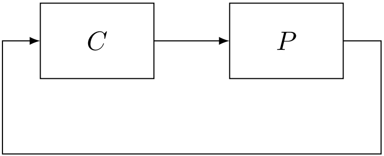
Preconditioning in First Order Optimization Methods

Pontus Giselsson

Stanford University
(joint work with Stephen Boyd)

Model predictive control (MPC)

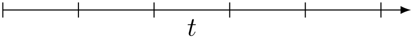


objective:

- steer system state to desired setpoint using (MPC-)controller C

procedure:

1. measure/estimate current state in P and send to C
2. compute control action by solving optimal control problem
3. go to 1



$\Rightarrow$ optimization algorithm efficiency is crucial

MPC features

what separates MPC optimization from standard optimization?

- many very similar optimization problems are solved
- there is often time for a lot of precomputations

this can be/has been utilized for/in

- explicit MPC
- code generation for specific problems (CVXGEN, FORCES...)
- code optimization

Our work

- use first-order methods to solve MPC optimization problem
- precondition problem data to improve performance

- MPC optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^T H x + \xi_t^T x \\ \text{subject to} & Bx = b\bar{x}_t \\ & \underline{d}_t \leq Cx \leq \bar{d}_t \end{array}$$

  where

  - H is positive definite
  - ξ_t, x_t, $\underline{d}_t$, and $\bar{d}_t$ may vary between optimization problems

Outline

- **operator theory**
- operators
  - gradient step operator
  - proximal operator
  - reflected proximal operator
- composite optimization algorithms
  - forward-backward splitting
    - linear convergence and preconditioning
  - Douglas-Rachford splitting
    - linear convergence and preconditioning
- preconditioning heuristics
- numerical results

Operators

- an operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ maps each point in $\mathbb{R}^n$ to a set in $\mathbb{R}^n$
- Ax (or $A(x)$) means A operates on x (and gives a set back)
- a fixed-point, $\mathrm{fix}A$, of A satisfies $\mathrm{fix}A = A(\mathrm{fix}A)$
- the graph of an operator A is defined as

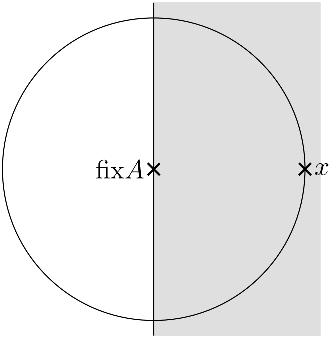
$$\mathrm{gph}A = \{(x, y) \mid y \in Ax\}$$

Monotone operators

- an operator A is monotone if

$$\langle x - y, u - v \rangle \geq 0$$

for all $(x, u) \in \text{gph}A$ and $(y, v) \in \text{gph}A$



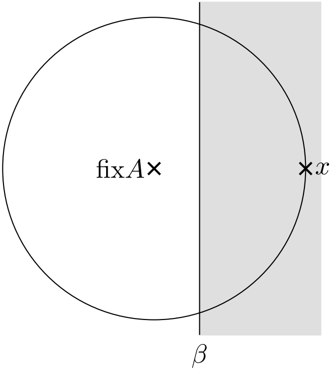
- it is maximal monotone if no extension of $\text{gph}A$ exists that preserves monotonicity

Strongly monotone operators

an operator A is β-strongly monotone if

$$\langle x - y, u - v \rangle \geq \beta \|x - y\|^2$$

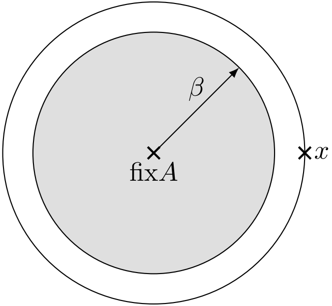
for all $(x, u) \in \mathrm{gph} A$ and $(y, v) \in \mathrm{gph} A$



