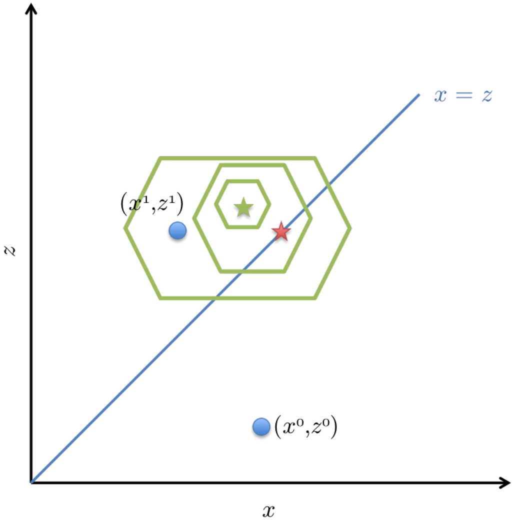
$\mathcal{L}_\mu(x, z; y^0)$

MM CARTOON



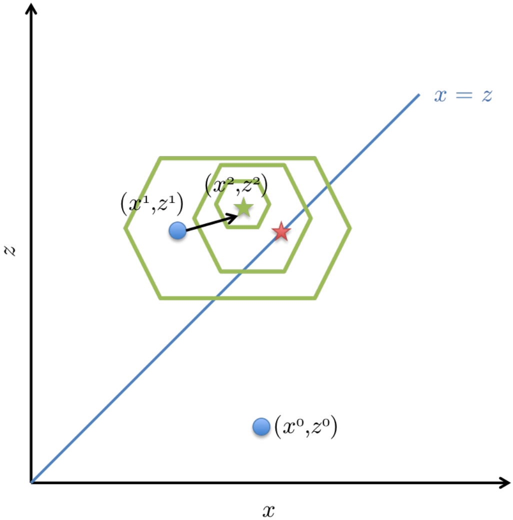
$\mathcal{L}_\mu(x, z; y^0)$

MM cartoon



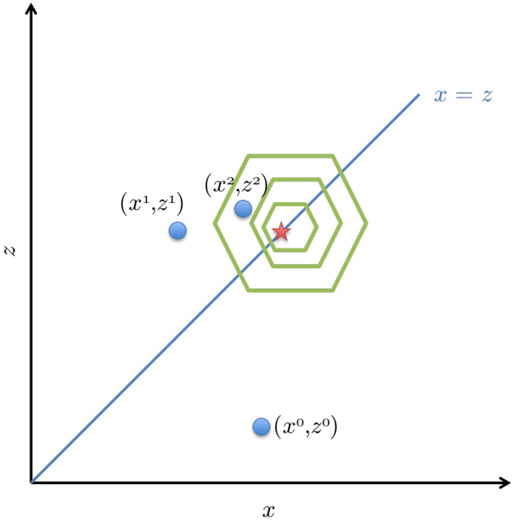
$\mathcal{L}_\mu(x, z; y^1)$

MM cartoon



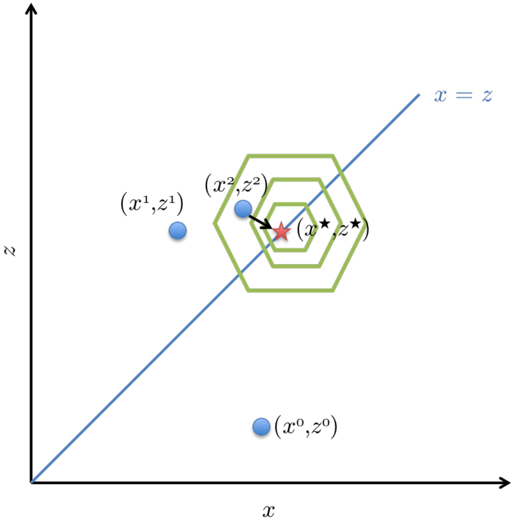
$\mathcal{L}_\mu(x, z; y^1)$

MM CARTOON



$\mathcal{L}_\mu(x, z; y^\star)$

MM cartoon



$\mathcal{L}_\mu(x, z; y^\star)$

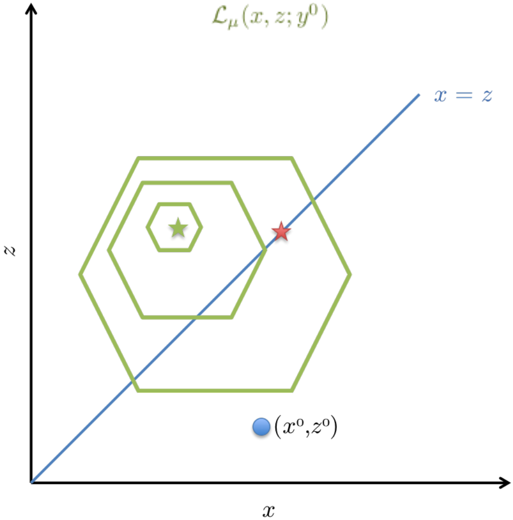
Alternating direction method of multipliers

$$
\begin{aligned}
x^{k+1} &= \underset{x}{\text{argmin}} \ \mathcal{L}_{\mu}(x, z^k; y^k) && \text{differentiable} \\
z^{k+1} &= \underset{z}{\text{argmin}} \ \mathcal{L}_{\mu}(x^{k+1}, z; y^k) && \mathbf{prox}_{\mu g}(\cdot) \\
y^{k+1} &= y^k + \frac{1}{\mu}(Tx^{k+1} - z^{k+1})
\end{aligned}
$$

- Convenient for distributed implementation
- Convergence speed influenced by μ
- **Challenge:** convergence for nonconvex f

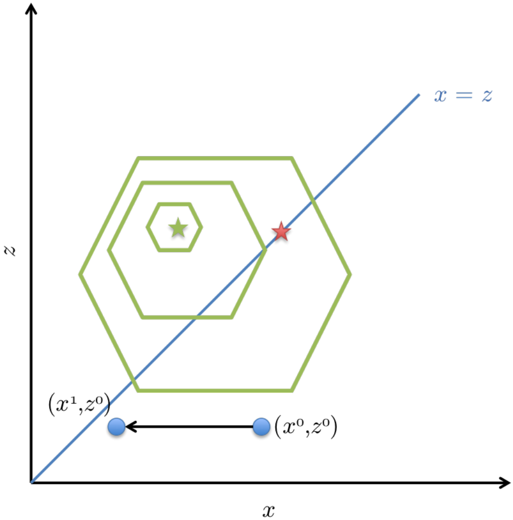
Hong, Luo, Razaviyayn, SIAM J. Optimiz. '16

