LSLQ: An iterative method for linear least-squares with an error minimization property

Michael Saunders
MS&E and ICME
Stanford University, USA

Joint work with Ron Estrin and Dominique Orban

ICIAM 2019, Valencia, Spain, July 15–19, 2019

Abstract

LSLQ uses the Golub-Kahan process to compute iterates equivalent to SYMMLQ applied to the normal equation. The norm of the approximate solution increases, and the error norm decreases. Bounds on the error norm lead to error bounds for the iterates of LSQR. For an inversion problem arising in geophysics, LSLQ allows approximate computation of the gradient of a penalty function that is to be minimized.

Outline

SPD $Ax = b$

SYMMLQ helps bound error $\left\| x - x_k^C \right\|$ **for CG**

Error bounds for CG via SYMMLQ

Theorem

$$\left\| x_k^L \right\| \leqslant \left\| x_k^C \right\|$$
$$\left\| x_* - x_k^C \right\| \leqslant \left\| x_* - x_k^L \right\| = |\tilde{\zeta}_k|$$

Improved bound:

Theorem

$$\left\| x_* - x_k^C \right\| \leqslant \sqrt{\tilde{\zeta}_k^2 - \bar{\zeta}_k^2}$$

(Estrin, Orban, and Saunders, 2019a)

Motivation:
Inexact Derivatives in Optimization

Inexact derivatives in optimization

Consider the unconstrained problem

$$\underset{x}{\text{minimize}} \;\; f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^2$, say.

Trust-region methods require (approximate) solution of subproblems

$$\underset{s}{\text{minimize}} \;\; m_k(x_k + s) \quad \text{subject to} \quad \|s\| \leqslant \Delta_k,$$

where m_k models f around x_k. Typically $B_k \approx \nabla^2 f(x_k)$ and

$$m_k(x_k + s) = f(x_k) + \nabla f(x_k)^T s + \tfrac{1}{2} s^T B_k s \approx f(x_k + s)$$

What do trust-region methods require of the model?

$$m_k(x_k + s) = f(x_k) + \nabla f(x_k)^T s + \tfrac{1}{2} s^T B_k s$$

The classic requirement is that $m_k(x_k) = f(x_k)$ and $\nabla m_k(x_k) = \nabla f(x_k)$,

but the TR convergence theory holds if we use $g_k \approx \nabla f(x_k)$, where

$$\|g_k - \nabla f(x_k)\| \leqslant c \, \|g_k\|$$

for some c fixed by the implementation.

Example 1: PDE-constrained optimization

Van Leeuwen and Herrmann (2013) describe a penalty method for seismic inversion:

$$\underset{m,u}{\text{minimize}} \ \tfrac{1}{2}\|r(u)\|^2 \quad \text{subject to} \ c(m,u) = 0,$$

where

- m is the control variable
- u is the state (wavefields)
- $c(m,u) = 0$ is a discretized PDE

They use a quadratic penalty approach

$$\underset{m,\,u}{\text{minimize}} \ \phi_\lambda(m,u) \qquad \phi_\lambda(m,u) := \tfrac{1}{2}\|r(u)\|^2 + \tfrac{1}{2}\lambda^2\|c(m,u)\|^2$$

and implicitly eliminate $u = u(m)$ from $\nabla_u \phi_\lambda(m, u(m)) = 0$.

Computing the states

In the full-wave inversion problems considered by Van Leeuwen and Herrmann (2013), $u(m)$ solves a linear least-squares problem.

For an inexact $\tilde{u} \approx u(m)$, it is possible to bound

$$\|\nabla_m \phi_\lambda(m, u(m)) - \nabla_m \phi_\lambda(m, \tilde{u})\| \leqslant \text{const}\, \|u(m) - \tilde{u}\|.$$

Conclusion: if we knew a least-squares method that allows us to control the error in the solution, we could iterate until $\|u(m) - \tilde{u}\| \leqslant c\, \|\nabla_m \phi_\lambda(m, \tilde{u})\|$.

