

Loud computations? Noise in iterative solvers

Stefan Wild

Argonne National Laboratory
Mathematics and Computer Science Division

Joint work with Jorge Moré

June 13, 2013

This Talk



- ◊ What is computational noise?
- ◊ How can noise be estimated efficiently?
- ◊ What insights can be provided for iterative solvers?
- ◊ How does noise affect numerical differentiation?

Computational Noise is not a Newcomer

From Hamming's 1971 Introduction to Numerical Analysis:

Where does this noise come from? ... infinite processes in mathematics which of necessity must be approximated by finite processes.

Truncation vs. roundoff Finite number length leads to roundoff. Finite processes lead to truncation.



Competing errors Smaller steps usually reduce truncation error and may increase roundoff error.

Deterministic In practice, the same input, barring machine failures, gives the same result.

Computational Noise is not a Newcomer

From Hamming's 1971 Introduction to Numerical Analysis:

Where does this noise come from? ... infinite processes in mathematics which of necessity must be approximated by finite processes.

Truncation vs. roundoff Finite number length leads to roundoff. Finite processes lead to truncation.



Competing errors Smaller steps usually reduce truncation error and may increase roundoff error.

Deterministic In practice, the same input, barring machine failures, gives the same result. ← changing!

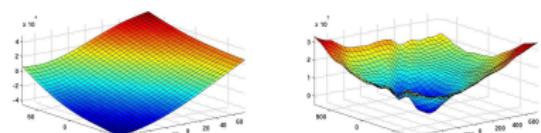
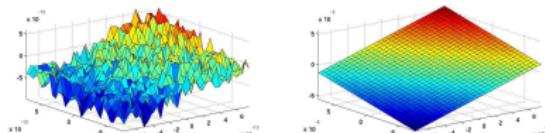
Computational Noise in Deterministic Simulations

Difference $|f(x) - f(x + Z\omega)|$,

Finite precision + finite processes

- ◊ Iteratively solving systems of PDEs or estimating eigenvalues
- ◊ Adaptively computing integrals
- ◊ Discretizations/meshes

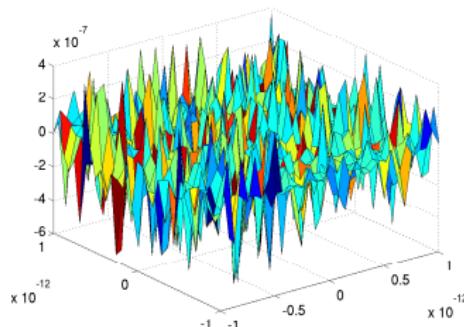
destroy underlying smoothness



X-ray microscopy simulation

Goal: estimate the “variation” in $f(\mathbf{x})$

- ◊ a few f evaluations
- ◊ deterministic and stochastic noise



Sparse linear large-scale system

The Noise Level ϵ_f

Simple model for the noise

$$f(t) = f_s(t) + \varepsilon(t), \quad t \in \mathcal{I}$$

f the computed function

f_s a smooth, deterministic function

ε is the noise with $\{\varepsilon(t) : t \in \mathcal{I}\}$ iid

← only assumption

The noise level of f is $\varepsilon_f = (\text{Var} \{\varepsilon(t)\})^{1/2}$

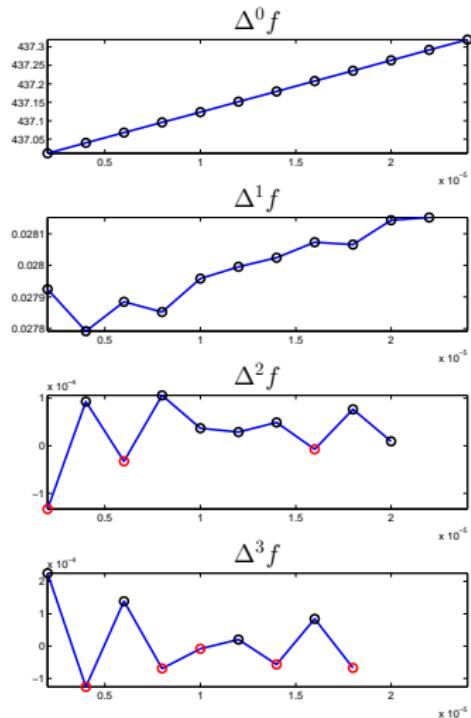
k -th Order Difference $\Delta^k f(t) = \Delta^k f_s(t) + \Delta^k \varepsilon(t)$

$$\begin{aligned}\Delta^{k+1} f(t) &= \Delta^k f(t+h) - \Delta^k f(t), \\ \Delta^0 f(t) &= f(t)\end{aligned}$$

Observe:

1. Differences of smooth f_s tend to zero rapidly
2. Differences of noise are bounded away from zero
3. If f_s is k -times differentiable,
$$\Delta^k f(t) = f_s^{(k)}(\xi_k)h^k + \Delta^k \varepsilon(t),$$
$$\xi_k \in (t, t+kh)$$

Idea: Choose h, k to remove smooth component



Theory Underlying the ECNoise Algorithm

For $\{\varepsilon(t + ih) : i = 0, \dots, m\}$ iid and $k \leq m$:

1. $\mathbf{E}\{\Delta^k \varepsilon(t)\} = 0$
2. $\gamma_k \mathbf{E}\{[\Delta^k \varepsilon(t)]^2\} = \varepsilon_f^2 \quad \gamma_k = \frac{(k!)^2}{(2k)!}$
3. If f_s is continuous at t , then

$$\lim_{h \rightarrow 0} \gamma_k \mathbf{E}\{[\Delta^k f(t)]^2\} = \varepsilon_f^2$$

4. If f_s is k -times continuously differentiable at t , then

$$\lim_{h \rightarrow 0} \frac{\gamma_k \mathbf{E}\{[\Delta^k f(t)]^2\} - \varepsilon_f^2}{h^{2k}} = \gamma_k [f_s^{(k)}(t)]^2$$

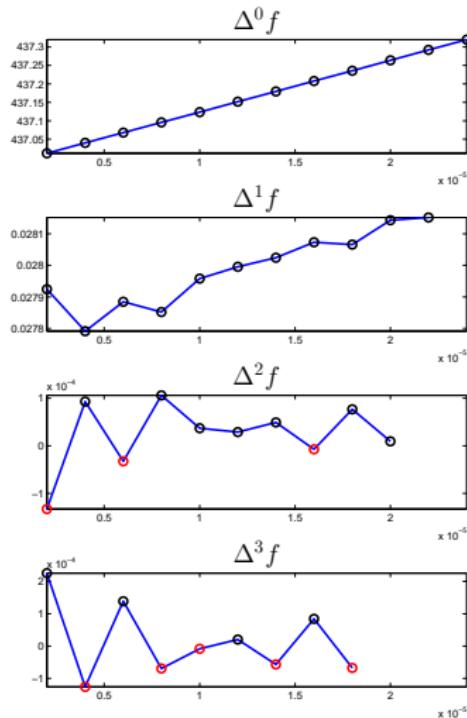
$$\Rightarrow \varepsilon_f^2 \approx \gamma_k \mathbf{E}\{[\Delta^k f(t)]^2\},$$

when the sampling distance h is sufficiently small

The ECNoise Algorithm

Uses $\sigma_k = \left(\frac{\gamma_k}{m+1-k} \sum_{i=0}^{m-k} [\Delta^k f(t + ih)]^2 \right)^{1/2}$

1. Chooses k
 2. Verifies h is small enough
-
- ◊ Random direction p for multivariate $f(x_b + tp) =: g(t)$
 - ◊ Works for deterministic f
 - ◊ Target: correct order of magnitude



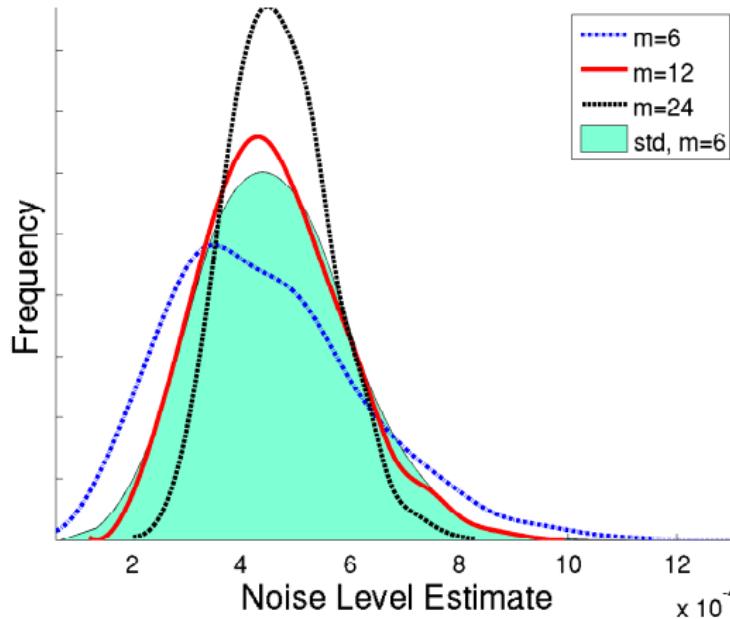
[Estimating Computational Noise. Moré & W., SISC 2011]

$$\text{ECNoise Estimator } \sigma_k = \left(\frac{\gamma_k}{m+1-k} \sum_{i=0}^{m-k} [\Delta^k f(t_i)]^2 \right)^{1/2}$$

For $f(t) = \cos(t) + \sin(t) + 10^{-3}U_{[0, 2\sqrt{3}]}$ ($m = 6, t_i = \frac{i}{100}$)

$f(t_i)$	$\Delta f(t_i)$	$\Delta^2 f(t_i)$	$\Delta^3 f(t_i)$	$\Delta^4 f(t_i)$	$\Delta^5 f(t_i)$	$\Delta^6 f(t_i)$
1.003	7.54e-3	2.15e-3	1.87e-4	-5.87e-3	1.46e-2	-2.49e-2
1.011	9.69e-3	2.33e-3	-5.68e-3	8.73e-3	-1.03e-2	
1.021	1.20e-2	-3.35e-3	3.05e-3	-1.61e-3		
1.033	8.67e-3	-2.96e-4	1.44e-3			
1.041	8.38e-3	1.14e-3				
1.050	9.52e-3					
1.059						
σ_k	6.78e-3	8.96e-4	9.02e-4	9.93e-4	1.10e-3	1.14e-3

