Numerical Linear System Solving With Parametric Entries By Error Correction

Brice BOYER
bboyer@ncsu.edu
Symbolic computation group, North Carolina State University

Erich L. KALTOFEN
kaltofen@ncsu.edu
Symbolic computation group, North Carolina State University

Abstract
We consider the problem of solving a full rank consistent linear system \( A(u)x = b(u) \) where \( A \in \mathbb{K}[u]^{m \times n} \) and \( b \in \mathbb{K}[u]^{m} \). Our algorithm computes the unique solution \( x = f(u)/g(u) \) [a vector of rational functions] by evaluating the parameter \( u \) at distinct points. Those points \( \Delta_u \) where the matrix \( A \) evaluates to \( A(\Delta_u) \) of lower rank, or in the numeric setting to an ill-conditioned matrix, are not identified but accounted for by error-correcting code techniques. We also correct true errors where the evaluation at some \( u \) equals \( \hat{u} \) results in an erroneous, possibly full rank consistent and well-conditioned scalar linear system.

We have implemented our algorithm with floating point arithmetic. For the determination of the exact numerator and denominator degrees and number of errors we use SVD-based numeric rank computations. The arising linear systems for the error-corrected parametric solution are shown to be well-conditioned even when the scalar inputs have noise.

Exact Vector-Of-Functions Solving (no errors)

Let bounds on the degrees be \( d_\ell = \max_{1 \leq j \leq 1+2m} \deg(f(j)) \) and \( d_\varphi = \deg(g) \). Let \( \hat{E}_\ell \) be a bound on the number of \( k \)-columns of indices where the rank in an evaluation drops, i.e., \( \ell \in A_1, \ldots, A_n \), iff \( \text{rank}(A(\hat{E}_\ell)) < \text{rank}(A(u)) \). Bounds given or estimated on input.

We select \( L = d_k + d_\ell + 1 \) pair-wise distinct elements \( \hat{E}_k \in K \), where \( 0 \leq \ell \leq L-1 \) and compute \( \hat{E} \) by solving the homogeneous linear system \( (1) \) in the unknown coefficients of \( \Phi(\ell)(u) \) and \( \Psi(\ell)(u) \) that is \( n(d_k + d_\ell + 1) \) unknown coefficients and \( m \) linear equations:

\[
A(\hat{E}_k) \begin{bmatrix} \hat{\Phi}(\ell)(\hat{E}_k) \\ \hat{\Psi}(\ell)(\hat{E}_k) \end{bmatrix} = \begin{bmatrix} g(\hat{E}_k) \\ \deg(g) \end{bmatrix} \\\end{bmatrix} \delta \leq d_k \| \hat{d} \| \leq d_k, \| \hat{E} \| \leq L - 1.
\]

\[ \implies \text{we need to indentify the \"unlucky\" evaluation } \hat{E}_k. \]

Numerical Algorithm
Input: Two \"black boxes\" \( A \in \mathbb{C}[u]^{m \times n} \) and \( b \in \mathbb{C}[u]^{m} \): Bounds \( \hat{E}_k, \hat{E}_\ell, d_k, d_\ell, d_\varphi \).
Output: Polynomials \( f \in \mathbb{C}[u]^{m \times n} \) and \( g \in \mathbb{C}[u] \) s.t. \( f/g \) is a solution to the system \( A(u)x = b(u) \).

1. Initialization.
   a) Create a set \( \Delta_u = d_k + d_\ell + 1 + (\hat{E}_k + 2\ell + 1) \) random evaluation points.
   b) Set up the vectors \( x \) and \( y \) corresponding to the unknowns in \( f \) and \( g \), resp. length \( n(d_k + d_\ell + 1) + d_k + d_\ell + 1 \).
   c) Set up a linear system \( W \) from evaluations of \( A \) and \( b \) at \( \Delta_u \) such that \( W(x, y) = 0 \).
2. SVD step.
   a) Compute a SVD of \( W \) and find its numeric rank \( \rho \).
   b) Construct a reduced linear system \( W \) by removing the \( \rho - 1 \) unknowns of highest degree to \( x \) and \( y \).
3. Error removal.
   a) From the evaluations of \( \Phi \) and \( \Psi \) at \( \Delta_u \), construct \( \Phi(\ell) \), \( \Psi(\ell) \), the error locating polynomial.
   b) From a least squares fit, compute the approximate division \( f/\Phi \) and \( g/\Psi \).

Notes: Instead of \( C \), we can also use \( R \) as the base field, and we itself have made experiments on \( R \).

Numerical Experiments (on \( C \))

<table>
<thead>
<tr>
<th>( \ell )</th>
<th>( n )</th>
<th>Rel. noise</th>
<th>( \deg(A) )</th>
<th>( \deg(b) )</th>
<th>( \deg(f) )</th>
<th>( \deg(g) )</th>
<th>( \hat{E}(1) )</th>
<th>( \hat{E}(2) )</th>
<th>Time (s)</th>
<th>Rel. Error</th>
<th>( \hat{\kappa}_\text{solv} )</th>
<th>( \hat{\kappa}_\text{solve} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>5, 15</td>
<td>2 (1), 2 (1)</td>
<td>0 (0), 1 (1)</td>
<td>&lt; 1</td>
<td>4.6e-7</td>
<td>3.8e-2</td>
<td>3.4e-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>5, 15</td>
<td>2 (1), 2 (1)</td>
<td>1 (1), 1 (1)</td>
<td>&lt; 1</td>
<td>6.8e-9</td>
<td>6.4e-10</td>
<td>4.5e-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>5, 15</td>
<td>2 (1), 2 (1)</td>
<td>1 (1), 1 (1)</td>
<td>&lt; 1</td>
<td>5.9e-6</td>
<td>7.3e-10</td>
<td>4.5e-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>5, 15</td>
<td>3 (3), 3 (3)</td>
<td>3 (3), 5 (5)</td>
<td>&lt; 5</td>
<td>1.7e-6</td>
<td>1.4e-10</td>
<td>1.2e-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5, 15</td>
<td>3 (3), 3 (3)</td>
<td>5 (5), 5 (5)</td>
<td>&lt; 5</td>
<td>1.3e-6</td>
<td>1.3e-10</td>
<td>1.2e-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>4, 19</td>
<td>3 (2), 3 (2)</td>
<td>6 (6), 6 (6)</td>
<td>&lt; 35</td>
<td>1.2e-6</td>
<td>4.8e-11</td>
<td>2.6e-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>4, 19</td>
<td>3 (2), 3 (2)</td>
<td>6 (6), 6 (6)</td>
<td>&lt; 35</td>
<td>1.3e-6</td>
<td>4.8e-11</td>
<td>2.6e-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>4, 19</td>
<td>3 (2), 3 (2)</td>
<td>15 (13), 15 (13)</td>
<td>&lt; 1000</td>
<td>6.9e-6</td>
<td>1.8e-11</td>
<td>7.3e-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>4, 19</td>
<td>3 (2), 3 (2)</td>
<td>15 (13), 15 (13)</td>
<td>&lt; 1200</td>
<td>6.5e-6</td>
<td>1.8e-11</td>
<td>7.3e-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: \( \hat{\kappa}_\text{solv} \) is the smallest singular value among all \( A(\hat{E}) \), \( \hat{E} \in C \) and \( \hat{\kappa}_\text{solve} \) is the smallest non-zero one for the non-singular system \( W \).

Bibliography


This research was supported in part by the National Science Foundation under Grant CCF-1115772.