ADMM cartoon



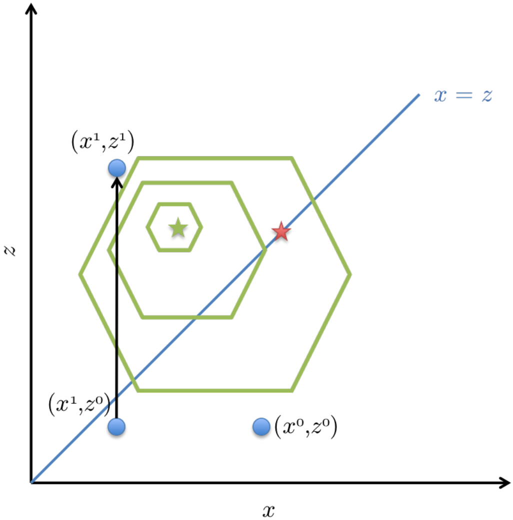
$\mathcal{L}_\mu(x, z; y^0)$

ADMM cartoon



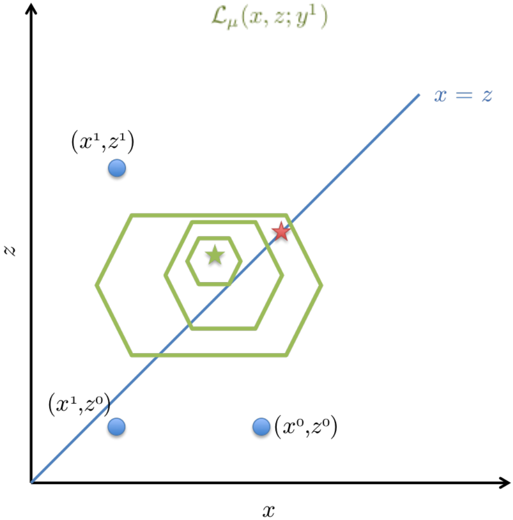
$\mathcal{L}_\mu(x, z; y^0)$

ADMM cartoon



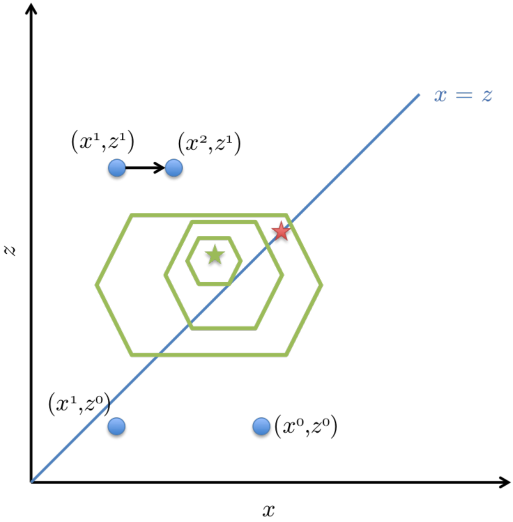
$\mathcal{L}_\mu(x, z; y^0)$

ADMM cartoon



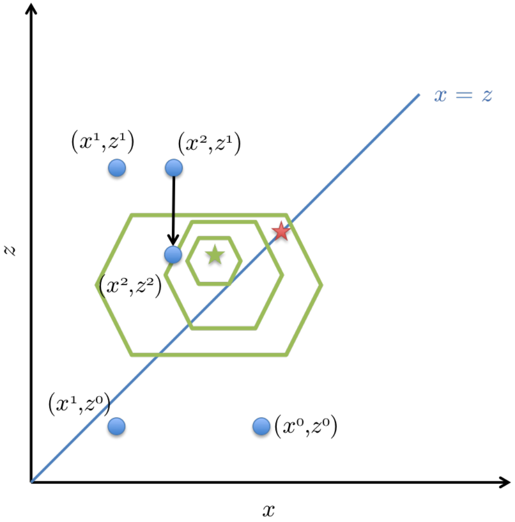
$\mathcal{L}_\mu(x, z; y^1)$

ADMM cartoon



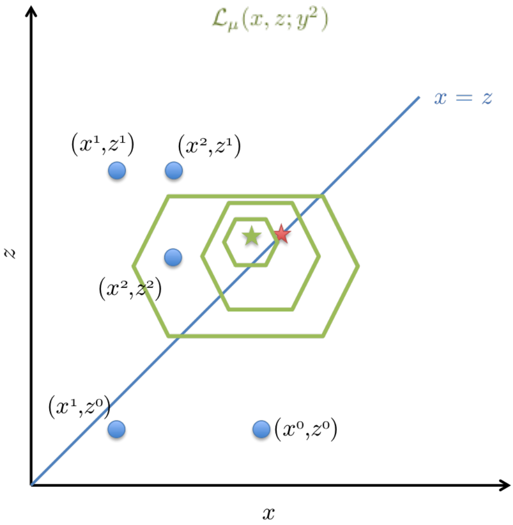
$\mathcal{L}_\mu(x, z; y^1)$

ADMM cartoon



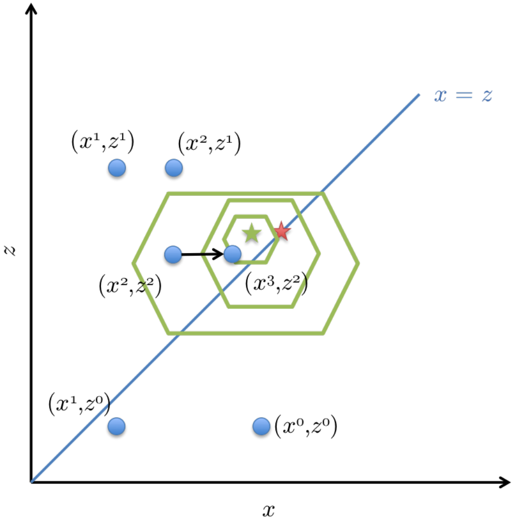
$\mathcal{L}_\mu(x, z; y^1)$

ADMM cartoon



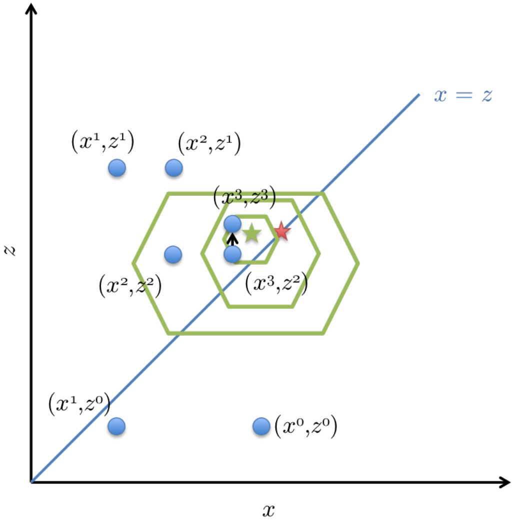
$\mathcal{L}_\mu(x, z; y^2)$

ADMM cartoon



$$\mathcal{L}_\mu(x, z; y^2)$$

ADMM CARTOON



$\mathcal{L}_\mu(x, z; y^2)$

Alternating direction method of multipliers

$$
\begin{aligned}
x^{k+1} &= \underset{x}{\operatorname{argmin}} \; \mathcal{L}_\mu(x, z^k; y^k) && \text{differentiable} \\
z^{k+1} &= \underset{z}{\operatorname{argmin}} \; \mathcal{L}_\mu(x^{k+1}, z; y^k) && \mathbf{prox}_{\mu g}(\cdot) \\
y^{k+1} &= y^k + \frac{1}{\mu}(Tx^{k+1} - z^{k+1})
\end{aligned}
$$

