Nonsmooth composite minimization:

an exponentially convergent primal-dual algorithm

Dongsheng Ding

joint work with

Bin Hu, Neil K. Dhingra, Mihailo R. Jovanović

57th IEEE Conference on Decision and Control, Miami Beach, FL, December 19, 2018

Nonsmooth composite minimization

$$\underset{x}{\text{minimize}} \qquad \underset{\textbf{performance}}{\underset{\Big\downarrow}{f(x)}} \qquad + \qquad \underset{\textbf{structure}}{\underset{\Big\downarrow}{g(Tx)}}$$

Nonsmooth composite minimization

$$\underset{x}{\text{minimize}} \qquad \underset{\substack{\big\downarrow \\ \textbf{performance}}}{f(x)} \qquad + \qquad \underset{\substack{\big\downarrow \\ \textbf{structure}}}{g(Tx)}$$

- T – select certain coordinates to impose structure
- f – strongly convex; Lipschitz cts gradient
- g – non-differentiable; convex

  $e.g., I_{\mathcal{C}}(\cdot), \|\cdot\|_1, \|\cdot\|_*$, easy to evaluate proximal operator

Proximal gradient descent

$$\underset{x}{\text{minimize}} \qquad \underset{\substack{\downarrow \\ \text{smooth}}}{f(x)} \qquad + \qquad \underset{\substack{\downarrow \\ \text{nonsmooth}}}{g(x)}$$

Beck & Teboulle, SIAM J. Imag. Sci. ’09
Parikh & Boyd, FnT in Optimization ’14

Proximal gradient descent

$$\underset{x}{\text{minimize}} \qquad \underset{\text{smooth}}{\underset{\downarrow}{f(x)}} \qquad + \qquad \underset{\text{nonsmooth}}{\underset{\downarrow}{g(x)}}$$

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

Beck & Teboulle, SIAM J. Imag. Sci. ’09
Parikh & Boyd, FnT in Optimization ’14

Proximal gradient descent

$$\underset{x}{\text{minimize}} \qquad \underset{\text{smooth}}{\underset{\downarrow}{f(x)}} \qquad + \qquad \underset{\text{nonsmooth}}{\underset{\downarrow}{g(x)}}$$

- **Gradient descent plus proximal operator**

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

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

Beck & Teboulle, SIAM J. Imag. Sci. ’09
Parikh & Boyd, FnT in Optimization ’14

Proximal gradient descent

$$\underset{x}{\text{minimize}} \qquad \underset{\substack{\downarrow \\ \text{smooth}}}{f(x)} \qquad + \qquad \underset{\substack{\downarrow \\ \text{nonsmooth}}}{g(x)}$$

- **Gradient descent plus proximal operator**

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

  - $\mathbf{prox}_{\mu g}(v) \;:=\; \underset{z}{\text{argmin}}\; g(z) + \frac{1}{2\mu}\|z - v\|^2$
  - explicit formula for $\mathbf{prox}_g \longrightarrow$ efficient implementation
  - does not apply to $g(Tx)$

Beck & Teboulle, SIAM J. Imag. Sci. ’09

Parikh & Boyd, FnT in Optimization ’14

$$\begin{array}{ll} \underset{x,z}{\text{minimize}} & f(x) + g(z) \\ \text{subject to} & T\,x - z = 0 \end{array}$$

Boyd, Parikh, Chu, Peleato, Eckstein, FnT in Optimization ’11

Augmented Lagrangian method

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & f(x) + g(z) \\ \text{subject to} \quad & T\,x - z = 0 \end{aligned}$$

Boyd, Parikh, Chu, Peleato, Eckstein, FnT in Optimization '11

Augmented Lagrangian method

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & f(x) + g(z) \\ \text{subject to} \quad & T\,x - z = 0 \end{aligned}$$

- **Augmented Lagrangian**

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

Boyd, Parikh, Chu, Peleato, Eckstein, FnT in Optimization ’11

Augmented Lagrangian method

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

- **Augmented Lagrangian**

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

- **ADMM**

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

Boyd, Parikh, Chu, Peleato, Eckstein, FnT in Optimization ’11

Proximal augmented Lagrangian

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

Dhingra, Khong, Jovanović, IEEE TAC ’18

Proximal augmented Lagrangian

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

Dhingra, Khong, Jovanović, IEEE TAC ’18

Proximal augmented Lagrangian

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

- **Minimizer of $\mathcal{L}_\mu(x, z; y)$ over z**

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

Dhingra, Khong, Jovanović, IEEE TAC '18

Proximal augmented Lagrangian

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

- **Minimizer of $\mathcal{L}_\mu(x,z;y)$ over z**

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

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

$$\mathcal{L}_\mu(x;y) \quad := \quad \mathcal{L}_\mu(x,z;y)\Big|_{z\,=\,z_\mu^\star}$$

Dhingra, Khong, Jovanović, IEEE TAC ’18

Proximal augmented Lagrangian

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

- **Minimizer of $\mathcal{L}_\mu(x,z;y)$ over z**

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

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

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

continuously differentiable in x and y

Dhingra, Khong, Jovanović, IEEE TAC '18

Primal-dual gradient flow dynamics

- **Primal-descent dual-ascent**

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

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

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

- Lipschitz cts RHS
- $\dot{x} = 0, \dot{y} = 0$ – optimality condition

$$\left.\begin{array}{l} \bullet \; f \text{ – strongly convex; Lipschitz cts gradient} \\ \bullet \; T \text{ – row full rank} \end{array}\right\} \rightarrow \text{exponentially stable}$$

Dhingra, Khong, Jovanović, IEEE TAC '18

Implementation issues

- **Key issue**
  - how do we implement it? discretization
- **Simple discretization**

$$\begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} x^k - \alpha \nabla_x \mathcal{L}_\mu(x^k; y^k) \\ y^k + \alpha \nabla_y \mathcal{L}_\mu(x^k; y^k) \end{bmatrix}$$

- **Key challenge**
  - CT convergence rate analysis ⟶ DT algorithm

Wibisono, Wilson, Jordan, PNAS '16
Wilson, Recht, Jordan, arXiv:1611.02635
Hu, Lessard, ICML '17

Proposed primal-dual algorithm

- **Forward Euler discretization**

$$\begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} x^k - \alpha \nabla_x \mathcal{L}_\mu(x^k; y^k) \\ y^k + \alpha \nabla_y \mathcal{L}_\mu(x^k; y^k) \end{bmatrix}$$

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

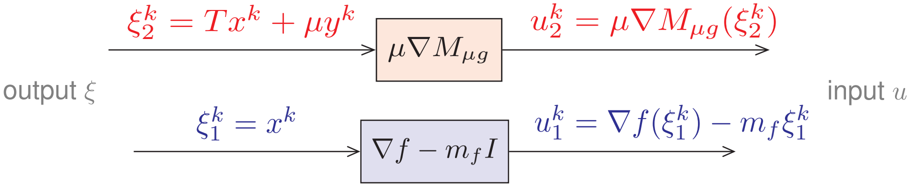
- **Contributions**
  - an automated tool ⟶ exp. converg. of PD algorithm
  - LMI condition ⟶ rate certificate
  - a range of step size values ⟶ exp. converg.

Lessard, Recht, Packard, SIAM J. Optim. ’16
Hu, Seiler, Rantzer, PMLR ’17
Fazlyab, Ribeiro, Morari, Preciado, SIAM J. Optim. ’18

Control-theoretic viewpoint

- **Feedback connection**

$$\begin{aligned}\begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} \;&=\; \begin{bmatrix} (1-\alpha m_f)x^k \\ (1-\alpha\mu)y^k \end{bmatrix} \\ &\quad -\alpha \begin{bmatrix} \nabla f(x^k) - m_f x^k \\ 0 \end{bmatrix} \\ &\quad -\alpha \begin{bmatrix} T^T \nabla M_{\mu g}(Tx^k + \mu y^k) \\ -\mu \nabla M_{\mu g}(Tx^k + \mu y^k) \end{bmatrix}\end{aligned}$$



$\xi_2^k = Tx^k + \mu y^k$

$\mu \nabla M_{\mu g}$

$u_2^k = \mu \nabla M_{\mu g}(\xi_2^k)$

output ξ

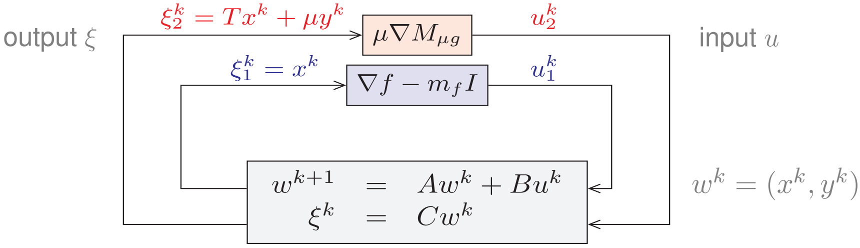
input u

$\xi_1^k = x^k$

$\nabla f - m_f I$

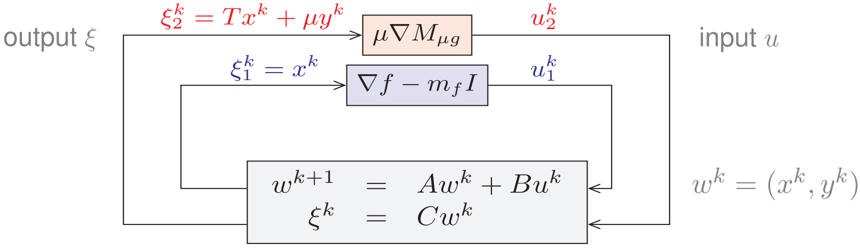
$u_1^k = \nabla f(\xi_1^k) - m_f \xi_1^k$

Control-theoretic viewpoint



$u_1^k \;=\; \nabla f(\xi_1^k) - m_f \xi_1^k \qquad\qquad u_2^k \;=\; \mu \nabla M_{\mu g}(\xi_2^k)$

Control-theoretic viewpoint



output ξ

$\xi_2^k = Tx^k + \mu y^k$

$\mu \nabla M_{\mu g}$

u_2^k

input u

$\xi_1^k = x^k$

$\nabla f - m_f I$

u_1^k

$$\begin{aligned} w^{k+1} &= Aw^k + Bu^k \\ \xi^k &= Cw^k \end{aligned}$$

$w^k = (x^k, y^k)$

$$u_1^k = \nabla f(\xi_1^k) - m_f \xi_1^k \qquad u_2^k = \mu \nabla M_{\mu g}(\xi_2^k)$$

- **LTI system:** (A, B, C)

$$A = \begin{bmatrix} (1 - \alpha m_f)I & 0 \\ 0 & (1 - \alpha\mu)I \end{bmatrix}$$

$$B = \begin{bmatrix} -\alpha I & -\frac{\alpha}{\mu} T^T \\ 0 & \alpha I \end{bmatrix}, C = \begin{bmatrix} I & 0 \\ T & \mu I \end{bmatrix}$$

Exponential convergence

- **Exponential/linear convergence with rate $r \in (0,1)$**

$$\begin{aligned} \begin{bmatrix} A^T P A - r^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} + \begin{bmatrix} C^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} &\prec 0 \\ P &\succ 0 \end{aligned}$$

Lessard, Recht, Packard, SIAM J. Optim. ’16

Exponential convergence

- **Exponential/linear convergence with rate** $r \in (0,1)$

$$\begin{aligned} \begin{bmatrix} A^T P A - r^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} + \begin{bmatrix} C^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} &\prec 0 \\ P &\succ 0 \end{aligned}$$

$w^k = (x^k, y^k)$

$$\Downarrow$$

$$\|w^k - \bar{w}\| \;\leq\; \sqrt{\mathsf{cond}(P)}\, r^k \, \|w^0 - \bar{w}\|$$

Lessard, Recht, Packard, SIAM J. Optim. '16

Exponential convergence

- **Exponential/linear convergence with rate $r \in (0,1)$**

$$\begin{bmatrix} A^T P A - r^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} + \begin{bmatrix} C^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \prec 0$$

$$P \succ 0$$

$$\Downarrow$$

$w^k = (x^k, y^k)$

$$\|w^k - \bar{w}\| \leq \sqrt{\text{cond}(P)}\, r^k \, \|w^0 - \bar{w}\|$$

- (A, B, C) – algorithm parameters (α, μ, T, m_f)
- Π – problem parameters (L_f, m_f)
- $(\alpha, r) \longrightarrow$ decision variable P
- $r \in (0,1) \longrightarrow$ decision variables (α, P)

Lessard, Recht, Packard, SIAM J. Optim. '16

Exponential convergence

- **Exponential/linear convergence with rate** $r \in (0,1)$

$$\begin{aligned} \begin{bmatrix} G(r\mathrm{e}^{\mathrm{j}\theta}) \\ I \end{bmatrix}^* \Pi \begin{bmatrix} G(r\mathrm{e}^{\mathrm{j}\theta}) \\ I \end{bmatrix} &\prec 0, \ \forall\, \theta \in [0, 2\pi) \\ G(r\mathrm{e}^{\mathrm{j}\theta}) &\in \mathcal{RH}_\infty \end{aligned}$$

Rantzer, Syst. Control Lett. '96

Exponential convergence

- **Exponential/linear convergence with rate $r \in (0,1)$**

$$\begin{aligned} \begin{bmatrix} G(r\mathrm{e}^{\mathrm{j}\theta}) \\ I \end{bmatrix}^* \Pi \begin{bmatrix} G(r\mathrm{e}^{\mathrm{j}\theta}) \\ I \end{bmatrix} &\prec 0, \ \forall\, \theta \in [0, 2\pi) \\ G(r\mathrm{e}^{\mathrm{j}\theta}) &\in \mathcal{RH}_\infty \end{aligned}$$

$w^k = (x^k, y^k)$

$$\Downarrow$$

$$\|w^k - \bar{w}\| \ \leq \ c\, r^k \, \|w^0 - \bar{w}\|$$

Rantzer, Syst. Control Lett. '96

Exponential convergence

- **Exponential/linear convergence with rate $r \in (0,1)$**

$$\begin{bmatrix} G(r\mathrm{e}^{\mathrm{j}\theta}) \\ I \end{bmatrix}^* \Pi \begin{bmatrix} G(r\mathrm{e}^{\mathrm{j}\theta}) \\ I \end{bmatrix} \prec 0, \; \forall\, \theta \in [0, 2\pi)$$

$$G(r\mathrm{e}^{\mathrm{j}\theta}) \in \mathcal{RH}_\infty$$

$$w^k = (x^k, y^k)$$

$$\Downarrow$$

$$\|w^k - \bar{w}\| \leq c\, r^k \, \|w^0 - \bar{w}\|$$

- $G(r\mathrm{e}^{\mathrm{j}\theta}) = C(r\mathrm{e}^{\mathrm{j}\theta} I - A)^{-1} B$ – transfer function
- Π – problem parameters (L_f, m_f)
- no need to find P

Rantzer, Syst. Control Lett. ’96

Sketch of the derivation

- **Choose $\mu = L_f - m_f$ and rewrite**

$$\begin{bmatrix} a(\zeta)I & b(\zeta)T^T \\ b(\zeta)T & c(\zeta)I + d(\zeta)TT^T \end{bmatrix} \succ 0, \forall \zeta \in [-1, 1]$$

  - functions $a(\zeta), b(\zeta), c(\zeta)$, and $d(\zeta)$ are parameterized by (α, r)

Sketch of the derivation

- **Choose $\mu = L_f - m_f$ and rewrite**

$$\begin{bmatrix} a(\zeta)I & b(\zeta)T^T \\ b(\zeta)T & c(\zeta)I + d(\zeta)TT^T \end{bmatrix} \succ \; 0, \, \forall \zeta \in [-1, 1]$$

  - functions $a(\zeta), b(\zeta), c(\zeta)$, and $d(\zeta)$ are parameterized by (α, r)

- **Take Schur complement and impose stability**
  - $a(\zeta) > 0, \; c(\zeta) + \left(d(\zeta) - \frac{b^2(\zeta)}{a(\zeta)} \right) TT^T \succ 0, \; \forall \, \zeta \in [-1, 1]$

Sketch of the derivation

- **Choose $\mu = L_f - m_f$ and rewrite**

$$\begin{bmatrix} a(\zeta)I & b(\zeta)T^T \\ b(\zeta)T & c(\zeta)I + d(\zeta)TT^T \end{bmatrix} \succ 0, \forall \zeta \in [-1, 1]$$

  - functions $a(\zeta), b(\zeta), c(\zeta)$, and $d(\zeta)$ are parameterized by (α, r)

- **Take Schur complement and impose stability**
  - $a(\zeta) > 0,\ c(\zeta) + \left(d(\zeta) - \frac{b^2(\zeta)}{a(\zeta)}\right) TT^T \succ 0,\ \forall\, \zeta \in [-1, 1]$
  - stability of transfer function $G(r\mathrm{e}^{\mathrm{j}\theta})$

Sketch of the derivation

- **Choose $\mu = L_f - m_f$ and rewrite**

$$\left[\begin{array}{cc} a(\zeta)I & b(\zeta)T^T \\ b(\zeta)T & c(\zeta)I + d(\zeta)TT^T \end{array}\right] \succ 0, \forall \zeta \in [-1, 1]$$

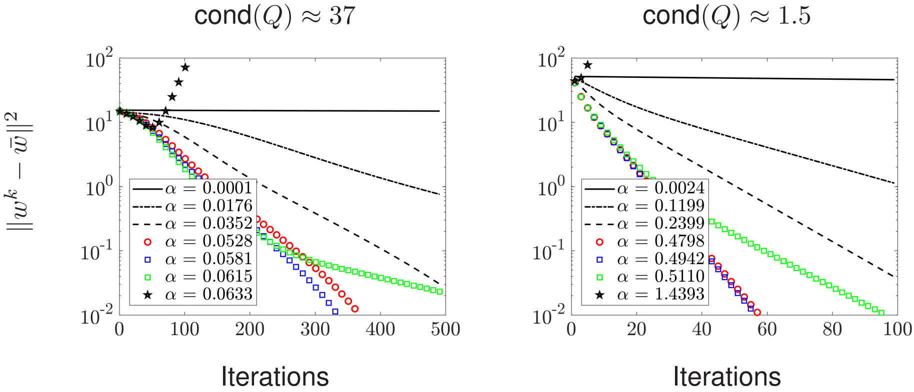
  - functions $a(\zeta), b(\zeta), c(\zeta)$, and $d(\zeta)$ are parameterized by (α, r)

- **Take Schur complement and impose stability**
  - $a(\zeta) > 0, \; c(\zeta) + \left(d(\zeta) - \frac{b^2(\zeta)}{a(\zeta)}\right) TT^T \succ 0, \; \forall \, \zeta \in [-1, 1]$
  - stability of transfer function $G(r\mathrm{e}^{\mathrm{j}\theta})$

$$\left.\begin{array}{l} \cdot \; f \text{ -- } m_f\text{-strongly convex} \\ \cdot \; \nabla f \text{ -- } L_f\text{-Lipschitz cts} \\ \cdot \; T \text{ -- row full rank} \end{array}\right\} \rightarrow \begin{array}{l} \text{Exponentially convergent if} \\ \quad \cdot \; \mu \geq L_f - m_f \\ \quad \cdot \; 0 < \alpha < \bar{\alpha}(\mu, m_f, L_f, \lambda_m(TT^T)) \end{array}$$

Quadratic programming

$$\begin{array}{ll} \underset{x,\,z}{\text{minimize}} & \frac{1}{2}\,x^T Q x \,+\, q^T x \,+\, g(z) \\ \text{subject to} & Tx \,-\, z \,=\, 0. \end{array}$$


cond$(Q) \approx 37$

cond$(Q) \approx 1.5$

Nishihara, Lessard, Recht, Packard, Jordan, ICML ’15

Summary

- **Results**
  - an iterative primal-dual algorithm
  - an automated tool
  - a region of step size with exp. converg.

- **Ongoing works**
  - reduce the conservative
  - other discretization methods

THANK YOU!