Using ILU to estimate the diagonal of the inverse of a matrix

Andreas Stathopoulos, Lingfei Wu, Jesse Laeuchli
College of William and Mary

Vasilis Kalantzis, Stratis Gallopoulos
University of Patras, Greece

Acks: NSF, DOE SciDAC
The problem

Given a large, $N \times N$ matrix $A$ and a function $f$

find trace of $f(A)$: $\text{Tr}(f(A))$

Common functions:

- $f(A) = A^{-1}$
- $f(A) = \log(A)$
- $f(A) = R_i^T A^{-1} R_j$

Applications: UQ, Data Mining, Quantum Monte Carlo, Lattice QCD

Our focus: $f(A) = A^{-1}$
The methods

Currently all methods are based on Monte Carlo (Hutchinson 1989)

If $x$ is a vector of random $\mathbb{Z}_2$ variables

$$x_i = \begin{cases} 
1 & \text{with probability } 1/2 \\
-1 & \text{with probability } 1/2 
\end{cases}$$

then

$$E(x^T A^{-1} x) = \text{Tr}(A^{-1})$$

Monte Carlo Trace
for $i=1:n$

$$x = \text{randZ2}(N,1)$$
$$\text{sum} = \text{sum} + x^T A^{-1} x$$

$$\text{trace} = \text{sum}/n$$
The methods

Currently all methods are based on Monte Carlo (Hutchinson 1989)

If $x$ is a vector of random $\mathbb{Z}_2$ variables

$$x_i = \begin{cases} 
1 & \text{with probability } 1/2 \\
-1 & \text{with probability } 1/2 
\end{cases}$$

then

$$E(x^T A^{-1} x) = \text{Tr}(A^{-1})$$

Monte Carlo Trace

for $i=1:n$

- $x = \text{randZ2}(N,1)$
- $x = \text{randZ2}(N,1)$
- $\text{sum} = \text{sum} + x^T A^{-1} x$
- $\text{sum} = \text{sum} + x^T A^{-1} x$

trace = sum/n

2 problems

Large number of samples

How to compute $x^T A^{-1} x$
The methods

Currently all methods are based on Monte Carlo (Hutchinson 1989)

If $x$ is a vector of random $\mathbb{Z}_2$ variables

$$x_i = \begin{cases} 
1 \text{ with probability } 1/2 \\
-1 \text{ with probability } 1/2 
\end{cases}$$

then

$$E(x^T A^{-1} x) = \text{Tr}(A^{-1})$$

Monte Carlo Trace

for $i=1:n$

$x = \text{randZ2}(N,1)$

sum = sum + $x^T A^{-1} x$    Solve $Ay = x$ with CG

compute $y^T x$    Find quadrature $x^T A^{-1} x$

trace = sum / $n$

(Golub’69, Bai’95, Meurant’06,’09, Strakos’11, ...)
The methods

Currently all methods are based on Monte Carlo (Hutchinson 1989)

If $x$ is a vector of random $Z_2$ variables

$$x_i = \begin{cases} 
1 & \text{with probability 1/2} \\
-1 & \text{with probability 1/2}
\end{cases}$$

then

$$E(x^T A^{-1} x) = \text{Tr}(A^{-1})$$

Monte Carlo Trace
for i=1:n
x = randZ2(N,1)
sum = sum + $x^T A^{-1} x$

$O(100 - 1000s)$ statistically independent RHS

trace = sum/n

Recycling (de Sturler), Deflation (Morgan, AS‘07, ...) speed up Krylov methods
Selecting the vectors in $x^T A^{-1} x$

Random

$x \in Z_2^N$

best variance for real matrices (Hutchinson 1989)

$x = \text{randn}(N, 1)$

worse variance than $Z_2$

$x = e_i$

variance depends only on $\text{diag}(A^{-1})$

single large element?

$x = F^T e_i$

mixing of diagonal elements: (Toledo et al. 2010)

$F = \text{DFT}$ or $F = \text{Hadamard}$

Deterministic

$x = H^T e_i, \ i = 1, \ldots, 2^k$

Hadamard in natural order (Bekas et al. 2007)

$x^m_i = \begin{cases} 
1 & i \in C_m \\
0 & \text{else}
\end{cases}$

Probing. Assumes multicolored graph (Tang et al. 2011)

Random-deterministic

Hierarchical Probing for lattices (A.S, J.L. 2013)
Variance of the estimators

Rademacher vectors $x_i \in \mathbb{Z}_2^N$

$$\overline{\text{Tr}} = \frac{1}{s} \sum_{i=1}^{s} x_i^T A^{-1} x_i$$

$$\text{Var}(\overline{\text{Tr}}) = \frac{2}{s} \|\tilde{A}^{-1}\|_F^2 = \frac{2}{s} \sum_{i \neq j} (A_{ij}^{-1})^2$$

Diagonal $x = e_{j(i)}$

$$\overline{\text{Tr}} = \frac{N}{s} \sum_{i=1}^{s} A_{j(i),j(i)}^{-1}$$

$$\text{Var}(\overline{\text{Tr}}) = \frac{N^2}{s} \text{Var}(\text{diag}(A^{-1}))$$

[Diagram showing the magnitude and variance of $A^{-1}$]
Approximating $\text{diag}(A^{-1})$ from ILU

Consider an incomplete LU of $A$: $[L, U] = ILU(A)$

If $U^{-1}L^{-1}$ good approximation to $A^{-1}$ then compute trace from:

$$M = \text{diag}(U^{-1}L^{-1})$$

Computing $M$ needs only one pass over $L, U$ (Erisman, Tienny, ’75)

$$E = U^{-1}L^{-1} - A^{-1}$$

In some cases, $\text{Tr}(E)$ can be sufficiently close to zero

However, what if $|\text{Tr}(E)|$ is not small?
Observation: even if $\text{Tr}(E)$ large, $M$ may approximate the pattern of $\text{diag}(A^{-1})$ and/or $E$ may have smaller variance

Ex. small Laplacian and DW2048
Capture pattern better by fitting $p(M)$ to $\text{diag}(A^{-1})$

Find $p()$: $\min \|p(M) - \text{diag}(A^{-1})\|$ on a set of $m$ indices

- Induce smoothness on $M$ by sorting
- Use $m$ equispaced indices to capture the range of values
- Compute $A_{jj}^{-1}$ of these indices
- Fit $M_j$ to $A_{jj}^{-1}$ using MATLAB’s $\text{LinearModel.stepwise}$

When ILU is a good preconditioner, $\text{Tr}(p(M))$ can be accurate to $O(1E-3)$!
Examples of fitting

Comparing $\text{diag}(A^{-1})$, $\text{diag}((LU)^{-1})$ and $\text{diag}((LU)^{-1})_\text{fit}$

TOLS4000, DW2048, af23560, conf6

Comparing $\text{diag}(A^{-1})$, $\text{diag}((LU)^{-1})$ and $\text{diag}((LU)^{-1})_\text{fit}$
Improving on the $\text{Tr}(M)$ and $\text{Tr}(p(M))$

- **MC on** $E = M - \text{diag}(A^{-1})$
  - potentially smaller variance on the diagonal
- **MC on** $E2 = p(M) - \text{diag}(A^{-1})$
  - $m$ inversions for fitting, $s - m$ inversions for MC
  - further variance improvement
- **MC with importance sampling** based on $M$ or $p(M)$

Or is traditional Hutchinson better?

- **MC with** $Z_2^N$ on $A^{-1}$
- **MC with** $Z_2^N$ on $E$

Depends on approximation properties of ILU
Experiments

AF23560, conf6 \((k_c - 10^{-8})\)
Experiments

TOLS4000 ILU\((A + 10I)\), DW2048 ILU\((A + 0.01I)\)
Experiments

Fitting allows ILU($A + \sigma I$)

Often ILU on $A$ is not possible, ill-conditioned, or too expensive

Better results if we use a better conditioned ILU($A + \sigma I$) and allow the fitting to fix the diagonal
Experiments

Sometimes $Z_2^N$ better

QCD matrix (49K) close to $k_c$

Monte carlo, Matrix: matb5

<table>
<thead>
<tr>
<th>Number of samples</th>
<th>Trace estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC on diag(E2)</td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td></td>
</tr>
<tr>
<td>800</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

Monte carlo, Matrix: epb3

<table>
<thead>
<tr>
<th>Number of samples</th>
<th>Trace estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC on diag(E2)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here $E$ had smaller off diagonal variance — Not easily predictable by theory
Dynamically identifying smallest variance

For every fitting point $i = 1, \ldots, m$

\[
\text{Compute } a_i = A^{-1}e_i
\]

- Based on $a_{ii} = A_{ii}^{-1}$ update estimates for
  \[
  \text{var(diag}(A))
  \]
  \[
  \text{var(diag}(E)) \ (a_{ii} - M_i)
  \]
  \[
  \text{var(diag}(E^2)) \ (a_{ii} - p(M_i))
  \]
- Use $\|a_i\|^2 - a_{ii}^2$ to update estimate for
  \[
  \text{var(MC on } A) = \|\overline{A}\|^2_F
  \]
- Compute $\mu_i = U^{-1}L^{-1}e_i$ and update estimate for
  \[
  \text{var(MC on } E) \ (\|a_i - \mu_i\|^2 - a_{ii}^2 - \mu_{ii}^2)
  \]

Large differences in various methods would show after a few points
Dynamically identifying smallest variance

Estimated variance converges to actual variance

Relative differences apparent almost immediately

Matrix cage8: Variance Estimation

Matrix bwm200: Variance Estimation
Dynamically identifying smallest variance

Given a total $s$ of allowed steps, ask what method will give the smallest error at $s$

Eg., the matb5 QCD matrix:

After 10 steps, excellent match between estimated and observed variances
Conclusions

A method to approximate $\text{Tr}(A^{-1})$ based on ILU

- Negligible additional computational cost
- Very good accuracy, if ILU is effective
- Fitting improves accuracy
- MC on the fitted diagonal improves speed too

Easy to monitor and choose the most appropriate MC estimator