Proximal augmented Lagrangian

$$\mathcal{L}_\mu(x,\, z;\, y) \;=\; f(x) \;+\; \underbrace{g(z) \;+\; \frac{1}{2\mu}\,\|z \;-\; (Tx \;+\; \mu y)\|^2}_{} \;-\; \frac{\mu}{2}\,\|y\|^2$$

Minimize over z

$$z_\mu^\star(x,y) \;=\; \mathbf{prox}_{\mu g}(Tx \;+\; \mu y)$$

Evaluate $\mathcal{L}_\mu(x,z;y)$ at $z^\star$

$$\begin{aligned} \mathcal{L}_\mu(x;\, y) \;&:=\; \mathcal{L}_\mu(x,\, z_\mu^\star(x,y);\, y) \\ &=\; f(x) \;+\; M_{\mu g}(Tx \;+\; \mu y) \;-\; \frac{\mu}{2}\,\|y\|^2 \end{aligned}$$

continuously differentiable in x and y

Dhingra, Khong, Jovanović, arXiv:1610.04514

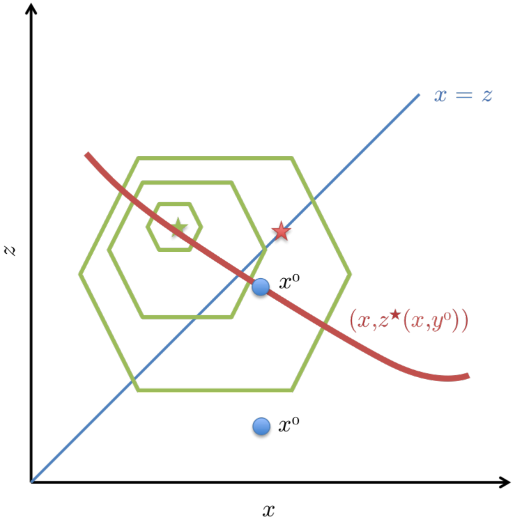
Proximal augmented Lagrangian MM

$$
\begin{aligned}
x^{k+1} &= \underset{x}{\operatorname{argmin}} \; \mathcal{L}_\mu(x;\, y^k) \\
y^{k+1} &= y^k \;+\; \frac{1}{\mu}\,(Tx^{k+1} \;-\; \mathbf{prox}_{\mu g}(Tx^{k+1} \;+\; \mu y^k))
\end{aligned}
$$

- Nonconvex f: convergence to local minimum
- x-minimization step: differentiable problem

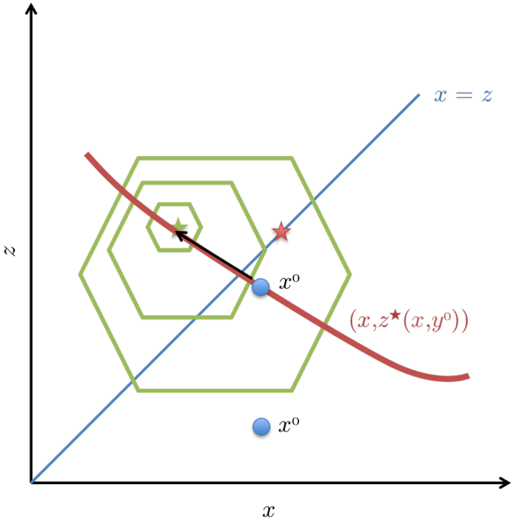
Dhingra, Khong, Jovanović, arXiv:1610.04514

Proximal augmented Lagrangian MM cartoon



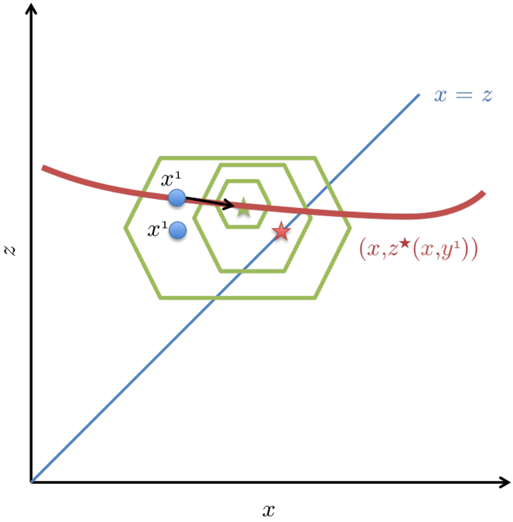
$\mathcal{L}_\mu(x, z; y^0)$, $\mathcal{L}_\mu(x; y^0)$

Proximal augmented Lagrangian MM cartoon

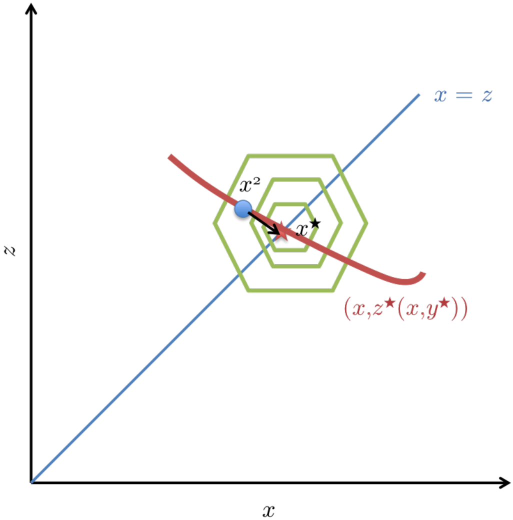


$\mathcal{L}_\mu(x, z; y^0)$, $\mathcal{L}_\mu(x; y^0)$

Proximal augmented Lagrangian MM cartoon

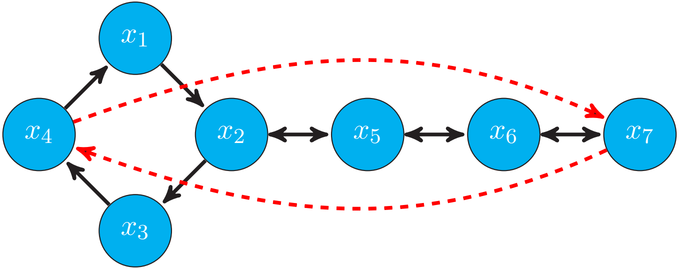


$\mathcal{L}_\mu(x, z; y^1)$, $\mathcal{L}_\mu(x; y^1)$



$\mathcal{L}_\mu(x, z; y^\star)$, $\mathcal{L}_\mu(x; y^\star)$

Edge addition in directed consensus networks



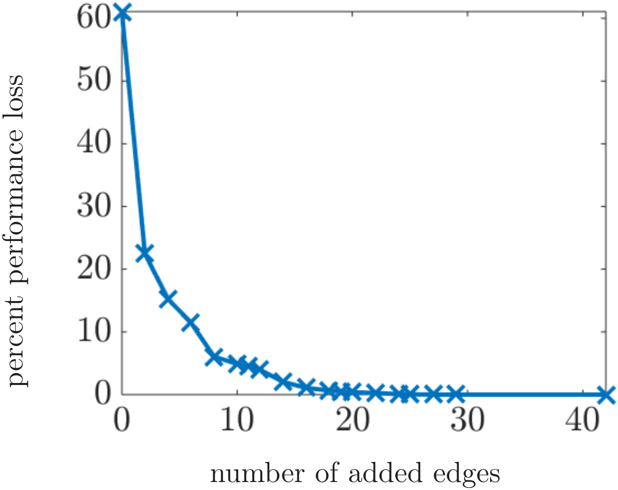
z are edges, columns of T are basis for space of balanced graphs