Ex.- ECNoise on Stochastic MC Function



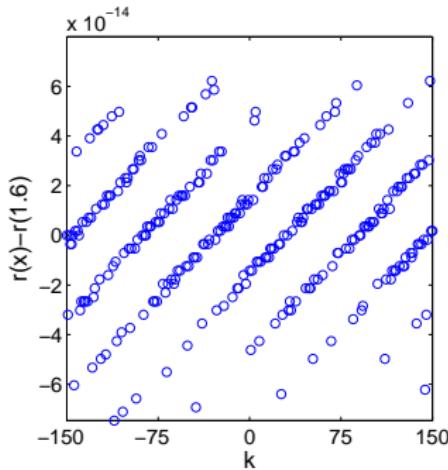
10000 noise estimates

- ◊ ECNoise produces reliable estimates
- ◊ 99.6% within a factor 4 for $m = 6$
- ◊ correct order of magnitude

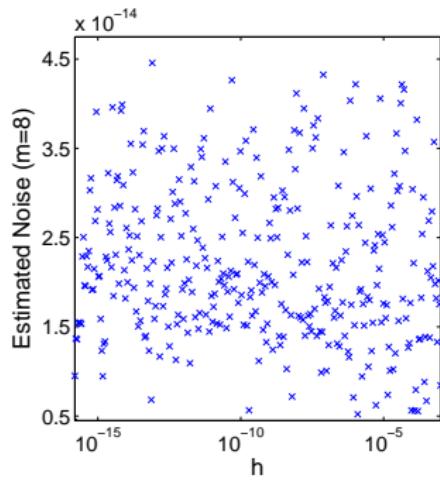
$$f(x) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \prod_{i=0}^n \frac{e^{-\frac{\|u\|^2}{2}}}{1+r_i(u,x)} du, \quad r_i(u,x) = \begin{cases} \frac{1}{10} & i=0 \\ r_{i-1}(u,x) e^{x_i u_i - x_i^2/2} & i \geq 1 \end{cases}$$

Transition to Non-IID & Deterministic Noise

Kahan's $r(x) = \frac{622-x(751-x(324-x(59-4x)))}{112-x(151-x(72-x(14-x)))}$ violates iid assumption



Kahan's r at $x = 1.6 + 2^{-52}k$



Noise estimates for $r(1.6)$

- ◊ All noise estimates within factor 4 of $2 \cdot 10^{-14}$
- ◊ (Unlikely) $m+1$ points solely on one line $\Rightarrow \epsilon_f \approx 2 \cdot 10^{-15}$

Deterministic Test Problems

“Convex” UF Quadratics

$$f_\tau(t) = \|y_\tau(x_0 + tp)\|_2^2,$$

y_τ from iterative solver for $Ay_\tau(x) = x$ with tolerance $\tau > 0$

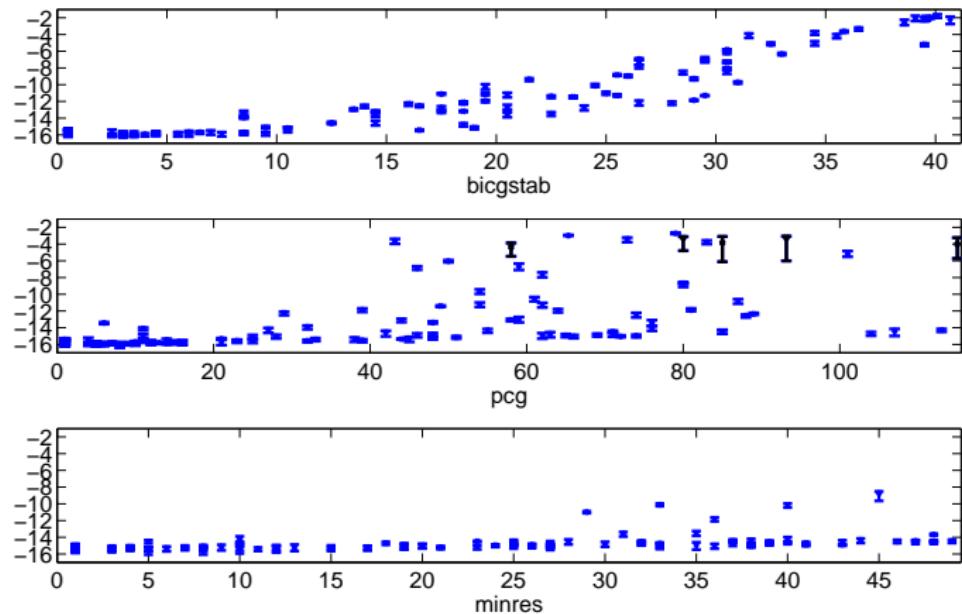
- ◊ 116 spd UF matrices ($n < 10^4$), scaled by diagonal
- ◊ 28 with $\kappa(A) \leq 10$, 10 with $\kappa(A) \geq 10^{10}$
- ◊ random direction $p \in \mathbb{R}^n$
- ◊ variety of tolerances τ in $[10^{-8}, 10^{-2}]$
- ◊ only $m = 8$ additional evaluations
- ◊ tested bicgstab, gmres, idr(s), pcg, minres, minresqlp, symmlq

