Principles and Analysis of Krylov Subspace Methods

Zdeněk Strakoš

Institute of Computer Science,
Academy of Sciences, Prague

www.cs.cas.cz/~strakos

Ostrava, February 2005

With special thanks to C.C. Paige, A. Greenbaum, J. Liesen, P. Tichý, J. Duintjer Tebbens, M. Rozložník, M. Tůma and to many others . . .

IDEALLY

$\equiv$ IN EXACT ARITHMETIC

COMPUTATIONALLY

$\equiv$ IN FINITE PRECISION ARITHMETIC

Introduction

Analysis of Krylov subspace methods (KSM)

Goal: To get some *understanding* when and why things work, and when and why they do not.

This is different from *solving* some particular problem, though, ultimately, the goal is to solve practical problems, and the analysis should serve this goal.

Preconditioning example

We do not analyze preconditioning, though it is the key part of practical KSM computations

- Ideally: preconditioning is included (analysis applied to the preconditioned system) . . . to some extent
- In order to understand preconditioning, we must understand the basic method first
- Preconditioning . . . better acceleration
  ([Hageman, Young - 81])

Limitations

I will attempt to present a picture, to the best of my abilities consistent and compact.

- It will definitely be incomplete.
- It will definitely be a personal view, regardless how fair I wish to be to the work of all distinguished developers and contributors.

Your role

To overcome the limitations by being critical to any point of view, position, argument and result presented in my contribution. Please judge their relevance, importance and correctness. Take what is good, and tell us what you doubt or disagree with.

Content

1. Principles and tools

2. Convergence (behavior) in exact arithmetic

3. Numerical behavior - general considerations

4. Numerical behavior - short recurrences

5. Numerical behavior of GMRES

6. Open problems

Chapter one

Principles and Tools

1. The principle of Krylov Subspace methods

2. Direct and iterative

3. . . . a linear problem?

4. Minimal polynomial idea

5. Orthogonality

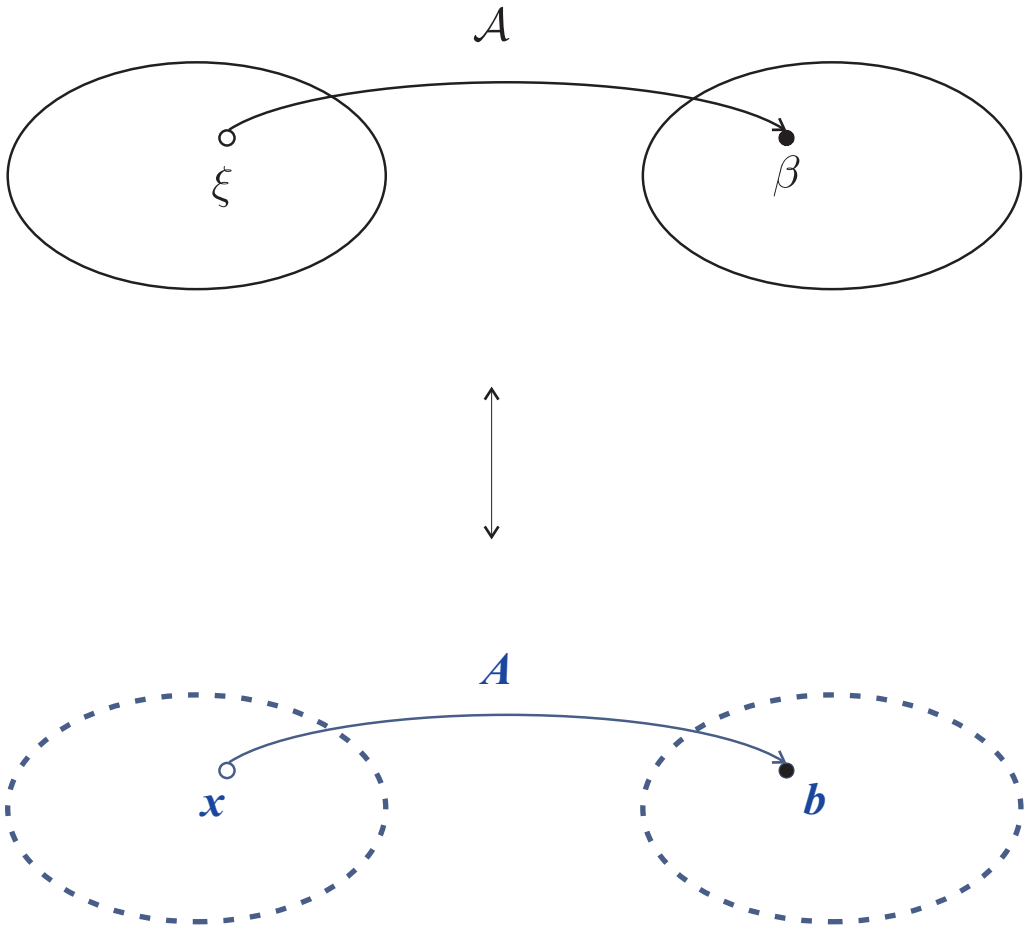
6. Initial guess and measuring convergence