Identify edges

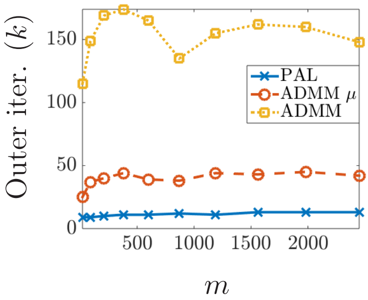
$$x(\gamma) \;=\; \underset{x}{\text{minimize}} \;\; f_2(x) \;+\; \gamma \|Tx\|_1$$

Design edge weights

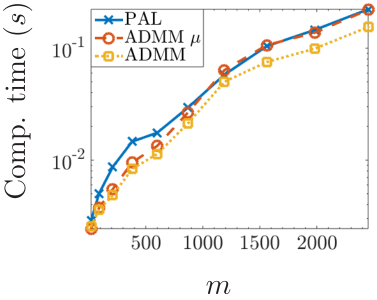
$$\begin{array}{rll} x^\star(\gamma) \;=\; & \underset{x}{\text{minimize}} & f_2(x) \\ & \text{subject to} & \text{sp}(Tx) \;\in\; \text{sp}(Tx(\gamma)) \end{array}$$


percent performance loss
60
50
40
30
20
10
0
0
10
20
30
40
number of added edges

Comparison with ADMM



Outer iterations



per outer iteration

- guaranteed convergence to local minimum
- computational savings from reduced outer iterations

Dhingra, Khong, Jovanović, arXiv:1610.04514

Outline

I Proximal augmented Lagrangian

- centralized approach – method of multipliers

II Primal-dual method

- distributable
- convergence for convex problems
- linear convergence for strongly convex problems

Primal-descent dual-ascent

Arrow-Hurwicz-Uzawa type gradient flow

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L} \\ \nabla_y \mathcal{L} \end{bmatrix}$$

- Existing methods use subgradients or projection
- Convenient for distributed implementation

Arrow, Hurwicz, Uzawa, '59
Nedic & Ozdaglar, TAC '09
Wang & Elia, CDC '11
Feijer & Paganini, AUT '10
Cherukuri, Gharesifard, Cortés, SCL '15

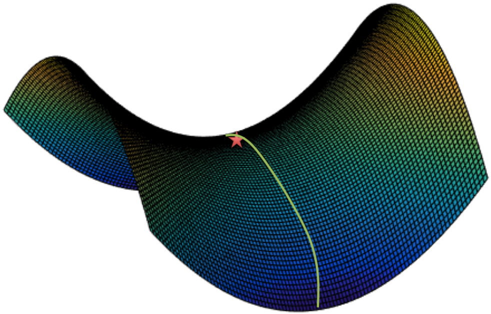
First-order primal-dual method

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L}_\mu(x;y) \\ \nabla_y \mathcal{L}_\mu(x;y) \end{bmatrix}$$

- Continuous rhs – even for non-differentiable $g(Tx)$
  - algorithmic implementation via forward Euler discretization
- Convex f – asymptotic convergence
  - Lyapunov function & LaSalle's invariance principle
- Strongly cvx, Lip. cts gradient – linear convergence
  - Integral Quadratic Constraints
  - extends to discrete-time

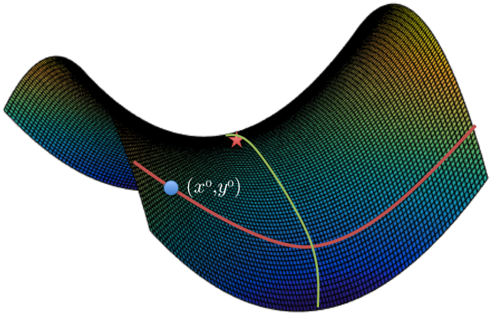
Dhingra, Khong, Jovanović, arXiv:1610.04514

Method of multipliers cartoon II



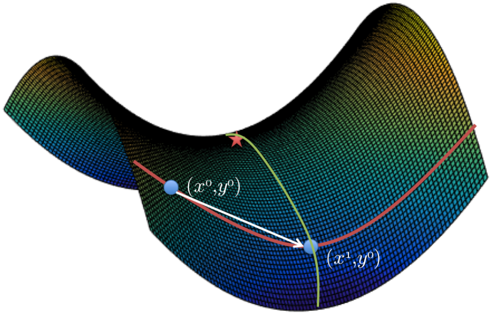
$$\mathcal{L}_\mu(x; y), \quad \min_x \mathcal{L}_\mu(x; y)$$

Method of multipliers cartoon II



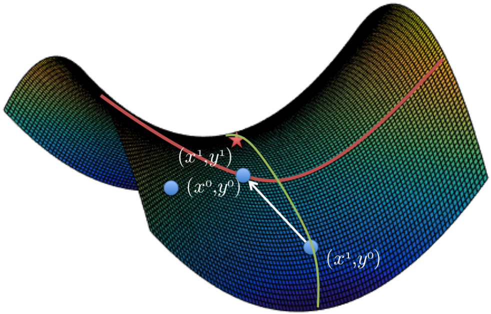
$$\mathcal{L}_\mu(x;y), \quad \min_x \mathcal{L}_\mu(x;y)$$

Method of multipliers cartoon II



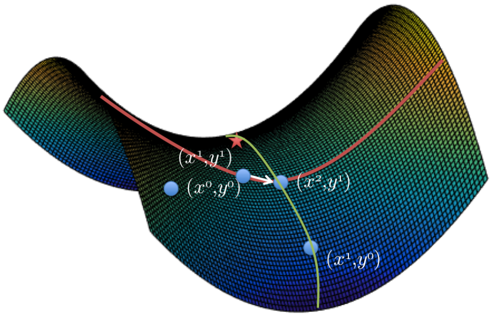
$$x^1 = \operatorname*{argmin}_x \mathcal{L}_\mu(x; y^0), \quad \min_x \mathcal{L}_\mu(x; y)$$

