

# Some Gradient Approximation Methods for Derivative Free Optimization

Albert S. Berahas, **Liyuan Cao**, Katya Scheinberg

Lehigh University

INFORMS Annual Meeting 2019

# Collaborators



Albert S. Berahas



Katya Scheinberg

# Black Box Optimization

## a.k.a. Derivative Free Optimization

### typical objective function

$$x \longrightarrow \boxed{f(x) = \sum_{i=1}^N \log(1 + \exp(y_i \cdot x^T \phi_i)))} \longrightarrow f(x)$$

use derivative based algorithms:

gradient descent , L-BFGS, Newton's method

### black box objective function



## Background

Let  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  be the objective function, and we are optimizing it with a gradient based algorithm, but with gradient estimates instead of the true gradients.

Theorem (Berahas, Cao, Scheinberg, 2019)

*Under ... assumptions, if for each iteration  $k$ , the gradient estimate  $g(x_k)$  is sufficiently accurate*

$$\|g(x_k) - \nabla\phi(x_k)\| \leq \theta \|\nabla\phi(x_k)\|$$

*with probability at least  $1 - \eta$ , then the expected number of iterations to reach  $\phi(X_k) - \phi^* \leq \epsilon$  is less than .... ( $\theta, \eta \in (0, 1)$ )*

# Finite Difference

Let  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ .

For each coordinate  $i = 1, 2, \dots, n$ , let  $e_i$  be the unit vector.

$$\frac{\partial \phi(x)}{\partial x_i} = \lim_{h \rightarrow 0} \frac{\phi(x + he_i) - \phi(x)}{h} \Rightarrow [g(x)]_i = \frac{\phi(x + he_i) - \phi(x)}{h}$$

# Finite Difference

Let  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ .

For each coordinate  $i = 1, 2, \dots, n$ , let  $e_i$  be the unit vector.

$$\frac{\partial \phi(x)}{\partial x_i} = \lim_{h \rightarrow 0} \frac{\phi(x + he_i) - \phi(x)}{h} \Rightarrow [g(x)]_i = \frac{\phi(x + he_i) - \phi(x)}{h}$$

If the gradient of  $\phi$  is  $L$ -Lipschitz continuous, then

$$\|g(x) - \nabla \phi(x)\| \leq \frac{\sqrt{n}Lh}{2}.$$

# Finite Difference

no noise:

$$\|g(x) - \nabla\phi(x)\| \leq \frac{\sqrt{n}Lh}{2}.$$

objective function with bounded noise:

$$f(x) = \phi(x) + \epsilon(x) \text{ and } |\epsilon(x)| < \epsilon_f$$

$$\|g(x) - \nabla\phi(x)\| \leq \frac{\sqrt{n}Lh}{2} + \frac{2\sqrt{n}\epsilon_f}{h}$$

# Interpolation

The sample set is  $\{x, x + hu_1, x + hu_2, \dots, x + hu_n\}$ , where  $\{u_1, u_2, \dots, u_n\} \subset \mathbb{R}^n$  with  $\|u_i\| \leq 1$  for all  $i$ .

$$\begin{pmatrix} hu_1^\top \\ hu_2^\top \\ \vdots \\ hu_n^\top \end{pmatrix} g(x) = \begin{pmatrix} f(x + hu_1) - f(x) \\ f(x + hu_2) - f(x) \\ \vdots \\ f(x + hu_n) - f(x) \end{pmatrix} \Rightarrow hQ_{\mathcal{X}}g(x) = F_{\mathcal{X}}$$

error bounds:

without noise:  $\|g(x) - \nabla\phi(x)\| \leq \|Q_{\mathcal{X}}^{-1}\| \frac{\sqrt{n}Lh}{2}$

with noise:  $\|g(x) - \nabla\phi(x)\| \leq \|Q_{\mathcal{X}}^{-1}\| \left( \frac{\sqrt{n}Lh}{2} + \frac{2\sqrt{n}\epsilon_f}{h} \right)$

# A Little Bit Summary

method	formula	bound
FD	$g_i(x) = \frac{f(x+he_i) - f(x)}{h}$	$\frac{\sqrt{n}Lh}{2} + \frac{2\sqrt{n}\epsilon_f}{h}$
interp	$hQ_{\mathcal{X}}g(x) = F_{\mathcal{X}}$	$\ Q_{\mathcal{X}}^{-1}\  \left( \frac{\sqrt{n}Lh}{2} + \frac{2\sqrt{n}\epsilon_f}{h} \right)$
GSG*	$g(x) = \frac{1}{m} \sum_{i=1}^m \frac{f(x+\sigma u_i) - f(x)}{\sigma} u_i$	

