Computational Noise & Noisy Derivatives

Stefan Wild

Argonne National Laboratory
Mathematics and Computer Science Division

January 31, 2014

III. Computational Noise

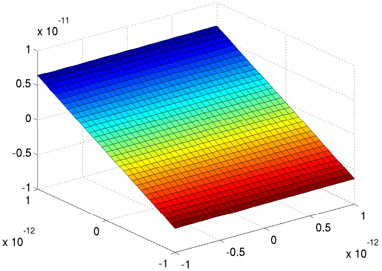
- What is computational noise?
- How can noise be estimated efficiently?
- How does noise affect numerical differentiation?
- How accurate are near-optimal finite-difference estimates?

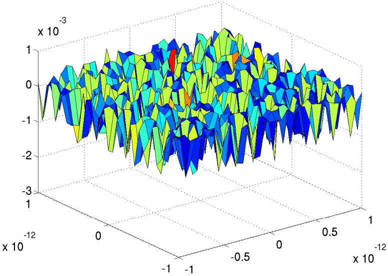
Two Questions To Ask Yourself

1. Do you know how accurate your derivatives are?
2. What do you do with this information?

Noise May Hurt You, Or It May Not

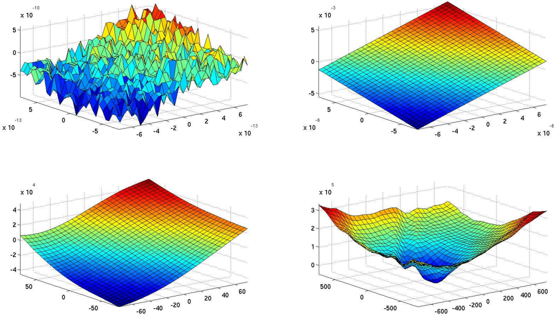
These are the same problem:





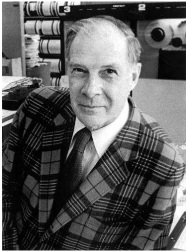
Noise May Hurt You, Or It May Not

So are these:



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.*

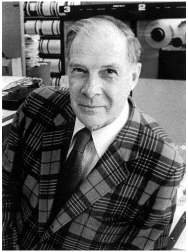
Truncation vs. roundoff *Finite number length leads to roundoff. Finite processes lead to truncation.*

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!

Living In A Finite-Precision World

Floating Point Arithmetic

Commutative:

$$A + B = B + A \qquad \text{and} \qquad A * B = B * A$$

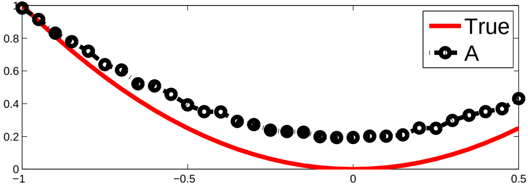
Non-associative:

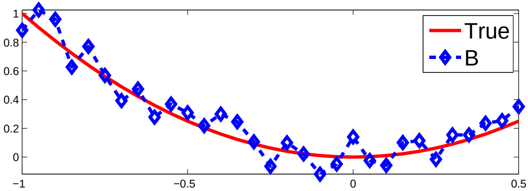
$$A + (B + C) \neq (A + B) + C$$

- This is likely to affect the reproducibility of your calculations in the future (for performance reasons)

Many details $\longrightarrow$ [What Every Computer Scientist Should Know About Floating-Point Arithmetic, Goldberg, 1991]

The Effects of Computational Noise





Noise is not truncation error

$$R_{m+1}(x) = f_a(x) - \sum_{i=0}^{m} P_i(x)$$

and is not roundoff error

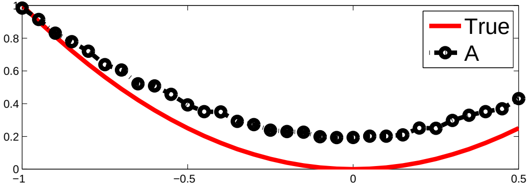
$$f_\infty(x) - f(x)$$

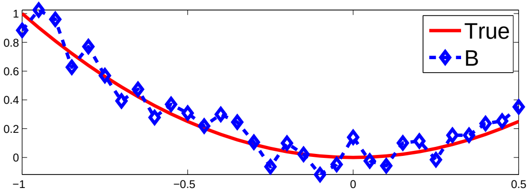
Which do you prefer?