Method of multipliers cartoon II



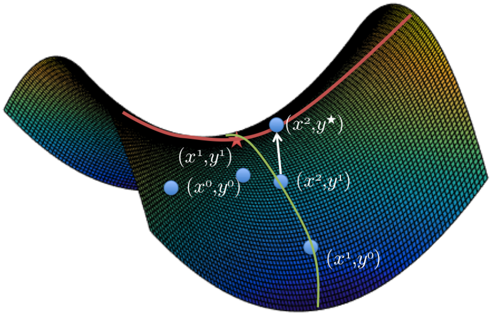
$$y^1 \ = \ y^0 \ + \ \frac{1}{\mu}\nabla_y \mathcal{L}_\mu(x^1; y^0), \quad \min_x \mathcal{L}_\mu(x; y)$$

Method of multipliers cartoon II



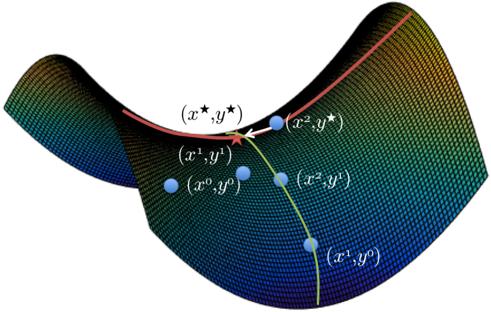
$$x^2 = \operatorname*{argmin}_x \mathcal{L}_\mu(x; y^1), \quad \min_x \mathcal{L}_\mu(x; y)$$

Method of multipliers cartoon II



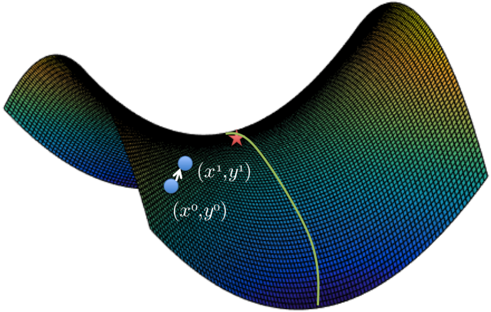
$$y^\star = y^1 + \frac{1}{\mu}\nabla_y \mathcal{L}_\mu(x^2; y^1), \quad \min_x \mathcal{L}_\mu(x; y)$$

Method of multipliers cartoon II



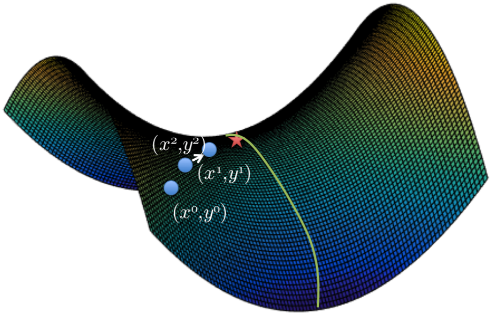
$$x^\star = \underset{x}{\operatorname{argmin}} \mathcal{L}_\mu(x; y^\star), \quad \min_x \mathcal{L}_\mu(x; y)$$

Primal-dual cartoon



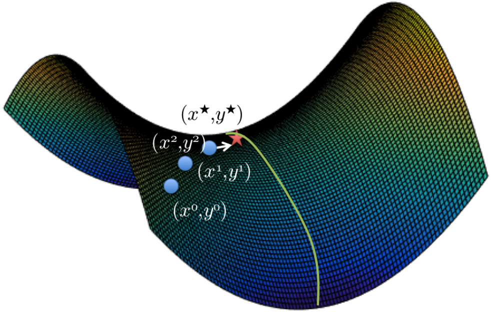
$$(x^1, y^1) \;=\; (x^0, y^0) \;-\; \alpha(\nabla_x \mathcal{L}_\mu(x^0; y^0),\, -\nabla_y \mathcal{L}_\mu(x^0; y^0)), \quad \min_x \mathcal{L}_\mu(x; y)$$

Primal-dual cartoon



$$(x^2, y^2) = (x^1, y^1) - \alpha(\nabla_x \mathcal{L}_\mu(x^1; y^1), -\nabla_y \mathcal{L}_\mu(x^1; y^1)), \quad \min_x \mathcal{L}_\mu(x; y)$$

Primal-dual cartoon



$$(x^\star, y^\star) = (x^2, y^2) - \alpha(\nabla_x \mathcal{L}_\mu(x^2; y^2), -\nabla_y \mathcal{L}_\mu(x^2; y^2)), \quad \min_x \mathcal{L}_\mu(x; y)$$

Distributed updates

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \;-\; T^T \nabla M_{\mu g}(Tx \;+\; \mu y) \\ \mu \nabla M_{\mu g}(Tx \;+\; \mu y) \;-\; \mu y \end{bmatrix}$$

- Recall $\nabla M_{\mu g}(v) = \frac{1}{\mu}(v - \mathbf{prox}_{\mu g}(v))$
- Distributed implementation if g separable and
  - $\nabla f \colon \mathbb{R}^n \to \mathbb{R}^n$ is a sparse mapping
  - $T^T T$ is sparse

Distributed updates

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \; - \; T^T \nabla M_{\mu g}(Tx \, + \, \mu y) \\ \mu \nabla M_{\mu g}(Tx \, + \, \mu y) \; - \; \mu y \end{bmatrix}$$

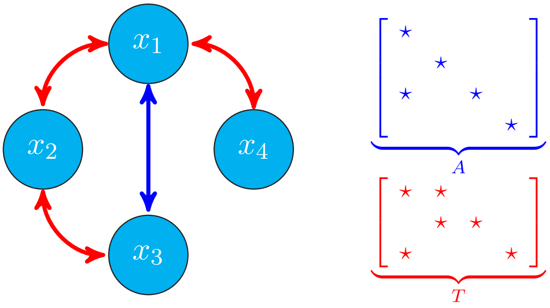
- Recall $\nabla M_{\mu g}(v) = \frac{1}{\mu}(v - \mathbf{prox}_{\mu g}(v))$
- Distributed implementation if g separable and
  - $\nabla f \colon \mathbb{R}^n \to \mathbb{R}^n$ is a sparse mapping
  - $T^T T$ is sparse
- Each node x_i
  - communicates according to ∇f and $T^T T$
  - stores y_i according to T^T

Overlapping group LASSO example

$$\text{minimize} \ \ \tfrac{1}{2}\|Ax-b\|_2^2 \ + \ \sum\|(Tx)_i\|_2$$

Gradient mapping: $\nabla f(x) = A^T(Ax-b)$

- communicate states x_i according to ∇f and $T^T T$
- store y_i corresponding to red edges



Reformulation of distributed optimization

$$\underset{x}{\text{minimize}} \quad \sum f_i(x) \quad \equiv \quad \begin{array}{ll} \underset{x_1, x_2, \ldots}{\text{minimize}} & \sum f_i(x_i) \\ \text{subject to} & Tx \; = \; 0 \end{array}$$

- T^T is Laplacian or incidence matrix of connected network

$$\equiv \quad \underset{x_1, x_2, \ldots}{\text{minimize}} \quad \sum f_i(x_i) \; + \; I_0(Tx)$$