Highly Nonlinear MINPACK-2 Problems

$$f_\tau(x) = \text{chop}(f(\text{chop}(x, \tau)), \tau)$$

→ similar

Consistency with Respect to the Sampling Distance h



For $h \leq 10^{-10}$, min and max ϵ_f within factor $\eta^2 = 16$

◊ `bicgstab` 116/116

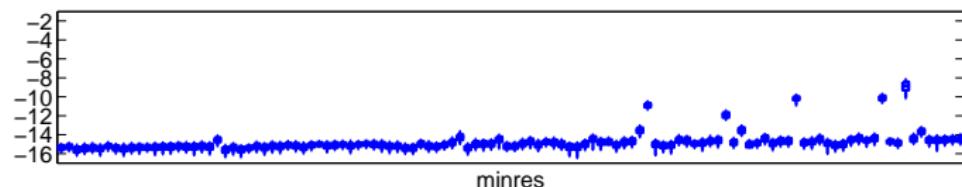
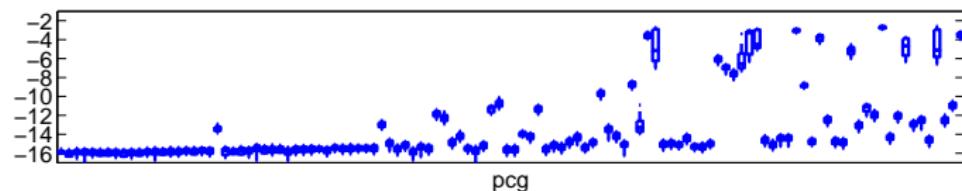
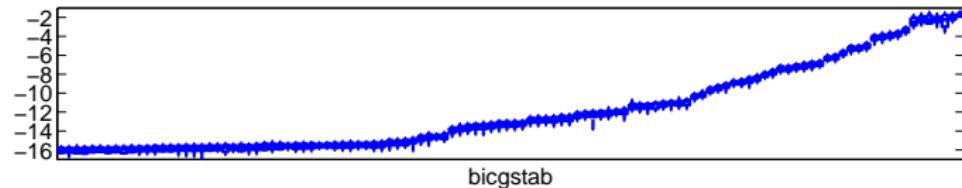
◊ `pcg` 111/116

◊ `minres` 116/116

x axis = mean number of iterations required to achieve tolerance

[116 UF matrices: $m = 8$; $1 p$; $\tau = 10^{-3}$; $h = 10^{-10}, \dots, 10^{-15}$]

Consistency with Respect to Sampling Direction p

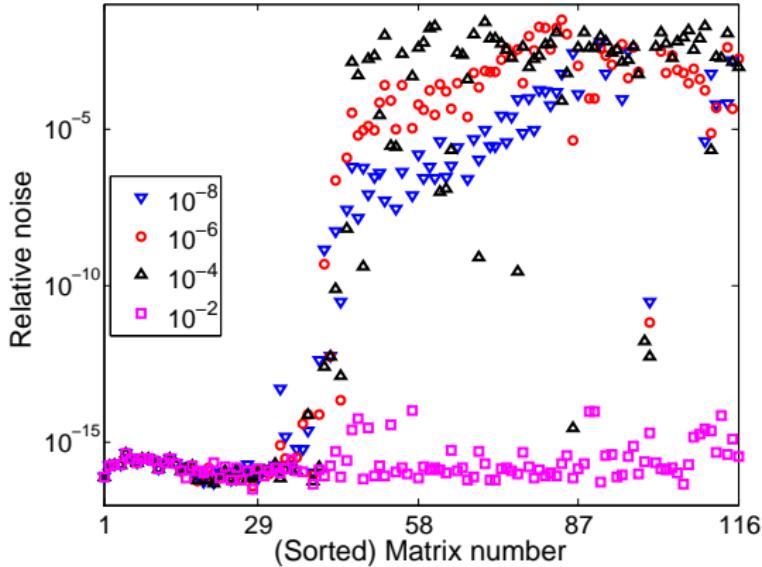


Estimates within factor
 $\eta = 4$ of median
◊ bicgstab 99.71%
◊ pcg 96.64%
◊ minres 99.81%

x axis = matrices sorted by bicgstab median noise

[116 UF matrices: $m = 8; 10^3$ p ; $\tau = 10^{-3}$; $h = 10^{-12}$]

ECNoise on Functions f_τ

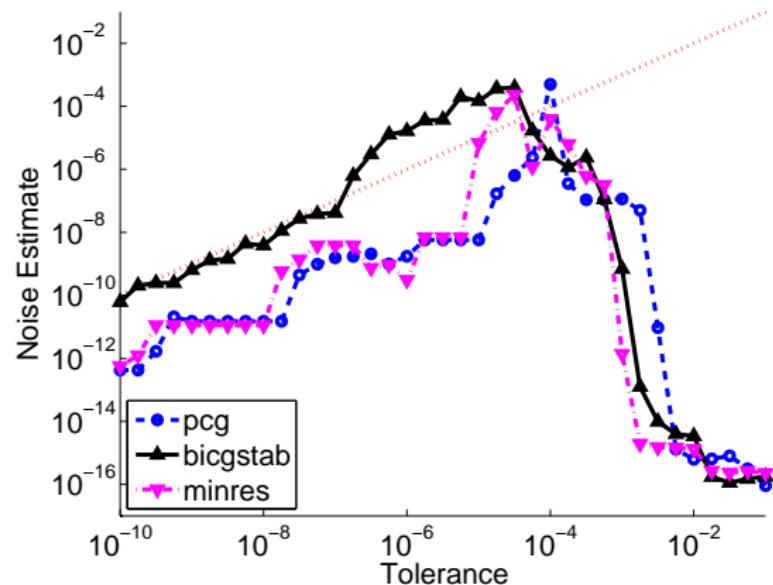


bicgstab, x axis sorted by $\kappa(A)$

Noisy UF Quadratics

- ◊ Reliable estimates, $m = 8$ additional evaluations
- ◊ Non-monotone relationship between the relative noise and tolerance τ

