Derivative-Free Optimization of Noisy Functions via Quasi-Newton Methods: Experiments and Theory

Richard Byrd
University of Colorado, Boulder

Albert Berahas
Northwestern University

Jorge Nocedal
Northwestern University

Huatulco, Jan 2018

My thanks to the organizers.

Saludos y Gracias a Don Goldfarb

Numerical Results

We compare our Finite Difference L-BFGS Method (FD-LM) to Model interpolation trust region method (MB) of Conn, Scheinberg, Vicente.

Their method, DFOtr, is:
a simple implementation not designed for fast execution
does not include a geometry phase

Our goal is not to determine which method "wins". Rather

1. Show that the FD-LM method is robust
2. Show that FD-LM is not wasteful in function evaluations

Adaptive Finite Difference L-BFGS Method

Estimate noise ϵ_f

Compute h by forward or central differences [(4-8) function evaluations]

Compute g_k

While convergence test not satisfied:

$d = -H_k g_k$ [L-BFGS procedure]

$(x_+, f_+, flag) =$ LineSearch$(x_k, f_k, g_k, d_k, f_s)$

IF flag=1 [line search failed]

$(\mathrm{x}_+, f_+, h) =$ Recovery$(x_k, f_k, g_k, d_k, max_{iter})$

endif

$x_{k+1} = x_+, \; f_{k+1} = f_+$

Compute g_{k+1} [finite differences using h]

$s_k = x_{k+1} - x_k, \; y_k = g_{k+1} - g_k$

Discard (s_k, y_k) if $s_k^T y_k \le 0$

$k = k + 1$

endwhile

Test problems

Test 58 problems from Hock-Schittkowski collection
Dimensions ranging from n=2 to n=100

4 different types of noise:
- Stochastic additive + multiplicative
- Deterministic additive + multiplicative
- 6 levels of noise: $10^{-8} - 10^{-1}$

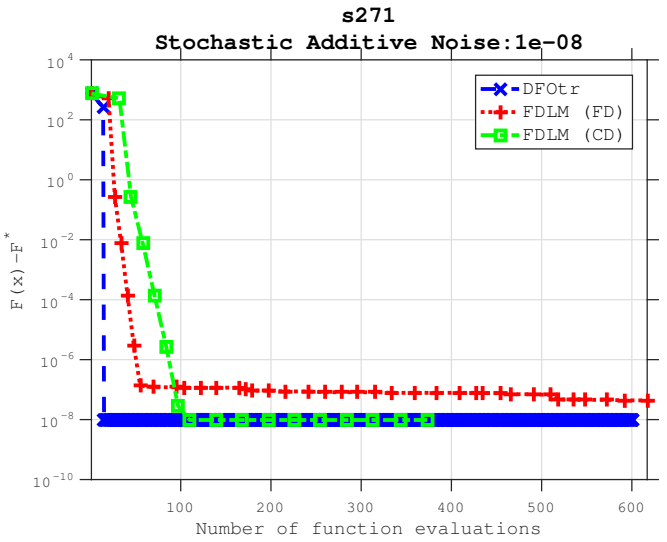
58 x 4 x 6 = 1392 runs

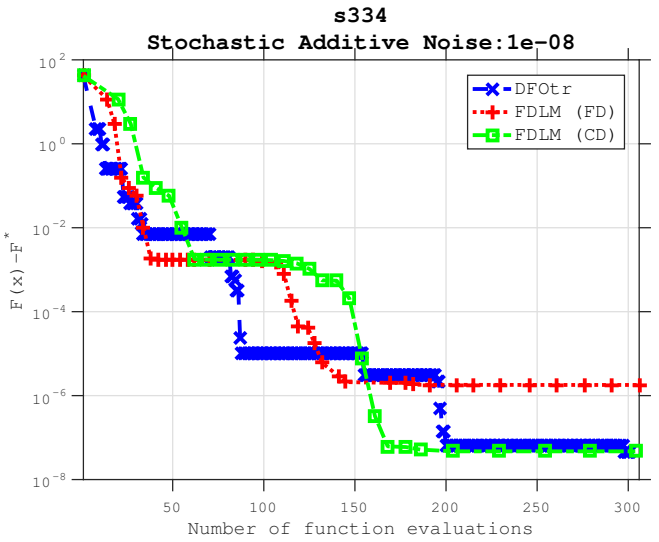
Plotting $f(x_k) - \phi^*$ vs no. of f evaluations

