A Theoretical and Empirical Comparison of Gradient Approximations Methods in Derivative-Free Optimization

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

Lehigh University

MOPTA 2021

Collaborators

Albert S. Berahas
University of Michigan

Katya Scheinberg
Cornell University

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

Finite Difference

Let $\phi : \mathbb{R}^n \to \mathbb{R}$.
For each coordinate $i = 1, 2, \ldots, n$, let e_i be the ith column of $I_{n\times n}$.

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

Finite Difference

Let $\phi : \mathbb{R}^n \to \mathbb{R}$.
For each coordinate $i = 1, 2, \ldots, n$, let e_i be the ith column of $I_{n\times n}$.

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

If the gradient of ϕ 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 $\{x, x+hu_1, x+hu_2, \ldots, x+hu_n\}$, where $\{u_1, u_2, \ldots, u_n\} \subset \mathbb{R}^n$ with $\|u_i\| \le 1$ for all i.

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

error bounds:

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

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

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

$*$ 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:

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

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

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

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 …

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,

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

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

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

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 $e(x) = 0$ *(no noise), if*

$$N \geq \frac{n}{(1-p)r^2}\left(3\|\nabla \phi(x)\|^2 + \frac{L^2\sigma^2}{4}(n+2)(n+4)\right);$$

or when $|e(x)| \leq \epsilon_f$*, if*

$$N \geq \frac{3n}{(1-p)r^2}\left(3\|\nabla \phi(x)\|^2 + \frac{L^2\sigma^2}{4}(n+2)(n+4) + \frac{4\epsilon_f^2}{\sigma^2}\right)$$

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* p.

Ball/Sphere Smooth Gradient

$$F(x) = \mathbb{E}_{u \sim \mathcal{U}(\mathcal{B}(0,1))}[f(x+\sigma u)] = \int_{\mathcal{B}(0,1)} f(x+\sigma u) \frac{1}{V_n(1)} du$$

$$\nabla F(x) = \frac{n}{\sigma} \mathbb{E}_{u \sim \mathcal{U}(\mathcal{S}(0,1))}[f(x+\sigma u)u]$$

With Bernstein inequality:

Theorem

When $|e(x)| \leq \epsilon_f$, *if*

$$N \geq \left[\frac{6n^2}{r^2} \left(\frac{\|\nabla \phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} \right) + \frac{2n}{3r} \left(2\|\nabla \phi(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma} \right) \right] \log \frac{n+1}{1-p},$$

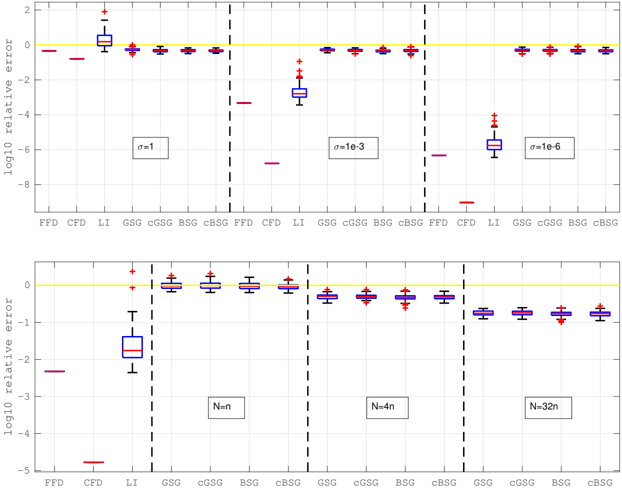
then for all $x \in \mathbb{R}^n$ *and* $r > 0$, $\|g(x) - \nabla f(x)\| \leq \sqrt{n}L\sigma + r$ *with probability at least* p.

Summary

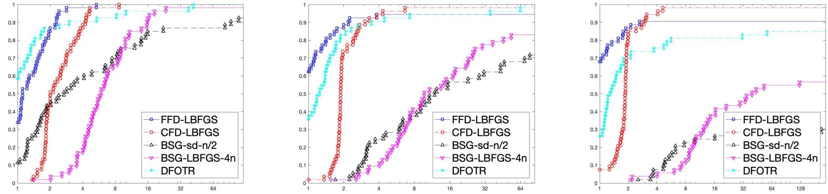
Table: Bounds on N, σ and $\|\nabla\phi(x)\|$ that ensure $\|g(x) - \nabla\phi(x)\| \leq \theta\|\nabla\phi(x)\|$ (* denotes result is with probability p).

Gradient Approximation	N	h or σ	$\|\nabla\phi(x)\|$
Forward Finite Differences	n	$2\sqrt{\frac{\epsilon_f}{L}}$	$\frac{2\sqrt{nL\epsilon_f}}{\theta}$
Central Finite Differences	n	$\sqrt[3]{\frac{6\epsilon_f}{M}}$	$\frac{\sqrt[3]{9}\sqrt[3]{n^{3/2}M\epsilon_f^2}}{2\theta}$
Linear Interpolation	n	$2\sqrt{\frac{\epsilon_f}{L}}$	$\frac{2\|Q_{\mathcal{X}}^{-1}\|\sqrt{nL\epsilon_f}}{\theta}$
Gaussian Smoothed Gradients*	$\frac{9n}{(1-p)\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{3(n+4)}{16(1-p)} + \frac{3}{n(1-p)}$	$\sqrt{\frac{\epsilon_f}{L}}$	$\frac{6n\sqrt{L\epsilon_f}}{\theta}$
Centered Gaussian Smoothed Gradients*	$\frac{9n}{(1-p)\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{n+6}{48(1-p)} + \frac{3}{4n(1-p)}$	$\sqrt[3]{\frac{\epsilon_f}{\sqrt{n}M}}$	$\frac{18\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\sqrt[3]{4}\theta}$
Sphere Smoothed Gradients*	$\left[\left(\frac{6n}{\theta^2}\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{4n}{3\theta}\right)\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{3n}{8} + \frac{6}{n} + \frac{\sqrt{n}}{3} + \frac{4}{3\sqrt{n}}\right]\log\frac{n+1}{(1-p)}$	$\sqrt{\frac{n\epsilon_f}{L}}$	$\frac{4n\sqrt{L\epsilon_f}}{\theta}$
Centered Sphere Smoothed Gradients*	$\left[\left(\frac{6n}{\theta^2}\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{4n}{3\theta}\right)\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{n}{24} + \frac{3}{2n} + \frac{\sqrt{n}}{9} + \frac{2}{3\sqrt{n}}\right]\log\frac{n+1}{(1-p)}$	$\sqrt[3]{\frac{n\epsilon_f}{M}}$	$\frac{6\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\sqrt[3]{4}\theta}$

On a test function with $n = 20$:



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: