A new Perspective on the Complexity of Interior Point Methods for LP

IFOR Seminar
ETH Zürich, Oct 13 2008

Raphael Hauser, Oxford
Coauthor: Coralia Cartis, Edinburgh

Interior Point Methods for Linear Programming

Interior-point methods for linear programming are based on replacing the constrained optimisation problem

$$\text{(P)} \quad \begin{aligned} &\min_x c^\mathsf{T} x \\ &\text{s.t. } Ax = b \\ &\qquad x \geq 0 \end{aligned}$$

with strictly feasible domain $\mathscr{F}^\circ = \{x : Ax = b, x > 0\}$ by a sequence of equality constrained optimisation problems

$$(\mathsf{P}_\mu) \quad \begin{aligned} &\min_x c^\mathsf{T} x + \mu f(x) \\ &\text{s.t. } Ax = b, \end{aligned}$$

where $f(x)$ is strictly convex on $\mathbb{R}^n_{++}$. The idea is to gradually decrease μ to zero from a positive value.

$\longrightarrow$ drawing on black board!

Let $x(\mu)$ be the $\arg\min$ of problem (P_μ). Then $\mu \mapsto x(\mu)$ is called the *primal central path*.

Let $F_\mu(x) = c^\mathsf{T} x - \mu f(x)$, and let

$$n_{F_\mu}(x) = -(\mathsf{D}^2 F_\mu(x))^{-1} \nabla F_\mu(x)$$

be the corresponding Newton step.

A generic idea for *short-step* path-following method is as follows:

- Choose radii $0 < r(\mu) < R(\mu)$ such that $\mathsf{B}_{R(\mu)}(x(\mu)) \subset \mathbb{R}^n_{++}$, and such that applying a fixed number $\widehat{\ell}$ iterative updates

$$x_+ = x + n_{F_\mu}(x)$$

to a starting point $x \in \mathsf{B}_{R(\mu)}(x(\mu))$ leads to a point in $\mathsf{B}_{r(\mu)}(x(\mu))$ ($\mathsf{B}_{R(\mu)}$ is contracted to $\mathsf{B}_{r(\mu)}$ in a fixed number of steps).

- Choose $\theta \in (0, 1)$ such that for $\mu_+ = \theta\mu$ we have,

$$x \in \mathsf{B}_{r(\mu)}(x(\mu)) \Rightarrow x \in \mathsf{B}_{R(\mu_+)}(x(\mu_+)).$$

- Self-similarity occurs when $R(\mu) = \rho\mu$, $r(\mu) = \xi\rho\mu$ for fixed $\rho > 0$ and $\xi \in (0, 1)$.

$\longrightarrow$ drawing on black board!

Which types of barrier terms can be used in this construction?

- The log-barrier $f(x) = -\sum_i \ln x_i$ leads to such a construction.

- The barrier $f(x) = \sum_i x_i^{-1}$ is not a good choice. Why?

- The most general $f(x)$ known to work are *self-concordant* barriers (Nesterov & Nemirovskii)

$$
\begin{array}{rl}
\text{(i)} & f \in \mathscr{C}^3(\mathscr{F}^\circ), \\
\text{(ii)} & \mathrm{D}^2 f(x) \succ 0, \quad (x \in \mathscr{F}^\circ), \\
\text{(iii)} & |\,\mathrm{D}^3 f(x)[h,h,h]| \leq 2\,\mathrm{D}^2 f(x)[h,h]^{3/2}, \quad (x \in \mathscr{F}^\circ,\, h \in \mathbb{R}^n), \\
\text{(iv)} & \mathrm{D}^2 f(x)[\nabla f(x), \nabla f(x)] \leq \eta, \quad (x \in \mathscr{F}^\circ).
\end{array}
$$

Theorem 1 (Nesterov-Nemirovskii). *Every convex domain* $\mathscr{C} \subset \mathbb{R}^n$ *admits a self-concordant barrier* $f(x)$ *(universal barrier).*

Thus, in principle all convex programming problems

$$\begin{aligned} &\min_x c^\mathsf{T} x \\ &\text{s.t. } x \in \mathscr{C} \end{aligned}$$

can be solved via interior-point methods.

Unfortunately, there are obstacles:

- $f(x)$ can only be efficiently evaluated for a restricted class of domains $\mathscr{C}$.

- If a primal-dual algorithm is desired, the domain needs to be of the form

$$\bar{\mathscr{C}} = \{Ax = b\} \cap K,$$

where K is a *symmetric cone*, and $f(x)$ needs to inherit the symmetry properties of K (Nesterov & Todd).

