Comparing Derivative Free Methods

Liyuan Cao, Albert Barahas, Katya Scheinberg

Lehigh University

INFORMS Annual Meeting 2018

Derivative Free Optimization

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

Problem: black-box function, derivative not available, maybe noisy

Assumption: Function evaluation is very expensive.

Finite Difference Methods

forward finite difference

$$g_i(x) = \frac{f(x + \epsilon e_i) - f(x)}{\epsilon} \quad \forall i = 1, \ldots, n$$

central finite difference

$$g_i(x) = \frac{f(x + \epsilon e_i) - f(x - \epsilon e_i)}{2\epsilon} \quad \forall i = 1, \ldots, n$$

Derivative Free Optimization

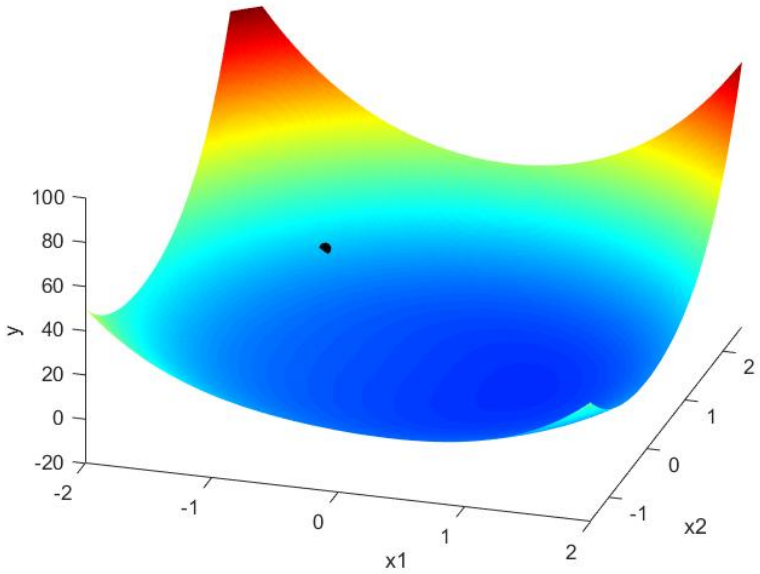
$$\min_{x \in \mathbb{R}^n} f(x) \tag{2}$$

Problem: black-box function, derivative not available, maybe noisy

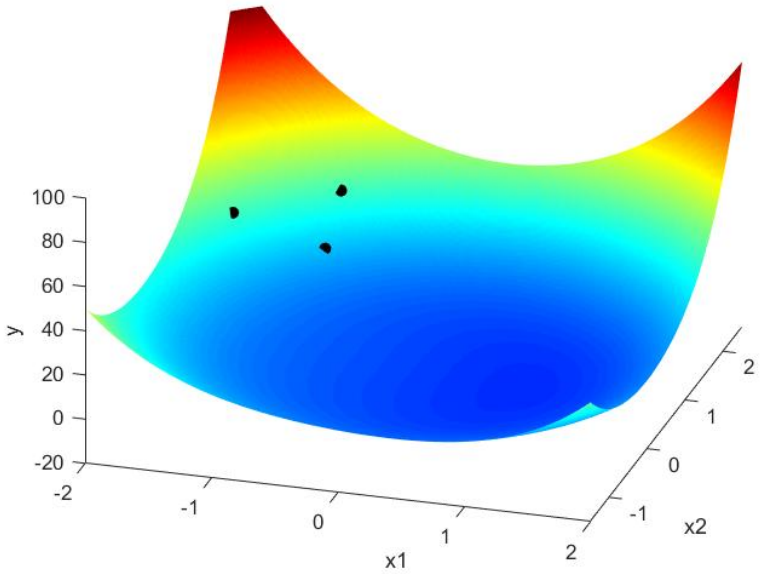
Assumption: Function evaluation is very expensive.

- finite difference methods (with steepest descent, L-BFGS, or . . .)
- model-base method (e.g. DFOTR)

DFOTR

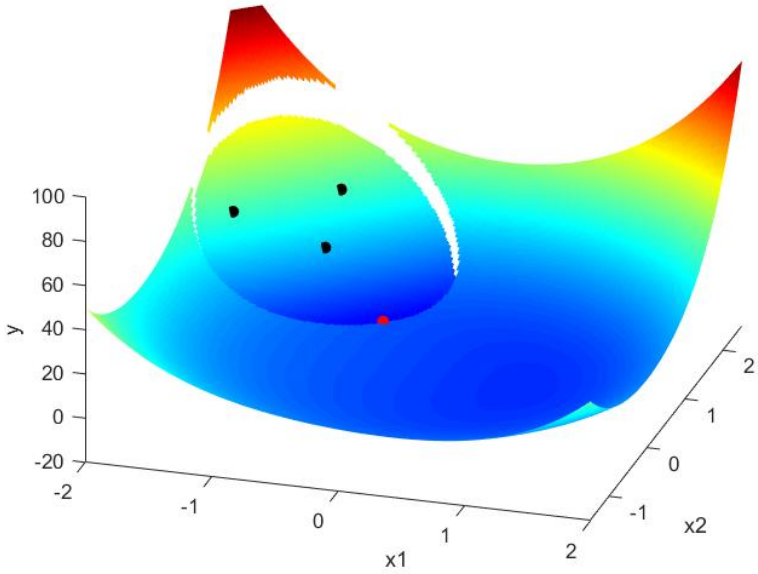


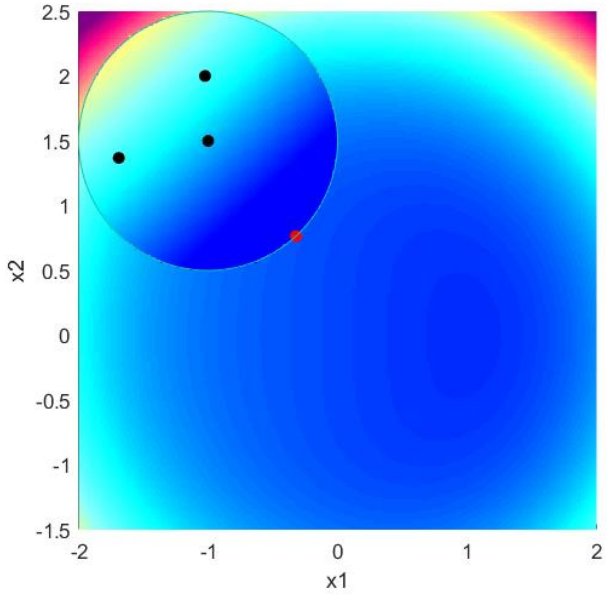
(a) starting point



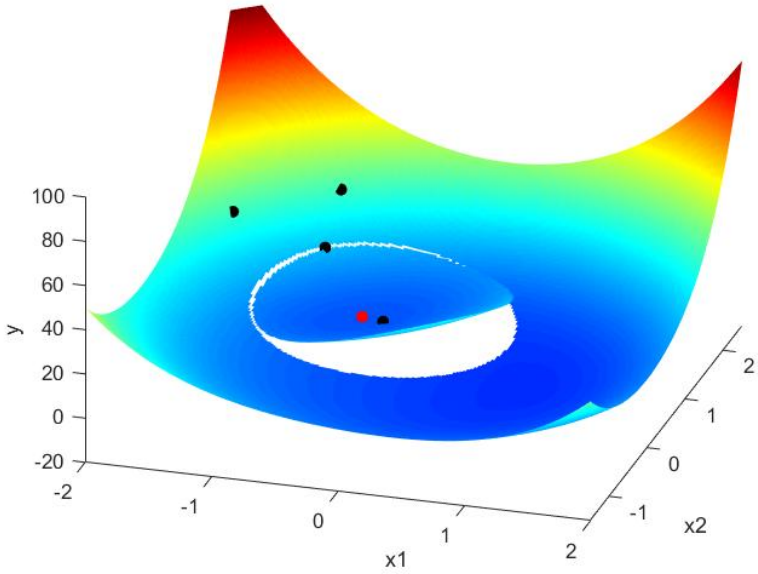
(b) initial sampling

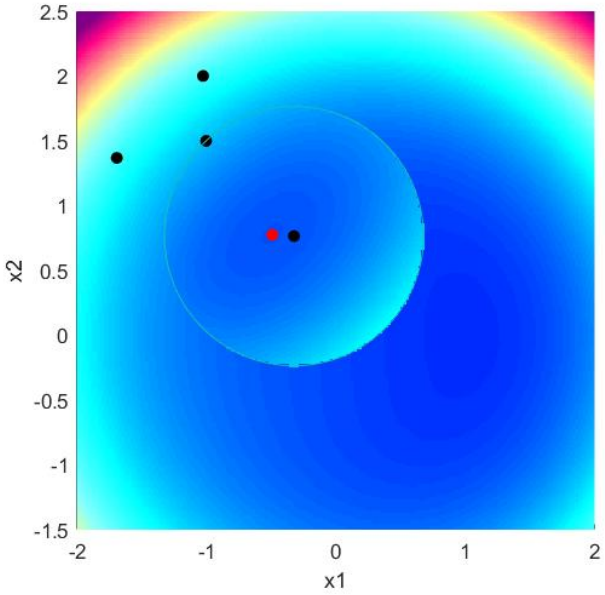
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

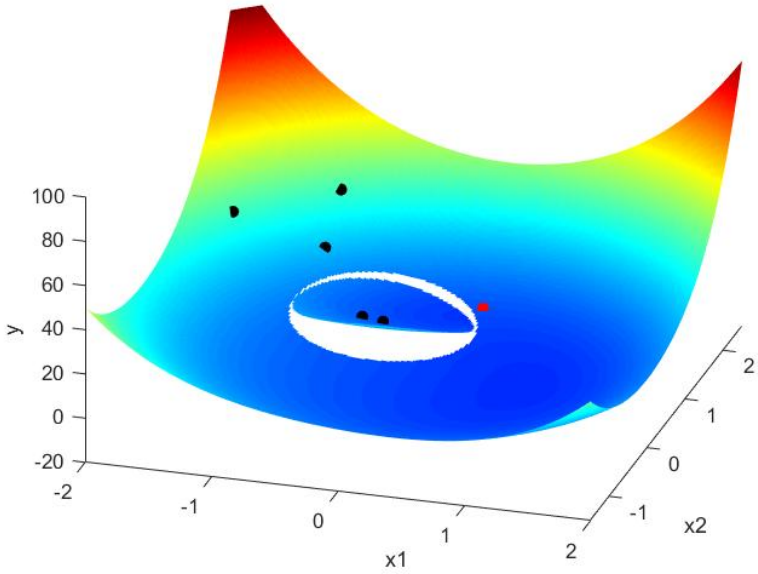

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

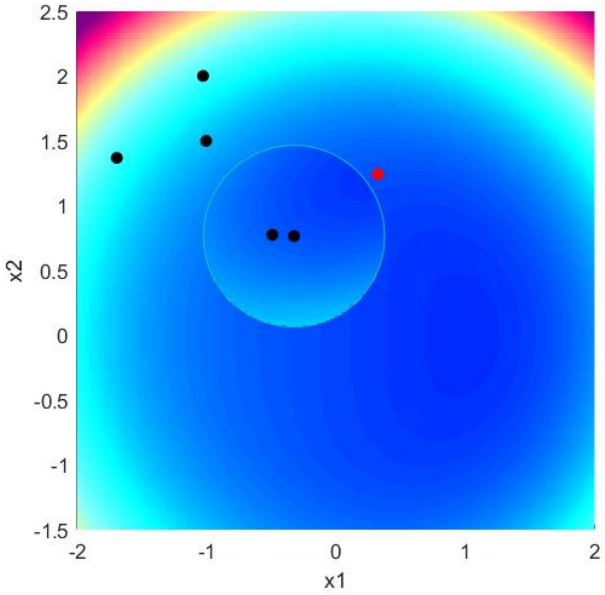
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