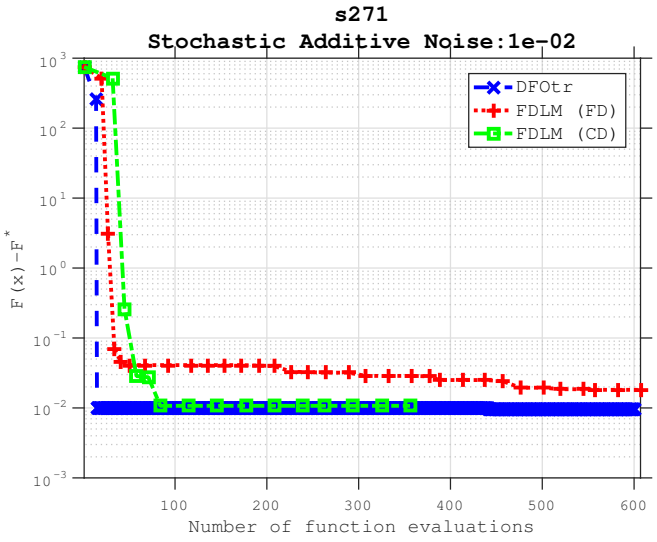
We show results for 4 representative problems

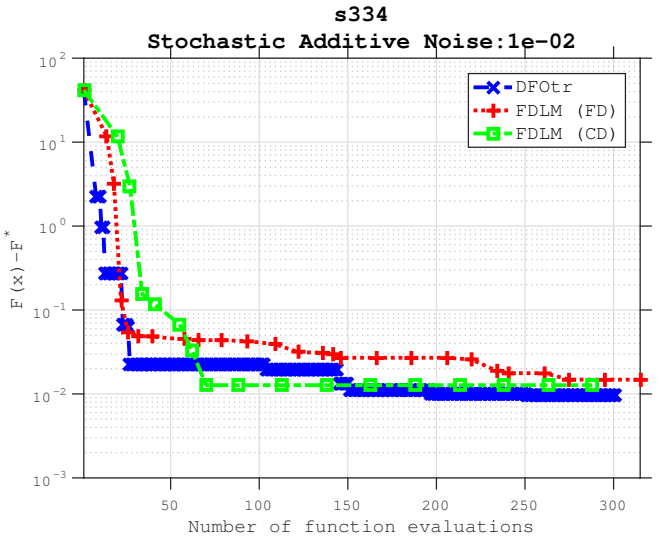
Numerical Results – Stochastic Additive Noise

$$f(x) = \phi(x) + \epsilon(x) \qquad \epsilon(x) \sim U(-\xi, \xi) \quad \xi \in [10^{-8}, \ldots, 10^{-1}]$$