Indicator function is $I_0(z) \; := \begin{cases} 0, & z = 0 \\ \infty, & z \neq 0 \end{cases}$

Reformulation of distributed optimization

$$\underset{x}{\text{minimize}} \quad \sum f_i(x) \quad \equiv \quad \begin{array}{rl} \underset{x_1, x_2, \ldots}{\text{minimize}} & \sum f_i(x_i) \\ \text{subject to} & Tx \;=\; 0 \end{array}$$

- T^T is Laplacian or incidence matrix of connected network

$$\equiv \quad \underset{x_1, x_2, \ldots}{\text{minimize}} \quad \sum f_i(x_i) \; + \; I_0(Tx)$$

$$\text{Indicator function is } I_0(z) \;:=\; \begin{cases} 0, & z = 0 \\ \infty, & z \neq 0 \end{cases}$$

- Let $\bar{y} := T^T y$ and $T^T T = L$

$$\begin{bmatrix} \dot{x} \\ \dot{\bar{y}} \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \; - \; \frac{1}{\mu} L x \; - \; \bar{y} \\ L x \end{bmatrix}$$

- Each agent stores x_i and $\bar{y}_i$, communicates across L

Reformulation of distributed optimization

- Discrete-time primal-dual

$$\begin{aligned} x^{k+1} &= x^k - \alpha\left(\nabla f(x^k) + \frac{1}{\mu} L x^k + \bar{y}^k\right) \\ \bar{y}^{k+1} &= \bar{y}^k + \alpha L x^k \end{aligned}$$

- EXTRA by Shi, Ling, Wu, Yin '15

$$x^{k+1} = W x^k - \alpha \nabla f(x^k) + \frac{1}{\mu} L x^k + \sum_{t=0}^{k-1} (W - \tilde{W}) x^t$$

Reformulation of distributed optimization

- Discrete-time primal-dual

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha\left(\nabla f(x^k) + \frac{1}{\mu}Lx^k + \bar{y}^k\right) \\
\bar{y}^{k+1} &= \bar{y}^k + \alpha L x^k
\end{aligned}
$$

- EXTRA by Shi, Ling, Wu, Yin '15

$$
x^{k+1} = Wx^k - \alpha\nabla f(x^k) + \frac{1}{\mu}Lx^k + \sum_{t=0}^{k-1}(W - \tilde{W})x^t
$$

Equivalent! $W = I - \frac{\alpha}{\mu}L$, $\tilde{W} = \frac{1}{2}(I + W)$, dual stepsize $\alpha_y = \frac{\alpha}{2\mu}$

$$
x^{k+1} = x^k - \alpha\left(\nabla f(x^k) + \frac{1}{\mu}Lx^k + \underbrace{\sum_{t=0}^{k-1} Lx^t}_{=\bar{y}^k}\right)
$$

Sketch of asymptotic convergence proof

- Introduce Lyapunov function with $\tilde{x} := x - x^\star$, $\tilde{y} := y - y^\star$

$$V(\tilde{x}, \tilde{y}) \;=\; \tfrac{1}{2}\|\tilde{x}\|^2 \;+\; \tfrac{1}{2}\|\tilde{y}\|^2$$

- Show $\dot{V} \leq 0$, thus by LaSalle's invariance principle,

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \to \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \,\middle|\, \dot{V}(\tilde{x}, \tilde{y}) \;=\; 0, \begin{bmatrix} \dot{\tilde{x}} \\ \dot{\tilde{y}} \end{bmatrix} \;=\; 0 \right\} \;=\; (x^\star, y^\star)$$

- **Convex $\to$ asymptotic convergence**

Dhingra, Khong, Jovanović, arXiv:1610.04514

Feedback representation

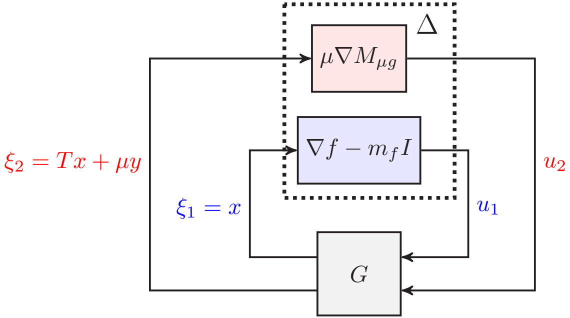
$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla f(x) - T^T \nabla M_{\mu g}(Tx + \mu y) \\ \mu \nabla M_{\mu g}(Tx + \mu y) - \mu y \end{bmatrix}$$

Feedback representation

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -(\nabla f(x) - m_f x) - T^T \nabla M_{\mu g}(Tx + \mu y) - m_f x \\ \mu \nabla M_{\mu g}(Tx + \mu y) - \mu y \end{bmatrix}$$

- 'borrow' m_f strong convexity from ∇f so G is stable

Feedback representation

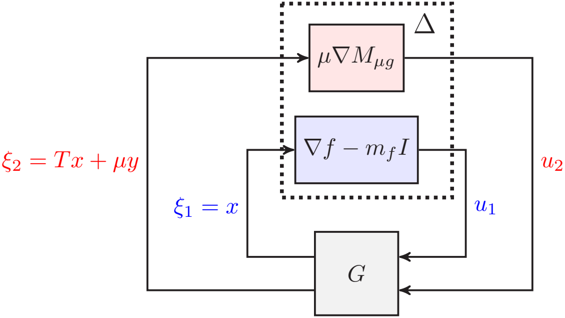


- Linear system G: $\dot{w} = Aw + Bu, \ \xi = Cw, \quad w := [x^T \ y^T]^T$

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -(\nabla f(x) - m_f x) - T^T \nabla M_{\mu g}(Tx + \mu y) - m_f x \\ \mu \nabla M_{\mu g}(Tx + \mu y) - \mu y \end{bmatrix}$$

  - ‘borrow’ m_f strong convexity from ∇f so G is stable

Feedback representation



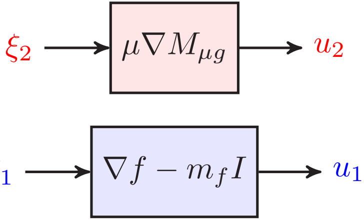
- Linear system G : $\dot{w} = Aw + Bu,\ \xi = Cw,\quad w := [x^T\ y^T]^T$

$$A = \begin{bmatrix} -m_f I & \\ & -\mu I \end{bmatrix},\ B = \begin{bmatrix} -I & -\frac{1}{\mu}T^T \\ & I \end{bmatrix},\ C = \begin{bmatrix} I & \\ T & \mu I \end{bmatrix}$$

$$u_1(\xi_1) = \nabla f(\xi_1) - m_f \xi_1,\ \ u_2(\xi_2) = \xi_2 - \mathbf{prox}_{\mu g}(\xi_2)$$