ECNoise on Functions f_τ



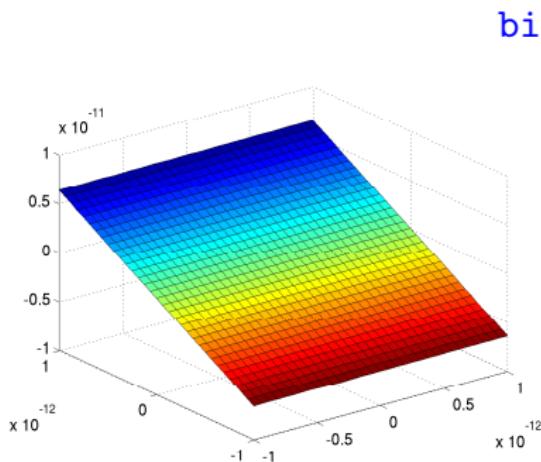
Noisy UF Quadratics

- ◊ Reliable estimates, $m = 8$ additional evaluations
- ◊ Non-monotone relationship between the relative noise and tolerance τ

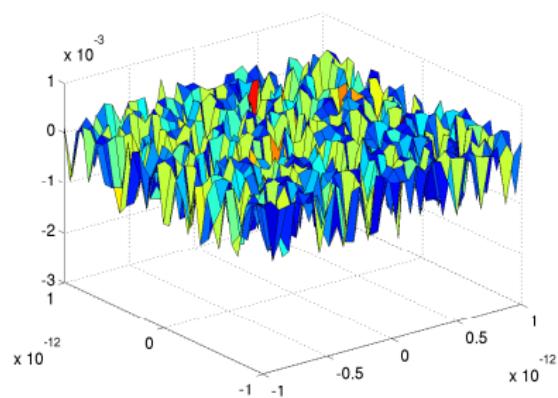
One quadratic (bcsstk02), multiple solvers

How does bcsstk02 Noise Change with the Tolerance?

2d slice of f for bcsstk02 ($n = 66, \kappa(A) = 1833$)



$$\tau = 10^{-2}, \text{ std} = 3.86e-12$$

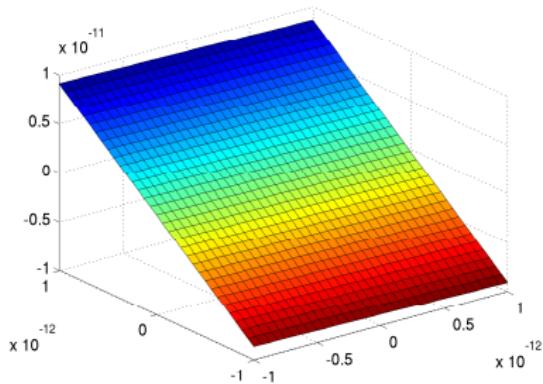


$$\tau = 10^{-5}, \text{ std} = 4.98e-04$$

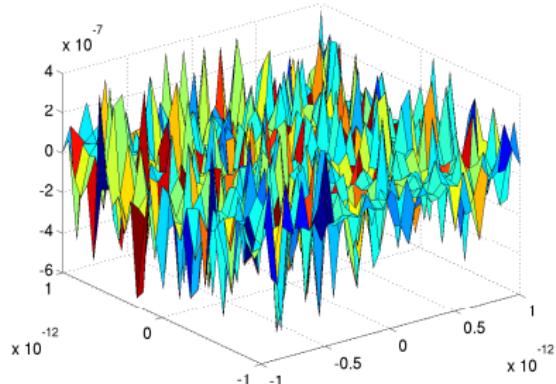
How does bcsstk02 Noise Change with the Tolerance?

2d slice of f for bcsstk02 ($n = 66, \kappa(A) = 1833$)

pcg

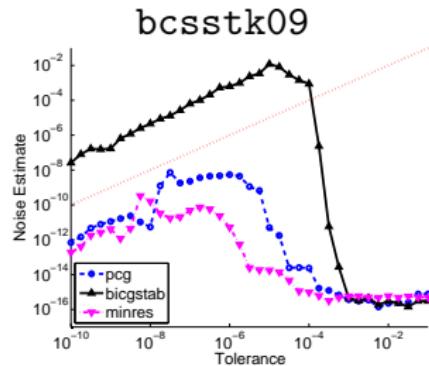
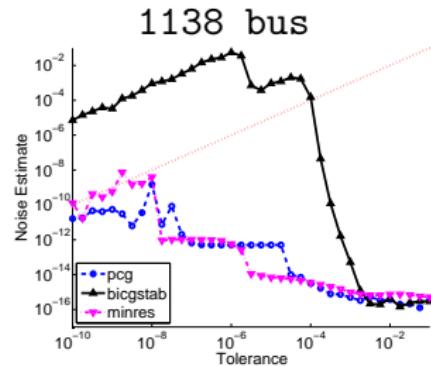
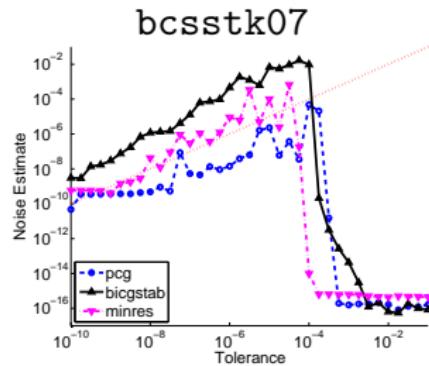
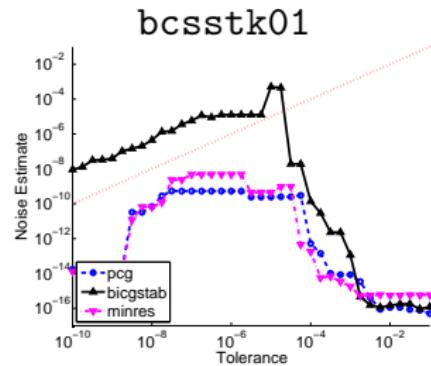


$$\tau = 10^{-2}, \text{ std} = 5.29e-12$$



$$\tau = 10^{-5}, \text{ std} = 1.90e-07$$

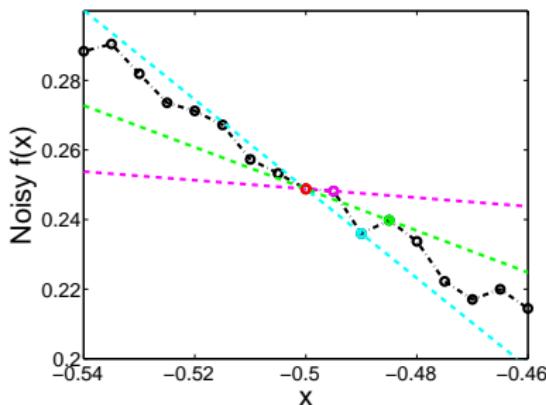
Noise Estimates for Different Tolerances



II. Noise Estimates in Finite Differences

Minimize the MSE

$$\mathbf{E} \{ \mathcal{E}(h) \} = \mathbf{E} \left\{ \left(\frac{f(t_0+h) - f(t_0)}{h} - f'_s(t_0) \right)^2 \right\}$$



Our h will depend on

- ◊ Loose estimate of noise
 - ◊ Stochastic theory
1. $f(t) = f_s(t) + \epsilon$ on
 $I = \{t_0 + h : 0 \leq h \leq h_0\}$
 2. f_s twice differentiable
 3. $\mu_L \leq |f''_s| \leq \mu_M$ on I !

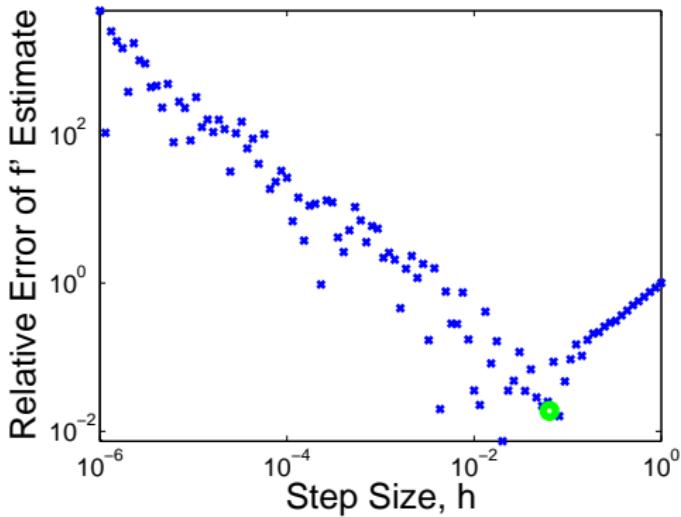
[*Estimating Noisy Derivatives. Moré & W., TOMS 2012*]

Near-Optimal Forward Difference Parameter h

$$\frac{1}{4}\mu_L^2 h^2 + 2\frac{\varepsilon_f^2}{h^2} \leq \mathbf{E} \{ \mathcal{E}(h) \} \leq \frac{1}{4}\mu_M^2 h^2 + 2\frac{\varepsilon_f^2}{h^2}$$

$h \downarrow$ Variance (noise) dominates

$h \uparrow$ Bias (f'') dominates



Near-Optimal Forward Difference Parameter h

$$\frac{1}{4}\mu_L^2 h^2 + 2\frac{\varepsilon_f^2}{h^2} \leq \mathbf{E} \{ \mathcal{E}(h) \} \leq \frac{1}{4}\mu_M^2 h^2 + 2\frac{\varepsilon_f^2}{h^2}$$