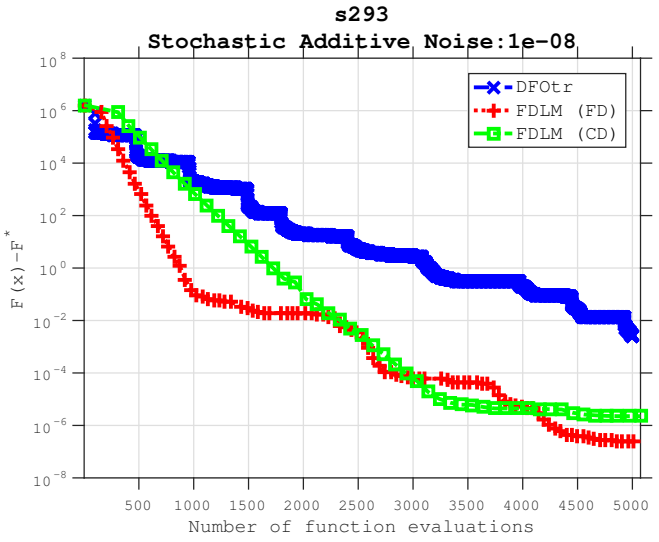


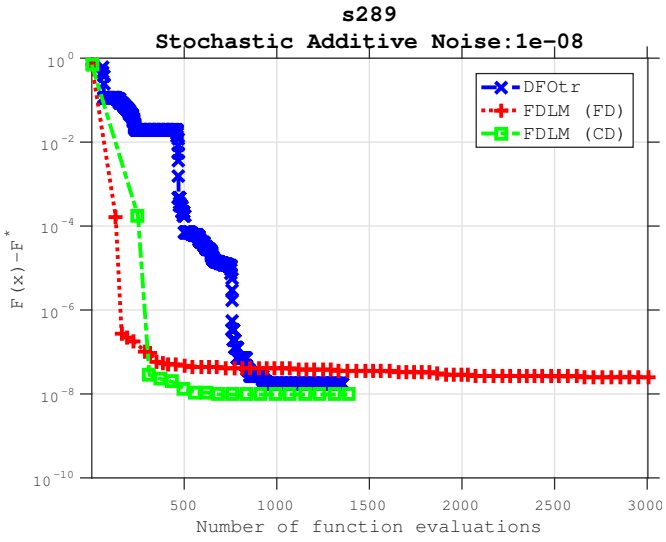


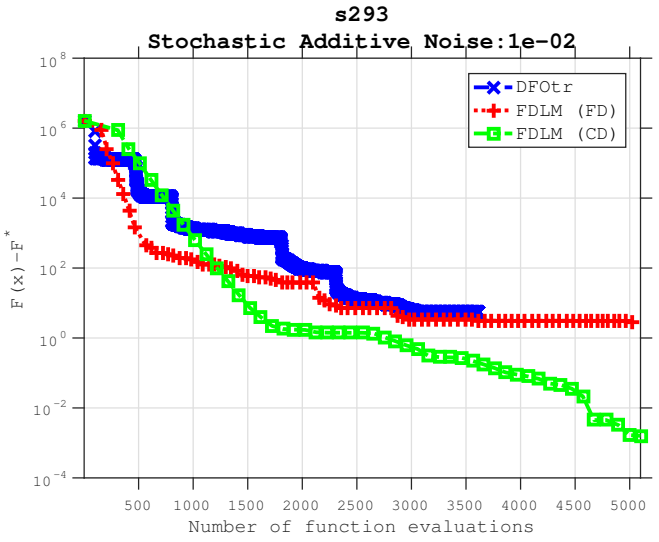


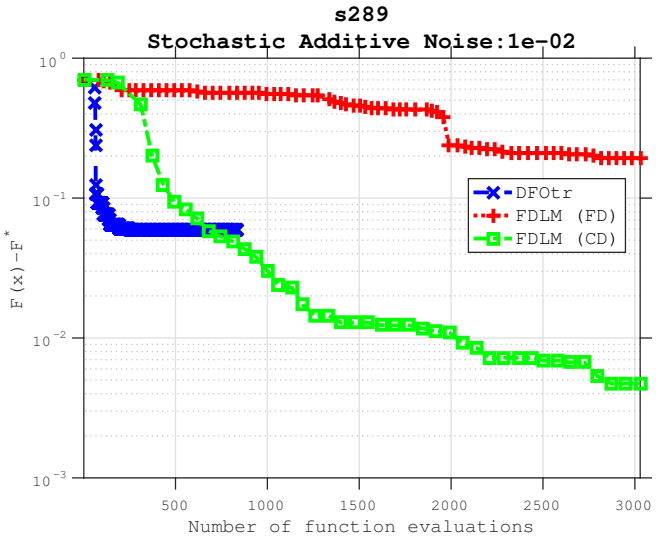
Numerical Results – Stochastic Additive Noise (continued)

$$f(x) = \phi(x) + \epsilon(x) \qquad \epsilon(x) \sim U(-\xi, \xi) \quad \xi \in [10^{-8}, \ldots, 10^{-1}]$$