\* Gaussian smooth gradient;  $u_i \in \mathbb{R}^n$ ,  $u_i \sim \mathcal{N}(0, I)$  for all  $i$  independently

# Gaussian Smooth Gradient

origin of the formula:

$$F(x) = \int_{\mathbb{R}^n} f(y) \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right) dy$$

$$\begin{aligned} \nabla F(x) &= \int_{\mathbb{R}^n} f(y) \frac{y-x}{\sigma^2} \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right) dy && y \sim \mathcal{N}(x, \sigma^2 I) \\ &= \int_{\mathbb{R}^n} \frac{f(x + \sigma u)}{\sigma} u \cdot \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{\|u\|^2}{2}\right) du && u \sim \mathcal{N}(0, I) \\ &= \int_{\mathbb{R}^n} \frac{f(x + \sigma u) - f(x)}{\sigma} u \cdot \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{\|u\|^2}{2}\right) du \end{aligned}$$

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

# Gaussian Smooth Gradient

[A derivative-free trust-region algorithm for the optimization of functions smoothed via gaussian convolution using adaptive multiple importance sampling](#)

[A Maggiar, A Wachter, IS Dolinskaya, J Staum - SIAM Journal on Optimization, 2018 - SIAM](#)

In this paper we consider the optimization of a functional  $F$  defined as the convolution of a function  $f$  with a Gaussian kernel. We propose this type of objective function for the optimization of the output of complex computational simulations, which often present some ...

 99 Cited by 7 Related articles All 3 versions Web of Science: 1

When  $f = \phi$  (no noise) and has  $L$ -Lipschitz continuous gradient,

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

Not bad comparing to  $\|Q_{\mathcal{X}}^{-1}\| \frac{\sqrt{n}Lh}{2}$ .

# Gaussian Smooth Gradient

[A derivative-free trust-region algorithm for the optimization of functions smoothed via gaussian convolution using adaptive multiple importance sampling](#)

[A Maggiar, A Wachter, IS Dolinskaya, J Staum - SIAM Journal on Optimization, 2018 - SIAM](#)

In this paper we consider the optimization of a functional  $F$  defined as the convolution of a function  $f$  with a Gaussian kernel. We propose this type of objective function for the optimization of the output of complex computational simulations, which often present some ...

☆ 99 Cited by 7 Related articles All 3 versions Web of Science: 1

When  $f = \phi$  (no noise) and has  $L$ -Lipschitz continuous gradient,

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

Not bad comparing to  $\|Q_{\mathcal{X}}^{-1}\| \frac{\sqrt{n}Lh}{2}$ .

However we don't have the expectation  $\nabla F(x)$ , only the finite sum  $g(x)$ .  
While  $\mathbb{E}g(x) = \nabla F(x)$ , its has large variance

$$\text{Var}\{g(x)\} = \frac{1}{N} \mathbb{E}_{u \sim \mathcal{N}(0, I)} \left[ \left( \frac{f(x + \sigma u) - f(x)}{\sigma} \right)^2 uu^\top \right] - \frac{1}{N} \nabla F(x) \nabla F(x)^\top.$$

# Gaussian Smooth Gradient

$$\|g(x) - \phi(x)\| \leq \|\nabla F(x) - \nabla \phi(x)\| + \|g(x) - \nabla F(x)\|$$

With Chebyshev inequality:

Theorem (Berahas, **Cao**, Scheinberg, 2019)

When  $f = \phi$  (no noise), if

$$N \geq \frac{3n\|\nabla f(x)\|^2}{\delta r^2} + \frac{n(n+2)(n+4)L^2\sigma^2}{4\delta r^2},$$

then for all  $x \in \mathbb{R}^n$  and  $r > 0$ ,  $\|g(x) - \nabla f(x)\| \leq \sqrt{n}L\sigma + r$  with probability at least  $1 - \delta$ .