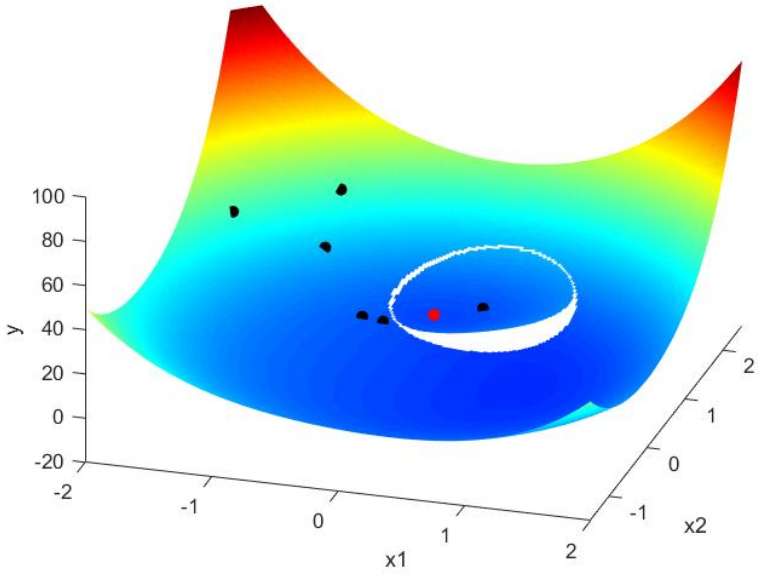

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

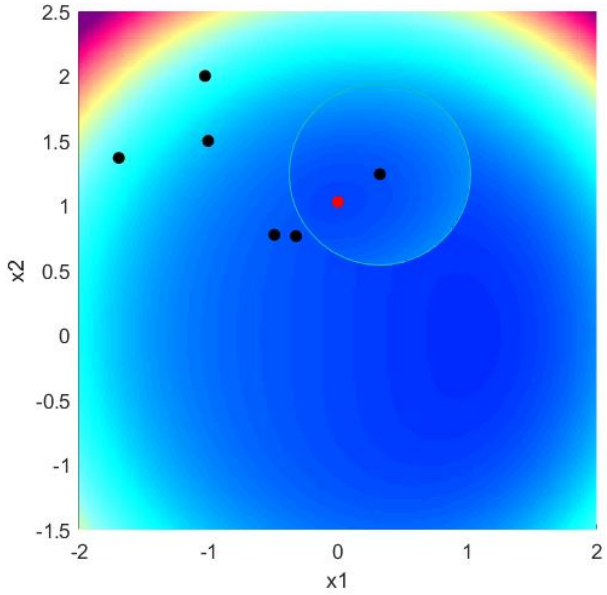
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

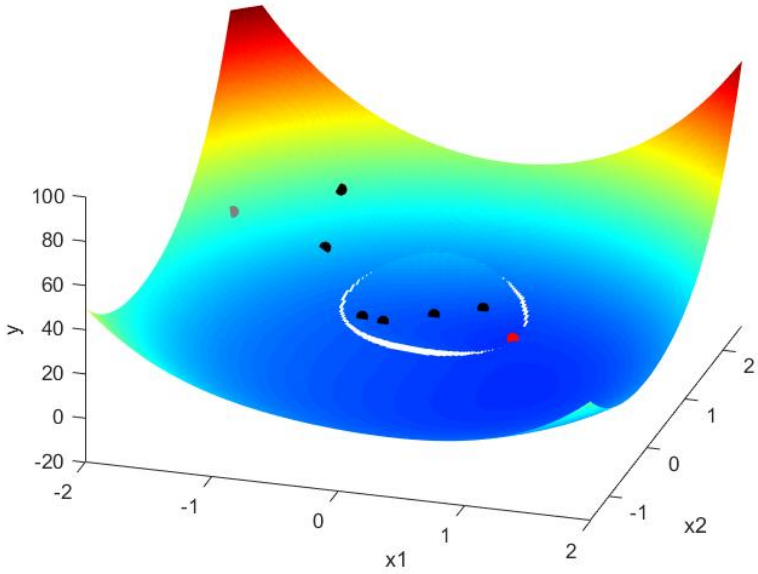

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

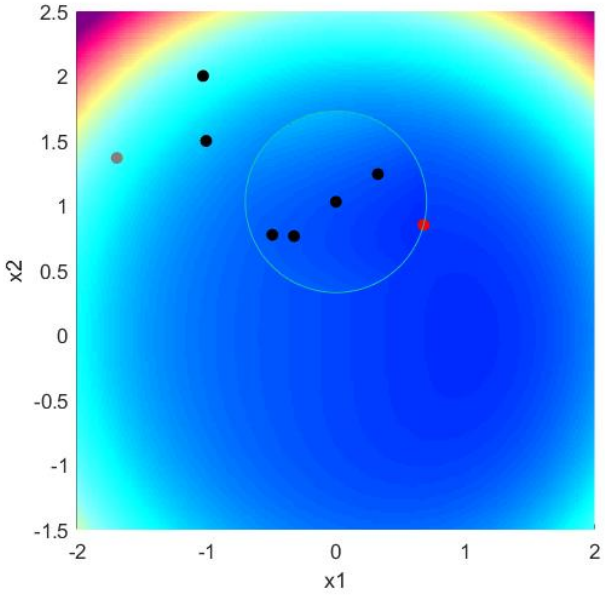
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

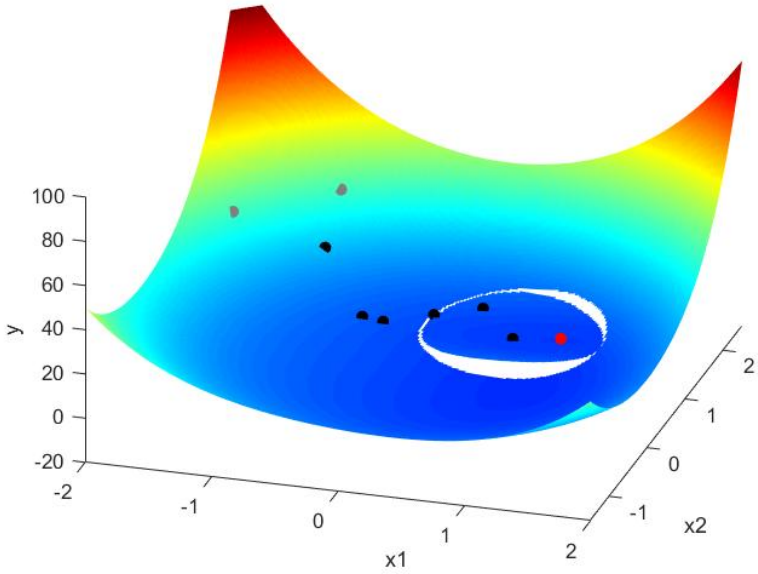

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

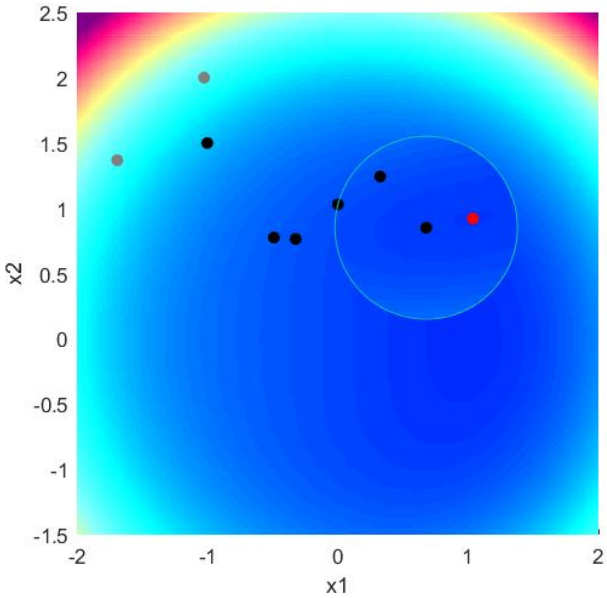
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

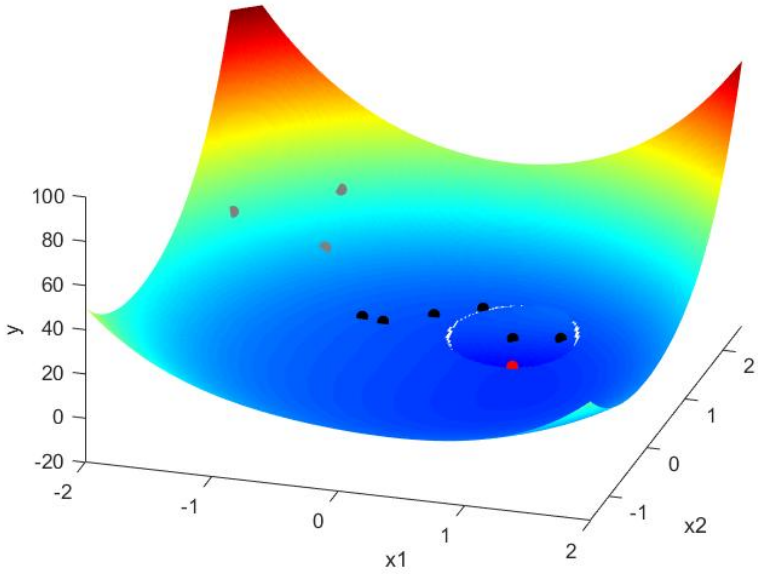

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

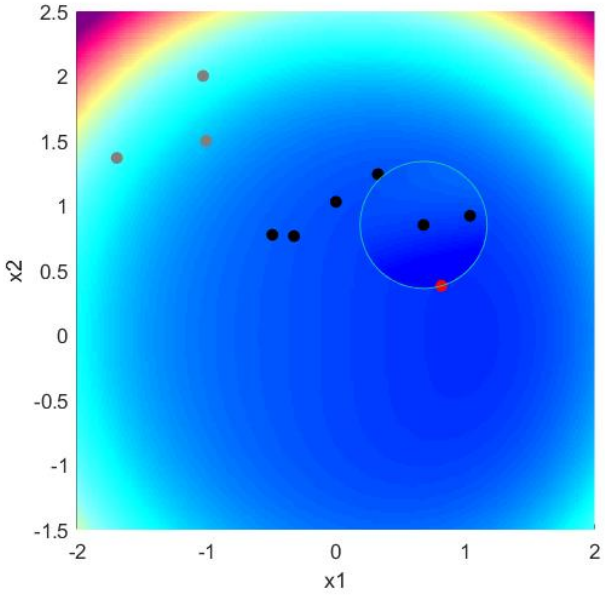
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

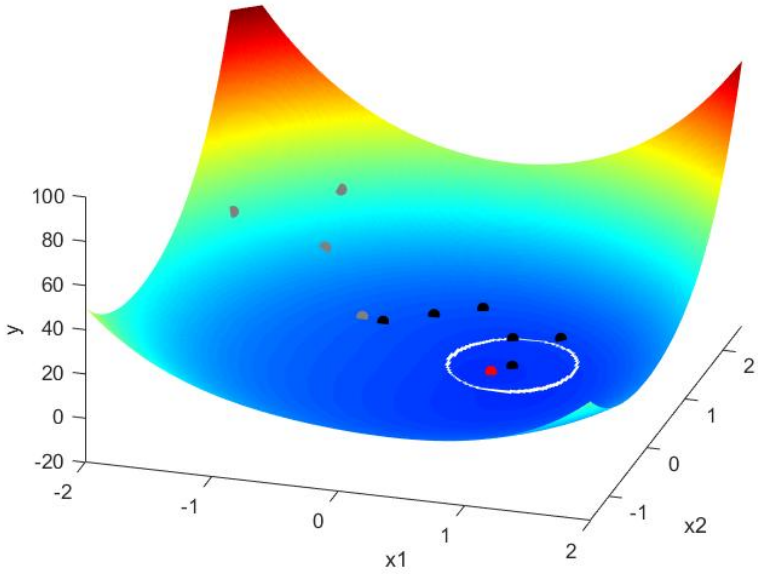

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

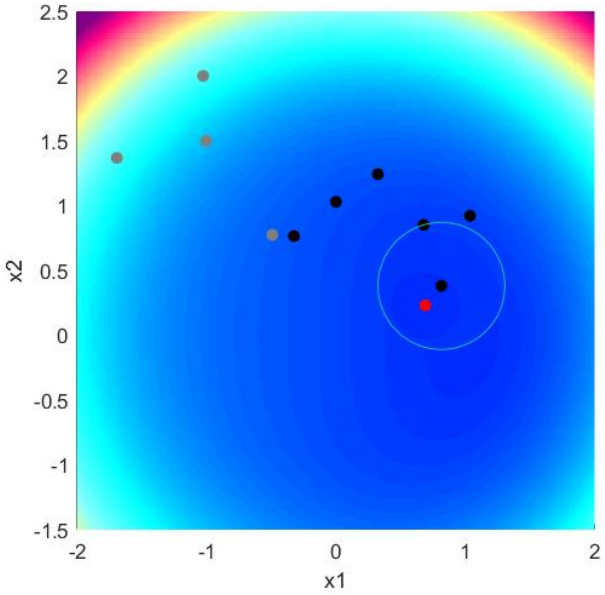
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

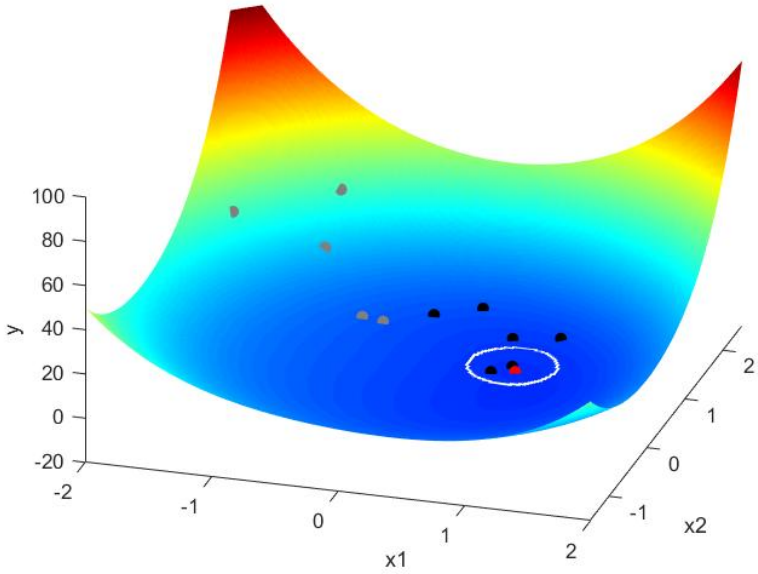

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

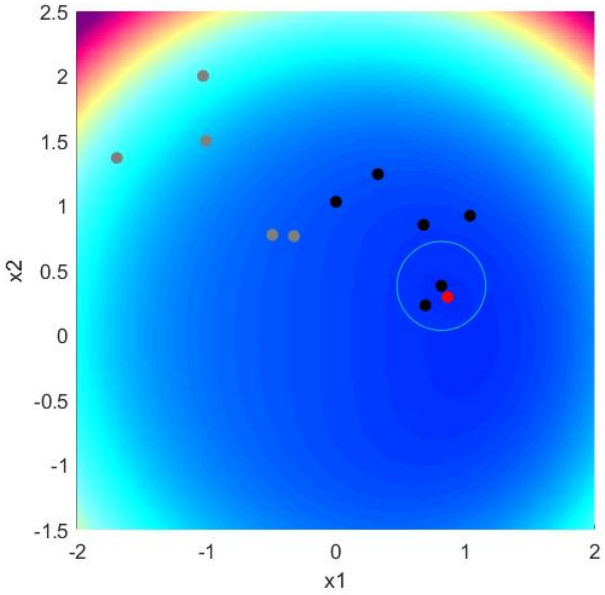
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

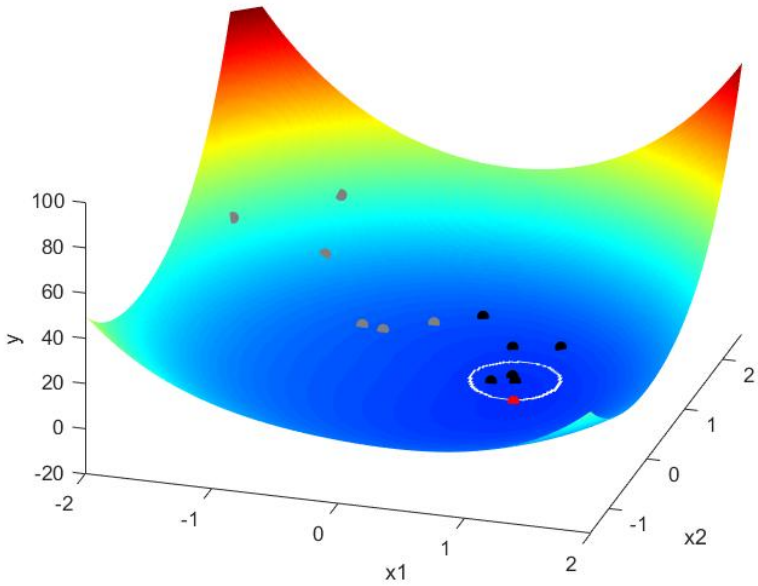

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

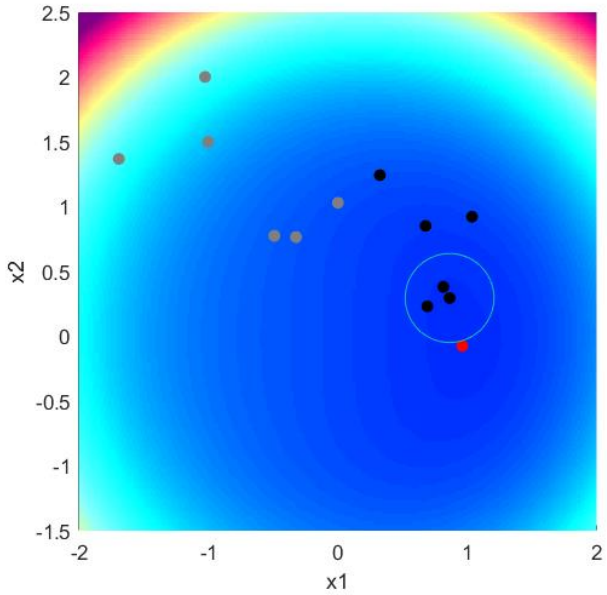
DFOTR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2