Lipschitz continuous operator

- an operator A is β-Lipschitz continuous if

$$\|Ax - Ay\| \leq \beta\|x - y\|$$

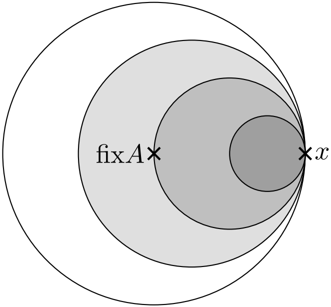


- $\beta < 1$: contractive
- $\beta = 1$: nonexpansive

Averaged operators

- an operator A is α-averaged if for some nonexpansive B and $\alpha \in (0,1)$:

$$A = (1-\alpha)I + \alpha B$$



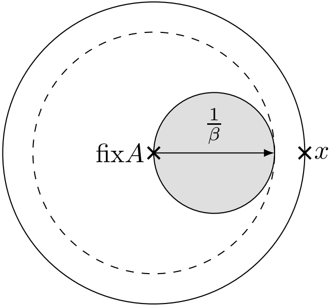
◯ – 0.75-averaged ◯ – 0.5-averaged ◯ – 0.25-averaged

- 0.5-averaged is called firmly nonexpansive

Cocoercive operators

- an operator A is β-cocoercive if

$$\langle Ax - Ay, x - y\rangle \geq \beta\|Ax - Ay\|^2$$



- β-cocoercivity implies $\frac{1}{\beta}$-Lipschitz continuity
- a 1-cocoercive operator is firmly nonexpansive

Subgradients and conjugate functions

suppose that f is proper, closed, and convex, then

- ∂f is a maximal monotone operator
- $f^*(y) \triangleq \sup_x \{\langle y, x\rangle - f(x)\}$ is proper, closed, and convex
- $\partial f^*(y) = \operatorname*{Argmax}_x \{\langle y, x\rangle - f(x)\}$

Dual properties

for proper, closed, and convex f, the following are equivalent:

(i) f is β-strongly convex w.r.t. $\|\cdot\|$

$$f(x) \geq f(y) + \langle u, x - y\rangle + \frac{\beta}{2}\|x - y\|^2$$

for all $u \in \partial f(y)$

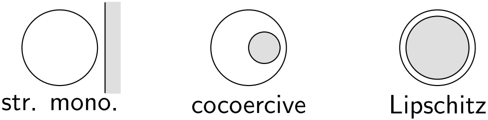
(ii) ∂f is β-strongly monotone w.r.t. $\|\cdot\|$

(iii) ∂f^* is β-cocoercive w.r.t. $\|\cdot\|$

(iv) ∂f^* is $\frac{1}{\beta}$-Lipschitz continuous w.r.t. $\|\cdot\|_*$

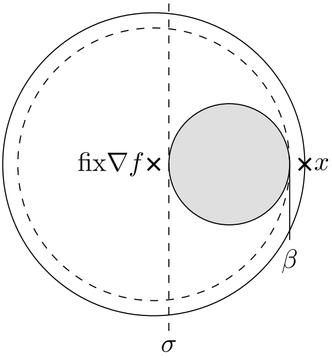
(v) f^* is $\frac{1}{\beta}$-smooth w.r.t. $\|\cdot\|_*$

$$f^*(x) \leq f^*(y) + \langle \nabla f^*(x), x - y\rangle + \frac{1}{2\beta}\|x - y\|_*^2$$



Additional property

- if ∇f β-Lipschitz continuous and σ-strongly monotone then $\nabla f - \sigma I$ is $\frac{1}{\beta-\sigma}$-cocoercive:



- call this σ-shifted $\frac{1}{\beta}$-cocoercivity

Outline

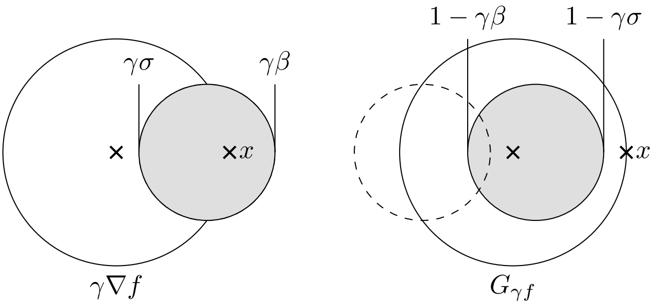
- operator theory
- **operators**
  - gradient step operator
  - proximal operator
  - reflected proximal operator
- composite optimization algorithms
  - forward-backward splitting
    - linear convergence and preconditioning
  - Douglas-Rachford splitting
    - linear convergence and preconditioning
- preconditioning heuristics
- numerical results

Gradient step operator

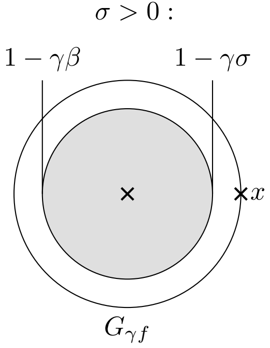
- the gradient step operator of f, denoted $G_{\gamma f}$, is defined as

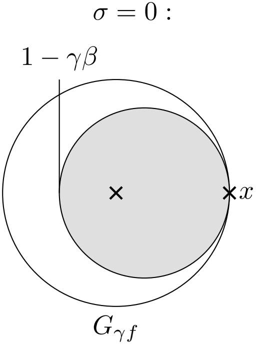
$$G_{\gamma f} := I - \gamma \nabla f$$

- assume f is β-smooth and σ-strongly convex
  $\Rightarrow \gamma \nabla f$ is $\gamma\beta$-Lipschitz and $\gamma\sigma$-strongly monotone
  (i.e. $\gamma\sigma$-shifted $\frac{1}{\gamma\beta}$-cocoercive)
  $\Rightarrow G_{\gamma f} = I - \gamma \nabla f$ is $\max(|1-\gamma\beta|, |1-\gamma\sigma|)$-Lipschitz



Gradient step operator





- $0 < \gamma < 2\beta \Rightarrow$ contractive
- optimal $\gamma = \frac{2}{\beta+\sigma} \Rightarrow$ factor $\frac{\beta/\sigma-1}{\beta/\sigma+1}(=\gamma\beta - 1 = 1 - \gamma\sigma)$

- $\gamma = 2\alpha/\beta, \alpha \in (0,1)$
  $\Rightarrow 1 - \gamma\beta = 1 - 2\alpha$
  $\Rightarrow G_{\gamma f}$ α-averaged

Proximal operator (resolvent)

- the proximal operator is defined as

$$\mathrm{prox}_{\gamma f}(y) = \underset{x}{\mathrm{argmin}} \left\{ \gamma f(x) + \tfrac{1}{2}\|x - y\|^2 \right\}$$

- define $h_\gamma = \frac{1}{2}\|\cdot\|^2 + \gamma f$, then:

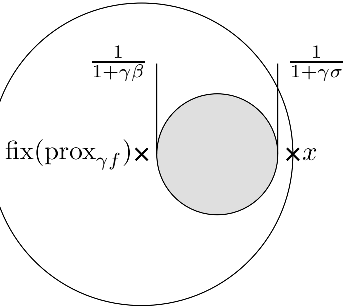
$$\mathrm{prox}_{\gamma f}(y) = \underset{x}{\mathrm{argmax}} \left\{ \langle x, y\rangle - \gamma f(x) - \tfrac{1}{2}\|x\|^2 \right\} = \nabla h_\gamma^*(y)$$

- proximal operator properties (f proper, closed, and convex)

f	∂f	$\partial h_\gamma = (I + \gamma \partial f)$	$\nabla h_\gamma^* = \mathrm{prox}_{\gamma f}$
cvx	mono.	1-str. mono.	1-cocoercive
σ-str. cvx	σ-str. mono.	$(1+\gamma\sigma)$-str.mono.	$\frac{1}{1+\gamma\sigma}$-Lipschitz
β-smooth	β-Lipschitz	$(1+\gamma\beta)$-Lipschitz	$\frac{1}{1+\gamma\beta}$-str. mono.