LSLQ for Least-Squares Problems

$$\min \|Ax - b\|$$

Least squares

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2}\|Ax-b\|^2 \qquad m\times n \text{ (any shape)}$$

$$A^TAx = A^Tb \text{ is spd (or semi-definite), consistent}$$

Krylov-type methods seek x_k in the k-th Krylov subspace

$$\mathcal{K}_k := \text{Span}\{A^Tb,\ (A^TA)A^Tb,\ \dots,\ (A^TA)^kA^Tb\}$$

On spd systems, CG produces monotonic $\|x_k - x^*\|_2$ (Hestenes and Stiefel, 1952)

Other methods do too: MINRES and SYMMLQ (Paige and Saunders, 1975)

Names

LSQR	is equivalent to	CG	applied to $A^TAx = A^Tb$
LSMR	is equivalent to	MINRES	" " "
LSLQ	is equivalent to	SYMMLQ	" " "

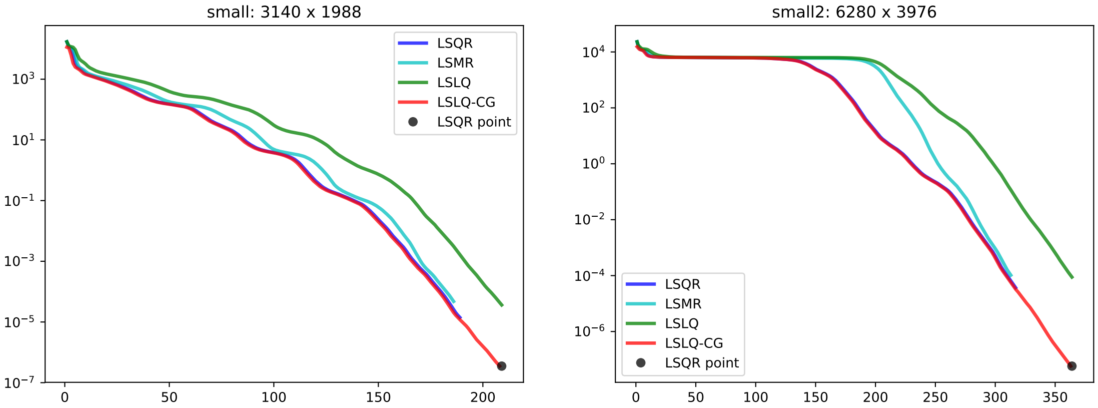
We describe LSLQ

Theorem

LSQR, LSMR, LSLQ converge to the min-length least-squares solution.

How does the error behave in least squares?

Problems from Hegland (1993)



LSQR seems best, but we need LSLQ to provide error bounds

The Golub and Kahan (1965) process

1: $\beta_1 u_1 = b$
2: $\alpha_1 v_1 = A^T u_1$
3: **for** $k = 1, 2, \ldots$ **do**
4: $\quad \beta_{k+1} u_{k+1} = A v_k - \alpha_k u_k$
5: $\quad \alpha_{k+1} v_{k+1} = A^T u_{k+1} - \beta_{k+1} v_k$

$$U_k := \begin{bmatrix} u_1 & \cdots & u_k \end{bmatrix}, \qquad V_k := \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix}, \qquad B_k := \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \\ & & & \beta_{k+1} \end{bmatrix}$$

Theoretically $\quad U_k^T U_k = I_k \quad$ and $\quad V_k^T V_k = I_k$

Golub-Kahan: Main identities

At iteration k,

$$AV_k = U_{k+1} B_k$$
$$A^T U_{k+1} = V_k B_k^T + \alpha_{k+1} v_{k+1} e_{k+1}^T$$

Seek $x_k = V_k y_k$

$$A^T A x_k = V_{k+1} H_k y_k \qquad H_k := \begin{bmatrix} B_k^T B_k \\ \alpha_{k+1}\beta_{k+1} e_k^T \end{bmatrix} = \begin{bmatrix} T_k \\ \alpha_{k+1}\beta_{k+1} e_k^T \end{bmatrix}$$

B_k bidiagonal $\qquad$ T_k symmetric tridiagonal

LSLQ: Main subproblems

$$x_k = V_k y_k$$

$$A^T A x_k - A^T b = V_{k+1}(H_k y_k - \alpha_1 \beta_1 e_1)$$

Small if $\qquad H_k y_k \approx \alpha_1 \beta_1 e_1$

LSQR $\quad x_k^C := V_k y_k^C \qquad T_k y_k^C = \alpha_1 \beta_1 e_1$

LSLQ $\quad x_k^L := V_k y_k^L \qquad$ minimize $\frac{1}{2}\|y_k^L\|^2$ subject to $H_{k-1}^T y_k^L = \alpha_1 \beta_1 e_1$

(H_{k-1}^T is T_k without its last row)

Optimality properties

x_k^C solves

$$\underset{x \in \mathcal{K}_k}{\text{minimize}} \; \|x_* - x\|_{A^T A} \tag{1}$$

x_k^L solves

$$\underset{x \in A^T A \mathcal{K}_{k-1}}{\text{minimize}} \; \|x_* - x\| \qquad \text{and} \qquad \underset{x \in \mathcal{K}_k}{\text{minimize}} \; \|x\| \; \text{subject to} \; r \perp \mathcal{K}_{k-1} \tag{2}$$

Theorem

Whether A has full column rank or not, $\|x_* - x_k^C\| \leqslant \|x_* - x_k^L\|$

- x_k^C is feasible for (2) $\Rightarrow \|x_k^L\| \leqslant \|x_k^C\|$
- $x_*^T x_k^C \geqslant 0$

Computing y_k^C and y_k^L

Each iteration of LSLQ updates a QR factorization

$$P_k^T B_k = \begin{bmatrix} R_k \\ 0 \end{bmatrix} \qquad T_k = B_k^T B_k = R_k^T R_k$$

and an LQ factorization

$$R_k = \bar{M}_k Q_k \qquad \bar{M}_k \text{ lower triangular}$$

We solve

$$R_k^T t_k = \alpha_1 \beta_1 e_1 \qquad \bar{M} \bar{z}_k = t_k, \qquad \bar{z}_k := \begin{bmatrix} z_{k-1} \\ \bar{\zeta}_k \end{bmatrix}$$

Then

$$y_k^L = Q_k^T \begin{bmatrix} z_{k-1} \\ 0 \end{bmatrix} \qquad y_k^C = Q_k^T \bar{z}_k$$

Computing x_k^C and x_k^L

x_k^L is updated along orthogonal directions:

$$\bar{W}_k := V_k Q_k^T = \begin{bmatrix} w_1 & \dots & w_{k-1} & \bar{w}_k \end{bmatrix} \qquad x_k^L = x_{k-1}^L + \zeta_{k-1} w_{k-1}$$

Hence

$$\|x_k^L\|^2 = \|x_{k-1}^L\|^2 + \zeta_{k-1}^2 \qquad \|x_* - x_k^L\|^2 = \|x_*\|^2 - \|x_k^L\|^2$$

We can estimate this error if we can estimate $\|x_*\|^2$ ...

LSLQ can transition cheaply to the LSQR point:

$$x_k^C = x_k^L + \bar{\zeta}_k \bar{w}_k \qquad \bar{w}_k \perp w_1, \dots, w_{k-1}$$

Upper bound on the LSLQ Error: Preliminaries

Write

$$\|x_*\|^2 = b^T A (A^T A)^{-2} A^T b = b^T A \; f(A^T A) \; A^T b, \qquad f(\xi) := \xi^{-2}, \quad \xi \in (0, \sigma_1^2]$$

where

$$f(A^T A) := \sum_{i=1}^{n} f(\sigma_i^2) p_i p_i^T \qquad (p_i = \text{eigenvector}).$$

Because $A^T b = \alpha_1 \beta_1 v_1$,

$$\|x_*\|^2 = (\alpha_1 \beta_1)^2 \sum_{i=1}^{n} f(\sigma_i^2) \mu_i^2, \qquad \mu_i := p_i^T v_1.$$

A tour de force

We found

$$\|x_*\|^2 = (\alpha_1\beta_1)^2 \sum_{i=1}^{n} f(\sigma_i^2)\mu_i^2$$

Golub and Meurant (1997) view the sum as the Riemann-Stieltjes integral

$$\sum_{i=1}^{r} f(\sigma_i^2)\mu_i^2 = \int_{\sigma_r}^{\sigma_1} f(\sigma^2)\,\mathrm{d}\mu(\sigma)$$

where the piecewise constant Stieltjes measure μ is defined as

$$\mu(\sigma) := \begin{cases} 0 & \text{if } \sigma < \sigma_n \\ \sum_{j=i}^{n} \mu_j^2 & \text{if } \sigma_i \leqslant \sigma < \sigma_{i+1} \\ \sum_{j=1}^{n} \mu_j^2 & \text{if } \sigma \geqslant \sigma_1 \end{cases}$$

Approximations to the integral via Gauss-related quadrature yield approximations to $\|x_\star\|^2$

Gauss-Radau quadrature yields an upper bound

Theorem

Suppose $f : \mathbb{R} \to \mathbb{R}$ is such that $f^{(2j+1)}(\xi) < 0$ for all $\xi \in (\sigma_n^2, \sigma_1^2)$ and all $j \geqslant 0$. Fix $\sigma_* \in (0, \sigma_n)$. Let B_k be the bidiagonal generated after k steps of GK and $\omega_k > 0$ be chosen so that the smallest singular value of

$$\widetilde{R}_k := \begin{bmatrix} R_{k-1} & \delta_k e_{k-1} \\ & \omega_k \end{bmatrix}$$

is precisely σ_*. Then $b^T A f(A^T A) A^T b \leqslant (\alpha_1 \beta_1)^2 e_1^T \; f(\widetilde{R}_k^T \widetilde{R}_k) \; e_1$.

Almost nothing changes if A is rank-deficient
because $A^T A x = A^T b$ is consistent and all iterations occur in Range(A^T)

Simply replace σ_n by σ_r

Upper bound on the LSLQ error

- ω_k can be determined from a few scalar operations
- $e_1^T(\widetilde{R}_k^T\widetilde{R}_k)^{-2}e_1$ is computed using a simple update of the LQ factorization of R_k:

$$\widetilde{R}_k = \widetilde{M}_k Q_k$$

- This yields

$$\widetilde{R}_k^T\widetilde{t}_k = \alpha_1\beta_1 e_1, \qquad \widetilde{M}_k\widetilde{z}_k = \widetilde{t}_k, \qquad \widetilde{t}_k = \begin{bmatrix} t_{k-1} \\ \tilde{\tau}_k \end{bmatrix}, \qquad \widetilde{z}_k = \begin{bmatrix} z_{k-1} \\ \tilde{\zeta}_k \end{bmatrix}$$

and finally

Theorem

$$\|x_* - x_k^L\| \leqslant |\tilde{\zeta}_k|$$

Regularization

$$
\begin{bmatrix} \alpha_1 & & \\ \beta_2 & \alpha_2 & \\ & \beta_3 & \alpha_3 \\ & & \beta_4 \\ \lambda & & \\ & \lambda & \\ & & \lambda \end{bmatrix}
\rightarrow
\begin{bmatrix} \alpha_1 & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & \\ & \beta_3 & \alpha_3 \\ & & \beta_4 \\ & \hat{\lambda}_2 & \\ & \lambda & \\ & & \lambda \end{bmatrix}
\rightarrow
\begin{bmatrix} \alpha_1 & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & \\ & \beta_3 & \alpha_3 \\ & & \beta_4 \\ & & \\ & \lambda_2 & \\ & & \lambda \end{bmatrix}
$$

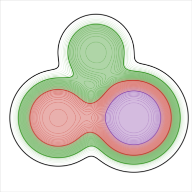
$$
\rightarrow
\begin{bmatrix} \alpha_1 & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & \\ & \hat{\beta}_3 & \hat{\alpha}_3 \\ & & \beta_4 \\ & & \\ & & \hat{\lambda}_3 \\ & & \lambda \end{bmatrix}
\rightarrow
\begin{bmatrix} \alpha_1 & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & \\ & \hat{\beta}_3 & \hat{\alpha}_3 \\ & & \beta_4 \\ & & \\ & & \\ & & \lambda_3 \end{bmatrix}
\rightarrow
\begin{bmatrix} \alpha_1 & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & \\ & \hat{\beta}_3 & \hat{\alpha}_3 \\ & & \hat{\beta}_4 \\ & & \\ & & \\ & & \end{bmatrix}
$$

Notes on the numerical experiments

- Each matrix A is scaled so its nonzero columns have unit norm
- Data and solutions are available in Rutherford-Boeing format from

  github.com/optimizers/animal

- Everything else is implemented in Julia:
  - github.com/JuliaSparse/HarwellRutherfordBoeing.jl: IO
  - github.com/JuliaSmoothOptimizers/LinearOperators.jl: abstract linear operators
  - github.com/JuliaSmoothOptimizers/Krylov.jl: iterative methods



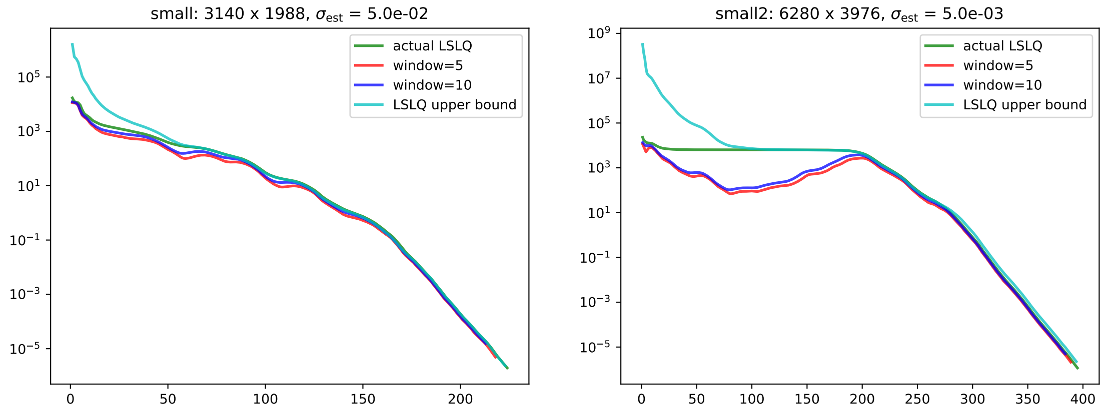
Notes on the numerical experiments

- The upper bounds require an estimate of σ_n or σ_r
- We run PROPACK.jl[1] to approximate the smallest nonzero singular value
- We use $\sigma_{\text{est}} := (1 - 10^{-10})\,\sigma_n \quad (\text{or } \sigma_r)$
- In the presence of regularization, $\sigma_n = \lambda$!

[1] github.com/JuliaSmoothOptimizers/PROPACK.jl

Numerical illustration (without regularization)

Stopping condition: $|\tilde{\zeta}_k| \leqslant 10^{-10} \|x_k^L\|$



Upper bound on the LSQR error

We can transition $x_k^L \rightsquigarrow x_k^C$. We can also get an improved error bound.