In the latter case, one can show that only the log-barrier satisfies all the required properties (H., H.-Lim, H.-Güler, Schmieta),

$$f(x) = c_0 + \oplus_{i=1}^{m} c_i \ln \det_{K_i}(x_i).$$

Stiff Vector Fields and Optimisation

Consider an unconstrained minimisation problem

$$x^* = \arg\min_{x \in \mathbb{R}^n} f(x),$$

or, more specifically, the toy model

$$x^* = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top B x,$$

where $B = \mathrm{Diag}(\lambda_1, \lambda_2)$ and $\lambda_1 \gg \lambda_2 > 0$. The *steepest descent* method for this problem proceeds as follows,

1. $v(x) = -\nabla f(x)$,

2. $h(x) = \arg\min_{t \geq 0} f(x + tv(x)) = \mathscr{O}(\lambda_1^{-1})$,

3. $x_+ = x + h(x)v(x)$, goto 1.

We wish to iterate this method until $f(x) < f(x^*) + \varepsilon$. The complexity of this method is then

$$\mathscr{O}\left(-\frac{\lambda_1 \log \varepsilon}{\lambda_2}\right) = \mathscr{O}\left(\kappa(B)|\log \varepsilon|\right)$$

iterations.

Interpreting the updating formula $x_+ = x + h(x)v(x)$ as discrete time-stepping for the dynamical system

$$\dot{x} = -v(x)$$

with time step $\Delta t = h(x)$, we are applying the forward Euler method.

Steepest descent behaves poorly because

- the "fast" eigenvalue λ_1 restricts the stepsize $h(x)$,

- the "slow" eigenvalue λ_2 imposes a long integration time.

Numerical analyst call this phenomenon *stiffness* of the vector field $v(x)$.

$\longrightarrow$ drawing on black board!

Numerical analysts know that *backward Euler steps*

$$x = x_+ - \Delta t \dot{x}_+$$

largely overcome this problem. In our optimisation context, the *linearised* backward Euler method

$$\begin{aligned} x = x_+ - \Delta t \left(-\nabla f(x) + (\mathsf{D}^2 f(x))(x - x_+)\right) \\ \approx -\nabla f(x_+), \end{aligned}$$

is a tractable alternative with similar stability properties,

$$x_+ = x - \left(\Delta t^{-1}\mathrm{I} + \mathsf{D}^2 f(x)\right) \nabla f(x).$$

($\rightarrow$ Levenberg-Marquart method, Newton's method.)

If $v(x)$ is however nonstiff, it is fine to use the forward Euler method in an optimisation context:

Theorem 2 (Cartis-H.)**.** *Let* $\beta \in (0, 1)$ *and*

$$\mathcal{N} := \{x \in \mathbb{R}^n : \|x - x^\dagger\| < R\},$$

for some $R > 0$ *and* $x^\dagger \in \mathbb{R}^n$*. Letting* $\overline{\mathcal{N}}$ *denote the closure of* $\mathcal{N}$*, we assume* $v : \overline{\mathcal{N}} \to \mathbb{R}^n$, $x \mapsto v(x)$ *is a* C^1 *vector field such that*

i) $v(x) = 0 \Leftrightarrow x = x^\dagger$,

ii) $\|I + Dv(x)\| \leq \beta$, *for all* $x \in \mathcal{N}$.

We consider the iterative process

$$x^{l+1} = x^l + v(x^l), \quad l \geq 0,$$

where x^0 is an arbitrary starting point in $\mathcal{N}$. Then

$$\|x^{l+1} - x^\dagger\| \leq \beta \|x^l - x^\dagger\|, \quad l \geq 0,$$

which implies

$$x^0 \in \mathcal{N} \quad \Longrightarrow \quad x^l \in \mathcal{N},\ l \geq 0,$$

and

$$x^l \to x^\dagger,\ \textit{as } l \to \infty,\ \textit{Q-linearly with convergence factor } \beta.$$

Furthermore, we have

$$\|v(x^l)\| \leq (1 + \beta)\|x^l - x^\dagger\|, \quad l \geq 0.$$

Thus

$$v(x^l) \to 0,\ \textit{as } l \to \infty,\ \textit{R-linearly with convergence factor } \beta.$$

Back to Short-Step IPMs for Linear Programming

Theorem 2 can explain the convergence of short-step IPMs:

$$\begin{aligned} \min_x \ & c^\top x \\ \text{s.t.} \ & Ax = b \\ & x \geq 0. \end{aligned}$$

- The central path $x(\mu)$, defined by

$$\begin{aligned} x(\mu) = \arg\min_x F_\mu(x) &= c^\top x - \mu \sum_i \ln x_i \\ \text{s.t.} \ Ax &= b, \end{aligned}$$

  satisfies

$$d(x(\mu), \partial \mathscr{F}^\circ) \begin{cases} = \Theta(\mu) \\ \leq C\mu, \end{cases}$$

where

$$C = \sup_{\mu \in (0,\mu_0]} \| \frac{d}{d\mu} x(\mu) \|$$

plays the role of a condition number.