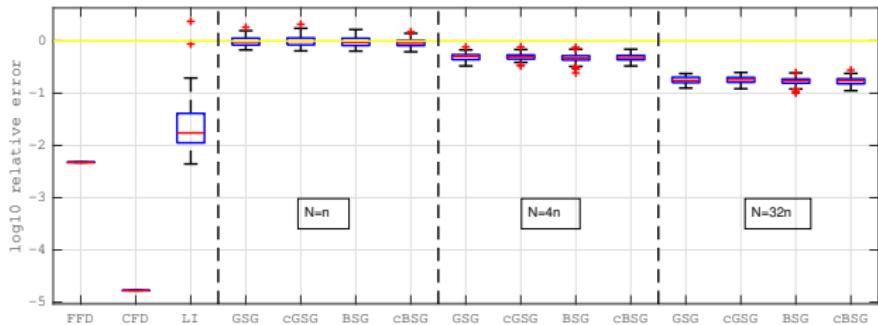
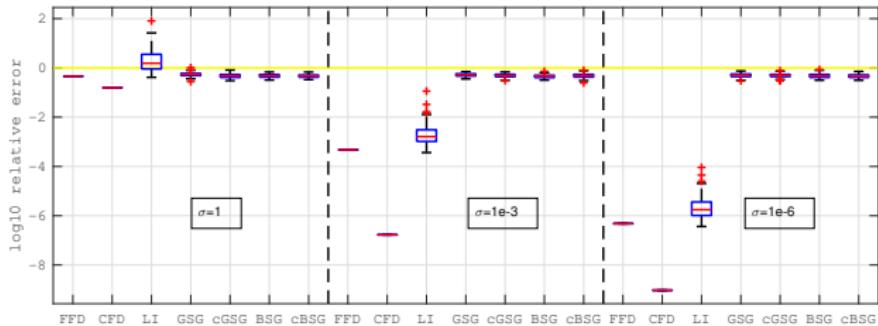
# Summary

**Table:** Bounds on  $N$  and  $\sigma$  which ensure  $\|g(x) - \nabla\phi(x)\| \leq \theta \|\nabla\phi(x)\|$  (possibly with probability  $1 - \delta$ ), for  $n > 12$

Gradient Approximation	# of Samples ( $N$ )	$\hbar$ or $\sigma$
Forward Finite Differences	$n$	$\frac{2\theta \ \nabla f(x)\ }{\sqrt{n}L}$
Central Finite Differences	$2n$	$\sqrt{\frac{6\theta \ \nabla f(x)\ }{\sqrt{n}M}}$
Linear Interpolation	$n$	$\frac{2\theta \ \nabla f(x)\ }{\sqrt{n}L \ Q^{-1}\ }$
Gaussian Smooth $g$	$\frac{6n}{\delta\theta^2} + \frac{(2n+13)}{4\delta}$	$\frac{\theta \ \nabla f(x)\ }{nL}$
Central GSG	$\frac{6n}{\delta\theta^2} + \frac{(2n+26)}{36\delta}$	$\sqrt{\frac{\theta \ \nabla f(x)\ }{n^{3/2}M}}$
Sphere Smooth $g$	$\left(\frac{4n}{\theta^2} + n + \frac{4\sqrt{2}n}{3\theta} + \frac{2\sqrt{2}\sqrt{n}}{3}\right) \log \frac{n+1}{\delta}$	$\frac{\theta \ \nabla f(x)\ }{\sqrt{n}L}$
Central SSG	$\left(\frac{4n}{\theta^2} + \frac{n}{9} + \frac{4\sqrt{2}n}{3\theta} + \frac{2\sqrt{2}\sqrt{n}}{9}\right) \log \frac{n+1}{\delta}$	$\sqrt{\frac{\theta \ \nabla f(x)\ }{\sqrt{n}M}}$

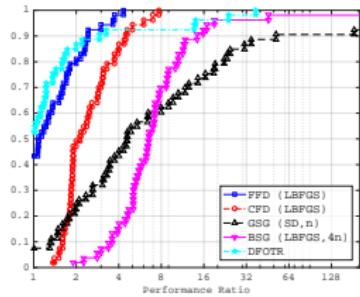
# Numerical Results

On a test function with  $n = 20$ :

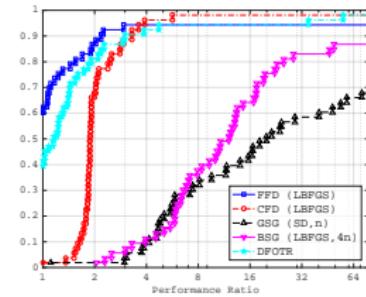


# Numerical Results

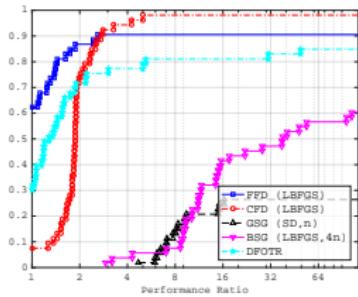
On Moré&Wild test set:



(a)  $\tau = 10^{-1}$



(b)  $\tau = 10^{-3}$



(c)  $\tau = 10^{-5}$

**Figure:** Performance profiles for best variant of each method.

# Numerical Results

On OpenAL Gym reinforcement learning problems:

