REGULARITY ILLUSTRATED

Adrian Lewis

Cornell

5 September, 2008

people.orie.cornell.edu/∼aslewis

Augsburg

Joint work with:
D. Leventhal, R. Luke
J. Malick, C.H.J. Pang. . .

1. OUTLINE

- Error bounds, well-posedness, and algorithm speed (Demmel. . .).
- Simple randomized algorithms for linear constraints.
- Regularity and a local analogue of Demmel's paradigm.
- Iterated averaged projections for intersection problems.
- Prevalence of regularity for semi-algebraic constraints..
- Pseudospectra.

2. EXAMPLE: ITERATIVE SOLVERS

Consider solving a positive-definite system for $x \in \mathbf{R}^n$:

$$Ax = b.$$

Equivalently,

$$\min_x \Big\{ x^T A x - 2b^T x \Big\}.$$

- Small **distance to ill-posedness** $\lambda_{\text{min}}(A)$.
- Weak **error bound**:

$$\|x - A^{-1}b\| \;\leq\; \frac{1}{\lambda_{\text{min}}} \|Ax - b\|.$$

- Slow **linear rate** of basic algorithms. Eg:

$$\left(\frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}} + \lambda_{\text{min}}} \right)^2$$

  for steepest descent. Conjugate gradients analogous.

How general is this pattern?

3. EXAMPLE: COORDINATE DESCENT

A randomized method to minimize

$$\begin{aligned} f(x) &= x^T A x - 2b^T x : \\ x^{r+1} &\leftarrow \operatorname*{argmin}_{x^r + \mathbf{R} e_i} f \quad \text{with probability } \frac{A_{ii}}{\text{trace}\, A}. \end{aligned}$$

Iterations are **cheap**: $O(n)$ operations, whereas steepest descent or conjugate gradients needs $O(n^2)$.

Then (Leventhal-Lewis '08) $f(x_r) \to \min f$ "in expectation" with **linear rate**

$$1 - \frac{\lambda_{\text{min}}(A)}{\text{trace}\, A} \;\leq\; 1 - \frac{1}{n} \frac{\lambda_{\text{min}}}{\lambda_{\text{max}}}.$$

(The speed of "cyclic" coordinate descent is harder to quantify.)

4. EXAMPLE: LINEAR INEQUALITIES

A random iterated projection method to find $x \in S = \cap_i H_i$,

$$\text{where} \quad H_i \;=\; \{x : a_i^T x \le b_i\} :$$

$$x^{r+1} \;\leftarrow\; \text{proj}_{H_i} x^r \quad \text{with probability } \frac{\|a_i\|^2}{\|A\|_F^2}.$$

Iterations are cheap: $O(n)$ operations. (Strohmer-Vershynin '07, L-L '08): $\text{dist}_S^2(x_r) \to 0$ in expectation, at **linear rate**

$$1 - \frac{1}{L^2 \|A\|_F^2},$$

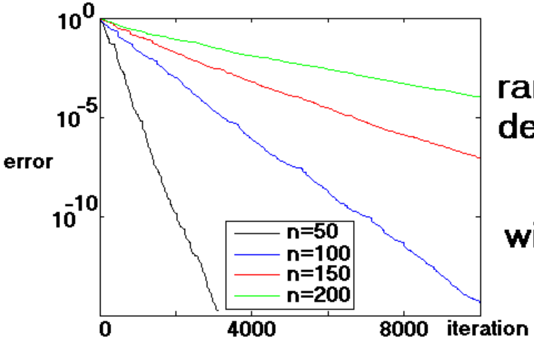
where L is the Hoffman constant for the **error bound**

$$\text{dist}_S(x) \;\le\; L\|((a_i^T x - b_i)^+)\| \quad \forall x,$$

also expressible via **distance to ill-posedness** for $(a_i^T x \le b_i)$.

Cf. (Renegar '95): speed of interior point methods.

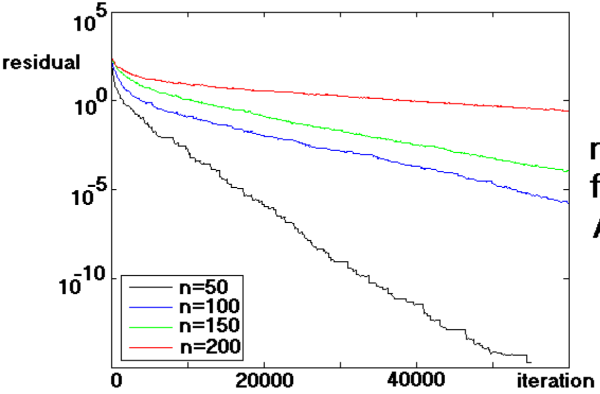
5. NUMERICAL EXAMPLES



random coordinate descent for

$$C^T C\, x = C^T b$$

with C 500-by-n



random projections for $Ax \leqslant b$ with A 500-by-n

6. DEMMEL'S PARADIGM (1987)

$$\begin{tikzcd}
 & \text{close to ill-posed} \arrow[dr] & \\
\text{slow algorithms} \arrow[ur] & & \text{weak error bound} \arrow[ll]
\end{tikzcd}$$

Other examples

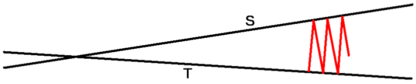
- Matrix inversion $A \mapsto A^{-1}$ depends on $\|A^{-1}\|$.
- Eigenvalue computation depends on the Wilkinson distance (to the defective matrices).
- Solving smooth $F(x) = 0$ locally depends on $\|\nabla F(x^*)^{-1}\|$.
- Alternating projections on subspaces depends on the "angle" between them (Aronszajn '50)...

7. ALTERNATING PROJECTIONS

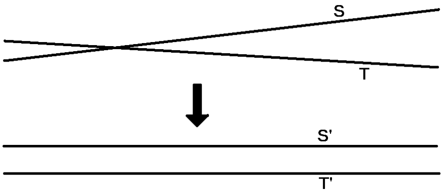
Find a common point of affine subspaces S, T. Equivalently, solve $d_S(x) + d_T(x) = 0$ (where d_S is distance to S.)

Three equivalent properties of instances:

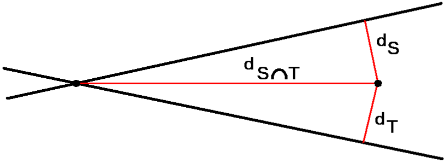
- Slow **linear rate** for alternating projection.



- Small perturbations cause **ill-posedness**.



- Weak **error bound**: $d_{S\cap T} \leq k(d_S + d_T)$ needs k large.



8. METRIC REGULARITY

Solve $y \in F(x)$ near $(\bar{x}, \bar{y}) \in \text{graph}\, F$.

$\nearrow$ $\searrow$

data vector unknown

modulus$(F)(\bar{x}, \bar{y})$ is the smallest k giving an **error bound**

$$\text{dist}_{F^{-1}(y)}(x) \leq k \cdot \text{dist}_{F(x)}(y) \quad \forall\ (x, y) \text{ near } (\bar{x}, \bar{y}).$$

radius$(F)(\bar{x}, \bar{y})$ is the smallest $\|$linear $G\|$ so

$$\text{modulus}(F + G)\Big(\bar{x}, \bar{y} + G(\bar{x})\Big) = \infty.$$

(A "local" **distance to ill-posedness**.)

Theorem (Dontchev-Lewis-Rockafellar '03)

$$\text{modulus} \times \text{radius} = 1.$$

(Two of Demmel's ingredients. . .)

9. EXAMPLE: AVERAGED PROJECTIONS

Find $x \in S = \cap_i S_i$ using the projections proj_{S_i}. (Even nonconvex projections may be easy: eg $\{$rank k matrices$\}$.) Equivalently, solve

$$0 \in F(x) \quad \text{where} \quad F(x) = \prod_{i=1}^{m} (S_i - x).$$

Averaged projection algorithm: $\quad x^{r+1} \leftarrow \frac{1}{m} \sum_i \text{proj}_{S_i} x^r$.

Theorem (Lewis-Luke-Malick '07) $\quad x^r \to S \quad$ at linear rate governed by $\text{modulus}(F)(\bar{x}, 0)$ (if finite), for x_0 near $\bar{x} \in S$.

Hence an algorithmic proof of:

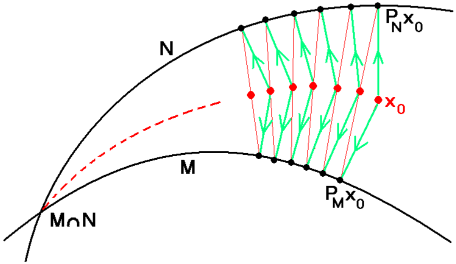
Extremal Principle (Mordukhovich '84)

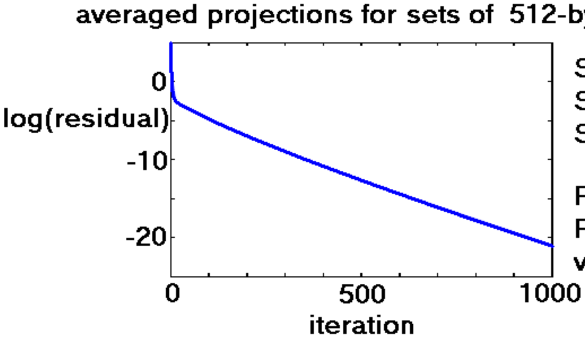
$$\bigcap_i (S_i - z_i) \neq \emptyset \quad \text{for small } z_i,$$

if nontrivial sums of "normals" to S_i at $\bar{x}$ cannot vanish. (Nonconvex local version of separating hyperplane theorem.)

10. METHOD OF AVERAGED PROJECTIONS

Following Auslender '69 in the convex case...




averaged projections for sets of 512-by-128 matrices

S_1= subspace
S_2= box
S_3= Stiefel manifold

Projections on S_1 and S_2 easy.
Projection on (nonconvex) S_3
via SVD: singular values = 1.

11. PREVALENCE OF REGULARITY

In $\mathbf{R}^n$, finite unions of sets defined by finitely many polynomial inequalities are called **semi-algebraic**.

Rich class, since preserved under projection (Tarski-Seidenberg).

If $\quad \text{modulus}(F)(x,y) = \infty \quad$ for some x, the value y is **critical**.

Theorem (Ioffe '07, Bolte-Daniilidis-Lewis-Shiota '05 $(m = 1)$)
Critical values of semi-algebraic $F \colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ are **rare** (dimension $< m$).

(An analogue of Sard's Theorem).

So semi-algebraic constraint systems

$$y \in F(x)$$

are well-behaved for "most" data vectors y.

12. PSEUDOSPECTRA

For square A, the **resolvent norm**

$$n_A(z) = \|(A - zI)^{-1}\| \qquad (z \in \mathbf{C})$$

is semi-algebraic, so has finitely many critical values ("Sard").

For $\epsilon > 0$, the ϵ-**pseudospectrum** mapping is

$$A \ \mapsto \ \Lambda_\epsilon(A) \ = \ \left\{ z \in \mathbf{C} : n_A(z) \geq \frac{1}{\epsilon} \right\}$$

(an "enlargement" of the spectrum).

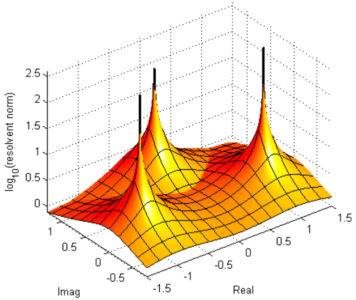
If ϵ is small enough,

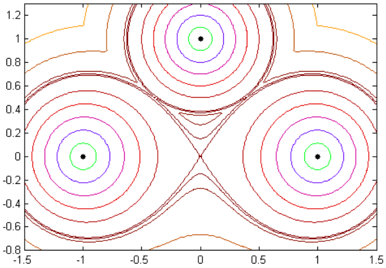
$\dfrac{1}{n_A(z)} = \epsilon \ \Rightarrow \ z$ noncritical $\Rightarrow \ \Lambda_\epsilon$ Lipschitz around A near z.

Theorem (Lewis-Pang '07)
Λ_ϵ is locally Lipschitz for all small $\epsilon > 0$. Whereas. . .

The **spectrum** is nonlipschitz around any defective matrix.

13. A PSEUDOSPECTRAL EXAMPLE





Smooth and nonsmooth critical points (saddlepoints and a local minimizer) of the resolvent norm for

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & i \end{bmatrix}.$$

Why are we interested in pseudospectra?

14. PSEUDOSPECTRA (Trefethen-Embree '05)

Asymptotics of $\dot{x} = Ax$ depend on spectrum $\Lambda(A)$.

But **transient** stability depends on the **pseudospectra**

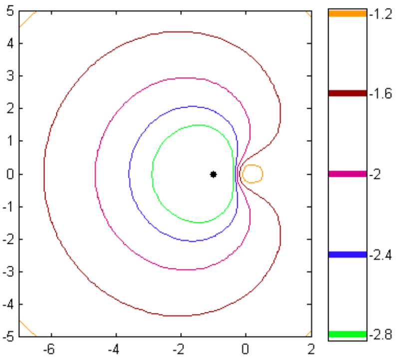
$$\Lambda_\epsilon(A) \;=\; \bigcup_{\|X-A\|\leq\epsilon} \Lambda(X).$$

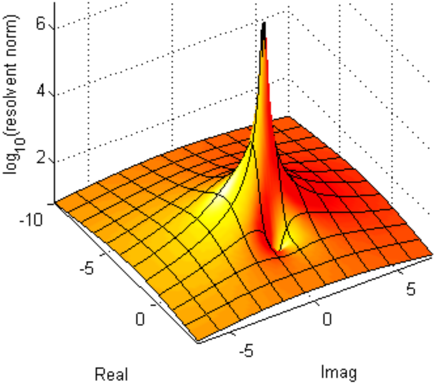
Kreiss Matrix Theorem (1962) Equivalent properties:

- $\dot{x} = Ax$ has big transient peaks
- $\Lambda_\epsilon(A)$ grows quickly into right halfplane as $0 < \epsilon \uparrow$

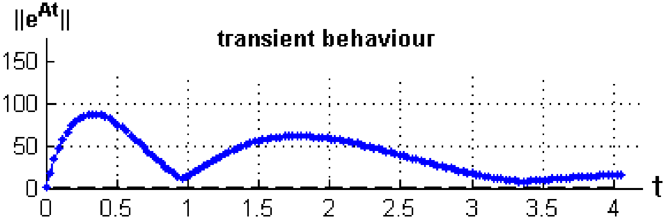
Demmel's example: $A = - \begin{bmatrix} 1 & 5 & 5^2 & 5^3 & 5^4 \\ 0 & 1 & 5 & 5^2 & 5^3 \\ 0 & 0 & 1 & 5 & 5^2 \\ 0 & 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \ldots$

15. EIGTOOL PLOTS (T. Wright)





$\Lambda_{.01}(A)$ intersects right halfplane, so
an unstable X satisfies $\|X - A\| \le .01$.



16. SUMMARY

- Demmel's paradigm:

  error bound ↔ ill-posedness ↔ speed.

  Eg: randomized algorithms for linear constraints.

- Local version:

  modulus of regularity ↔ radius ↔ speed.

  Eg: averaged projections.

- Metric regularity is typical for semi-algebraic constraints.
- Small pseudospectra are computable and locally Lipschitz, so we can optimize, and enhance transient system stability.

`people.orie.cornell.edu/∼aslewis`