Theorem

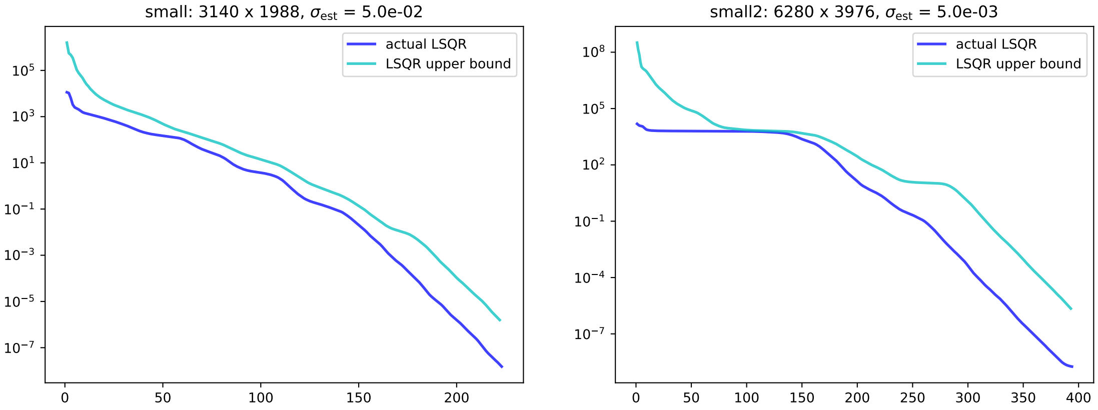
With the same value of ω_k,

$$\|x_* - x_k^C\|^2 \leqslant \tilde{\zeta}_k^2 - \bar{\zeta}_k^2$$

SYMMLQ/CG: (Estrin et al., 2019a)
LSLQ: (Estrin, Orban, and Saunders, 2019b)
LNLQ: (Estrin, Orban, and Saunders, submitted 2019)

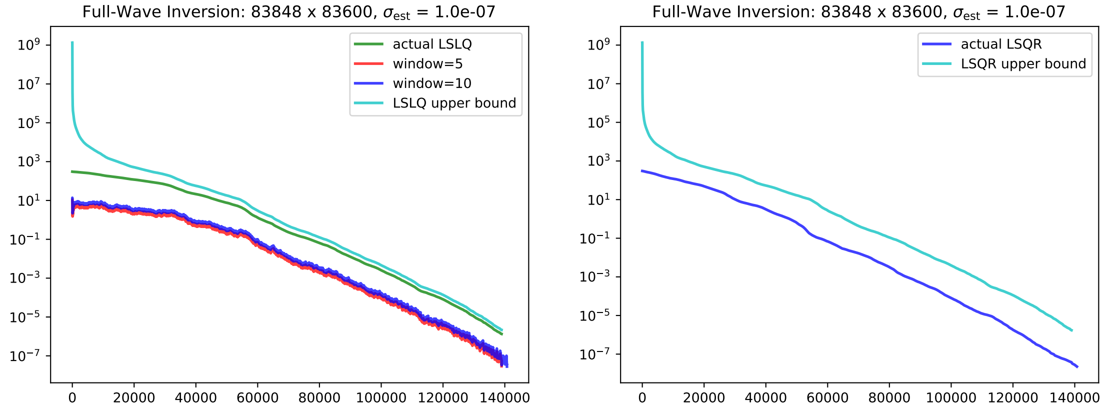
Numerical illustration (without regularization)

Stopping condition: $\sqrt{\tilde{\zeta}_k^2 - \bar{\zeta}_k^2} \leqslant 10^{-10} \, \|x_k^C\|$



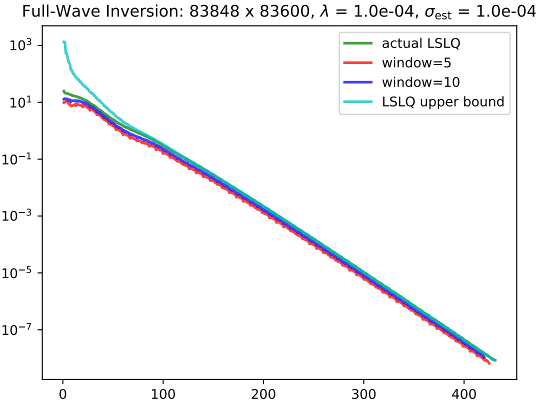
Seismic inversion problem (without regularization)