$h \downarrow$ Variance (noise) dominates

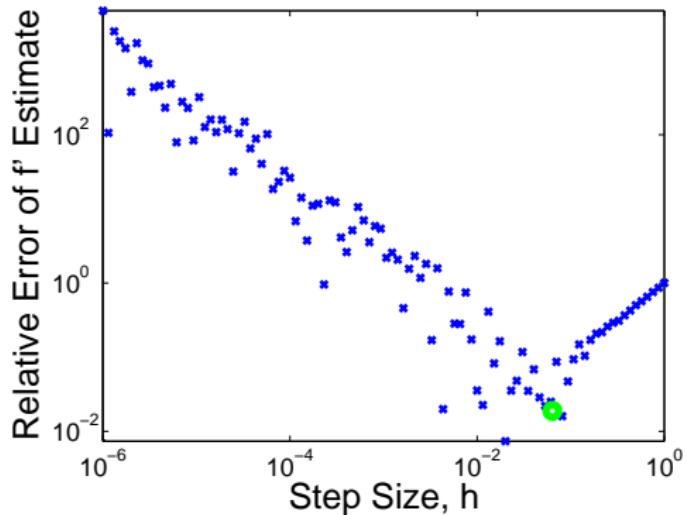
$h \uparrow$ Bias (f'') dominates

For h_0 sufficiently large

1. Upper bound minimized by

$$h^* = 8^{1/4} \left(\frac{\varepsilon_f}{\mu_M} \right)^{1/2}$$

2. When $\mu_L > 0$, h^* is near-optimal:

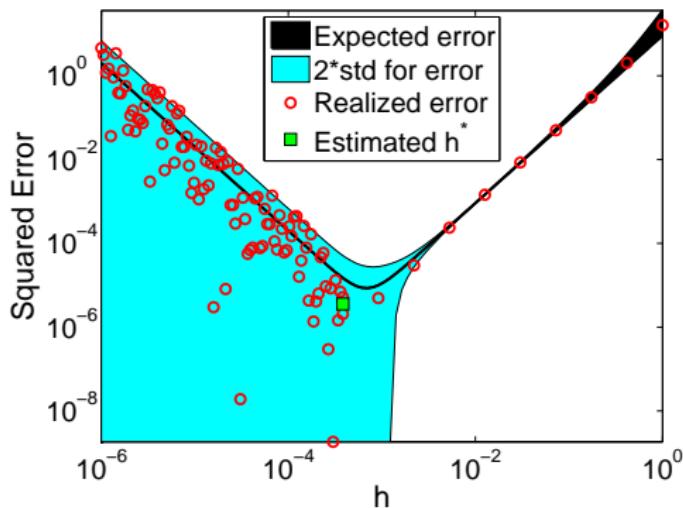


$$\mathbf{E} \{ \mathcal{E}(h^*) \} = \sqrt{2}\mu_M\varepsilon_f \leq \left(\frac{\mu_M}{\mu_L} \right) \min_{0 \leq h \leq h_0} \mathbf{E} \{ \mathcal{E}(h) \}$$

Stochastic Examples

Estimate $f'_s(t) = E\{f(t)\}'$ at $t = 1$ ($\varepsilon_f = 10^{-6}$)

Cubic, $t^3 + 10^{-6}U_{[-2\sqrt{3}, 2\sqrt{3}]}$



Log-log realizations of $\mathcal{E}(h) = \mathbf{E} \left\{ \left(\frac{f(t_0+h) - f(t_0)}{h} - f'_s(t_0) \right)^2 \right\}$

Expected error and uncertainty regions predicted by the theory

Extension: Central Differences

First derivatives, $\frac{f(t_0+h)-f(t_0-h)}{2h}$

- ◊ $|h^*| = \gamma_5 \left(\frac{\varepsilon_f}{\mu_M} \right)^{1/3}, \quad \gamma_5 = 3^{1/3} \approx 1.44$
- ◊ $\mu_L \leq |f_s^{(3)}| \leq \mu_M$
- ◊ $\mathbf{E}\{\mathcal{E}_c(h^*)\} \leq \left(\frac{\mu_M}{\mu_L} \right)^{2/3} \min_{|h| \leq h_0} \mathbf{E}\{\mathcal{E}_c(h)\}$

Second derivatives, $\frac{f(t_0+h)-2f(t_0)+f(t_0-h)}{h^2}$

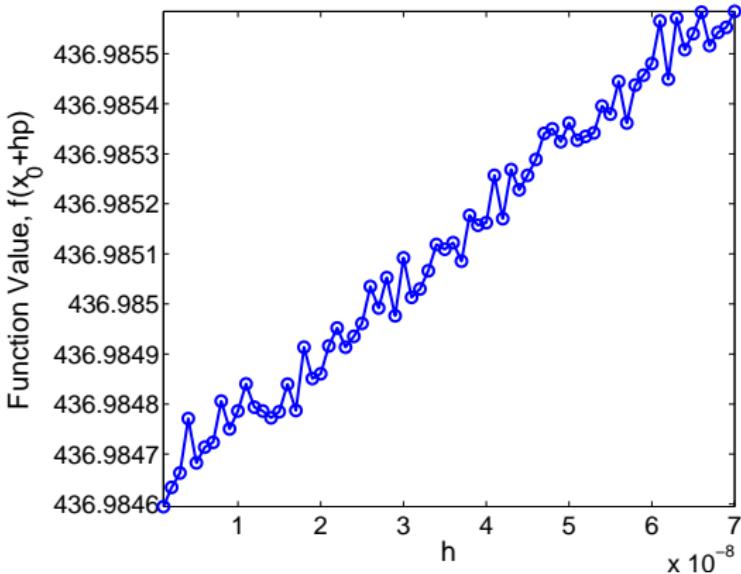
- ◊ $|h^*| = \gamma_7 \left(\frac{\varepsilon_f}{\mu_M} \right)^{1/4}, \quad \gamma_7 = 2^{5/8} 3^{1/8} \approx 2.33$
- ◊ $\mu_L \leq |f_s^{(4)}| \leq \mu_M$
- ◊ $\mathbf{E}\{\mathcal{E}_2(h^*)\} \leq \left(\frac{\mu_M}{\mu_L} \right) \min_{|h| \leq h_0} \mathbf{E}\{\mathcal{E}_2(h)\}$
- ◆ use to obtain rough estimate of $|f_s''|$ for forward-difference h