  - ‘borrow’ m_f strong convexity from ∇f so G is stable

Integral Quadratic Constraints



- $f - \frac{m_f}{2}\|\tilde{x}\|^2$ convex because f is m_f-strongly convex
- L_f Lipschitz continuous gradient of convex function

$$\begin{bmatrix} \xi & - & \xi_0 \\ u & - & u_0 \end{bmatrix}^T \underbrace{\begin{bmatrix} 0 & L_f I \\ L_f I & -2I \end{bmatrix}}_{\Pi_{L_f}} \begin{bmatrix} \xi & - & \xi_0 \\ u & - & u_0 \end{bmatrix} \geq 0$$

Linear convergence

- Linear convergence

$$\|w(t)\| \;\leq\; \tau\, \mathrm{e}^{-\rho t}\|w(0)\|$$

$w := [x^T \; y^T]^T$

if (after applying KYP Lemma)

$$\begin{bmatrix} G_\rho(\mathrm{j}\omega) \\ I \end{bmatrix}^* \Pi \begin{bmatrix} G_\rho(\mathrm{j}\omega) \\ I \end{bmatrix} \;\preceq\; 0, \quad \forall\, \omega \;\in\; \mathbb{R}$$

- transfer function $G_\rho(\mathrm{j}\omega) = C(\mathrm{j}\omega I - (A + \rho I))^{-1} B$
- Π describes IQC for u_1 and u_2

Lessard, Recht, Packard ‘16
Hu and Seiler, ‘16

Sketch of linear convergence proof

1. Set $\mu = L_f - m_f$ and evaluate

$$\begin{bmatrix} \dfrac{\mu\hat{m} + \hat{m}^2 + \omega^2}{\hat{m}^2 + \omega^2} I & \dfrac{\hat{m}}{\hat{m}^2 + \omega^2} T^T \\ * & \dfrac{\hat{m}/\mu}{\hat{m}^2 + \omega^2} T T^T + \dfrac{\omega^2 - \rho\hat{\mu}}{\hat{\mu}^2 + \omega^2} I \end{bmatrix} \succ 0$$

$\hat{m} := m_f - \rho, \; \hat{\mu} := \mu - \rho$

2. Take Schur complement and diagonalize
    - concave scalar function quadratic in ω^2
    - show absence of roots at $\omega^2 \geq 0$ for $\rho = 0$

Sketch of linear convergence proof

1. Set $\mu = L_f - m_f$ and evaluate

$$\begin{bmatrix} \dfrac{\mu\hat{m} + \hat{m}^2 + \omega^2}{\hat{m}^2 + \omega^2} I & \dfrac{\hat{m}}{\hat{m}^2 + \omega^2} T^T \\ * & \dfrac{\hat{m}/\mu}{\hat{m}^2 + \omega^2} TT^T + \dfrac{\omega^2 - \rho\hat{\mu}}{\hat{\mu}^2 + \omega^2} I \end{bmatrix} \succ 0$$

$\hat{m} := m_f - \rho, \; \hat{\mu} := \mu - \rho$

2. Take Schur complement and diagonalize
   - concave scalar function quadratic in ω^2
   - show absence of roots at $\omega^2 \geq 0$ for $\rho = 0$

- f is m_f strongly convex
- ∇f is L_f Lipschitz cts
- TT^T is full rank

$\Bigg\} \to$ linear convergence when $\mu \geq L_f - m_f$

Dhingra, Khong, Jovanović, arXiv:1610.04514

Sketch of linear convergence proof

1. Set $\mu = L_f - m_f$ and evaluate

$$\begin{bmatrix} \dfrac{\mu\hat{m} + \hat{m}^2 + \omega^2}{\hat{m}^2 + \omega^2}\, I & \dfrac{\hat{m}}{\hat{m}^2 + \omega^2}\, T^T \\ * & \dfrac{\hat{m}/\mu}{\hat{m}^2 + \omega^2}\, TT^T + \dfrac{\omega^2 - \rho\hat{\mu}}{\hat{\mu}^2 + \omega^2}\, I \end{bmatrix} \succ 0$$

$\hat{m} := m_f - \rho, \ \hat{\mu} := \mu - \rho$

2. Take Schur complement and diagonalize
   - concave scalar function quadratic in ω^2
   - show absence of roots at $\omega^2 \geq 0$ for $\rho = 0$

- f is m_f strongly convex
- ∇f is L_f Lipschitz cts
- TT^T is full rank

$\Bigg\} \to$ linear convergence when $\mu \geq L_f - m_f$ **conservative!**

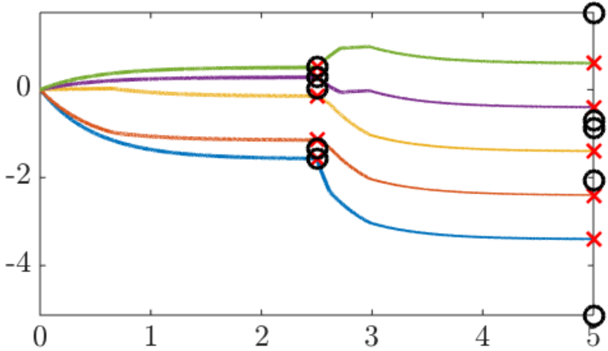
Dhingra, Khong, Jovanović, arXiv:1610.04514

Optimal placement

- Monitor targets and stay near neighbors

$$\underset{x}{\text{minimize}} \quad \sum \frac{1}{2}(x_i - b_i)^2 + I_{[-1,1]}(Tx)$$

Sampling speed of 1 kHz and a step-size of 1×10^{-3}.



Conclusions

Proximal augmented Lagrangian

- continuously differentiable
- enables MM

Distributed implementation

- primal-dual method
- connections with existing distributed optimization techniques

Ongoing work

- remove rank constraint for linear convergence
- second order methods

Extra slides

Asymptotic convergence for convex problems

At any (x, y) there is a $0 \preceq D \preceq I$ such that

$$D(T\tilde{x} + \mu\tilde{y}) \;=\; \mathbf{prox}_{\mu g}(Tx + \mu y) \;-\; \mathbf{prox}_{\mu g}(Tx^\star + \mu y^\star)$$

Derivative of V negative semidefinite

$$\dot{V}(\tilde{x}, \tilde{y}) \;=\; -\langle \tilde{x}, \nabla f(x) - \nabla f(x^\star) \rangle \;-\; \tfrac{1}{\mu} \langle T\tilde{x}, (I - D)\, T\tilde{x} \rangle \;-\; \mu \, \langle \tilde{y}, D\tilde{y} \rangle$$

Asymptotic convergence for convex problems

At any (x, y) there is a $0 \preceq D \preceq I$ such that

$$D(T\tilde{x} + \mu\tilde{y}) \; = \; \mathbf{prox}_{\mu g}(Tx + \mu y) \; - \; \mathbf{prox}_{\mu g}(Tx^\star + \mu y^\star)$$

Derivative of V negative semidefinite

$$\dot{V}(\tilde{x}, \tilde{y}) \; = \; -\langle \tilde{x}, \nabla f(x) - \nabla f(x^\star) \rangle \; - \; \tfrac{1}{\mu} \langle T\tilde{x}, (I - D)\, T\tilde{x} \rangle \; - \; \mu \, \langle \tilde{y}, D\tilde{y} \rangle$$

If $\dot{V} = 0$, $\nabla f(x) = \nabla f(x^\star)$, $\tilde{y} \in \ker\{D\}$, $T\tilde{x} \in \ker\{(I - D)\}$, thus

$$\begin{bmatrix} \dot{\tilde{x}} \\ \dot{\tilde{y}} \end{bmatrix} \; = \; \begin{bmatrix} -T^T \tilde{y} \\ 0 \end{bmatrix}$$

If additionally $\tilde{y} \in \ker\{T^T\}$, (x, y) is optimal

Linear convergence for strongly convex problems

Schur complement:

$$\frac{\hat{m}/\mu}{\mu\hat{m}+\hat{m}^2+\omega^2}\,TT^T \;+\; \frac{\omega^2-\rho\hat{\mu}}{\hat{\mu}^2+\omega^2}\,I \;\succ\; 0$$

Diagonalize where λ_i are eigenvalues of TT^T

$$\omega^4 \;+\; \left(\tfrac{\hat{m}\lambda_i}{\mu}+\hat{m}^2+\mu\hat{m}-\rho\hat{\mu}\right)\omega^2 \;+\; \tfrac{\hat{m}\hat{\mu}^2\lambda_i}{\mu} \;-\; \rho\hat{\mu}(\mu\hat{m}+\hat{m}^2) \;>\; 0$$

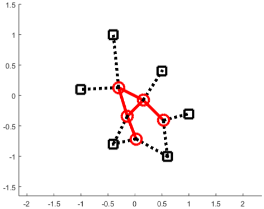
Set $\rho = 0$

$$\omega^4 \;+\; \left(\tfrac{m_f\lambda_i}{\mu}+m_f^2+\mu m_f\right)\omega^2 \;+\; \mu m_f\lambda_i \;>\; 0$$

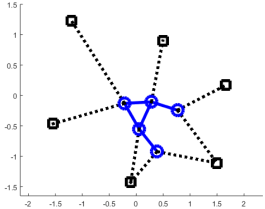
positive coefficients $\implies$ roots negative or complex

Optimal placement II

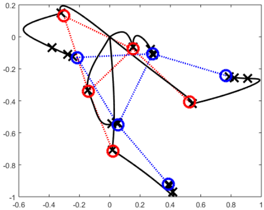
$$\text{minimize}\quad \frac{1}{2}\left\|\begin{bmatrix} A \\ T \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix}\right\|^2 + I_{[-c,c]}(Tx)$$



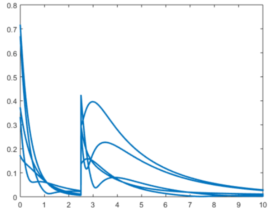
(a) Optimal configuration I



(b) Optimal configuration II

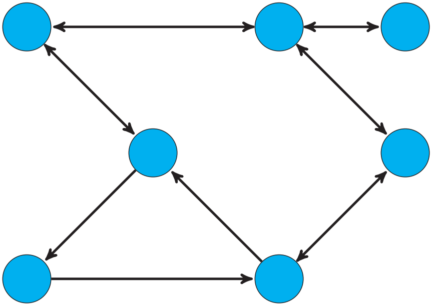


(c) Agent trajectories



(d) Distance from optimal

Directed consensus networks



- Distributed information exchange over edges z_{ij}

$$\dot{\psi}_i = \sum_j z_{ij}(\psi_j - \psi_i)$$

- Want nodes to compute average, $\psi_i(t) \to \frac{1}{n}\psi_i(0)$

Consensus networks

Aggregate dynamics

$$\dot{\psi} \; = \; -L_p \, \psi \, + \, d$$

- If L_p is balanced, nodes approach average

Penalize deviation from average

$$\zeta \; = \; \left[\, I \, - \, (1/n) \mathbb{1}\mathbb{1}^T \, \right] \psi$$

Aggregate dynamics

$$\dot{\psi} \;=\; -(L_p \,+\, L_c)\,\psi \,+\, d$$

- If $L_p + L_c$ is balanced, nodes approach average

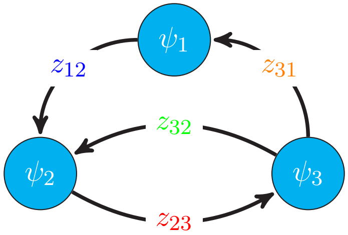
Penalize deviation from average

$$\zeta \;=\; \begin{bmatrix} I \,-\, (1/n)\mathbb{1}\mathbb{1}^T \\ -\, R^{1/2} L_c \end{bmatrix} \psi$$

Add edges to network

- $F(z) \,=\, L_c$ is graph Laplacian of added edges z

Balanced network



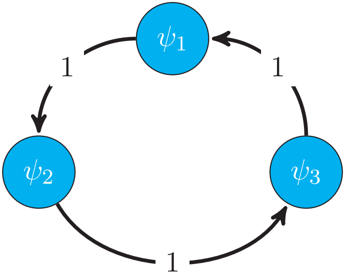
- For each node ψ_i, in-degree equals out-degree, $\sum_j z_{ij} = \sum_j z_{ji}$

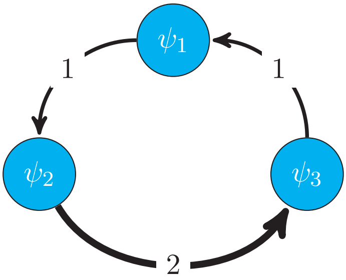
$$\begin{aligned} \psi_1 : &\quad z_{12} \qquad\qquad\qquad - z_{31} &= 0 \\ \psi_2 : &\quad -z_{12} + z_{23} - z_{32} &= 0 \\ \psi_3 : &\quad \qquad - z_{23} + z_{32} + z_{31} &= 0 \end{aligned}$$

- Linear constraint on added edges $Ez = 0$
- $z = Tx$ parametrizes balanced graphs, $Ez = E(Tx) = 0$

linear constraint in z if L_p balanced, affine if not

Balanced vs. unbalanced directed consensus networks





$$L_1 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \qquad L_2 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -2 & 0 & 2 \end{bmatrix}$$

$$v_1^T L_1 = 0, \quad v_1^T = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \qquad v_2^T L_2 = 0, \quad v_2^T = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & 2 & 1 \end{bmatrix}$$

- Nodes approach weighted avg. $\psi_i(t) \to v^T \psi(0) \mathbb{1}$
- Weighted avg. doesn't 'move', i.e., $\dot{(v^T \psi)} = -(v^T L)\psi = 0$

Edge addition in consensus networks

$$\underset{x}{\text{minimize}} \quad f_2(x) \; + \; \gamma \, \|Tx\|_1$$

Performance:

- $\mathcal{H}_2$ norm of deviations from average and control effort
- Nonconvex

Structure:

- Balanced L_c
- Minimize number of edges

Cannot use proximal gradient because T nondiagonal