More properties of prox operator

- assume ∂f is β-Lipschitz continuous and σ-strongly monotone
- then $\mathrm{prox}_{\gamma f}$ is $\frac{1}{1+\gamma\beta}$-shifted $(1+\gamma\sigma)$-cocoercive



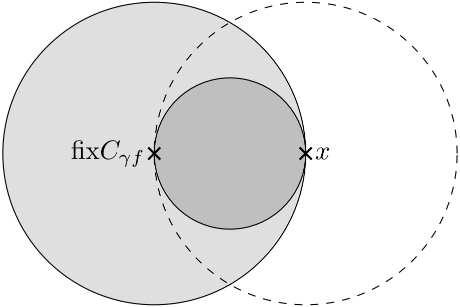
- (since $\mathrm{prox}_{\gamma f} = \nabla h_\gamma^*$ is $\frac{1}{1+\gamma\beta}$-str. mono and $\frac{1}{1+\gamma\sigma}$-Lipschitz)

Reflected proximal operator (reflected resolvent)

- the reflected proximal operator (or Cayley operator) is defined as

$$C_{\gamma f} := 2\text{prox}_{\gamma f} - I$$

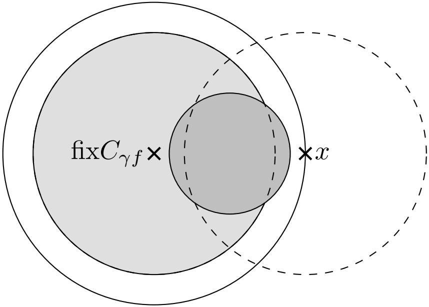
- $C_{\gamma f}$ is nonexpansive in the general case



- (fixed-points of $C_{\gamma f}$ coincide with fixed-points of $\text{prox}_{\gamma f}$)

More properties of reflected proximal operator

- if ∇f is σ-strongly monotone and β-Lipschitz then $C_{\gamma f}$ is $\max\left(\frac{1-\gamma\sigma}{1+\gamma\sigma}, \frac{\gamma\beta-1}{1+\gamma\beta}\right)$-contractive



- contraction factor optimized for $\gamma = \frac{1}{\sqrt{\sigma\beta}}$ (gives a contraction factor of $\frac{\sqrt{\beta/\sigma}-1}{\sqrt{\beta/\sigma}+1}$)

Outline

- operator theory
- operators
  - gradient step operator
  - proximal operator
  - reflected proximal operator
- **composite optimization algorithms**
  - forward-backward splitting
    - linear convergence and preconditioning
  - Douglas-Rachford splitting
    - linear convergence and preconditioning
- preconditioning heuristics
- numerical results

Composite optimization problems

- we consider composite optimization problems of the form

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

  where

  - f and g are proper, closed, and convex
  - A is a real matrix

- introduce $\hat{g} := g \circ A$ to get primal problem

$$\text{minimize} \quad f(x) + \hat{g}(x) \tag{P}$$

- introduce $\hat{f} := f^* \circ (-A^T)$ to get dual problem

$$\text{minimize} \quad \hat{f}(y) + g^*(y) \tag{D}$$

Optimality conditions

- primal (P) and dual (D) problems have form

$$\text{minimize} \quad \psi(x) + \phi(x)$$

- assume ψ is β-smooth
- x optimal solution to composite problem iff

$$x = \mathrm{prox}_{\gamma\phi}((I - \gamma\nabla\psi)x)$$

- algorithm: find fixed point to operator $\mathrm{prox}_{\gamma\phi} \circ (I - \gamma\nabla\psi)$

Forward-backward splitting

- FB-splitting obtained by iterating optimality conditions

$$x^{k+1} = \mathrm{prox}_{\gamma\phi}((I - \gamma\nabla\psi)x^k)$$

- (also known as proximal gradient method)
- convergence
  - $\mathrm{prox}_{\gamma\phi}$ is firmly nonexpansive ($\frac{1}{2}$-averaged)
  - $(I - \gamma\nabla\psi)$ is α-averaged if $\gamma = 2\alpha/\beta$
  - composition of averaged operators averaged $\Rightarrow$ convergence

