Part 6: Interior-point methods for inequality constrained optimization

Nick Gould (RAL)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \text{ subject to } c(x) \geq 0$$

Part C course on continuoue optimization

CONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \text{ subject to } c(x) \geq 0$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
and the **constraints** $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$

- assume that $f,\ c \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary

CONSTRAINTS AND MERIT FUNCTIONS

Two conflicting goals:

- minimize the objective function $f(x)$
- satisfy the constraints

Recall — overcome this by minimizing a composite **merit function** $\Phi(x,p)$ for which

- p are parameters
- (some) minimizers of $\Phi(x,p)$ wrt x approach those of $f(x)$ subject to the constraints as p approaches some set $\mathcal{P}$
- only uses **unconstrained** minimization methods

A MERIT F$^{\text{n}}$ FOR INEQUALITY CONSTRAINTS

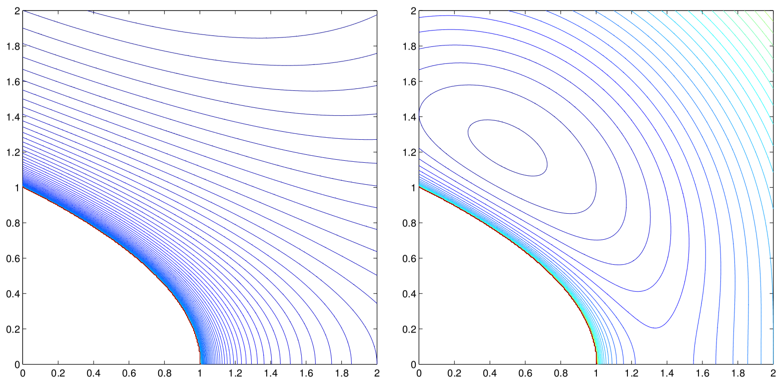
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \text{ subject to } c(x) \geq 0$$

Merit function **(logarithmic barrier function)**:

$$\Phi(x,\mu) = f(x) - \mu \sum_{i=1}^{m} \log c_i(x)$$

- required solution as μ approaches $\{0\}$ from above
- may have other useless stationary points
- requires a strictly interior point to start
- consequent points are interior

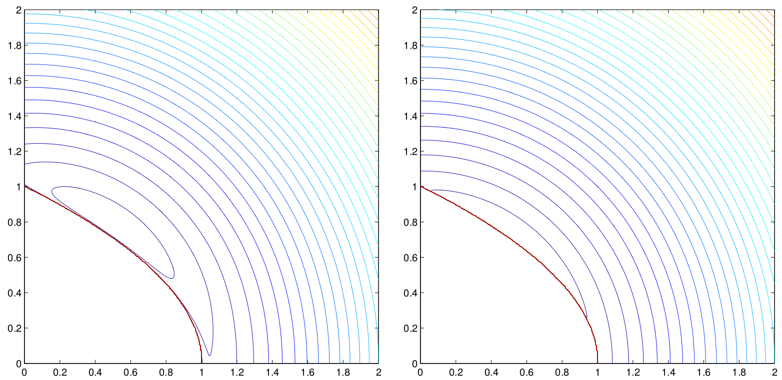
CONTOURS OF THE BARRIER FUNCTION



$\mu = 10$ $\mu = 1$

Barrier function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 \geq 1$

CONTOURS OF THE BARRIER FUNCTION (cont.)