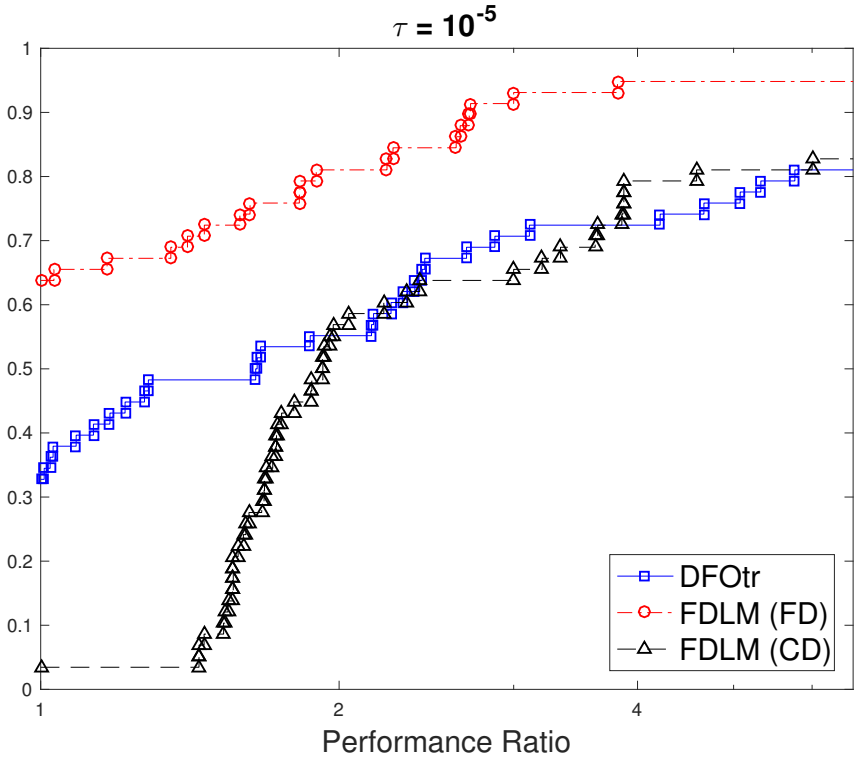




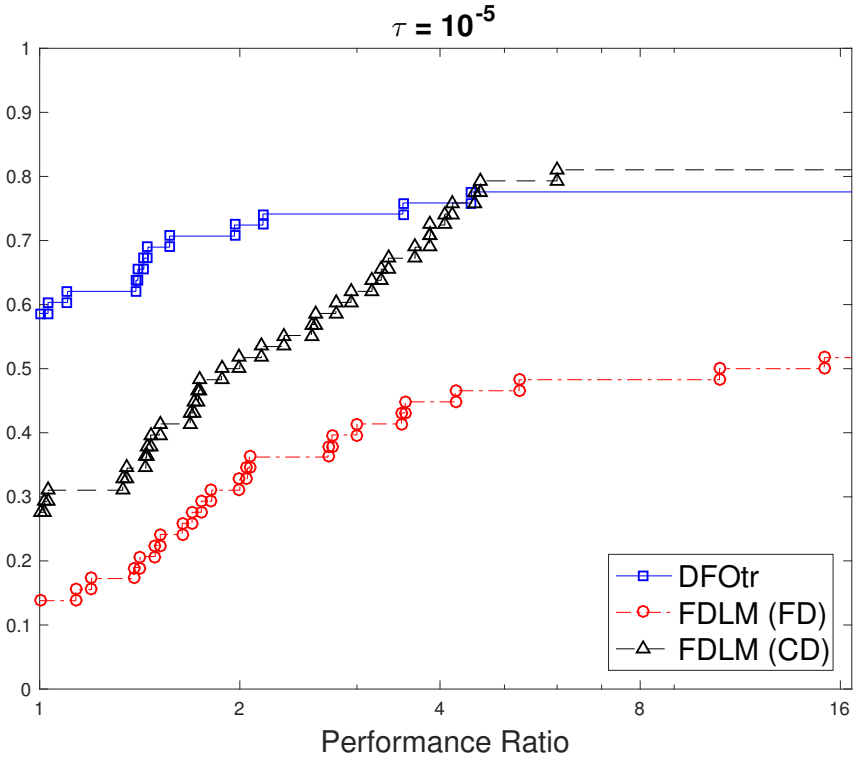


Numerical Results – Stochastic Additive Noise – Performance Profiles

$Noise = 10^{-8}$

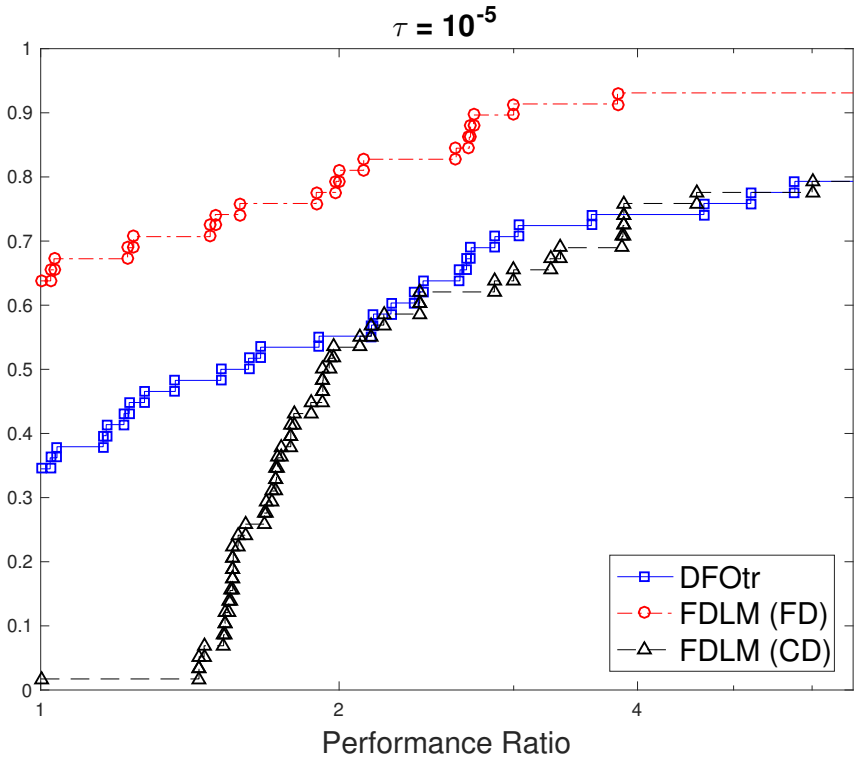


$Noise = 10^{-2}$

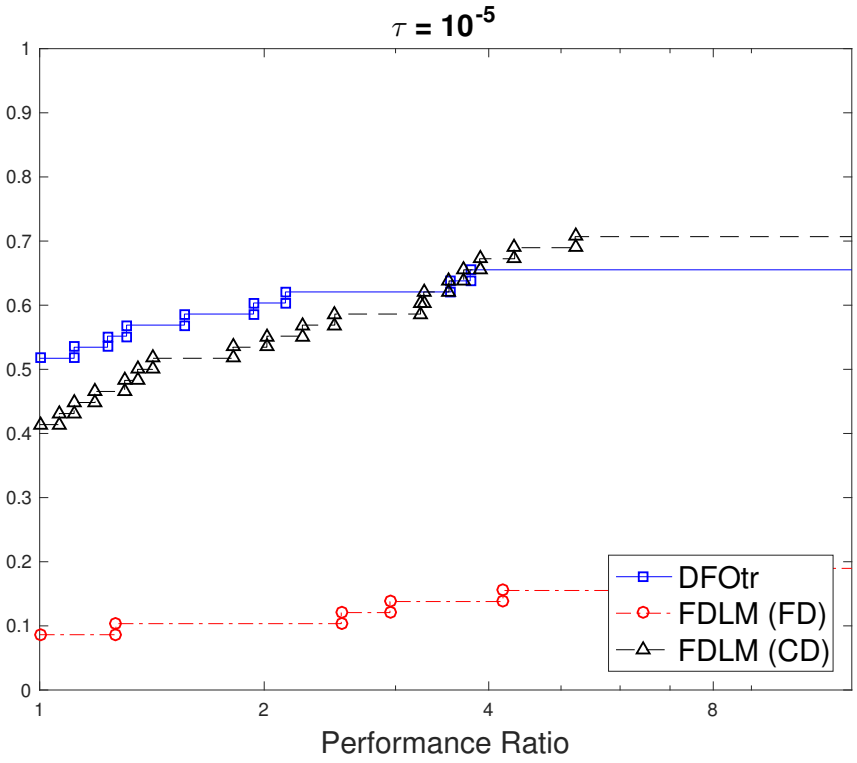


Numerical Results – Stochastic Multiplicative Noise – Performance Profiles

$Noise = 10^{-8}$



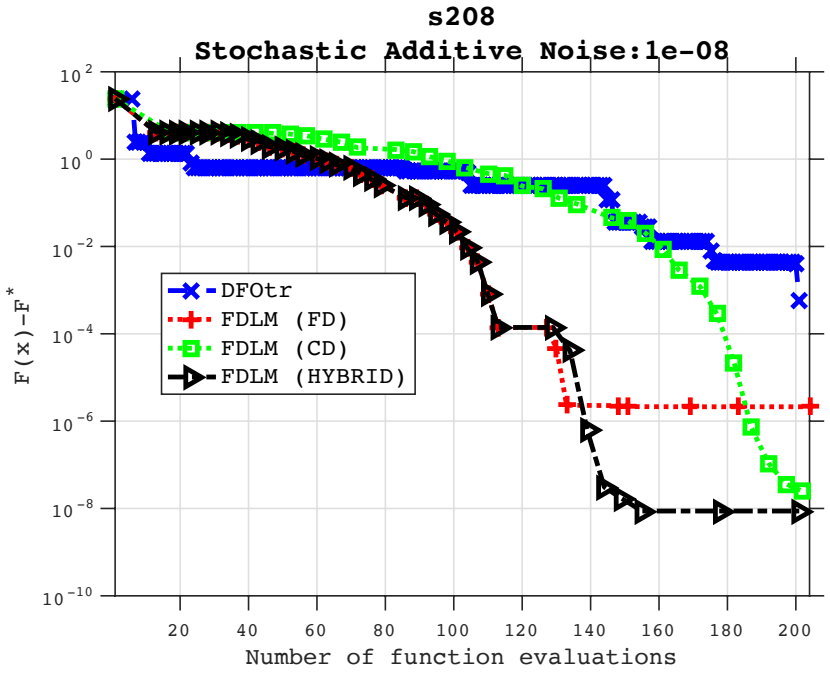
$Noise = 10^{-2}$

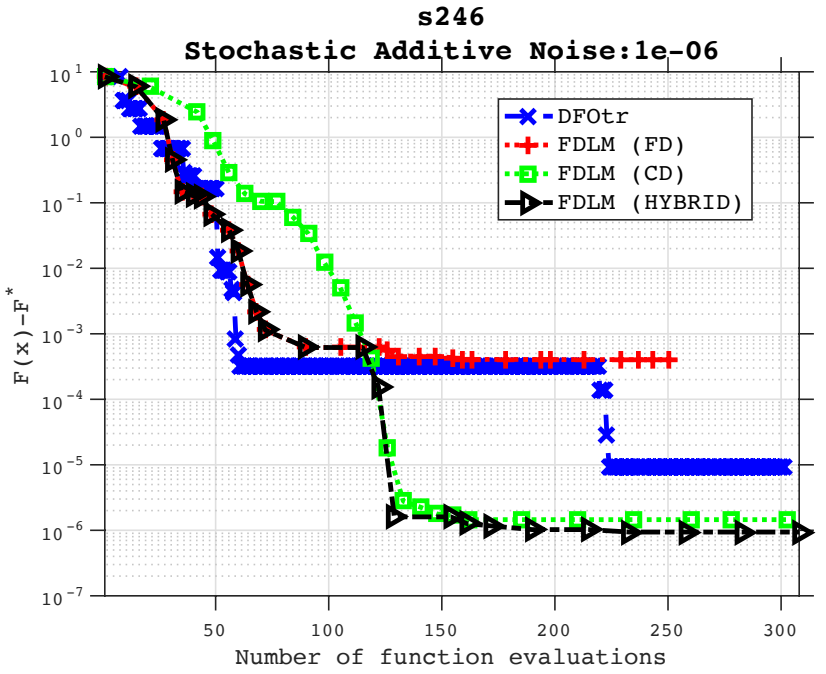


Numerical Results – Hybrid Method – Recovery Mechanism