1.1 The principle of KSM

We start with solving a real-world problem and assume a typical case modeled by integro-differential equations (in other cases the situation is similar).

Real problem –

modeling – discretization – computation.


$\mathcal{A}$
ξ
β
A
x
b

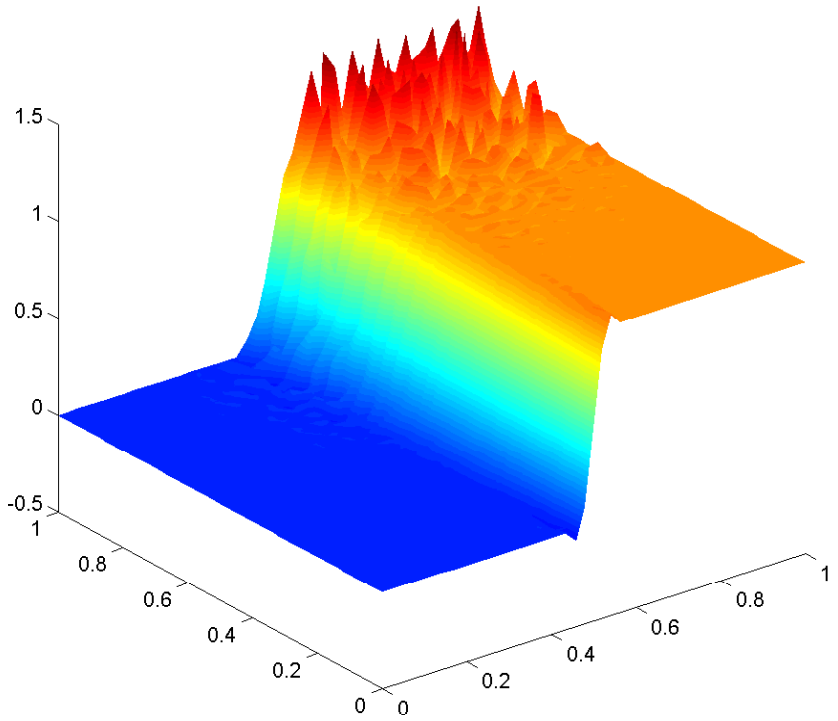
All steps should be considered a part of a single problem.

The numerical method can display behavior which is qualitatively wrong due to

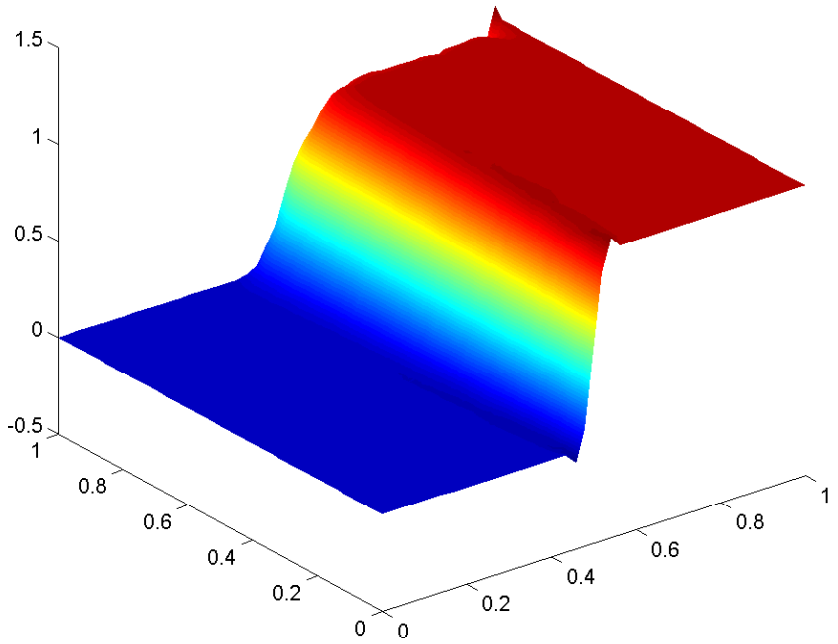
- discretization error
  (numerical chaos, see [Baxter, Iserles - 03, p. 19];
  convection-diffusion problems, see [Elman, Ramage - 01]),

- computational error.

Linear FEM discretization of the Raithby convection-diffusion model problem (plot from FEMLAB) gives qualitatively wrong solution



Linear FEM SUPG discretization of the Raithby convection-diffusion model problem (plot from FEMLAB) gives a stabilized solution, at the price of some artificial smoothing

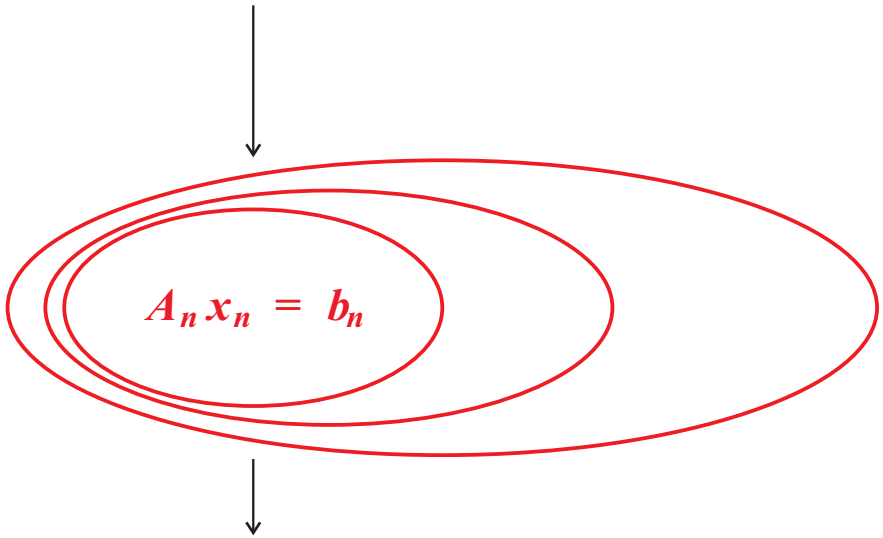


Principle of KSM

Projections of the N-dimensional problem onto nested Krylov subspaces of increasing dimension.

Step n: Model reduction from dimension N to dimension n, $n \ll N$.

$$A\,x \;=\; b$$



x_n approximates the solution x
using the subspace of small dimension.

Projection processes

$$x_n \in x_0 + \mathcal{S}_n, \quad r_0 \equiv b - Ax_o$$

where the constraints needed to determine x_n are given by

$$r_n \equiv b - Ax_n \in r_0 + A\mathcal{S}_n, \quad r_n \perp \mathcal{C}_n\,.$$

Here $\mathcal{S}_n$ is called the search space, $\mathcal{C}_n$ is called the constraint space.

r_0 decomposed to r_n + the part in $A\mathcal{S}_n$. It should be called orthogonal projection if $\mathcal{C}_n = A\mathcal{S}_n$, oblique otherwise.

Krylov subspace methods:

$\mathcal{S}_n \equiv \mathcal{K}_n \equiv \mathcal{K}_n(A, r_0) \equiv span\,\{r_0, \cdots, A^{n-1} r_0\}.$

$$
\begin{aligned}
x_n &\in x_0 + \mathcal{K}_n(A, r_0)\,, \\
x - x_n &= p_n(A)\,(x - x_0)\,,
\end{aligned}
$$

$$
\begin{aligned}
r_n &\equiv b - A x_n = p_n(A)\, r_0 \\
&\in r_0 + A\mathcal{K}_n(A, r_0)\,, \quad p_n(0) = 1\,.
\end{aligned}
$$

More general setting possible.

Krylov subspaces tend to contain the dominant information of A with respect to r_0. Unlike in the power method for computing the dominant eigenspace, here all the information accumulated along the way is used [Parlett - 80, Example 12.1.1].

Discretization means approximation of a continuous problem by a finite dimensional one. Computation using Krylov subspace methods means nothing but further model reduction. Well-tuned combination has a chance for being efficient.

The idea of Krylov subspaces is in a fundamental way linked with the problem of moments.

In Stieltjes' formulation, a sequence of numbers ξ_k, $k = 0, 1, \ldots$, is given and a non-decreasing distribution function $\omega(\lambda)$, $\lambda \geq 0$, is sought such that the Riemann-Stieltjes integrals satisfy

$$\int_0^\infty \lambda^k d\omega(\lambda) \;=\; \xi_k, \quad k = 0, 1, \ldots .$$

Here $\int_0^\infty \lambda^k d\omega(\lambda)$ represents the k-th moment of the distribution function $\omega(\lambda)$.

[Shohat, Tamarkin - 43], [Akhiezer - 65], [Karlin, Shapley - 53]

Vector moment problem of Vorobyev:

Find a linear operator A_n on $\mathcal{K}_n$ such that

$$\begin{aligned} A_n\, r_0 &= A r_0\,, \\ A_n\,(A r_0) &= A^2 r_0\,, \\ &\vdots \\ A_n\,(A^{n-2} r_0) &= A^{n-1} r_0\,, \\ A_n\,(A^{n-1} r_0) &= Q_n\,(A^n r_0)\,, \end{aligned}$$

where Q_n projects onto $\mathcal{K}_n$ orthogonally to $\mathcal{C}_n$.

[Vorobyev - 65], [Brezinski - 97]

Please notice that here $A^n r_0$ is decomposed into the part $Q_n\,(A^n r_0) \in \mathcal{K}_n$ and a part orthogonal to $\mathcal{C}_n$.

Therefore Q_n is the orthogonal projector if $\mathcal{C}_n = \mathcal{K}_n$, oblique otherwise.

1.2 Direct and iterative

- Some practitioners: when enough computer resources are available, then direct methods should be preferred to iterative. They compute accurately.

- If our goal is to improve methodology for solving given problem, then the questions about having or not having enough computer resources do not make sense.

- The statement should be read: "We wish to focus on a particular step and not to be disturbed by possible problems of the other steps."

- "and" means combination for eliminating the disadvantages and strengthening the advantages.

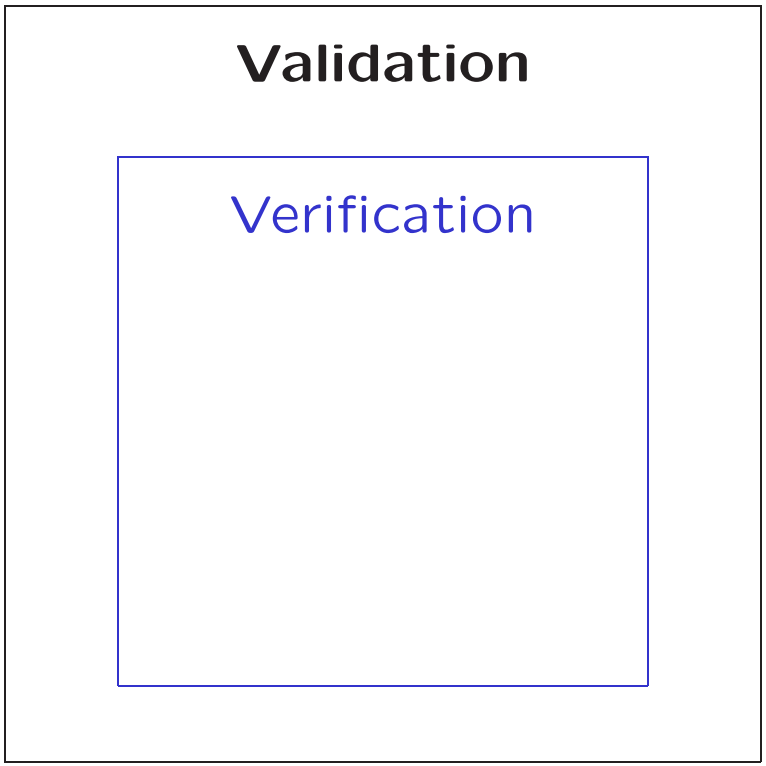
- Principal advantage of the iterative part is the possibility of stopping the computation at the desired accuracy level.