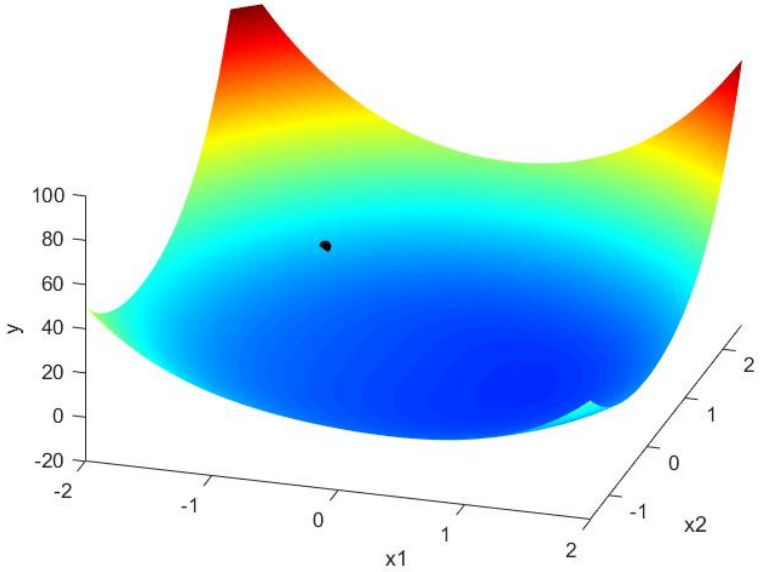

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFOTR

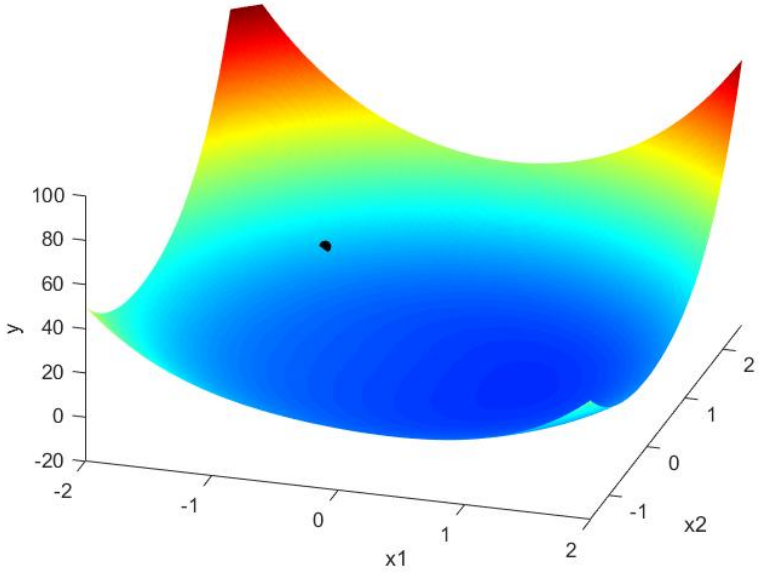

100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

FD

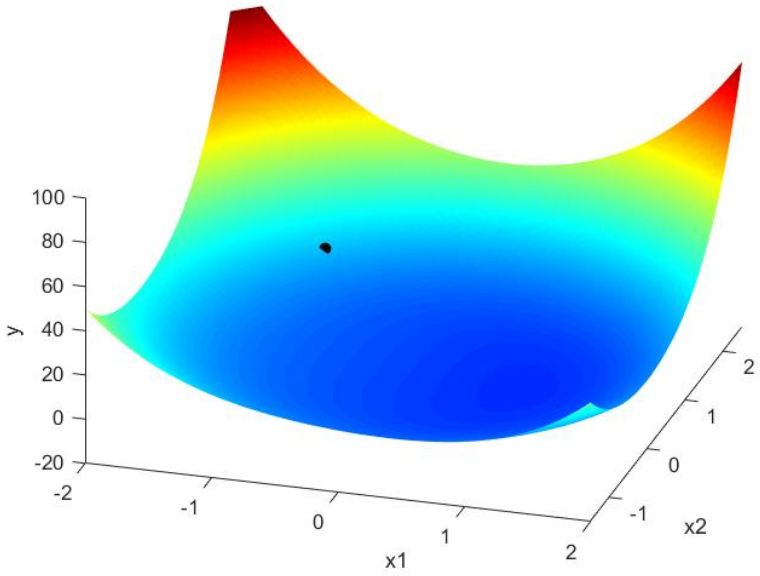


(a) starting point

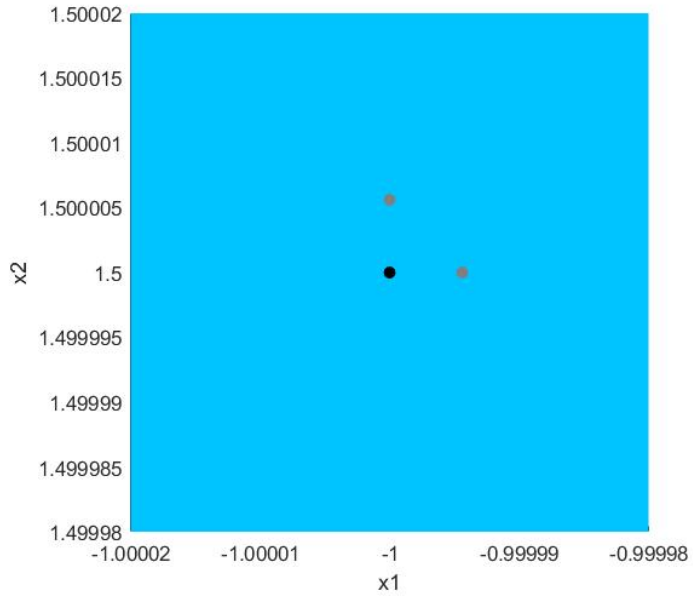


(b) initial sampling

FD

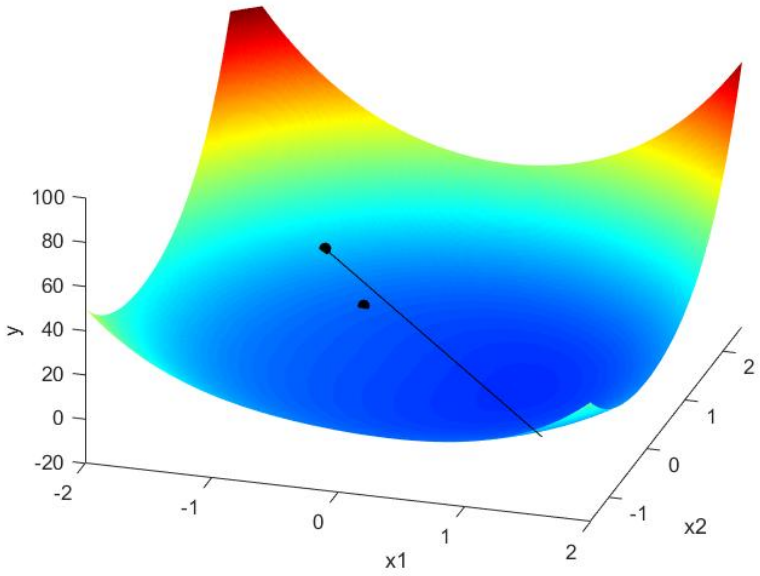


(a) starting point

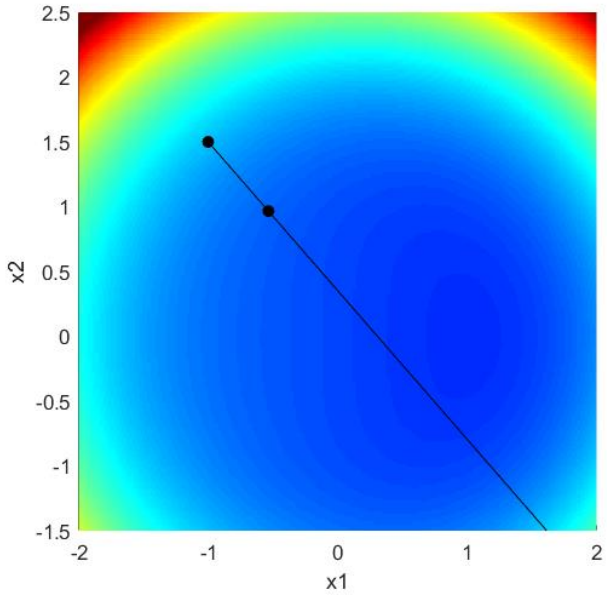


(b) initial sampling

FD

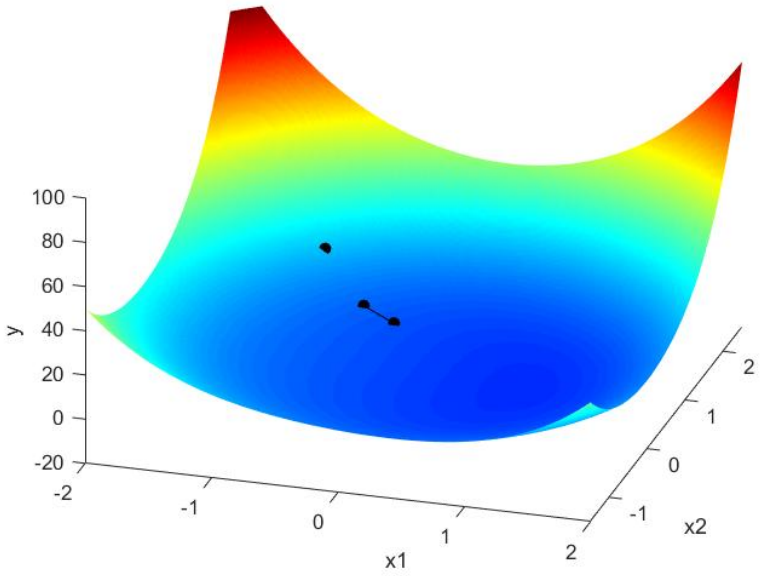


(a) starting point

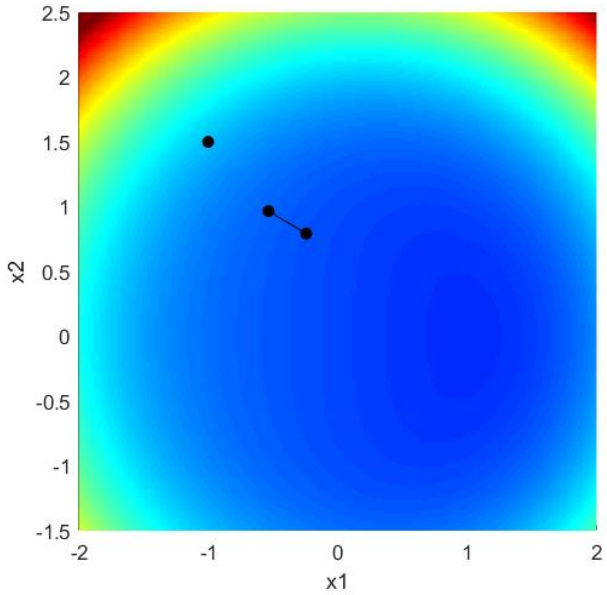


(b) initial sampling

FD

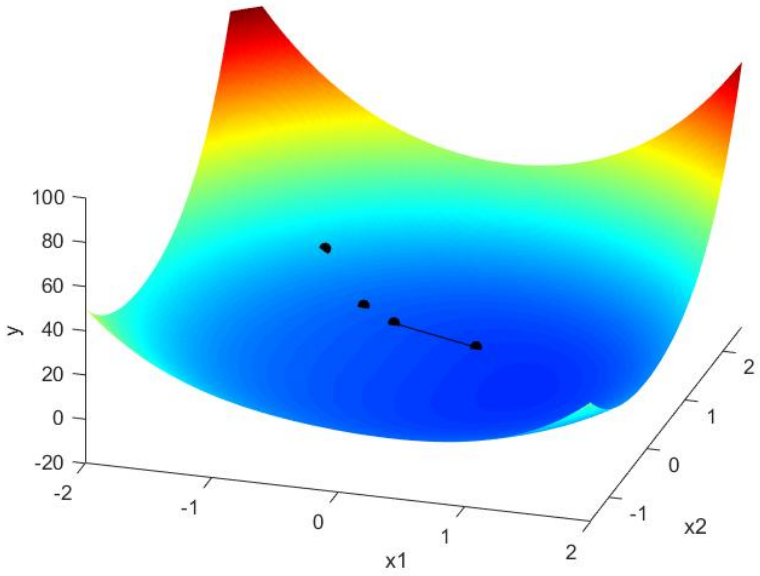


(a) starting point

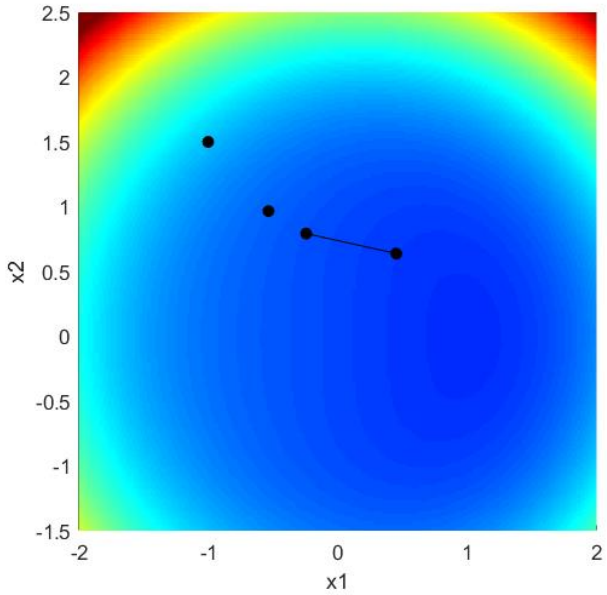


(b) initial sampling

FD

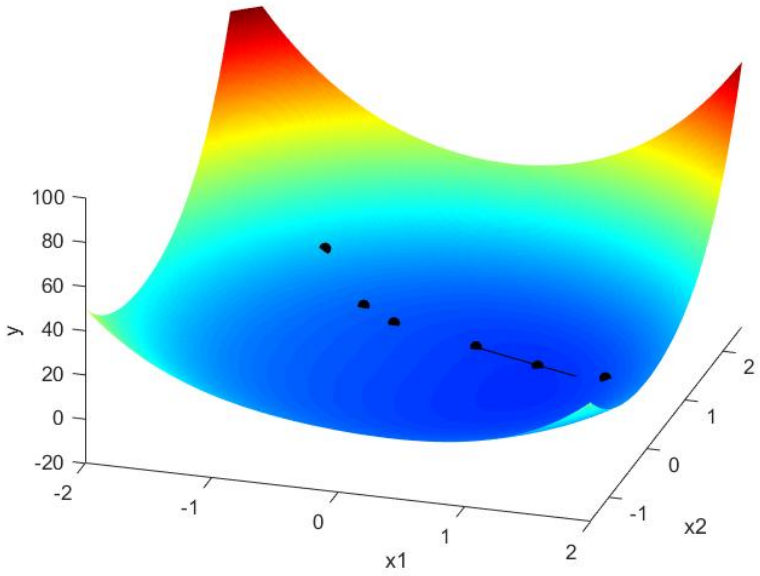


(a) starting point

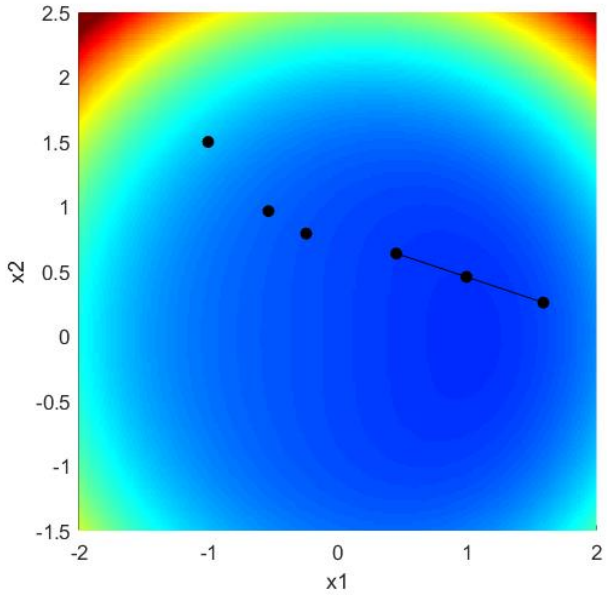


(b) initial sampling

FD

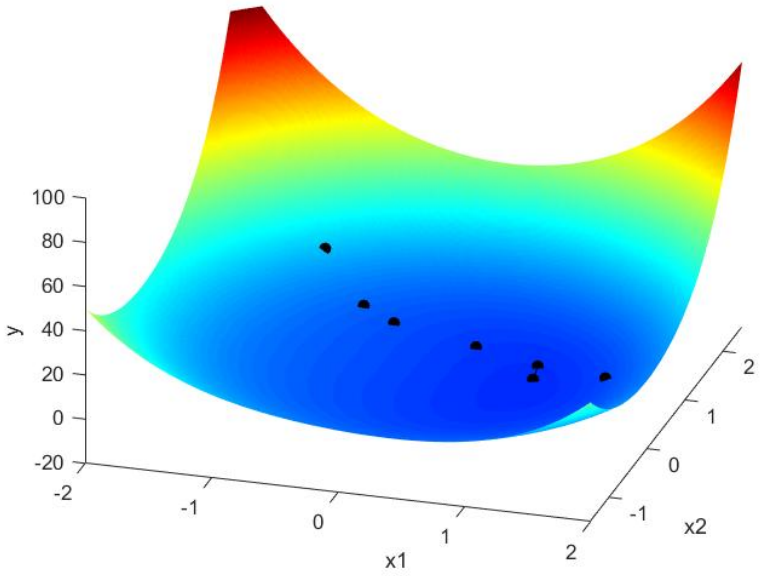


(a) starting point

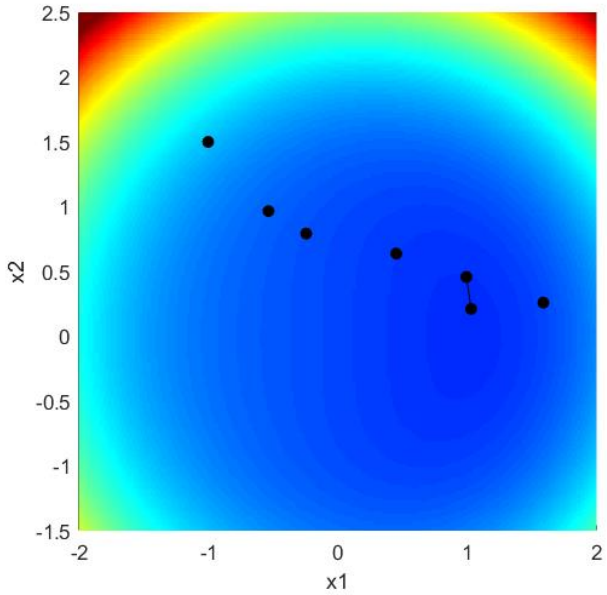


(b) initial sampling

FD

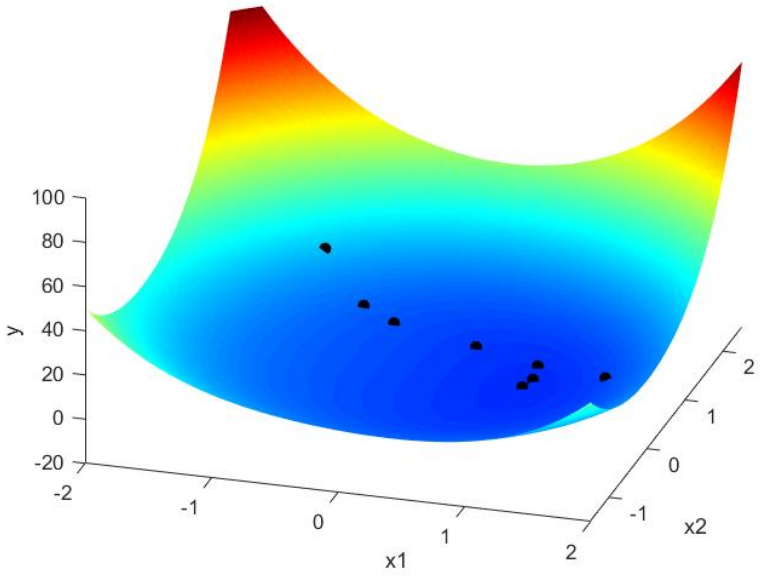


(a) starting point

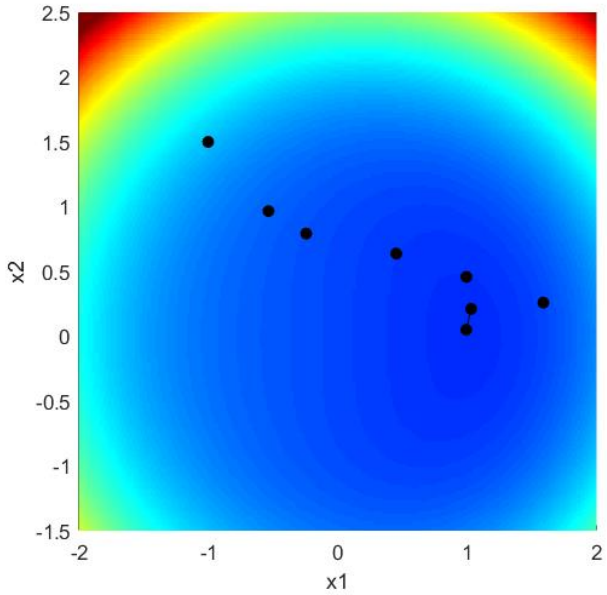


(b) initial sampling

FD

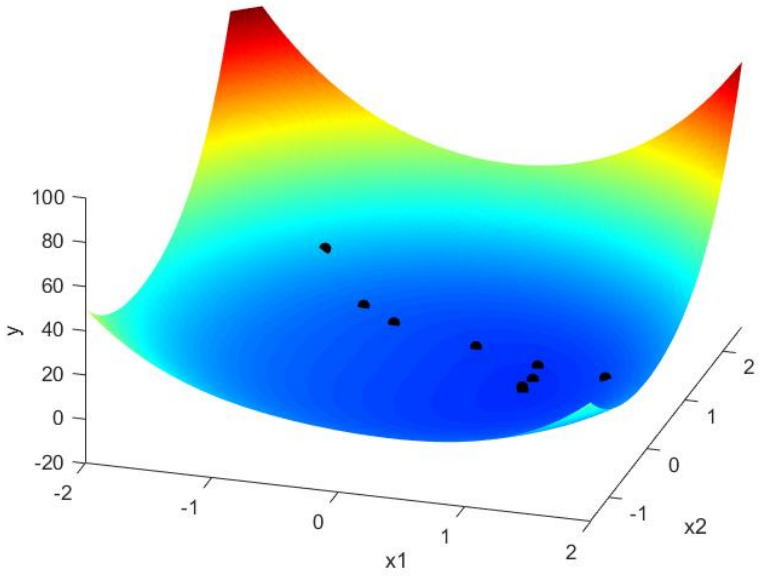


(a) starting point

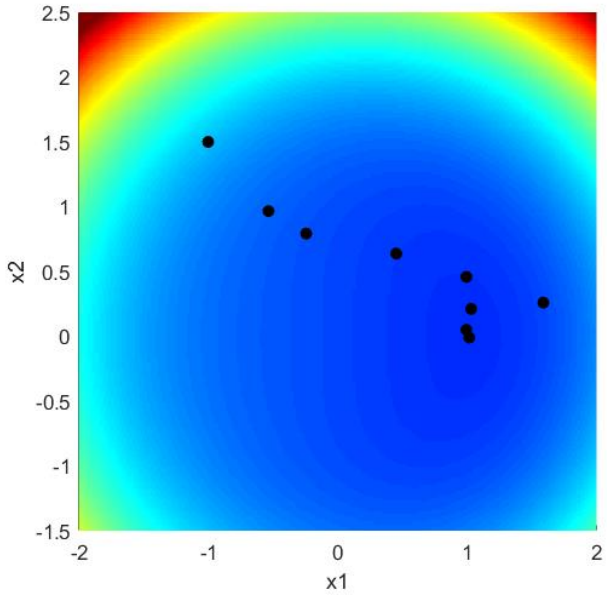


(b) initial sampling

FD



(a) starting point



(b) initial sampling

Derivative Free Optimization

$$\min_{x \in \mathbb{R}^n} f(x) \tag{3}$$

Problem: black-box function, derivative not available, maybe noisy

Assumption: Function evaluation is very expensive.

Derivative Free Optimization

$$\min_{x \in \mathbb{R}^n} f(x) \tag{3}$$

Problem: black-box function, derivative not available, maybe noisy

Assumption: Function evaluation is very expensive.

Idea: x be the policy and f be the (negative) reward

Derivative Free Optimization

$$\min_{x \in \mathbb{R}^n} f(x) \tag{3}$$

Problem: black-box function, derivative not available, maybe noisy

Assumption: Function evaluation is very expensive.