$\mu = 0.1$ $\mu = 0.01$

Barrier function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 \geq 1$

BASIC BARRIER FUNCTION ALGORITHM

Given $\mu_0 > 0$, set $k = 0$
Until "convergence" iterate:
  Find $x_k^{\rm s}$ for which $c(x_k^{\rm s}) > 0$
  Starting from $x_k^{\rm s}$, use an unconstrained minimization algorithm to find an "approximate" minimizer x_k of $\Phi(x, \mu_k)$
  Compute $\mu_{k+1} > 0$ smaller than μ_k such that $\lim_{k\to\infty} \mu_{k+1} = 0$ and increase k by 1

- often choose $\mu_{k+1} = 0.1\mu_k$ or even $\mu_{k+1} = \mu_k^2$
- might choose $x_{k+1}^{\rm s} = x_k$

MAIN CONVERGENCE RESULT

The **active set** $\mathcal{A}(x) = \{i \mid c_i(x) = 0\}$

Theorem 6.1. Suppose that f, $c \in \mathcal{C}^2$, that $(y_k)_i \stackrel{\rm def}{=} \mu_k / c_i(x_k)$ for $i = 1, \ldots, m$, that

$$\|\nabla_x \Phi(x_k, \mu_k)\|_2 \leq \epsilon_k$$

where ϵ_k converges to zero as $k \to \infty$, and that x_k converges to x_* for which $\{a_i(x_*)\}_{i\in\mathcal{A}(x_*)}$ are linearly independent. Then x_* satisfies the first-order necessary optimality conditions for the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \ f(x) \ \ \text{subject to} \ \ c(x) \geq 0$$

and $\{y_k\}$ converge to the associated Lagrange multipliers y_*.

PROOF OF THEOREM 6.1

Let $\mathcal{M} \stackrel{\text{def}}{=} \{1, \ldots, m\}$, $\mathcal{A} \stackrel{\text{def}}{=} \{i \mid c_i(x_*) = 0\}$ and $\mathcal{I} \stackrel{\text{def}}{=} \mathcal{M} \setminus \mathcal{A}$.
Generalized inv. $A^+_{\mathcal{A}}(x) \stackrel{\text{def}}{=} \left(A_{\mathcal{A}}(x)A^T_{\mathcal{A}}(x)\right)^{-1} A_{\mathcal{A}}(x)$ bounded near x_*.
Define

$$(y_k)_i = \frac{\mu_k}{c_i(x_k)}, i \in \mathcal{M}, \quad (y_*)_{\mathcal{A}} = A^+_{\mathcal{A}}(x_*)g(x_*) \ \text{ and } \ (y_*)_{\mathcal{I}} = 0.$$

$$\|(y_k)_{\mathcal{I}}\|_2 \leq 2\mu_k \sqrt{|\mathcal{I}|} / \min_{i\in\mathcal{I}} |c_i(x_*)| \tag{1}$$

(if $\mathcal{I} \neq \emptyset$) for all sufficiently large k. (1) + inner-it. termination $\Longrightarrow$

$$\begin{aligned}\|g(x_k) - A^T_{\mathcal{A}}(x_k)(y_k)_{\mathcal{A}}\|_2 \leq\ & \|g(x_k) - A^T(x_k)y_k\|_2 + \|A^T_{\mathcal{I}}(x_k)(y_k)_{\mathcal{I}}\|_2 \\ \leq\ & \bar{\epsilon}_k \stackrel{\text{def}}{=} \epsilon_k + \mu_k \frac{2\sqrt{|\mathcal{I}|}\|A_{\mathcal{I}}\|_2}{\min_{i\in\mathcal{I}} |c_i(x_*)|}\end{aligned} \tag{2}$$

$$\begin{aligned}\Longrightarrow \|A^+_{\mathcal{A}}(x_k)g(x_k) - (y_k)_{\mathcal{A}}\|_2 =\ & \|A^+_{\mathcal{A}}(x_k)(g(x_k) - A^T_{\mathcal{A}}(x_k)(y_k)_{\mathcal{A}})\|_2 \\ \leq\ & 2\|A^+_{\mathcal{A}}(x_*)\|_2 \bar{\epsilon}_k\end{aligned}$$

$$\begin{aligned}\Longrightarrow\ & \|(y_k)_{\mathcal{A}} - (y_*)_{\mathcal{A}}\|_2 \\ \leq\ & \|A^+_{\mathcal{A}}(x_*)g(x_*) - A^+_{\mathcal{A}}(x_k)g(x_k)\|_2 + \|A^+_{\mathcal{A}}(x_k)g(x_k) - (y_k)_{\mathcal{A}}\|_2\end{aligned}$$