- It requires a meaningful stopping criterion. The errors of the model, discretization error and the computational error should be of the same order.

- Due to difficulties with the previous point this (potential) principal advantage is often presented as a disadvantage (a need for a stopping criteria . . .).

The point was presented by the founding fathers, it is well understood in the context of multilevel methods. But it is not accepted in practical computational mathematics in general. See the following quote from a negative referee report:

"... the author give a misguided argument. The main advantage of iterative methods over direct methods does not primarily lie in the fact that the iteration can be stopped early [whatever this means], but that their memory (mostly) and computational requirements are moderate."



[Babuška - 03], [Oden et al - 03]

1.3 . . . a linear problem?

Methods and phenomena in NLA need not be linear!

How fast we get an acceptable approximate solution?

- In modern iterative methods we have to study transient phase (early stage of computation) which represents a nonlinear phenomenon in a finite dimensional space of small dimensionality. [Pták, Trefethen, Baxter and Iserles, ...]

- Operator approach can not describe the transient phase! Linearization by asymptotic tools is (even with adaptation) of limited use.

Another look

An information-based argument: A^{-1} uses global information from A. Consequently, good iterative solution requires that:

- Either the global information is taken care for by a good preconditioner. An extremely good preconditioner, transforming the system matrix almost to identity, reduces the number of iterations to $\mathcal{O}(1)$

- Or the global information is gathered together by many (close to N) matrix-vector multiplications.

What is wrong?

Solving $Ax = b$ for some particular meaningful b can be different from solving the system with the matrix A and some worst-case right hand side! The data have typically some meaning and are correlated.

For an acceptable accuracy we may not need the full global communication.

Operator approach (in analytical considerations working with an approximation of A^{-1}) leads to asymptotics. Solving $Ax = b$ for the particular meaningful $\{A, b\}$ is different from computing A^{-1}.

[Beckermann, Kuijlaars - 02]

1.4 Minimal polynomial idea

- $r_n = p_n(A)\, r_0$.
  If p_n is the minimal polynomial of A, then $x_n = x$.

- More thoughtfully: p_n minimal polynomial of A with respect to $x - x_0$ (or $r_0 = b - Ax_0$), then $x_n = x$.

- Should we focus on approximating the minimal polynomial? In general, no! Quantification is extremely difficult. We will see on an example of the conjugate gradient method how easily one can be misguided to a wrong conclusion.

- what else? Linear model reduction – extraction of the dominant information as fast as possible.

Question:

When does $K_n(A, r_0)$ contain enough information about the original N-dimensional linear algebraic problem in order to provide a good approximate solution x_n?

$\downarrow$

Convergence (better Behavior).

1.5 Orthogonality

- Refining the multidimensional information (unlike in the power method), and computing projections from $b, Ab, \ldots, A^{n-1}b$?

  $\downarrow$

  Mutual orthogonalization, or orthogonalization against some auxiliary vectors.

  Goal: getting in an affordable way a good basis ("good" does not necessarily mean "the best possible")

- Practical computations means limited accuracy.

- Computer science concept of analysis and synthesis of an algorithm with computing intermediate results to an arbitrary accuracy needed for the prescribed accuracy of the final solution does not work as a general approach.

  The intermediate quantities can be much less accurate than the final computed result! We will see an example of this surprising fact revealed by the Wilkinson work and pointed out by Parlett, later.

- (Fixed) limited accuracy means rounding errors.

Rounding errors in our calculations destroy orthogonality, mathematical structure of Krylov subspaces and the projection principle. Is there a chance to understand finite precision iterative processes and to estimate maximal attainable accuracy?

$\downarrow$

Numerical Stability Analysis

A complicated matter. For example, an algorithm can be inherently unstable due to the wrong choice of the subspaces involved, and more accurate computing of the bases can make the overall behaviour worse! We will see an example of a variant of GMRES.

1.6 Initial guess and measuring convergence

$Ax = b$, A square, nonsingular, N by N.

Consider $x_0 \equiv 0$, for a good reason. Ideally, no loss of generality; with a nonzero x_0 replace b by $b - Ax_0$.

Computationally, unless x_0 can be based on some relevant information ensuring $x_0 \approx x$ $(\|b - Ax_0\| \leq \|b\|)$, the choice $x_0 = 0$ should be preferred.