Idea: x be the policy and f be the (negative) reward

Evolution strategies as a scalable alternative to reinforcement learning

T **Salimans**, J Ho, X Chen, S Sidor… - arXiv preprint arXiv …, 2017 - arxiv.org

We explore the use of Evolution Strategies (ES), a class of black box optimization algorithms, as an alternative to popular MDP-based RL techniques such as Q-learning and Policy Gradients. Experiments on MuJoCo and Atari show that ES is a viable solution …

☆ 99 Cited by 168 Related articles All 8 versions »

evolutionary strategy

Let $F(x) = \int_{\mathbb{R}^n} f(y)\phi_\sigma(y;x)dy$ be a Gaussian smoothed version of the objective function. The function $\phi_\sigma(y;x) = \frac{1}{\sqrt{(2\pi)^n}\sigma^n}\exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right)$.

evolutionary strategy

Let $F(x) = \int_{\mathbb{R}^n} f(y)\phi_\sigma(y;x)dy$ be a Gaussian smoothed version of the objective function. The function $\phi_\sigma(y;x) = \frac{1}{\sqrt{(2\pi)^n}\sigma^n}\exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right)$.

$$\nabla_x F(x) = \frac{1}{\sigma}\mathbb{E}_Z f(x+\sigma Z)Z$$

with Z following multivariate normal distribution with mean zero and covariance matrix I.

$$g = \frac{1}{N\sigma}\sum_{i=1}^{N} f(x+\sigma Z_i)Z_i$$

Finite Difference Methods

evolutionary strategy

$$\nabla_x F(x) = \frac{1}{\sigma}\mathbb{E}_Z f(x+\sigma Z)Z, \quad g = \frac{1}{N\sigma}\sum_{i=1}^{N} f(x+\sigma Z_i)Z_i$$

forward finite difference evolutionary strategy

$$\nabla_x F(x) = \frac{1}{\sigma}\mathbb{E}_Z \left(f(x+\sigma Z) - f(x)\right) Z$$

$$g = \frac{1}{N\sigma}\sum_{i=1}^{N} \Big(f(x+\sigma Z_i) - f(x)\Big) Z_i$$

central finite difference evolutionary strategy

$$g = \frac{1}{2N\sigma}\sum_{i=1}^{N} \Big(f(x+\sigma Z_i) - f(x-\sigma Z_i)\Big) Z_i$$

Finite Difference Methods

Let $D \in \mathbb{R}^{n \times n}$ be a random orthonormal matrix; i.e. $D^\intercal D = I$.
Let d_i be the ith column of D.

rotated forward finite differencing

$$g = \frac{1}{\epsilon} D \begin{pmatrix} f(x + \epsilon d_1) - f(x) \\ f(x + \epsilon d_2) - f(x) \\ \vdots \\ f(x + \epsilon d_n) - f(x) \end{pmatrix}$$

rotated central finite differencing

$$g = \frac{1}{2\epsilon} D \begin{pmatrix} f(x + \epsilon d_1) - f(x + \epsilon d_1) \\ f(x + \epsilon d_2) - f(x + \epsilon d_2) \\ \vdots \\ f(x + \epsilon d_n) - f(x + \epsilon d_n) \end{pmatrix}$$

Accuracy of the Finite Difference Methods

relative error

$$\frac{\|\nabla f(x) - g\|}{\|\nabla f(x)\|}$$

Pick $\epsilon = 2\sqrt{10^{-12}}(1 + \|x\|) \approx 10^{-4}$.
Set $\sigma = \frac{\Gamma(n/2)}{\sqrt{2}\Gamma((n+1)/2)}\epsilon$, meaning $\mathbb{E}\|\sigma Z\| = \epsilon$.

method	nfeval	mean	variance
1. forward finite differencing	$n+1$	7.3672e-05	2.9948e-07
2. central finite defferencing	$2n$	1.9268e-07	3.5945e-12
3. rotated forward finite differencing	$n+1$	5.7320e-05	1.6836e-07
4. rotated central finite differencing	$2n$	1.6563e-07	2.8062e-12
5. evolutionary strategy	$n+1$	9.9697e+05	1.1644e+14
6. forward finite-difference evolution strategy	$n+1$	0.9921	0.2724
7. central finite-difference evolution strategy	$2n$	0.9785	0.2413
8. interpolation of Gaussian sample	$n+1$	3.1680e-04	3.6337e-05

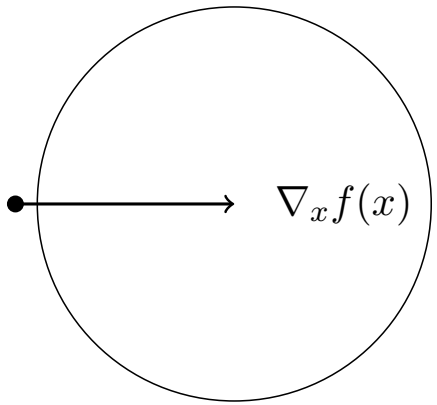
Table: mean and variance of the errors for each methods

Accuracy of the Finite Difference Methods

method	nfeval	mean	variance
5. evolutionary strategy	$n+1$	9.9697e+05	1.1644e+14
	$10n+1$	3.0004e+05	7.7568e+12
	$100n+1$	9.7531e+04	9.3396e+11
6. forward finite-difference evolution strategy	$n+1$	0.9921	0.2724
	$10n+1$	0.3280	0.0172
	$100n+1$	0.1041	0.0017
7. central finite-difference evolution strategy	$2n$	0.9785	0.2413
	$20n$	0.3284	0.0177
	$200n$	0.1044	0.0017

Accuracy of the Finite Difference Methods

method	nfeval	mean	variance
5. evolutionary strategy	$n+1$	9.9697e+05	1.1644e+14
	$10n+1$	3.0004e+05	7.7568e+12
	$100n+1$	9.7531e+04	9.3396e+11
6. forward finite-difference evolution strategy	$n+1$	0.9921	0.2724
	$10n+1$	0.3280	0.0172
	$100n+1$	0.1041	0.0017
7. central finite-difference evolution strategy	$2n$	0.9785	0.2413
	$20n$	0.3284	0.0177
	$200n$	0.1044	0.0017



Accuracy of the Finite Difference Methods

evolutionary strategy

$$g = \frac{1}{N\sigma} \sum_{i=1}^{N} f(x + \sigma Z_i) Z_i$$

If f has Lipschitz-continuous gradient with constant L, then

$$\|\nabla_x f(x) - \nabla_x F(x)\| \leq \sqrt{n} L \sigma.$$

(Similarly forward finite difference has $\|\nabla_x f(x) - g\| \leq \sqrt{n} L \epsilon$.)

$$\|\nabla_x f(x) - g\| \leq \|\nabla_x f(x) - \nabla_x F(x)\| + \|\nabla_x F(x) - g\|$$

Variance of Evolutionary Strategy

evolutionary strategy

$$g = \frac{1}{N\sigma} \sum_{i=1}^{N} f(x + \sigma Z_i) Z_i$$

$$\begin{aligned}
\mathsf{Var}\,(g) &= \tfrac{1}{N^2} \textstyle\sum_{i=1}^{N} \mathsf{Var}\left(\tfrac{1}{\sigma} f(x+\sigma Z_i) Z_i\right) \\
&= \tfrac{1}{N^2} \textstyle\sum_{i=1}^{N} \mathbb{E}\left(\tfrac{1}{\sigma} f(x+\sigma Z_i) Z_i - \nabla_x F(x)\right)\left(\tfrac{1}{\sigma} f(x+\sigma Z_i) Z_i - \nabla_x F(x)\right)^{\intercal} \\
&= \tfrac{1}{N^2} \textstyle\sum_{i=1}^{N} \mathbb{E} \begin{pmatrix} \tfrac{1}{\sigma^2} f(x+\sigma Z_i)^2 Z_i Z_i^{\intercal} \\ -\tfrac{1}{\sigma} f(x+\sigma Z_i) Z_i \nabla_x F(x)^{\intercal} \\ -\tfrac{1}{\sigma} f(x+\sigma Z_i) \nabla_x F(x) Z_i^{\intercal} \\ +\nabla_x F(x) \nabla_x F(x)^{\intercal} \end{pmatrix}.
\end{aligned}$$

Variance of Evolutionary Strategy

evolutionary strategy

$$g = \frac{1}{N\sigma} \sum_{i=1}^{N} f(x + \sigma Z_i) Z_i$$

When $f(x) = x^\intercal x$,

$$\text{Var}\,(g) = \frac{1}{N} \left(\frac{(x^\intercal x)^2}{\sigma^2} + \sigma^2(n^2 + 6n + 8) + 2(n+4)x^\intercal x \right) I + \frac{4}{N} x x^\intercal.$$

When $x = [1; 1]$, $\sigma = 0.001$, and sample size $N = 1000$

$$\text{Var}\,(g) = \begin{pmatrix} 4000.028 & 0.004 \\ 0.004 & 4000.028 \end{pmatrix}.$$

Experiments

method	nfeval	mean	variance
1. forward finite differencing	$n+1$	7.3672e-05	2.9948e-07
2. central finite defferencing	$2n$	1.9268e-07	3.5945e-12
3. rotated forward finite differencing	$n+1$	5.7320e-05	1.6836e-07
4. rotated central finite differencing	$2n$	1.6563e-07	2.8062e-12
5. evolutionary strategy	$n+1$	9.9697e+05	1.1644e+14
6. forward finite-difference evolution strategy	$n+1$	0.9921	0.2724
7. central finite-difference evolution strategy	$2n$	0.9785	0.2413
8. interpolation or regression of Gaussian sample	$n+1$	3.1680e-04	3.6337e-05