Linear convergence

- assume that ψ is σ-strongly convex and β-smooth
- $(I - \gamma\nabla\psi)$ is $\max(|1-\gamma\sigma|, |\gamma\beta - 1|)$-contractive
- optimal $\gamma = \frac{2}{\sigma+\beta} \Rightarrow (I - \gamma\nabla\psi)$ is $\frac{\beta/\sigma-1}{\beta/\sigma+1}$-contractive
- $\mathrm{prox}_{\gamma\phi}$ (firmly) nonexpansive

  $\Rightarrow$ FB-operator $\mathrm{prox}_{\gamma\phi} \circ (I - \gamma\nabla\psi)$ is $\frac{\beta/\sigma-1}{\beta/\sigma+1}$-contractive

  $\Rightarrow$ FB-splitting converges linearly with factor $\frac{\beta/\sigma-1}{\beta/\sigma+1}$

Optimal parameter selection and preconditioning

- convergence factor minimized by letting $\gamma = \frac{2}{\beta+\sigma}$
- FB splitting converges linearly with factor $\frac{\beta/\sigma-1}{\beta/\sigma+1}$
- precondition by minimizing β/σ (i.e., reduce conditioning)

Example – Quadratic case

precondition primal problem (P) (i.e., precond. f)

- $f(x) = \frac{1}{2}x^T Hx + \xi^T x$
  $\Rightarrow \beta_f = \lambda_{\max}(H)$ and $\sigma_f = \lambda_{\min}(H)$
- introduce $Tq = x$
- $f_T(q) = f(Tq) = \frac{1}{2}q^T T^T HTq + \xi^T Tq$
  $\Rightarrow \beta_{f_T} = \lambda_{\max}(T^T HT)$ and $\sigma_{f_T} = \lambda_{\min}(T^T HT)$
- choose T diagonal to not increase computational complexity
  $\Rightarrow$ minimize condition number of $T^T HT$ subject to T diagonal

Example – Quadratic case

precondition dual problem (D) (i.e., precond. $\hat{f} = f^* \circ (-A^T)$)

- $f(x) = \frac{1}{2}x^T Hx + \xi^T x$
- $\hat{f}(\mu) = \frac{1}{2}(\xi + A^T\mu)^T H^{-1}(\xi + A^T\mu)$
  $\Rightarrow \beta_{\hat{f}} = \lambda_{\max}(AH^{-1}A^T)$ and $\sigma_{\hat{f}} = \lambda_{\min}(AH^{-1}A^T)$
- introduce $E^T\nu = \mu$
- $\hat{f}_E(\nu) = f(E^T\nu) = \frac{1}{2}(\xi + A^T E^T\mu)^T H^{-1}(\xi + A^T E^T\mu)$
  $\Rightarrow \beta_{\hat{f}_E} = \lambda_{\max}(EAH^{-1}A^T E^T)$ and $\sigma_{\hat{f}_E} = \lambda_{\min}(EAH^{-1}A^T E^T)$
- choose E diagonal to not increase computational complexity
  $\Rightarrow$ minimize condition nbr of $EAH^{-1}A^T E^T$ s.t. E diagonal

Acceleration

- fast proximal gradient method

$$y^k = x^k + \theta^k(x^k - x^{k-1})$$
$$x^{k+1} = \mathrm{prox}_{\gamma\phi}((I - \gamma\nabla\psi)y^k)$$

- preconditioning improves performance of FB-operator
  $\Rightarrow$ same preconditioning can be used with acceleration

Optimality conditions

- composite problem

$$\text{minimize} \quad \psi(x) + \phi(x)$$

- x optimal solution to such problems iff

$$z = C_{\gamma\psi} C_{\gamma\phi} z \qquad\qquad x = \operatorname{prox}_{\gamma\phi}(z)$$