+ (1) $\Longrightarrow \{y_k\} \longrightarrow y_*$. Continuity of gradients + (2) $\Longrightarrow$

$$g(x_*) - A^T(x_*)y_* = 0$$

$c(x_k) > 0$, defs. of y_k and $y_* + c_i(x_k)(y_k)_i = \mu_k \Longrightarrow$
$c(x_*) \geq 0$, $y_* \geq 0$ and $c_i(x_*)(y_*)_i = 0$.
$\Longrightarrow$ (x_*, y_*) satisfies the first-order optimality conditions.

ALGORITHMS TO MINIMIZE $\Phi(x, \mu)$

Can use

- ⊙ linesearch methods
  - ⋄ should use specialized linesearch to cope with singularity of log
- ⊙ trust-region methods
  - ⋄ need to reject points for which $c(x_k + s_k) \not> 0$
  - ⋄ (ideally) need to "shape" trust region to cope with contours of the singularity

DERIVATIVES OF THE BARRIER FUNCTION

- ⊙ $\nabla_x \Phi(x, \mu) = g(x, y(x))$
- ⊙ $$\begin{aligned}\nabla_{xx}\Phi(x, \mu) &= H(x, y(x)) + \mu A^T(x)C^{-2}(x)A(x) \\ &= H(x, y) + A^T(x)C^{-1}(x)Y(x)A(x) \\ &= H(x, y) + \tfrac{1}{\mu}A^T(x)Y^2(x)A(x)\end{aligned}$$

where

- ⊙ **Lagrange multiplier estimates**: $y(x) = \mu C^{-1}(x)e$ where e is the vector of ones
- ⊙ $C(x) = \mathrm{diag}(c_1(x), \ldots, c_m(x))$
- ⊙ $Y(x) = \mathrm{diag}(y_1(x), \ldots, y_m(x))$
- ⊙ $g(x, y(x)) = g(x) - A^T(x)y(x)$: **gradient of the Lagrangian**
- ⊙ $H(x, y(x)) = H(x) - \sum_{i=1}^{m} y_i(x)H_i(x)$: **Lagrangian Hessian**

LIMITING DERIVATIVES OF Φ

Let $\mathcal{I}$ = inactive set at $x_* = \{1,\ldots,m\} \setminus \mathcal{A}$
For small μ: roughly

$$\nabla_x \Phi(x,\mu) = \underbrace{g(x) - A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^{-1}(x) e}_{\text{moderate}} - \underbrace{\mu A_{\mathcal{I}}^T(x) C_{\mathcal{I}}^{-1}(x) e}_{\text{small}}$$
$$\approx g(x) - A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^{-1}(x) e$$

$$\nabla_{xx} \Phi(x,\mu) = \underbrace{H(x,y(x))}_{\text{moderate}} + \underbrace{\mu A_{\mathcal{I}}^T(x) C_{\mathcal{I}}^{-2}(x) A_{\mathcal{I}}(x)}_{\text{small}} + \underbrace{\frac{1}{\mu} A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^2(x) A_{\mathcal{A}}(x)}_{\text{large}}$$
$$\approx \frac{1}{\mu} A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^2(x) A_{\mathcal{A}}(x)$$
$$= A_{\mathcal{A}}^T(x) C_{\mathcal{A}}^{-1}(x) Y_{\mathcal{A}}(x) A_{\mathcal{A}}(x)$$
$$= \mu A_{\mathcal{A}}^T(x) C_{\mathcal{A}}^{-2}(x) A_{\mathcal{A}}(x)$$

GENERIC BARRIER NEWTON SYSTEM

Newton correction s from x for barrier function is

$$\big(H(x,y(x)) + A^T(x) C^{-1}(x) Y(x) A(x)\big)\, s = -g(x,y(x))$$

LIMITING NEWTON METHOD

For small μ: roughly

$$\mu A_{\mathcal{A}}^T(x) C_{\mathcal{A}}^{-2}(x) A_{\mathcal{A}}(x) s \approx -\big(g(x) - A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^{-1}(x) e\big)$$

POTENTIAL DIFFICULTIES I

Ill-conditioning of the Hessian of the barrier function:
roughly speaking (non-degenerate case)

- m_a eigenvalues $\approx \lambda_i \left[A_{\mathcal{A}}^T Y_{\mathcal{A}}^2 A_{\mathcal{A}}\right] / \mu_k$
- $n - m_a$ eigenvalues $\approx \lambda_i \left[N_{\mathcal{A}}^T H(x_*, y_*) N_{\mathcal{A}}\right]$

where
m_a = number of active constraints
$\mathcal{A}$ = active set at x_*
Y = diagonal matrix of Lagrange multipliers
$N_{\mathcal{A}}$ = orthogonal basis for null-space of $A_{\mathcal{A}}$

$\Longrightarrow$ condition number of $\nabla_{xx}\Phi(x_k,\mu_k) = O(1/\mu_k)$
$\Longrightarrow$ may not be able to find minimizer easily

POTENTIAL DIFFICULTIES II

Value $x_{k+1}^{\mathrm{s}} = x_k$ is a poor starting point: Suppose

$$0 \approx \nabla_x \Phi(x_k,\mu_k) = g(x_k) - \mu_k A^T(x_k) C^{-1}(x_k) e$$
$$\approx g(x_k) - \mu_k A_{\mathcal{A}}^T(x_k) C_{\mathcal{A}}^{-1}(x_k) e$$

Roughly speaking (non-degenerate case) Newton correction satisfies

$$\mu_{k+1} A_{\mathcal{A}}^T(x_k) C_{\mathcal{A}}^{-2}(x_k) A_{\mathcal{A}}(x_k) s \approx (\mu_{k+1} - \mu_k) A_{\mathcal{A}}^T(x_k) C_{\mathcal{A}}^{-1}(x_k) e$$

$\Longrightarrow$ (full rank)

$$A_{\mathcal{A}}(x_k) s \approx \left(1 - \frac{\mu_k}{\mu_{k+1}}\right) c_{\mathcal{A}}(x_k)$$

$\Longrightarrow$ (Taylor expansion)

$$c_{\mathcal{A}}(x_k + s) \approx c_{\mathcal{A}}(x_k) + A_{\mathcal{A}}(x_k) s \approx \left(2 - \frac{\mu_k}{\mu_{k+1}}\right) c_{\mathcal{A}}(x_k) < 0$$

if $\mu_{k+1} < \frac{1}{2}\mu_k \Longrightarrow$ Newton step infeasible $\Longrightarrow$ slow convergence

PERTURBED OPTIMALITY CONDITIONS

First order optimality conditions for

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x)\ \text{ subject to }\ c(x) \geq 0$$

are:

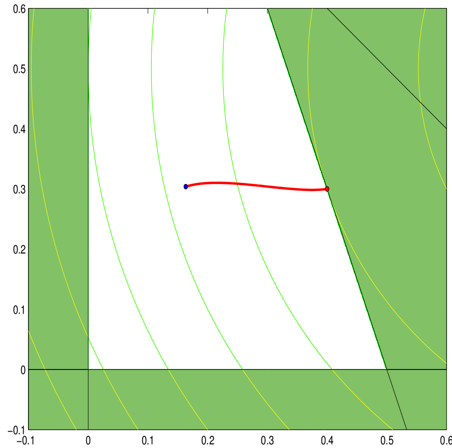
$$\begin{array}{cc} g(x) - A^T(x)y = 0 & \text{dual feasibility} \\ C(x)y = 0 & \text{complementary slackness} \\ c(x) \geq 0 \ \text{ and } \ y \geq 0 & \end{array}$$

Consider the "perturbed" problem

$$\begin{array}{cc} g(x) - A^T(x)y = 0 & \text{dual feasibility} \\ C(x)y = \mu e & \textbf{perturbed}\ \text{comp. slkns.} \\ c(x) > 0 \ \text{ and } \ y > 0 & \end{array}$$

where $\mu > 0$

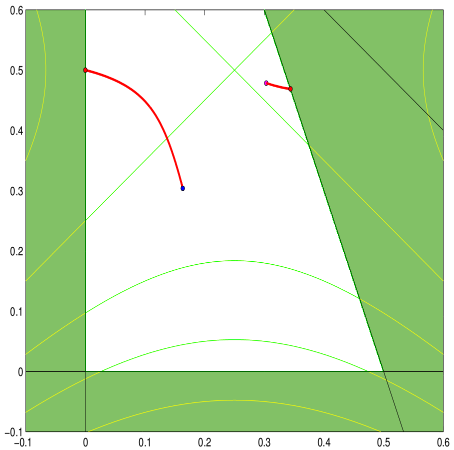
CENTRAL PATH TRAJECTORY



$$\begin{array}{c} \min (x_1 - 1)^2 + (x_2 - 0.5)^2 \\ \text{subject to } \ x_1 + x_2 \leq 1 \\ 3x_1 + x_2 \leq 1.5 \\ (x_1, x_2) \geq 0 \end{array}$$

Trajectory $x(\mu)$ of perturbed optimality conditions
as μ ranges from infinity down to zero

TRAJECTORIES FOR THE NON-CONVEX CASE



$$\begin{array}{c} \min -2(x_1 - 0.25)^2 + 2(x_2 - 0.5)^2 \\ \text{subject to } \ x_1 + x_2 \leq 1 \\ 3x_1 + x_2 \leq 1.5 \\ (x_1, x_2) \geq 0 \end{array}$$

Trajectories $x(\mu)$ of perturbed optimality conditions
as μ ranges from infinity down to zero

PRIMAL-DUAL PATH-FOLLOWING METHODS

Track roots of

$$g(x) - A^T(x)y = 0 \ \text{ and } \ C(x)y - \mu e = 0$$

as $0 < \mu \rightarrow 0$, while maintaining $c(x) > 0$ and $y > 0$

⊙ nonlinear system $\Longrightarrow$ use Newton's method

Newton correction (s, w) to (x, y) satisfies

$$\begin{pmatrix} H(x,y) & -A^T(x) \\ YA(x) & C(x) \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} g(x) - A^T(x)y \\ C(x)y - \mu e \end{pmatrix}$$

Eliminate $w \Longrightarrow$

$$\big(H(x,y) + A^T(x)C^{-1}(x)YA(x)\big)\, s = - \big(g(x) - \mu A^T(x)C^{-1}(x)e\big)$$

c.f. Newton method for barrier minimization!

PRIMAL VS. PRIMAL-DUAL

Primal:

$$\left(H(x,y(x)) + A^T(x)C^{-1}(x)Y(x)A(x)\right) s^{\mathrm{P}} = -g(x,y(x))$$

Primal-dual:

$$\left(H(x,y) + A^T(x)C^{-1}(x)YA(x)\right) s^{\mathrm{PD}} = -g(x,y(x))$$

where

$$y(x) = \mu C^{-1}(x)e$$

What is the difference?