A less noise, more error

B less error, more noise

The Effects of Computational Noise





It matters how noisy your simulation is!

Noise is not truncation error

$$R_{m+1}(x) = f_a(x) - \sum_{i=0}^{m} P_i(x)$$

and is not roundoff error

$$f_\infty(x) - f(x)$$

Which do you prefer?

- A less noise, more error
  - $\rightarrow$ Optimization
  - $\rightarrow$ Sensitivity Analysis
- B less error, more noise
  - $\rightarrow$ Physics

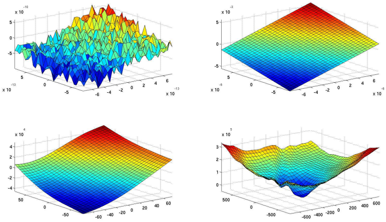
Computational Noise in Deterministic Simulations

Finite precision + finite processes

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

destroy underlying smoothness

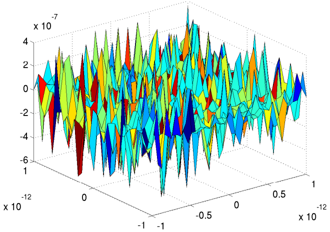
Difference $|f(x) - f(x + Z\omega)|$,



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 $\leftarrow$ only assumption

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

(independent of t)

The k-th Order Difference $\Delta^k f(t)$

$$\Delta^{k+1} f(t) = \Delta^k f(t+h) - \Delta^k f(t), \qquad \Delta^0 f(t) = f(t)$$

$$\Delta^k f(t) = \Delta^k f_s(t) + \Delta^k \varepsilon(t)$$

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

Goal: make h small enough to remove smooth component

Theory Underlying the ECNoise Algorithm

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

1. $\mathrm{E}\left\{\Delta^k \varepsilon(t)\right\} = 0$
2. $\gamma_k \mathrm{E}\left\{[\Delta^k \varepsilon(t)]^2\right\} = \varepsilon_f^2 \qquad \gamma_k = \frac{(k!)^2}{(2k)!}$
3. If f_s is continuous at t, then
$$\lim_{h \to 0} \gamma_k \mathrm{E}\left\{\left[\Delta^k f(t)\right]^2\right\} = \varepsilon_f^2$$
4. If f_s is k-times continuously differentiable at t, then
$$\lim_{h \to 0} \frac{\gamma_k \mathrm{E}\left\{[\Delta^k f(t)]^2\right\} - \varepsilon_f^2}{h^{2k}} = \gamma_k \left[f_s^{(k)}(t)\right]^2$$

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

when the sampling distance h is sufficiently small

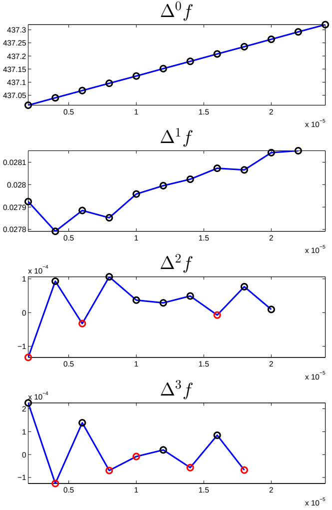
The ECNoise Algorithm

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

◇ Works for deterministic f

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



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

$f(t_i)$	$\Delta f(t_i)$	$\Delta^2 f(t_i)$	$\Delta^3 f(t_i)$	$\Delta^4 f(t_i)$	$\Delta^5 f(t_i)$	$\Delta^6 f(t_i)$
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