- find fixed-point to $C_{\gamma\psi} C_{\gamma\phi}$ to solve problem

Generalized Douglas-Rachford splitting

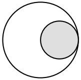
- iterate $C_{\gamma\psi} C_{\gamma\phi}$ to find fixed-point (Peaceman-Rachford splitting)

$$z^{k+1} = C_{\gamma\psi} C_{\gamma\phi} z^k$$

  - $C_{\gamma\psi}$ and $C_{\gamma\phi}$ are nonexpansive, so is composition
    $\Rightarrow$ not guaranteed to converge in general case
- introduce averaging with $\alpha \in (0, 1)$:

$$z^{k+1} = ((1 - \alpha)I + \alpha C_{\gamma\psi} C_{\gamma\phi}) z^k$$

  - $\alpha = 0.5$: Douglas-Rachford splitting
  - $\alpha = 0.5$ applied to (D) : ADMM
  - iteration of averaged operator converges to fixed-point



0.5-averaged

Linear convergence

- assume that ψ is σ-strongly convex and β-smooth
- $C_{\gamma\psi}$ is $\max\left(\frac{1-\gamma\sigma}{1+\gamma\sigma}, \frac{\gamma\beta-1}{1+\gamma\beta}\right)$-contractive (so is $C_{\gamma\psi}C_{\gamma\phi}$)
- D-R operator $((1-\alpha)I + \alpha C_{\gamma\psi}C_{\gamma\phi})$ is $(1-\alpha) + \alpha \max\left(\frac{1-\gamma\sigma}{1+\gamma\sigma}, \frac{\gamma\beta-1}{1+\gamma\beta}\right)$-contractive

  $\Rightarrow$ D-R algorithm converges linearly with same factor

Optimal parameter selection and preconditioning

- convergence factor $(1-\alpha)+\alpha\max\left(\frac{1-\gamma\sigma}{1+\gamma\sigma},\frac{\gamma\beta-1}{1+\gamma\beta}\right)$
- optimal parameters
  - $\alpha=1$ (i.e. Peaceman-Rachford splitting)
  - selection $\gamma=\frac{1}{\sqrt{\sigma\beta}}$

    $\Rightarrow$ convergence rate $\frac{\sqrt{\beta/\sigma}-1}{\sqrt{\beta/\sigma}+1}$
- precondition by minimizing β/σ

Outline

- operator theory
- operators
  - gradient step operator
  - proximal operator
  - reflected proximal operator
- composite optimization algorithms
  - forward-backward splitting
    - linear convergence and preconditioning
  - Douglas-Rachford splitting
    - linear convergence and preconditioning
- **preconditioning heuristics**
- numerical results

Preconditioning heuristics

- assumption that ψ both strongly convex and smooth is rare
- can do heuristic extensions to cover wider classes
- here: focus on preconditioning heuristic for MPC problems

MPC problem on composite form

- MPC optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^T H x + \xi_t^T x \\ \text{subject to} & Bx = b\bar{x}_t \\ & \underline{d}_t \leq Cx \leq \bar{d}_t \end{array}$$

can be cast on the form

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

- splitting 1:

$$\begin{aligned} f(x) &= \frac{1}{2}x^T H x + \xi_t^T x + I_{Bx = b\bar{x}_t}(x) \\ g(y) &= I_{\underline{d}_t \leq y \leq \bar{d}_t}(y) \\ A &= C \end{aligned}$$

- splitting 2:

$$\begin{aligned} f(x) &= \frac{1}{2}x^T H x + \xi_t^T x + I_{\underline{d}_t \leq Cx \leq \bar{d}_t}(x) \\ g(y) &= I_{y = bx_t}(y) \\ A &= B \end{aligned}$$

Properties

- for splitting 1 and 2:
  - f is 1-strongly convex w.r.t. $\|\cdot\|_H$
  - f^* is 1-smooth w.r.t. $\|\cdot\|_{H^{-1}}$
  - $\hat{f} = f^* \circ (-A^T)$ is 1-smooth w.r.t. $\|\cdot\|_{AH^{-1}A^T}$
- implications:
  - primal formulation (P) has a nonsmooth strongly convex term
    $\Rightarrow$ can be solved by DR-splitting but not FB-splitting
  - dual formulation (D) has a non-strongly convex smooth term
    $\Rightarrow$ can be solved by DR-splitting and FB-splitting

Heuristic preconditioning

- for both splitting 1 and 2:

$$f(x) = \frac{1}{2}\|x\|_H^2 + \xi^T x + I_{x \in \mathcal{X}}(x)$$

  where $I_{x \in \mathcal{X}}$ is indicator function for different sets

- heuristic: do preconditioning and parameter selection for quadratic part (i.e., assume $I_{x \in \mathcal{X}} = 0$)

Primal and dual preconditioning

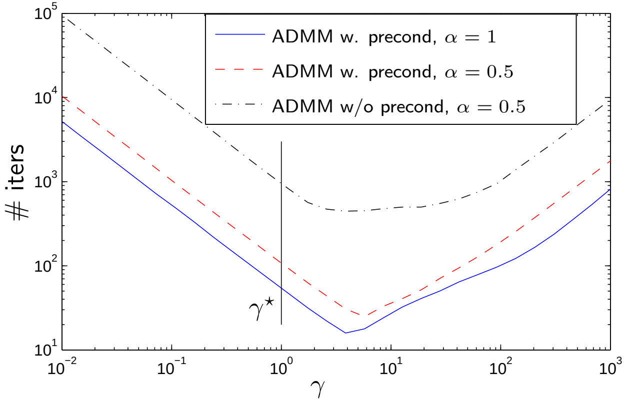
- preconditioning of primal formulation
  - precondition quadratic part ($\frac{1}{2}\|x\|_H^2 + \xi^T x$)
  - minimize condition number of $T^T HT$ subject to T diagonal
- preconditioning of dual formulation
  - precondition quadratic part ($\frac{1}{2}(\xi + A^T\mu)^T H^{-1}(\xi + A^T\mu)$)
  - minimize condition number of $EAH^{-1}A^T E^T$ s.t. E diagonal
- if matrix positive semi definite only
  - minimize ratio between largest and smallest nonzero eigenvalues

Application - MPC of pitch angle in aircraft

- 4 states
- 2 outputs
- 2 inputs
- control horizon 10
- hard input constraints
- soft output constraints
- 100 decision variables
- diagonal quadratic positive definite cost matrices
- condition number of Hessian: 10^{10}

Numerical evaluation – ADMM

Figure: Average number of iterations for different γ-values, with and without preconditioning, and for different relaxation α.



- theoretical optimal: $\gamma^\star = 1$, $\alpha = 1$
- empirical optimal: $\gamma = 4$, $\alpha = 1$

Numerical evaluation

- fast dual FB-splitting with and without preconditioning
- ADMM with and without preconditioning
- MATLAB implementation

			exec time (ms)		nbr iters	
alg.	precond	split./param	avg.	max	avg.	max
FDFBS	y	1/-	1.2	5.8	20.0	105
FDFBS	n	1/-	98.9	679.4	1850.1	12783
FDFBS	y	2/-	2.3	12.1	21.7	102
FDFBS	n	2/-	4713.9	28411	50845	308210
ADMM	y	1/th. opt.	4.5	15.3	54.2	197
ADMM	y	1/emp. opt.	1.6	3.6	15.6	43
ADMM	n	1/emp. opt.	17.2	82.5	224.3	1127

Numerical evaluation – C

C implementation comparison to FORCES and MOSEK

		exec time (ms)	
algorithm	splitting	avg.	max
FDFBS	1	0.061	0.196
FDFBS	2	0.079	0.232
FORCES	-	0.347	0.592
MOSEK	-	4.9	5.4

- FDFBS: preconditioned fast dual forward-backward splitting
- FORCES: code generator for model predictive control problems based on interior point methods
- MOSEK: commercial QP solver

Thank you

Questions?