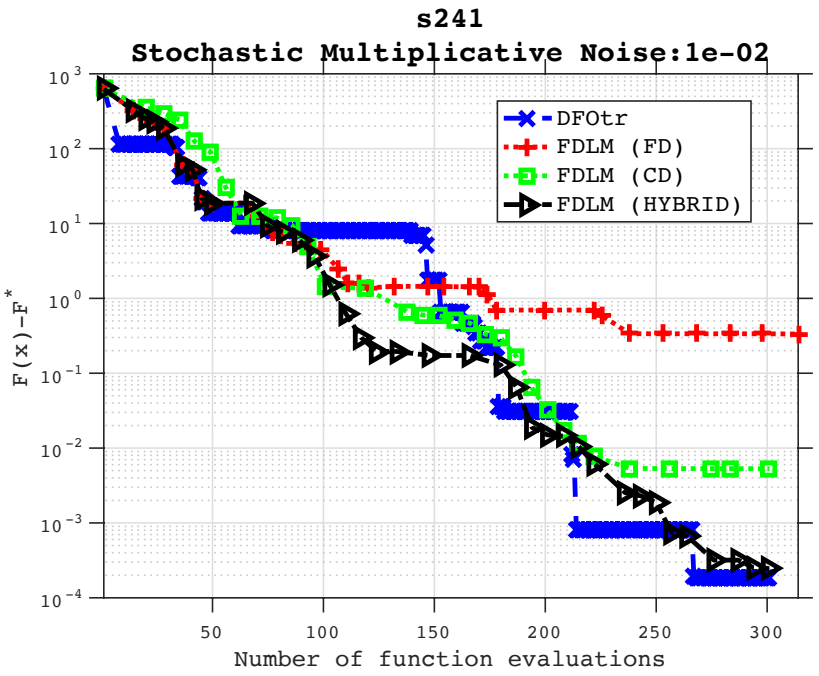
- As Jorge mentioned in Part I, our algorithm has a recovery mechanism
- This procedure is very important for the stable performance of the method
- Principle recovery mechanism is to re-estimate h
- HYBRID METHOD: If h is acceptable, then we switch from Forward to Central differences

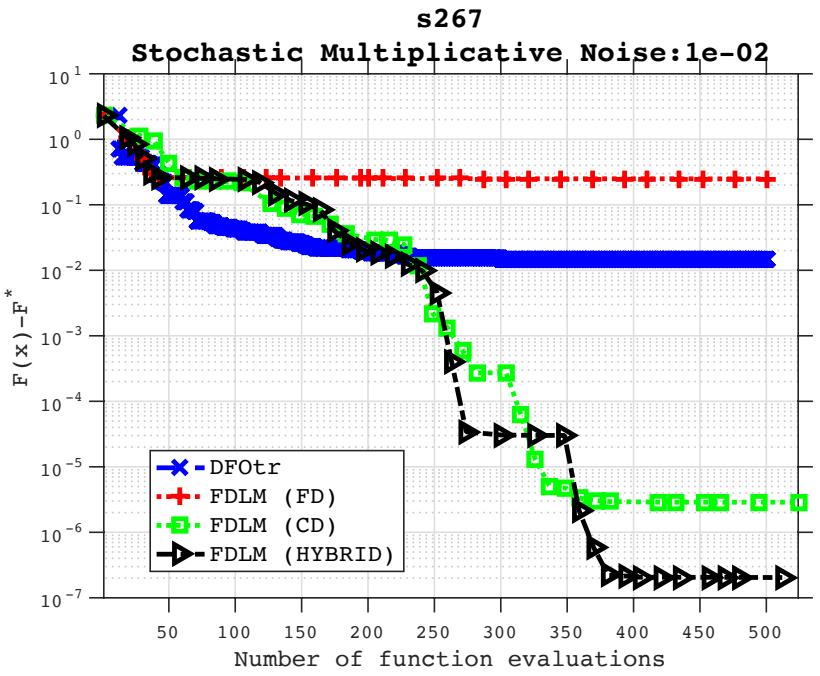
Numerical Results – Hybrid FC Method – Stochastic Additive Noise





Numerical Results – Hybrid Method FC – Stochastic Multiplicative Noise





Numerical Results – Conclusions

- Both methods are fairly reliable
- FD-LM method not wasteful in terms of function evaluations
- No method dominates
- Central difference appears to be more reliable, but is twice as expensive per iteration
- Hybrid approach shows promise

Convergence analysis

1. What can we prove about the algorithm proposed here?
2. We first note that there is a theory for the Implicit Filtering Method of Kelley – which is a finite difference BFGS method
   - He establishes deterministic convergence guarantees to the solution
   - Possible because it is assumed that noise can be diminished as needed at every iteration
   - Similar to results on Sampling methods for stochastic objctives
3. In our analysis we assume that noise does not go to zero
   - We prove convergence to a neighborhood of the solution whose radius depends on the noise level in the function
   - Results of this type were pioneered by Nedic-Bertsekas for incremental gradient method with constant steplengths
4. We prove two sets of results for strongly convex functions
   - Fixed steplength
   - Armijo line search
5. Up to now, little analysis of line search with noise