Table: mean and variance of the errors for each methods

Experiments: Smooth Problems

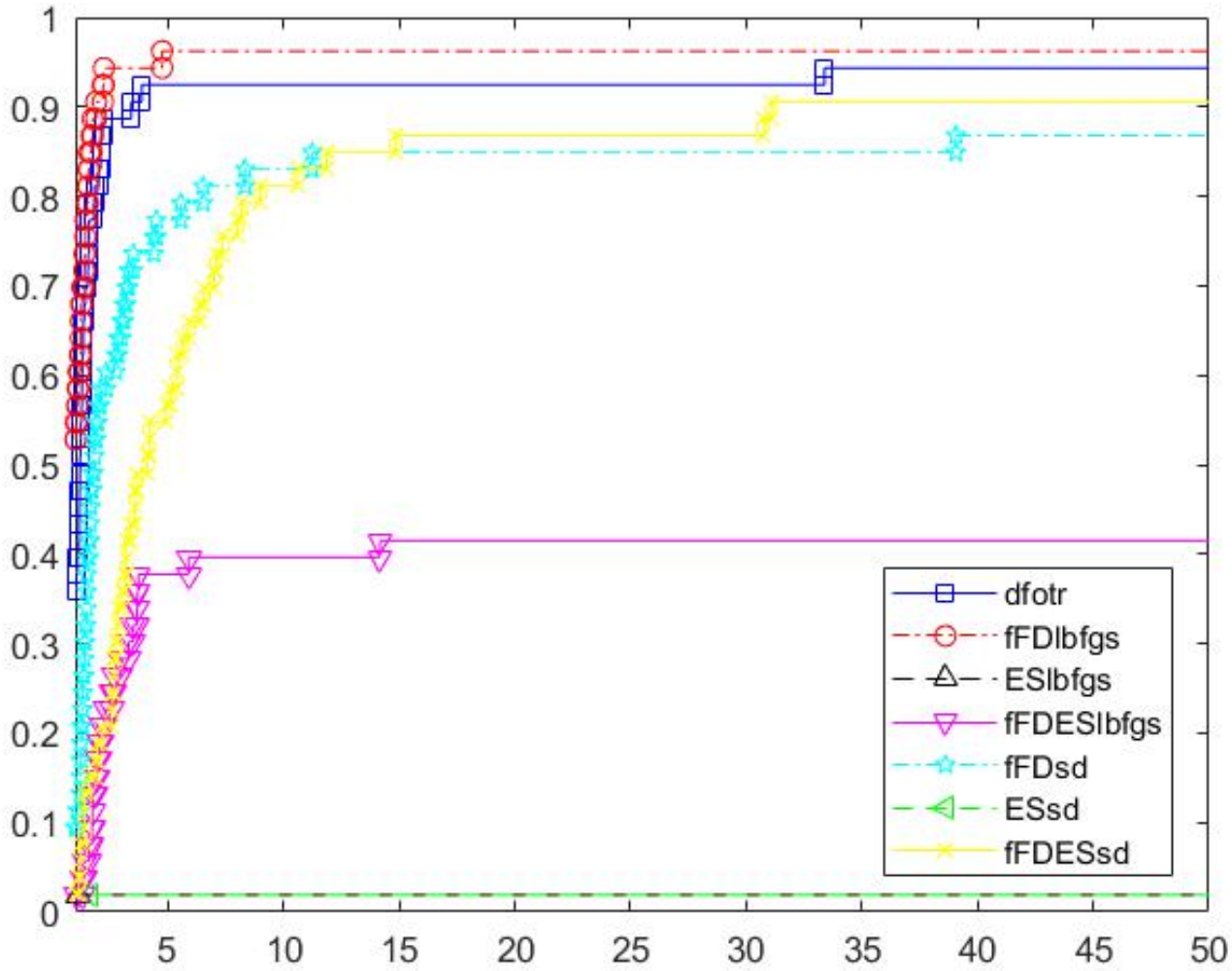


Figure: performance profile on More&Wild smooth, 1%

Experiments: Smooth Problems

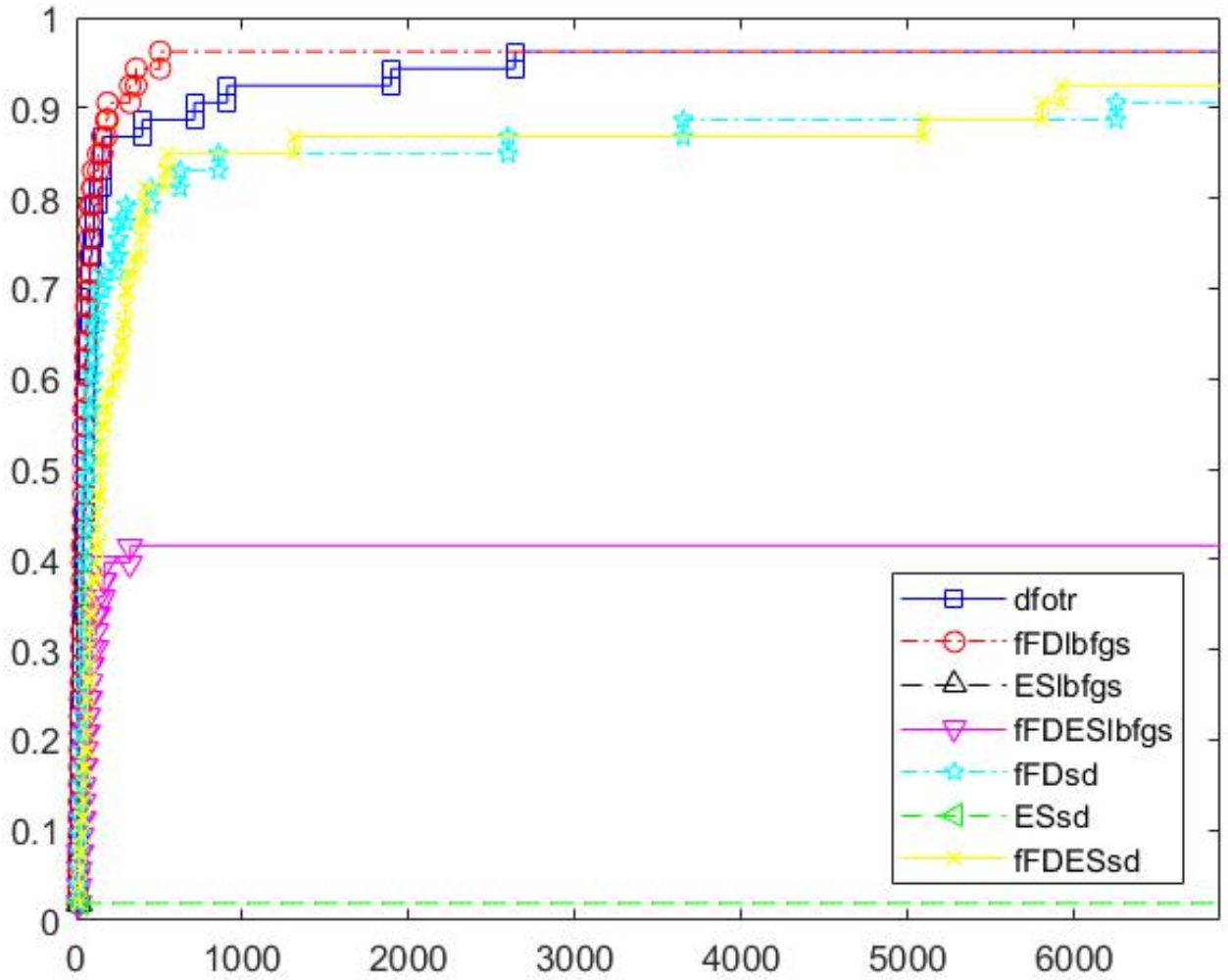
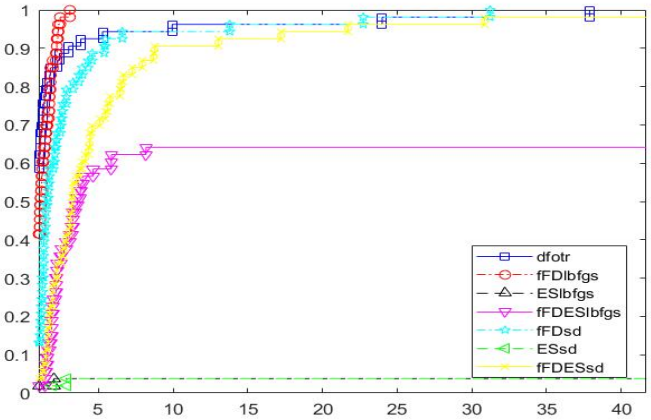
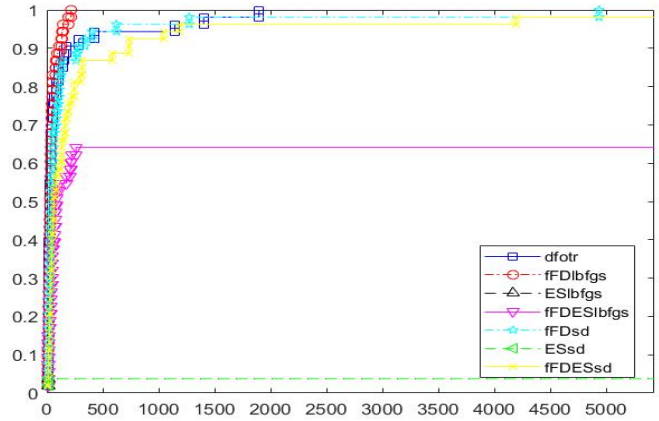


Figure: data profile on More&Wild smooth, 1%

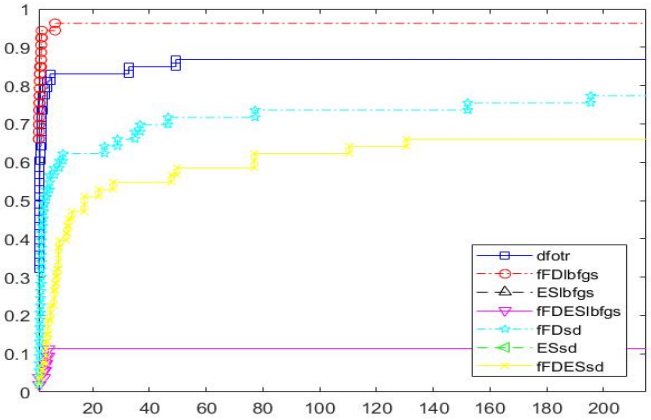
Experiments: Smooth Problems



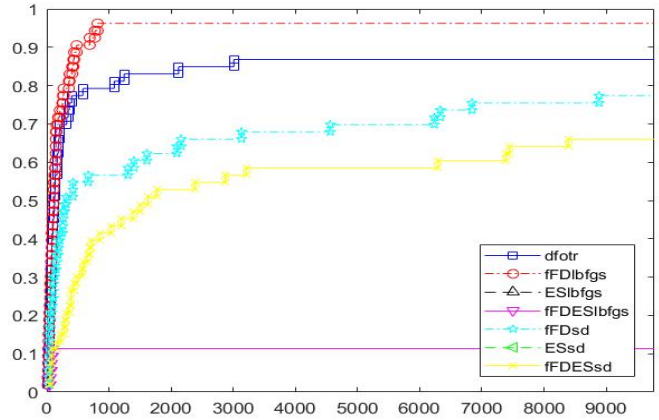
(a) performance profile on More&Wild smooth, 10%



(b) data profile on More&Wild smooth, 10%

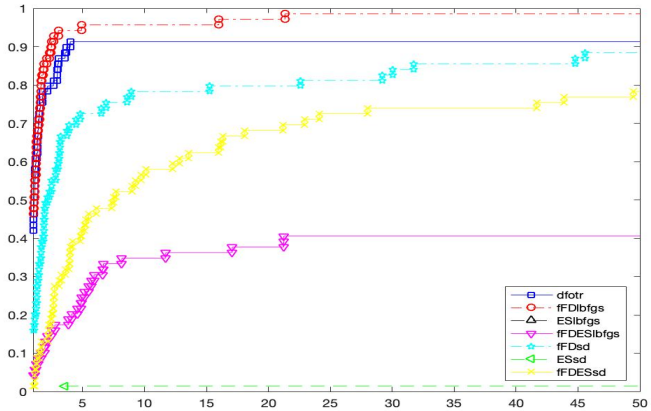


(c) performance profile on More&Wild smooth, 0.001%

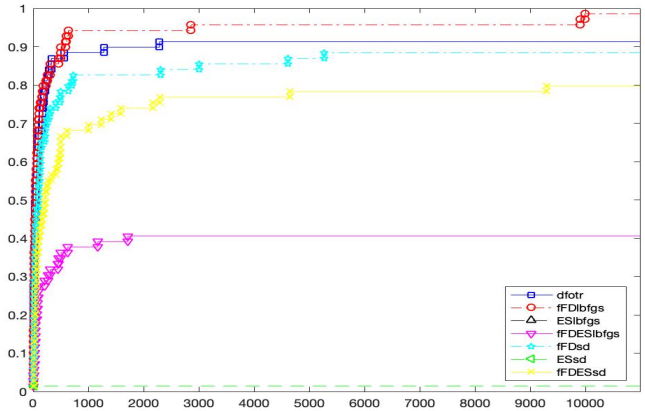


(d) data profile on More&Wild smooth, 0.001%

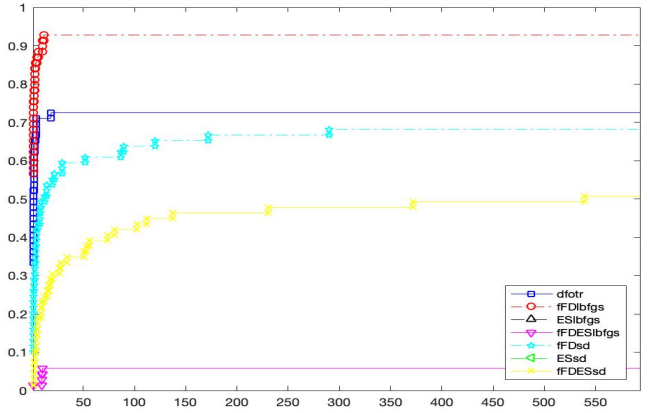
Experiments: Smooth Problems



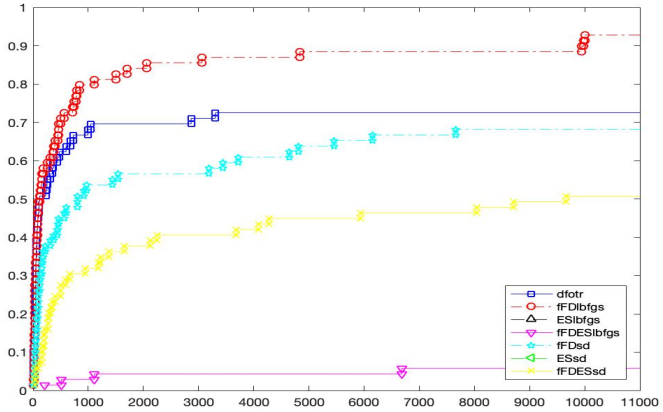
(a) performance profile on Schittkowski, 1%