If a nonzero x_0 is used, then the possible illusion of (entirely artificial) fast convergence should be avoided by the following step:

Given a preliminary guess x_p, determine the scaling parameter

$$\|r_0\| = \|b - A(x_p\zeta_{\min})\| = \min_{\zeta} \|b - A(x_p\zeta)\|, \quad \zeta_{\min} = \frac{b^* A x_p}{\|Ax_p\|^2},$$

and set $x_0 = x_p\zeta_{\min}$.

[Hegedüs], [Paige, S - 02]

How good is an approximate solution x_n ?

Consider a computed approximation x_n. Then

$$Ax_n = b - r_n\,, \quad r_n = b - Ax_n\,.$$

Thus, $-r_n$ represents the (unique) perturbation Δb of the right hand side b such that x_n is the exact solution of the perturbed system.

A simple one-sided example of the perturbation theory – backward error approach.

[Goldstine, Von Neumann - 47], [Turing], [Wilkinson - 63, 65]

How good is an approximate solution $\widehat{x}$ of the linear algebraic problem $A\,x \;=\; b$?

Perturbation theory: $\quad (A + \Delta A)\,\widehat{x} \;=\; b + \Delta b$.

Normwise relative backward error: Given $\widehat{x}$, construct ΔA , Δb such that both $\|\Delta A\|/\|A\|$ and $\|\Delta b\|/\|b\|$ are minimal;

$$\widehat{x} \;\longrightarrow\; \frac{\|\Delta A\|}{\|A\|} = \frac{\|\Delta b\|}{\|b\|} = \frac{\|b - A\widehat{x}\|}{\|b\| + \|A\|\|\widehat{x}\|} \;.$$

Measuring convergence: $\|r_n\| \,/\, (\,\|b\| + \|A\|\,\|x_n\|\,)$.

We ask and answer the question

"How close is the problem $(A + \Delta A)\, x_n = b + \Delta b$, which is solved by x_n accurately, to the original problem $A\, x = b$?"

Perhaps this is what we need – the matrix A and the right hand side b are inaccurate anyway.

Is the computed convergence curve close to the exact one?

Please notice the difference between the normwise relative backward error and the role of the relative residual norm.

Backward error restricted to the right hand side only is given by

$$\|r_n\|/\|b\| \, .$$

Moreover, for an unwise choice of x_0 this may differ greatly from the frequently used relative residual norm

$$\|r_n\|/\|r_0\| \, .$$

Example: Liesen, Tichý

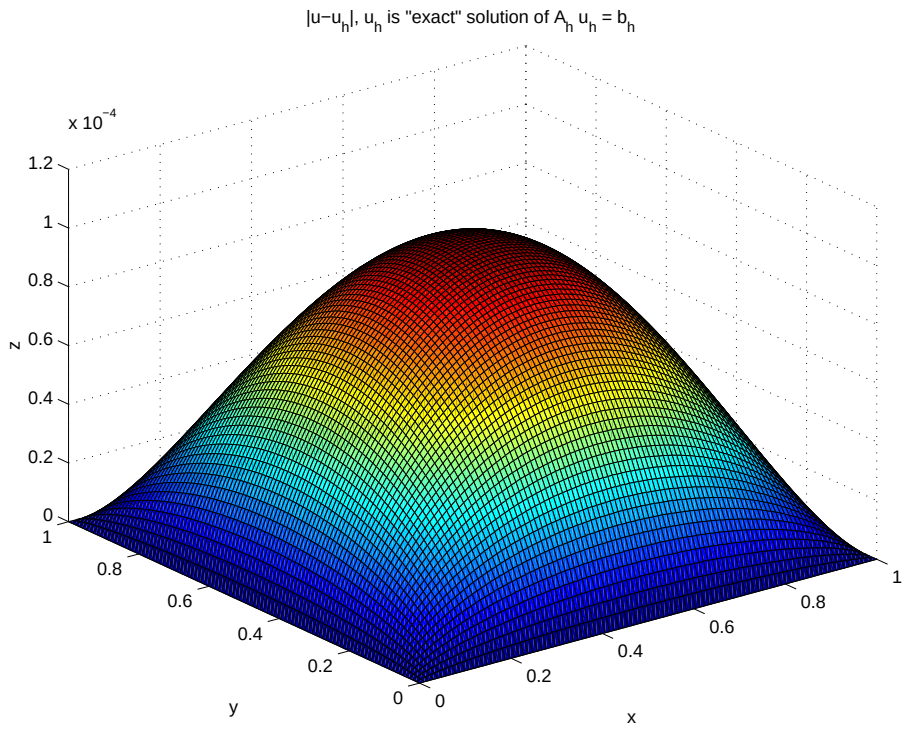
$$-\Delta\, u \;=\; 32(\eta_1 - \eta_1^2 + \eta_2 - \eta_2^2)$$

on a unit square with zero Dirichlet boundary conditions. Exact solution is $u(\eta_1, \eta_2) \;=\; 16(\eta_1\eta_2 - \eta_1\eta_2^2 - \eta_1^2\eta_2 + \eta_1^2\eta_2^2)$.