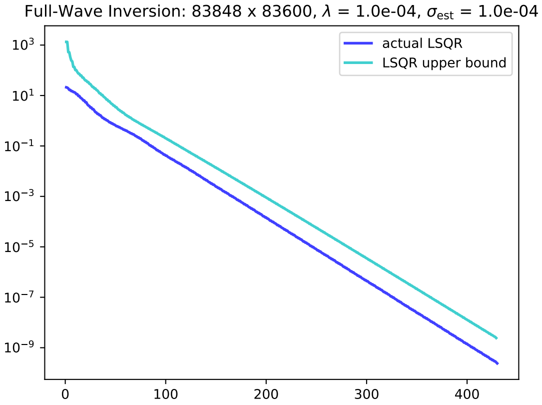
Stopping condition: $\sqrt{\tilde{\zeta}_k^2 - \bar{\zeta}_k^2} \leqslant 10^{-10} \|x_k^C\|$



Seismic inversion problem (with regularization)

Stopping condition: $\sqrt{\tilde{\zeta}_k^2 - \bar{\zeta}_k^2} \leqslant 10^{-10} \|x_k^C\|$





Conclusions so far

- Monitor the error upper bound and transition to the LSQR point
- Can regularize cheaply, and that yields an obvious σ_{est} in the rank-deficient case
- A low-memory approach can be used to tighten the upper bound at moderate cost

LNLQ for Least-Norm Problems

$$\min \|x\|^2 \textbf{ subject to } Ax = b$$

Normal equations of the second kind

$$\underset{x}{\text{minimize}} \ \tfrac{1}{2}\|x\|^2 \quad \text{subject to} \ Ax = b$$

Optimality conditions:

$$\begin{bmatrix} I & A^T \\ A & \end{bmatrix} \begin{bmatrix} x \\ -y \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}$$

$$\equiv \ AA^T y = b, \qquad x = A^T y \qquad \text{(NE2)}$$

If we assume $Ax = b$ is consistent, we can do much the same as LSLQ.

CG applied to (NE2) is sometimes known as CRAIG's method or CGNE.

LNLQ

Main idea: Apply SYMMLQ to (NE2) with possible transition to the CRAIG point.

Main iterates: $y_k^L \rightsquigarrow y_k^C$, with cheap updates of $x_k^L := A^T y_k^L$ and $x_k^C := A^T y_k^C$.

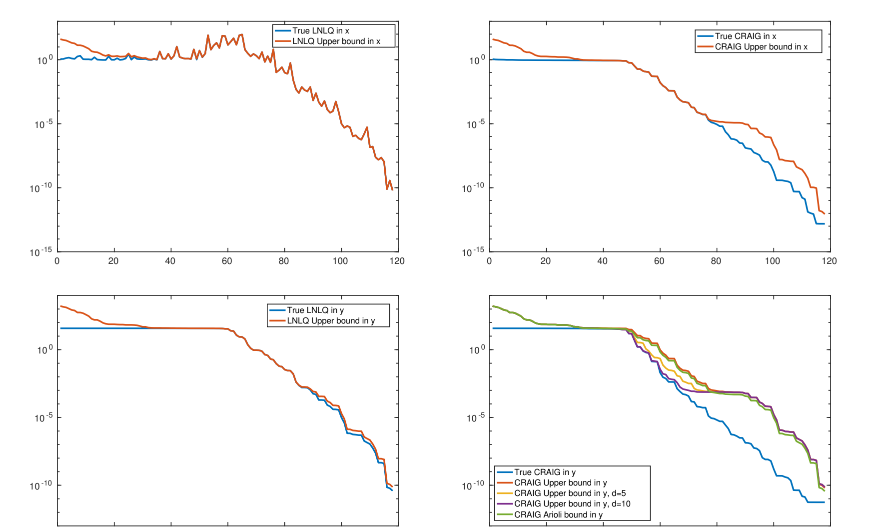
Benefits:

- Far simpler implementation than LSLQ
- Access to both x and y
- x_k^L and x_k^C solve minimum-norm and minimum-error problems
- y_k^L and x_k^C are updated along orthogonal directions
- LNLQ also computes y_k^C as an orthogonal update of y_k^L (unlike CRAIG)
- Access to $\|x_\star - x_k^L\|$, $\|x_\star - x_k^C\|$, $\|y_\star - y_k^L\|$ and $\|y_\star - y_k^C\|$

Arioli (2013) used Gauss-Radau to bound $\|x_\star - x_k^C\|$
and gave the crude bound $\|y_\star - y_k^C\| \leqslant \|x_\star - x_k^C\|/\sigma_n$

Meszaros/scagr7-2c from UFL

2447 rows, 3479 columns, full rank, $\sigma_* = (1 - 10^{-10})\sigma_{\min}$



Fletcher's merit function: a PDE-constrained problem

$$\begin{aligned} \underset{u,z}{\text{minimize}} \quad & \tfrac{1}{2}\int_\Omega \|u - u_d\|^2 \mathrm{d}x + \tfrac{1}{2}\alpha \int_\Omega z^2 \mathrm{d}x \\ \text{subject to} \quad & -\nabla\cdot(z\nabla u) = f \qquad \text{in } \Omega \\ & u = 0 \qquad \text{on } \partial\Omega \end{aligned}$$

$\Omega = [-1, 1]^2, \quad \alpha = 10^{-4}$ Discretized problem has $n = 2050$, $m = 961$

η	Iterations	# Hv	# Jprod	# Adj Jprod
10^{-2}	22	878	3448	3672
10^{-4}	21	896	4251	4459
10^{-6}	20	744	4651	4928
10^{-8}	20	746	5611	5887
10^{-10}	20	746	6595	6871

η = accuracy of solving KKT system for search directions

Estrin et al., SIMAX (2019)

Summary

- LSLQ and LNLQ fill gaps in the family of Krylov methods

SYMMLQ	provides error bounds for	CG
LSLQ	" " "	LSQR
LNLQ	" " "	Craig's method

- Application: optimization with inexact gradient

M. Arioli. Generalized Golub-Kahan bidiagonalization and stopping criteria. *SIAM J. Mat. Anal. & Appl.*, 34(2): 571–592, 2013. DOI: 10.1137/120866543.

R. Estrin, D. Orban, and M. A. Saunders. Euclidean-norm error bounds for CG via SYMMLQ. *SIAM J. Mat. Anal. & Appl.*, 40(1):235–253, 2019a. DOI: 10.1137/16M1094816.

R. Estrin, D. Orban, and M. A. Saunders. LSLQ: An iterative method for linear least-squares with an error minimization property. *SIAM J. Mat. Anal. & Appl.*, 40(1):254–275, 2019b. DOI: 10.1137/17M1113552.

R. Estrin, D. Orban, and M. A. Saunders. LNLQ: An iterative method for least-norm problems with an error minimization property. *SIAM J. Mat. Anal. & Appl.*, submitted 2019.

G. H. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *SIAM J. Numer. Anal.*, 2(2):205–224, 1965. DOI: 10.1137/0702016.

G. H. Golub and G. Meurant. Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods. *BIT Num. Math.*, 37(3):687–705, 1997. DOI: 10.1007/BF02510247.

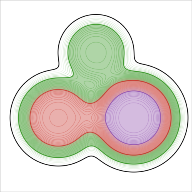
M. Hegland. Description and use of animal breeding data for large least squares problems. Technical Report TR/PA/93/50, CERFACS, Toulouse, France, 1993.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. N.B.S.*, 49(6): 409–436, 1952.

C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975. DOI: 10.1137/0712047.

T. Van Leeuwen and F. J. Herrmann. A penalty method for PDE-constrained optimization. Technical Report TR-EOAS-2013-6, University of British Columbia, 2013.

Julia version of LSLQ and other Krylov solvers:
github.com/JuliaSmoothOptimizers/Krylov.jl
Dominique Orban: dominique.orban@gerad.ca



Matlab version of LSLQ and LNLQ:
github.com/restrin/LinearSystemSolvers
Ron Estrin: ronestrin756@gmail.com