(b) data profile on Schittkowski, 1%



(c) performance profile on Schittkowski, 0.001%



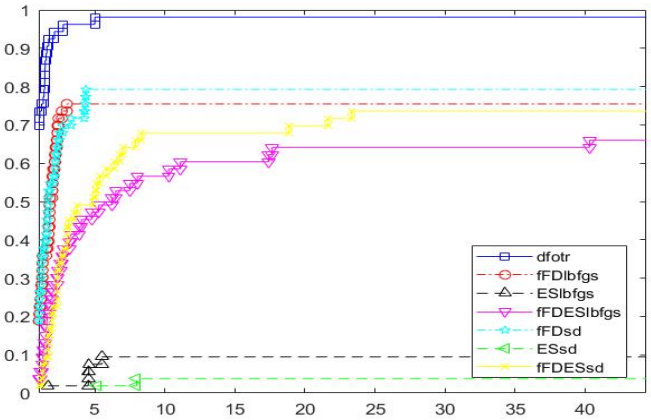
(d) data profile on Schittkowski, 0.001%

Dealing with Noise

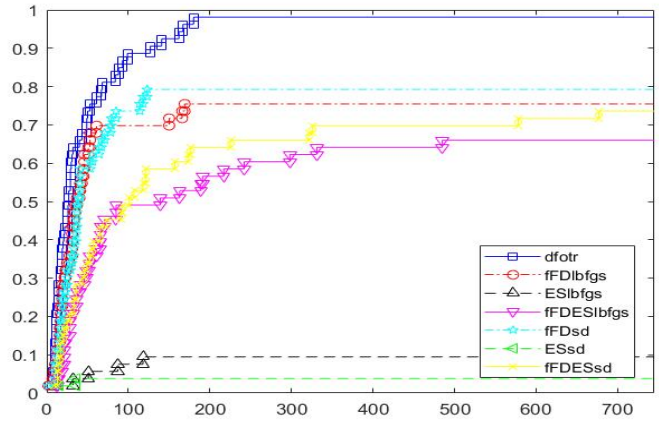
Moré, Jorge J., and Stefan M. Wild. "Estimating computational noise." *SIAM Journal on Scientific Computing* 33.3 (2011): 1292-1314.

Moré, Jorge J., and Stefan M. Wild. "Estimating derivatives of noisy simulations." *ACM Transactions on Mathematical Software (TOMS)* 38.3 (2012): 19.

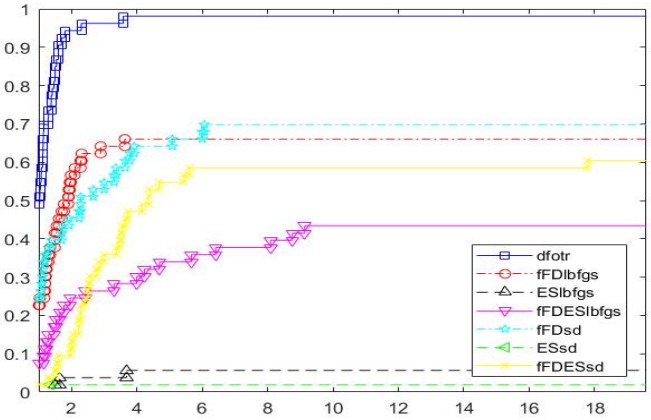
Experiments: Noisy Problems



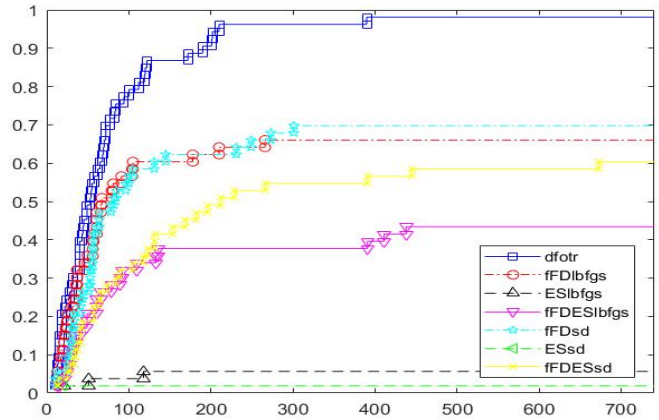
(a) performance profile on More&Wild noisy, 10%



(b) data profile on More&Wild noisy, 10%



(c) performance profile on More&Wild noisy, 1%

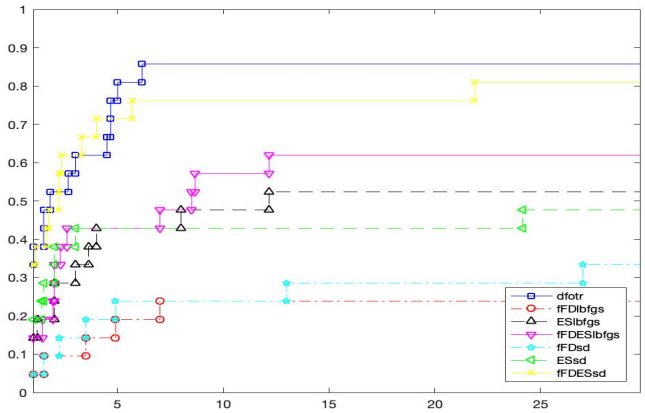


(d) data profile on More&Wild noisy, 1%

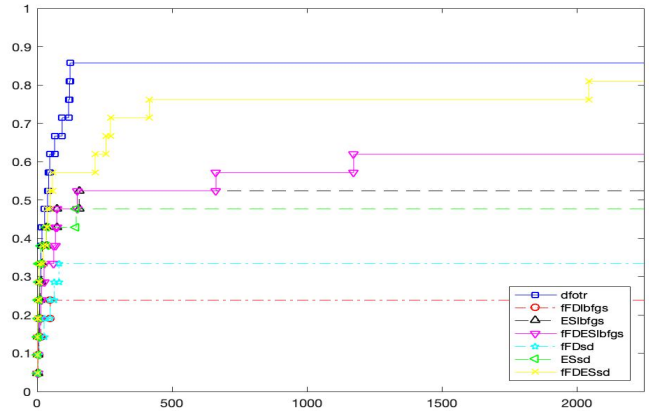
Experiments: Classification Problems

$$\min_{w,w_0} \quad \frac{1}{m}\sum_{i=1}^{m} 1_{y_i(w^\intercal x_i + w_0)<0}$$

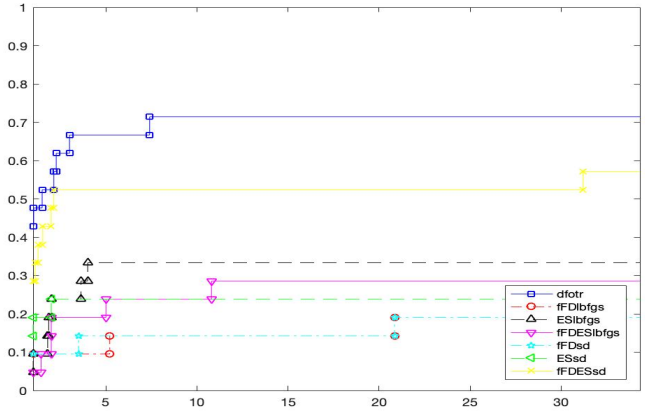
Experiments: Classification Problems



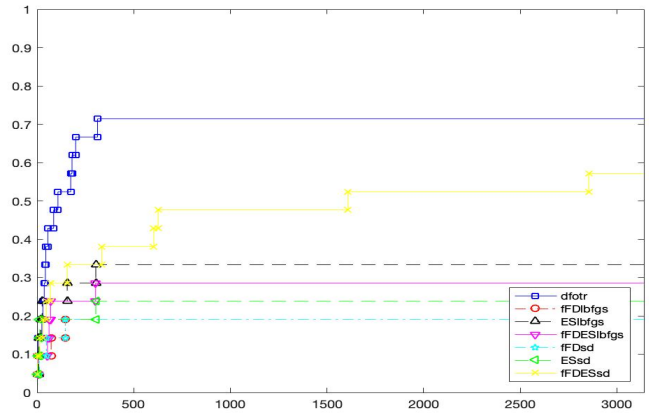
(a) performance profile on More&Wild noisy, 10%



(b) data profile on More&Wild noisy, 10%



(c) performance profile on More&Wild noisy, 1%



(d) data profile on More&Wild noisy, 1%

Experiments: Classification Problems

name	1	2	3	4	5	6	7	LR
a9a	83.3113	24.081	76.7298	80.5903	24.081	77.8139	77.5836	84.9114
australian	69.7101	44.4928	68.5507	68.2609	44.4928	68.4058	69.4203	87.5362
banknote	99.344	91.7638	56.9971	88.8484	95.9184	84.0379	99.2711	99.1983
breast-cancer	65.0073	34.9927	65.0073	65.0073	34.9927	65.0073	65.0073	65.0073
climate	92.037	91.4815	91.6667	91.6667	91.4815	91.4815	91.6667	97.037
covtype	57.0635	55.3431	57.2076	55.6538	55.3431	56.3575	56.6487	75.0336
diabetes	78.9062	66.4062	67.7083	69.7917	76.1719	66.276	68.3594	78.5156
fourclass	79.3503	75.058	74.478	77.2622	75.058	79.0023	79.1183	76.1021
german	78.1	70.8	70.5	70.9	70.8	69.4	70.6	78.5
ijcnn1	90.3561	90.2921	90.3481	90.3041	90.2921	90.2921	90.3041	92.4825
ionosphere	87.7493	76.9231	80.3419	86.6097	76.9231	77.7778	87.1795	93.7322
liver-disorders	66.3768	44.9275	59.1304	60.5797	44.9275	60.8696	57.6812	65.7971
madelon	56.95	50	55.15	61	50	57.5	61.95	74
mushrooms	98.1905	51.7971	77.2895	91.4205	51.7971	85.8813	95.6179	100
phishing	90.4116	55.6943	71.6327	75.45	55.6943	75.251	72.9444	94.1384
skin nonskin	94.4658	79.2461	90.2953	91.502	79.2461	91.4722	93.0041	91.8827
splice	73.0079	51.9055	61.6063	66.5827	51.9055	58.8661	68.5669	85.9528
svm1	83.2632	82.292	79.9935	64.7459	82.292	81.4179	76.2706	95.1764
svm3	79.5656	77.313	76.2671	76.7498	77.313	77.0716	76.6693	83.5076
transfusion	76.2032	23.7968	76.2032	76.2032	23.7968	76.2032	76.2032	77.139
w5a	97.1582	2.8418	97.2189	97.3604	2.8418	97.2391	97.3301	99.0898

Table: classification accuracy: 1DFOTR, 2fFDlbfgs, 3ESlbfgs, 4fFDESlbfgs, 5fFDsd, 6ESsd, 7dFDESsd

Thank you!