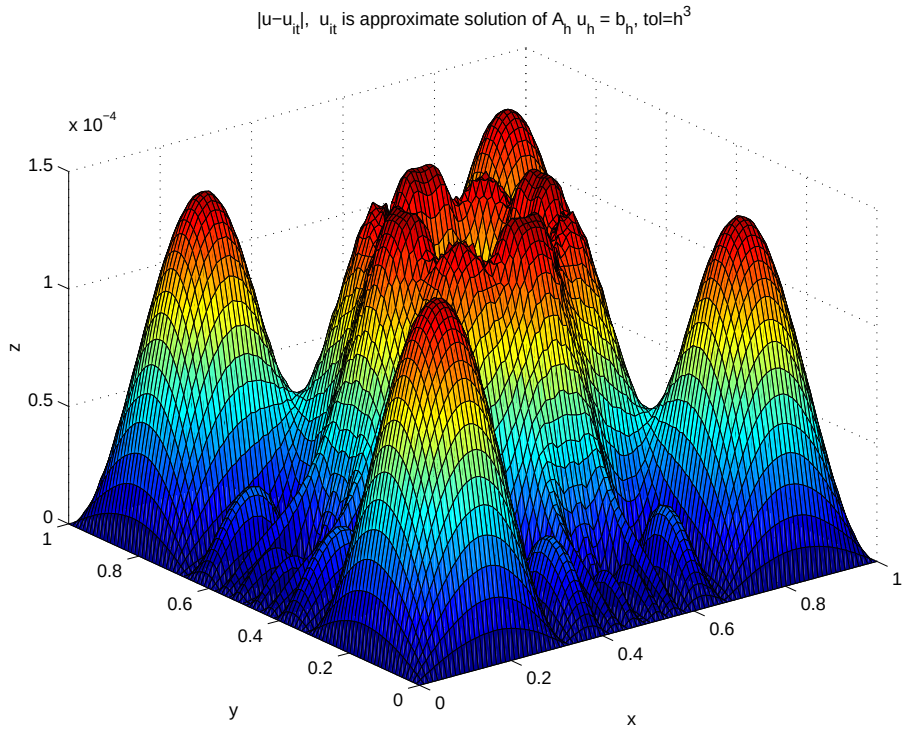
Linear FEM, discretization error in forming the linear algebraic system $\approx h^{-2}$. For illustration, the stopping criterion for the CG computation was based on the normwise relative backward error and set to

$$\frac{\|r_n\|}{\|b\| + \|A\|\,\|x_n\|} < h^{-3}.$$

Discretization error $u - x$ for $h = 1/101$. The exact solution is approximated sufficiently accurately by the MATLAB direct solver.



Total error $u - x_n$ with stopping tolerance for the normwise relative backward error in CG computation set to h^{-3}.



- Then the computational error does not contribute significantly to the total error measured by the accuracy of the computed approximate solution

- The criterion is cost–efficient. A similar reasoning based on the relative residual norm approximately doubles the number of iterations

- A simple example - gradient is not well approximated.

- It is desirable that the evaluation of the computational error (and stopping criteria) is based on a physically meaningful quantity.

Consistency of the whole solution process

Elliptic PDE can serve as a nice example. The weak formulation leads to a SPD bilinear form, with the energy as the quantity in charge. The Galerkin FEM discretized problem is again SPD, and, consequently, an algebraic iterative method consistent with the whole solution process should minimize the energy norm of the error of the finite-dimensional approximate solution at each iteration step. The world makes a sense - the conjugate gradient method represents such consistency.

For some pioneering work on "cascadic CG" see [Deuflhard - 93 (94)]. Recently, a general theory has been built by M. Arioli and his co-workers, see [Arioli et al. - 04].