- ⊙ freedom to choose y in $H(x,y) + A^T(x)C^{-1}(x)YA(x)$ for primal-dual . . . vital
- ⊙ Hessian approximation for small μ

$$H(x,y) + A^T(x)C^{-1}(x)YA(x) \approx A_{\mathcal{A}}^T(x)C_{\mathcal{A}}^{-1}(x)Y_{\mathcal{A}}A_{\mathcal{A}}(x)$$

POTENTIAL DIFFICULTY II . . . REVISITED

Value $x_{k+1}^{\mathrm{S}} = x_k$ can be a good starting point:

- ⊙ primal method has to choose $y = y(x_k^{\mathrm{S}}) = \mu_{k+1}C^{-1}(x_k)e$
  - ⋄ factor μ_{k+1}/μ_k too small for a good Lagrange multiplier estimate
- ⊙ primal-dual method can choose $y = \mu_k C^{-1}(x_k)e \rightarrow y_*$

Advantage: roughly (non-degenerate case) correction s^{PD} satisfies

$$\mu_k A_{\mathcal{A}}^T(x_k)C_{\mathcal{A}}^{-2}(x_k)A_{\mathcal{A}}(x_k)s^{\mathrm{PD}} \approx (\mu_{k+1} - \mu_k)A_{\mathcal{A}}^T(x_k)C_{\mathcal{A}}^{-1}(x_k)e$$

$\Longrightarrow$ (full rank)

$$A_{\mathcal{A}}(x_k)s^{\mathrm{PD}} \approx \left(\frac{\mu_{k+1}}{\mu_k} - 1\right) c_{\mathcal{A}}(x_k)$$

$\Longrightarrow$ (Taylor expansion)

$$c_{\mathcal{A}}(x_k + s^{\mathrm{PD}}) \approx c_{\mathcal{A}}(x_k) + A_{\mathcal{A}}(x_k)s^{\mathrm{PD}} \approx \frac{\mu_{k+1}}{\mu_k} c_{\mathcal{A}}(x_k) > 0$$

$\Longrightarrow$ Newton step allowed $\Longrightarrow$ fast convergence

PRIMAL-DUAL BARRIER METHODS

Choose a search direction s for $\Phi(x,\mu_k)$ by (approximately) solving the problem

$$\underset{s\in\mathbb{R}^n}{\text{minimize}} \;\; g(x,y(x))^T s + \tfrac{1}{2}s^T\left(H(x,y) + A^T(x)C^{-1}(x)YA(x)\right)s$$

possibly subject to a trust-region constraint

- ⊙ $y(x) = \mu C^{-1}(x)e \Longrightarrow g(x,y(x)) = \nabla_x\Phi(x,\mu)$
- ⊙ $y = \ldots$
  - ⋄ $y(x) \Longrightarrow$ primal Newton method
  - ⋄ occasionally $(\mu_{k-1}/\mu_k)y(x) \Longrightarrow$ good starting point
  - ⋄ $y^{\mathrm{OLD}} + w^{\mathrm{OLD}} \Longrightarrow$ primal-dual Newton method
  - ⋄ $\max(y^{\mathrm{OLD}} + w^{\mathrm{OLD}}, \epsilon(\mu_k)e)$ for "small" $\epsilon(\mu_k) > 0$ (e.g., $\epsilon(\mu_k) = \mu_k^{1.5}$) $\Longrightarrow$ practical primal-dual method

POTENTIAL DIFFICULTY I . . . REVISITED

Ill-conditioning $\not\Longrightarrow$ we can't solve equations accurately:

roughly (non-degenerate case, $\mathcal{I}$ = inactive set at x_*)

$$\begin{pmatrix} H & -A^T \\ YA & C \end{pmatrix}\begin{pmatrix} s \\ w \end{pmatrix} = -\begin{pmatrix} g - A^T y \\ Cy - \mu e \end{pmatrix} \Longrightarrow$$

$$\begin{pmatrix} H & -A_{\mathcal{A}}^T & -A_{\mathcal{I}}^T \\ Y_{\mathcal{A}}A_{\mathcal{A}} & C_{\mathcal{A}} & 0 \\ Y_{\mathcal{I}}A_{\mathcal{I}} & 0 & C_{\mathcal{I}} \end{pmatrix}\begin{pmatrix} s \\ w_{\mathcal{A}} \\ w_{\mathcal{I}} \end{pmatrix} = -\begin{pmatrix} g - A_{\mathcal{A}}^T y_{\mathcal{A}} - A_{\mathcal{I}}^T y_{\mathcal{I}} \\ C_{\mathcal{A}}y_{\mathcal{A}} - \mu e \\ C_{\mathcal{I}}y_{\mathcal{I}} - \mu e \end{pmatrix} \Longrightarrow$$