Ex.- Noisy Deterministic Functions (bicgstab, $\tau = 10^{-3}$)

Subset of 100 UF
matrices

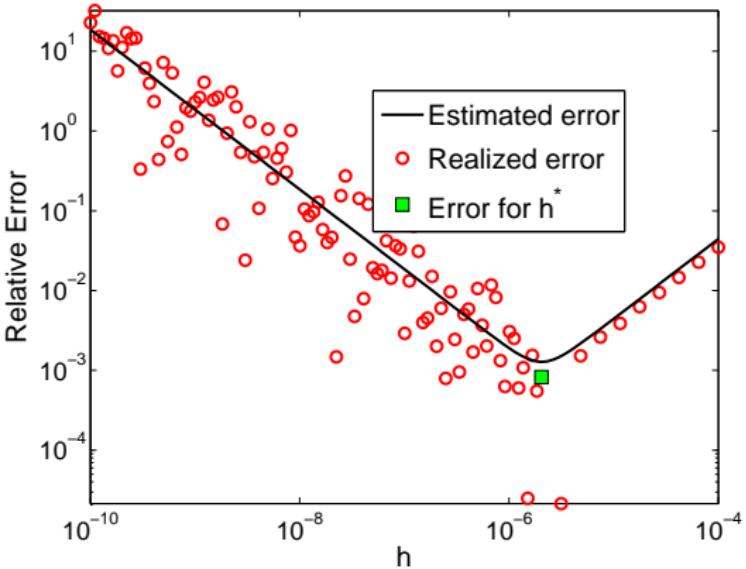
- ◊ FD sensitive to noise



Ex.- Noisy Deterministic Functions (bicgstab, $\tau = 10^{-3}$)

Subset of 100 UF matrices

- ◊ FD sensitive to noise
- ◊ Exhibits behavior similar to stochastic FD

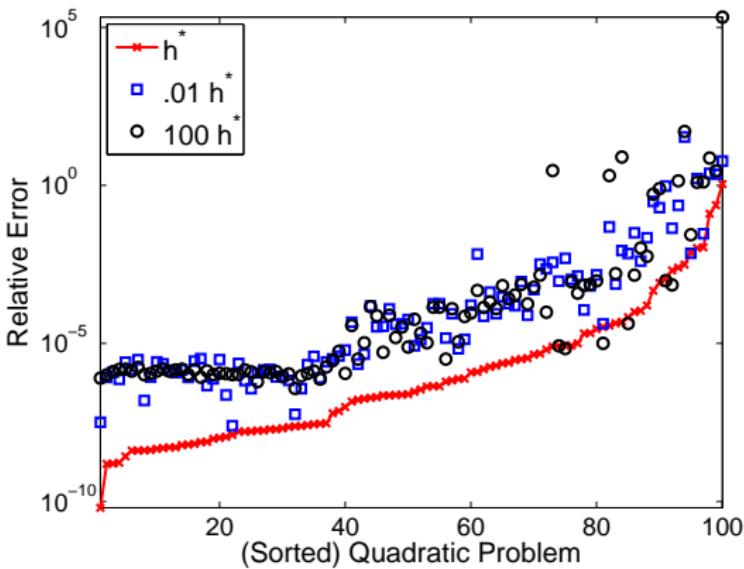


Compared with AD (INTLAB) derivative

Ex.- Noisy Deterministic Functions (bicgstab, $\tau = 10^{-3}$)

Subset of 100 UF matrices

- ◊ FD sensitive to noise
- ◊ Exhibits behavior similar to stochastic FD
- ◊ h_M obtains 2 more correct digits than $10^{\pm 2} h_M$
- ◊ h_M significantly better than $\sqrt{\epsilon_{\text{mach}}}$



Compared with AD (INTLAB) derivative

Summary: How Loud Are Your Simulations?

- ◊ Computational noise complicates analysis of simulation-based functions, worst-case bounds overly pessimistic (see Baudouin talk)
- ◊ With a few (6-8) additional evaluations, ECNoise reliably estimates the noise
- ◊ Stochastic theory for near-optimal difference parameters
- ◊ Coarse estimates of $|f''|$ (2-4 evaluations) yield more accurate directional derivatives
- ◊ Both work on deterministic functions in practice



Some refs <http://mcs.anl.gov/~wild>:

Estimating Computation Noise, SISC 2011.

Estimating Derivatives of Noisy Simulations, TOMS 2012.

Do You Trust Derivatives or Differences? Preprint, 2013.

Obtaining Quadratic Models of Noisy Functions, Preprint, 2013.

Computing <http://mcs.anl.gov/~wild/cnoise>

Merci!