Discussion

1. The algorithm proposed here is complex, particularly if the recovery mechanism is included
2. The effect that noisy function evaluations and finite difference gradient approximations have on the line search are difficult to analyze
3. In fact: the study of stochastic line searches is one of our current research projects
4. How should results be stated:
   - in expectation?
   - in probability?
   - what assumptions on the noise are realistic?
   - some results in the literature assume the true function value $\phi(x)$ is available

- This field is emerging

Context of our analysis

1. We will bypass these thorny issues by assuming that
   - Noise in the function and gradient are bounded

$$\| \epsilon(x) \| \leq C_f \qquad \| e(x) \| \leq C_g$$

   - And consider a general gradient method with errors

$$x_{k+1} = x_k - \alpha_k H_k g_k$$

   - g_k is any approximation to the gradient
   - could stand for a finite difference approximation or some other
   - treatment is general
   - to highlight the novel aspects of this analysis we assume $H_k = I$

Fixed Steplength Analysis

Recall $f(x) = \phi(x) + \epsilon(x)$

Define $g_k = \nabla\phi(x_k) + e(x_k)$

Iteration $x_{k+1} = x_k - \alpha g_k$

Assume $\mu I \prec \nabla^2\phi(x_k) \prec LI$

$\| e(x) \| \leq C_g$

Theorem. If $\alpha < 1/L$ then for all k

$$\phi(x_{k+1} - \phi^N) \leq (1 - \alpha\mu)[\phi(x_k) - \phi^N]$$

$\phi^N \equiv \phi^* + \dfrac{C_g^2}{2\mu}$ best possible objective value

Therefore,

$$\phi_k - \phi^* \leq (1 - \alpha\mu)^k(\phi_0 - \phi^N) + \frac{C_g^2}{2\mu}$$

Idea behind the proof

$$\phi(x_{k+1}) \leq \phi(x_k) - \alpha \nabla \phi(x_k)^T \left(\nabla \phi(x_k) + e(x_k)\right) + \frac{\alpha^2 L}{2} \| \nabla \phi(x_k) + e(x_k) \|^2$$

$$\phi(x_{k+1}) \leq \phi(x_k) - \frac{\alpha}{2} \| \nabla \phi(x_k) \|^2 + \frac{\alpha}{2} \| e(x_k) \|^2$$

Line Search

Our algorithm uses a line search
Move away from fixed steplengths and exploit the power of line searches
Very little work on noisy line searches
How should sufficient decrease be defined?

Introduce new Armijo condition:

$$f(x_k + \alpha d_k) \leq f(x_k) + c_1 \alpha g_k^T d_k + \epsilon_A$$

where $\alpha = \max\{1, \tau, \tau^2, \ldots\}$
and $\epsilon_A > 2C_f$

Line Search Analysis

New Armijo condition:

$$f(x_k + \alpha d_k) \leq f(x_k) + c_1 \alpha g_k^T d_k + \epsilon_A$$

where $\alpha = \max\{1, \tau, \tau^2, \ldots\}$
and $\epsilon_A > 2C_f$

Because of relaxation term Armijo is always satisfied for alpha <<1.
But how long will the step be?

Consider 2 sets of iterates:

Case 1: Gradient error is small relative to gradient.
Step of 1/L is accepted, and good progress is made.

Case 2: Gradient error is large relative to gradient.
Step could be poor, but size of step is only of order C_g

Line Search Analysis

Iteration $x_{k+1} = x_k - \alpha_k g_k$

Assume $\mu I \prec \nabla^2 \phi(x_k) \prec LI$ $\quad\quad$ $\| e(x) \| \leq C_g$ and $\| \epsilon(x) \| \leq C_f$

Theorem: Above algorithm with relaxed Armijo with $c_1 < 1/2$ gives

$$\phi(x_{k+1}) - \phi^N \leq \rho[\phi(x_k) - \phi^N]$$

where $\rho = 1 - \dfrac{2\mu c_1 \tau (1-\beta)^2}{L}$

and $\phi^N = \phi_* + \dfrac{1}{1-\rho}[\dfrac{c_1 \tau (1-\beta)^2 C_g^2}{L\beta^2} + \epsilon_A + 2C_f]$

Here β is a free parameter in $(0, \dfrac{1-2c_1}{1+2c_1}]$

THANK YOU.