$$\begin{pmatrix} H + A_{\mathcal{I}}^T C_{\mathcal{I}}^{-1} Y_{\mathcal{I}} A_{\mathcal{I}} & -A_{\mathcal{A}}^T \\ A_{\mathcal{A}} & C_{\mathcal{A}}Y_{\mathcal{A}}^{-1} \end{pmatrix}\begin{pmatrix} s \\ w_{\mathcal{A}} \end{pmatrix} = -\begin{pmatrix} g - A_{\mathcal{A}}^T y_{\mathcal{A}} - \mu A_{\mathcal{I}}^T C_{\mathcal{I}}^{-1} e \\ c_{\mathcal{A}} - \mu Y_{\mathcal{A}}^{-1} e \end{pmatrix}$$

- ⊙ potentially bad terms $C_{\mathcal{I}}^{-1}$ and $Y_{\mathcal{A}}^{-1}$ bounded
- ⊙ in the limit becomes well-behaved

$$\begin{pmatrix} H & -A_{\mathcal{A}}^T \\ A_{\mathcal{A}} & 0 \end{pmatrix}\begin{pmatrix} s \\ w_{\mathcal{A}} \end{pmatrix} = -\begin{pmatrix} g - A_{\mathcal{A}}^T y_{\mathcal{A}} \\ 0 \end{pmatrix}$$

PRACTICAL PRIMAL-DUAL METHOD

Given $\mu_0 > 0$ and feasible $(x_0^{\mathrm{s}}, y_0^{\mathrm{s}})$, set $k = 0$
Until "convergence" iterate:
  Inner minimization: starting from $(x_k^{\mathrm{s}}, y_k^{\mathrm{s}})$, use an unconstrained minimization algorithm to find (x_k, y_k) for which
   $\|C(x_k)y_k - \mu_k e\| \leq \mu_k$ and $\|g(x_k) - A^T(x_k)y_k\| \leq \mu_k^{1.00005}$
  Set $\mu_{k+1} = \min(0.1\mu_k, \mu_k^{1.9999})$
  Find $(x_{k+1}^{\mathrm{s}}, y_{k+1}^{\mathrm{s}})$ using a primal-dual Newton step from (x_k, y_k)
  If $(x_{k+1}^{\mathrm{s}}, y_{k+1}^{\mathrm{s}})$ is infeasible, reset $(x_{k+1}^{\mathrm{s}}, y_{k+1}^{\mathrm{s}})$ to (x_k, y_k)
  Increase k by 1

OTHER ISSUES

- polynomial algorithms for many convex problems
  - linear programming
  - quadratic programming
  - semi-definite programming …
- excellent practical performance
- globally, need to keep away from constraint boundary until near convergence, otherwise very slow
- initial interior point:

$$\underset{(x,c)}{\text{minimize}} \;\; e^T c \;\; \text{subject to} \;\; c(x) + c \geq 0$$

FAST ASYMPTOTIC CONVERGENCE

Theorem 6.2. Suppose that f, $c \in \mathcal{C}^2$, that a subsequence $\{(x_k, y_k)\}$, $k \in \mathcal{K}$, of the practical primal-dual method converges to (x_*, y_*) satisfying second-order sufficiency conditions, that $A_{\mathcal{A}}(x_*)$ is full-rank, and that $(y_*)_{\mathcal{A}} > 0$. Then the starting point satisfies the inner-minimization termination test (i.e., $(x_k, y_k) = (x_k^{\mathrm{s}}, y_k^{\mathrm{s}})$) and the whole sequence $\{(x_k, y_k)\}$ converges to (x_*, y_*) at a superlinear rate (Q-factor 1.9998).