- The directed family of vector fields $(n_{F_\mu}(x))_{(0,\mu_0]}$ satisfies

$$\begin{aligned} &\text{(i)} \quad n_{F_\mu}(x) = 0 \Leftrightarrow x = x(\mu), \\ &\text{(ii)} \quad \| \mathrm{I} + \mathrm{D}\, n_{F_\mu}(x) \| \leq 2\sqrt{n}(1 + \bar{\chi}(A)) \frac{\rho}{\rho_0} =: \beta, \end{aligned}$$

where

$$\bar{\chi}(A) = \sup\{ \|A^\top (ADA^\top)^{-1} AD\| : D \succ 0 \text{ diagonal}\}$$

is the Vavasis-Ye condition number,

$$\rho_0 = (C\mu_0 + (1 + \bar{\chi}(A))\|c\|)^{-1},$$

and $\rho \in (0, \rho_0)$ is chosen so that $\beta < 1$.

Hence, Theorem 2 is applicable and shows that the following generic short-step ipm has inner complexity (number of Newton steps per outer iteration)

$$\hat{\ell} = \left\lceil \frac{\log \xi}{\log\left((2\sqrt{n}(1+\bar{\chi}))(C\mu_0 + (1+\bar{\chi})\|c\|)\rho\right)} \right\rceil,$$

and outer complexity (number of outer iterations until $\mu < \varepsilon$)

$$\hat{k} = \left\lceil \left(\log \frac{1 + C\rho^{-1}}{\xi + C\rho^{-1}}\right)^{-1} \log \frac{\mu_0}{\varepsilon} \right\rceil.$$

Step 1 If $\mu_k < \varepsilon$, stop.

Step 2 $\mu_{k+1} = \theta\mu_k$, $x^{k,0} = x_k$.

for $\ell = 0, \ldots, \widehat{\ell} - 1 = \lfloor \log \xi / \log \beta \rfloor$

$$x^{k,\ell+1} = x^{k,\ell} + n_{F_\mu}(x^{k,\ell})$$

$x^{k+1} = x^{k,\widehat{\ell}}$.

Step 3 $k \leftarrow k + 1$, goto Step 1.

Theorem 3 (Cartis-H.). *Choosing* $\theta \in (\theta_0, 1)$, *where*

$$\theta_0 = \frac{\xi\rho + C}{\rho + C},$$

the above ipm takes at most

$$\hat{k} = \left\lceil \left(\log \frac{1 + C\rho^{-1}}{\xi + C\rho^{-1}}\right)^{-1} \log \frac{\mu_0}{\varepsilon} \right\rceil$$

outer iterations.

More generally, the same algorithm and theorem apply for any pair $(x(\mu), (v_\mu(x))_{(0,\mu_0]})$ that satisfies the LSDA properties

- $\mathscr{U} \subset \mathbb{R}^n$ an open convex domain,

- $x : \mu \to x(\mu) \in \mathscr{U}$ a differentiable path such that

$$\sup_{\mu \in (0,\mu_0]} \|\dot{x}(\mu)\| = C < \infty,$$

- $\rho > 0$ a constant such that $\mathsf{B}_{\rho\mu}(x(\mu)) \subset \mathscr{U}$ for all μ,

- $(v_\mu(x))_{(0,\mu_0]}$ a directed set of vector fields such that

  (i) $v_\mu(x) = 0 \Leftrightarrow x = x(\mu)$,

  (ii) $\|x + v_\mu(x) - x(\mu)\| \leq \beta \|x - x(\mu)\|, \quad (x \in \mathsf{B}_{\rho\mu}(x(\mu)))$

for some fixed $\beta \in (0,1)$.

Practical implication: If $(x(\mu), (v_\mu(x))_{(0,\mu_0]})$ has the LSDA property, then so does

$$(x(\mu), (v_\mu(x) + w_\mu(x))_{(0,\mu_0]}),$$

where $w_\mu(x)$ is any vector field that satisfies

$$\|w_\mu(x)\| < (1-\beta)\|x - x(\mu)\|.$$

Thus, if the linear system in each inner iteration of a short-step ipm is solved only to constant relative accuracy, the algorithm stays on track!

Thanks for listening!