Extension to Multivariate $g : \mathbb{R}^n \mapsto \mathbb{R}$

Given base point $x_b \in \mathbb{R}^n$, unit direction $p \in \mathbb{R}^n$, consider

$$f_p(t) = g(x_b + tp), \quad t \geq 0$$

Apply univariate theory

- Directional differences, directional derivatives
- ε_f may now depend on a direction $p \in \mathbb{R}^n$
- `ECnoise` uses $T_{i,0} = f(x_b + ihp)$ with random unit direction $p \in \mathbb{R}^n$

Computational Experience with Stochastic Noise

Validate `ECnoise` and empirical properties of

$$\sigma_k^2 = \frac{\gamma_k}{m+1-k} \sum_{i=0}^{m-k} T_{i,k}^2$$

under known conditions:

- Known noise level ε_f
- Theory directly applies

Target: every estimate within a factor $\eta = 4$ of the mean

Noisy Quadratic, $f(x) = (x^T x)(1 + R), \quad x \in \mathbb{R}^{10}$

Estimate relative noise
$\frac{\sigma_k}{f(x_b)} \approx \sqrt{\mathrm{Var}\left\{R\right\}} = 10^{-3}$

x_b random base point

p 10000 random unit directions

m evaluations

Noisy Quadratic, $f(x) = (x^T x)(1 + R), \quad x \in \mathbb{R}^{10}$

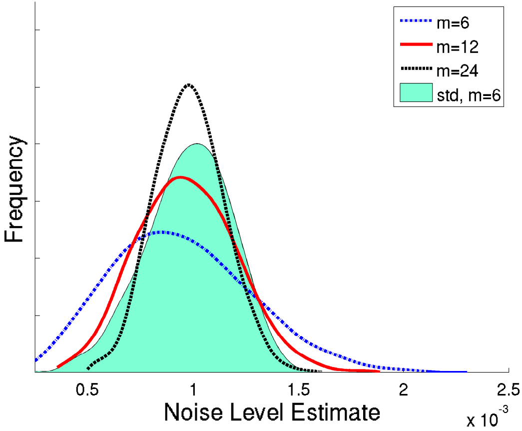
$$R \sim \text{Uniform}\left[-\sqrt{3}\cdot 10^{-3}, \sqrt{3}\cdot 10^{-3}\right]$$

Estimate relative noise

$\frac{\sigma_k}{f(x_b)} \approx \sqrt{\mathrm{Var}\{R\}} = 10^{-3}$

- x_b random base point
- p 10000 random unit directions
- m evaluations

99.2% within a factor $\eta = 4$ for $m = 6$



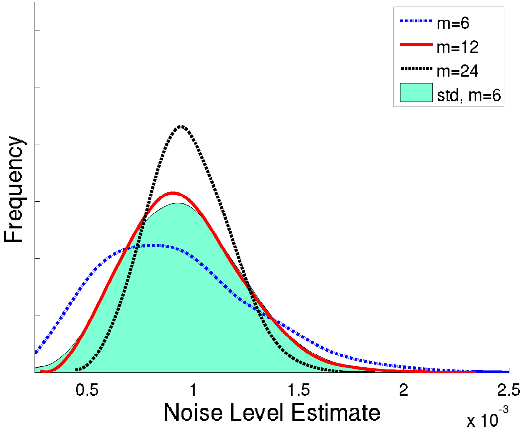
Noisy Quadratic, $f(x) = (x^T x)(1 + R), \quad x \in \mathbb{R}^{10}$

$R \sim$Normal$\left(0, 10^{-6}\right)$

Estimate relative noise
$\frac{\sigma_k}{f(x_b)} \approx \sqrt{\mathrm{Var}\left\{R\right\}} = 10^{-3}$

- x_b random base point
- p 10000 random unit directions
- m evaluations

98.9% within a factor $\eta = 4$ for $m = 6$



MC Finance Example with Higher Order Derivatives

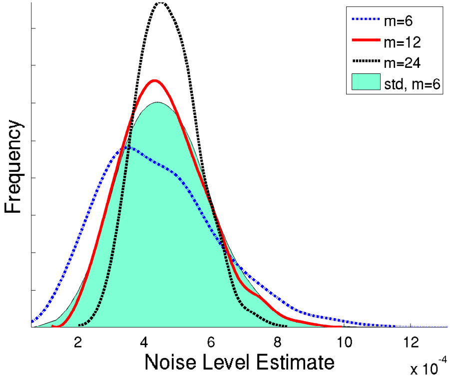
Today's value of a \$1 payment n years from now rates [Caflisch]:

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

10000 MC integrations (directions p) with

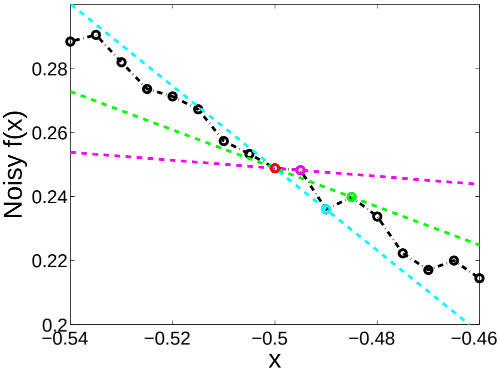
- $n = 3$ years, $x_b = [.1,\ .1,\ .1]$
- $tol = 5000$ standard normal random variables
- no variance reduction

99.6% within a factor 4 for $m = 6$



Finite Differences Sensitive to Choice of h

$$\frac{f(t_0 + h) - f(t_0)}{h} \approx f_s'(t_0)$$



Noisy Forward Differences

Minimize $\quad \mathrm{E}\left\{\mathcal{E}(h)\right\} = \mathrm{E}\left\{\left(\frac{f(t_0+h)-f(t_0)}{h} - f_s'(t_0)\right)^2\right\}$

Our h will depend on

- Loose estimate of noise
- Loose estimate of $|f''|$ **!**
- 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]]

Optimal Forward Difference Parameter h

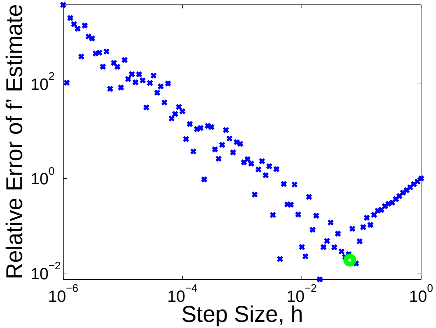
$$\frac{1}{4}\mu_L^2 h^2 + 2\frac{\varepsilon_f^2}{h^2} \le \mathrm{E}\left\{\mathcal{E}(h)\right\} \le \frac{1}{4}\mu_M^2 h^2 + 2\frac{\varepsilon_f^2}{h^2}$$

$h \downarrow$ Variance (noise) dominates

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

For h_0 sufficiently large

1. Upper bound minimized by $h_M = 8^{1/4}\left(\frac{\varepsilon_f}{\mu_M}\right)^{1/2}$
2. When $\mu_L > 0$, h_M is near-optimal:



$$\mathrm{E}\left\{\mathcal{E}(h_M)\right\} = \sqrt{2}\mu_M\varepsilon_f \le \left(\frac{\mu_M}{\mu_L}\right) \min_{0\le h\le h_0} \mathrm{E}\left\{\mathcal{E}(h)\right\}.$$

Alternative FD Step Sizes

[Gill, Murray, Saunders, Wright; 1983]

Given uniform bound on roundoff error,

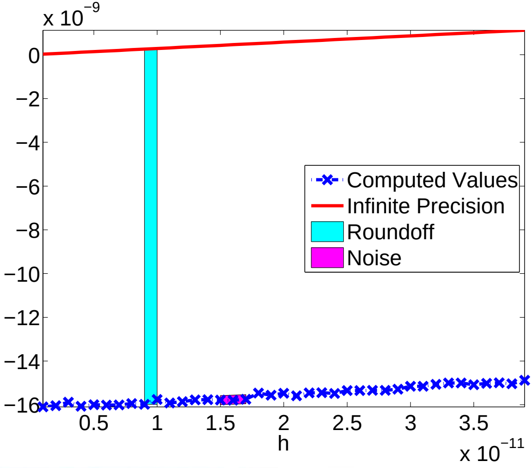
$$|f(t) - f_\infty(t)| \leq \varepsilon_A \qquad t \in I,$$

Minimizer of (upper bound on) l_1 error is

$$h_A = 2\left(\frac{\varepsilon_A}{\mu_M}\right)^{1/2}$$

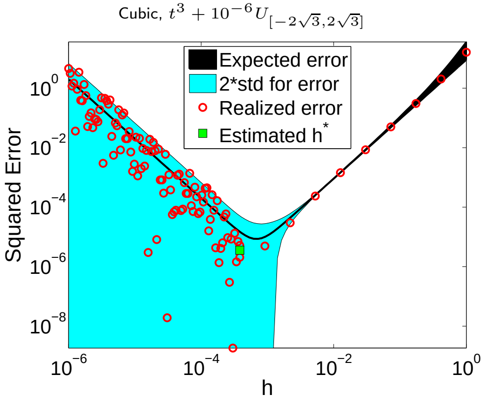
Assumes:

- $h_A \leq h_0$
- Estimate of ε_A available



Stochastic Examples

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



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

Expected error and uncertainty regions predicted by the theory

Extension: Central Differences

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

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

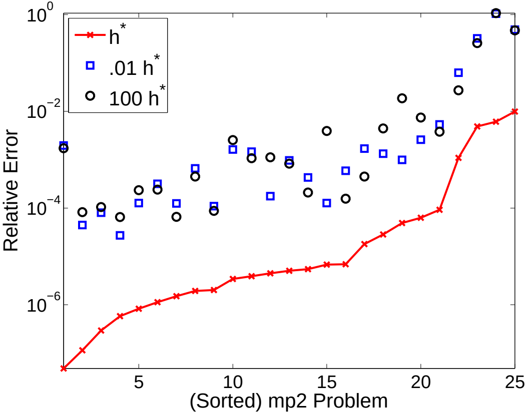
Second derivatives, $\frac{f(t_0+h)-2f(t_0)+f(t_0-h)}{h^2}$

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

Ex.- Highly Nonlinear MINPACK-2 Problems

25 problems, $n \leq 64 \cdot 10^4$

- Accurate estimates obtained even when f'' not constant



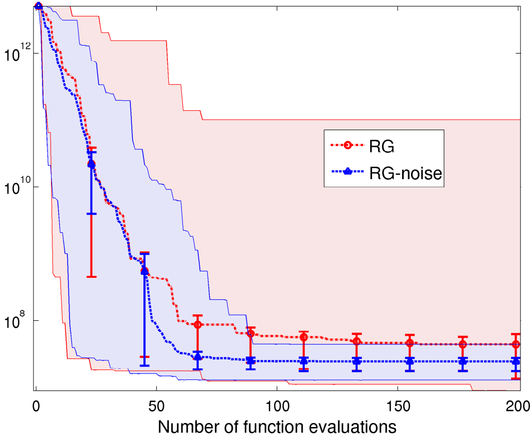
Compared with hand-coded derivative

Using the Noise in Nesterov's Random Gradient Method

General RG iteration

1. Generate direction d_k
2. Evaluate gradient-free oracle $g(x_k; h_k) = \frac{f(x_k + h_k\, d_k) - f(x_k)}{h} d_k$
3. Compute $x_{k+1} = x_k - \delta_k g(x_k; h_k)$, evaluate $f(x_{k+1})$

bicgstab quadratic: tol$= 10^{-2}$, $\frac{\varepsilon_f}{|f|} \approx$ 5e-3



Summary: How Loud Are Your Functions?

- Computational noise complicates analysis of real-world functions, worst-case bounds overly pessimistic
- 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!