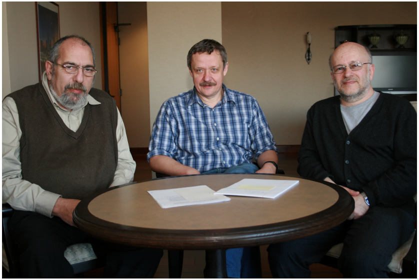
Small Errors in Zeroth Order Optimization are *Imaginary**



SIAM Conference on Optimization | July 19-23, 2021,

Wouter Jongeneel (RAO, EPFL)

with Man-Chung Yue and Daniel Kuhn

*Based on: WJ, Man-Chung Yue, and Daniel Kuhn (2021). "*Small Errors in Random Zeroth Order Optimization are Imaginary*". `arXiv:2103.05478`

The basic question

For $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, how to find

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{D}} f(x)\,?$$

The basic question

For $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, how to find

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{D}} f(x)\,?$$

Common approach: ***gradient*** *descent*

$$x_{k+1} = x_k - \mu_k \nabla f(x_k). \tag{1}$$

The basic question

For $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, how to find

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{D}} f(x)\,?$$

Common approach: ***gradient*** *descent*

$$x_{k+1} = x_k - \mu_k \nabla f(x_k). \tag{1}$$

Let f be convex and differentiable with a L-Lipschitz **gradient**

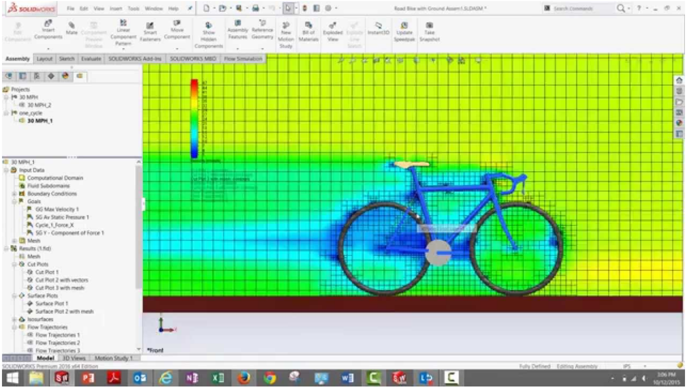
$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathcal{D},$$

then for $\mu_k = \frac{1}{L}$ and $x_0, x_1, \ldots, x_{K-1}$ generated by (1) one obtains

$$f(x_{K-1}) - f(x^\star) \leq \mathcal{O}\left(\frac{L \cdot \|x_0 - x^\star\|_2^2}{K}\right).$$

Do we always have the gradient?

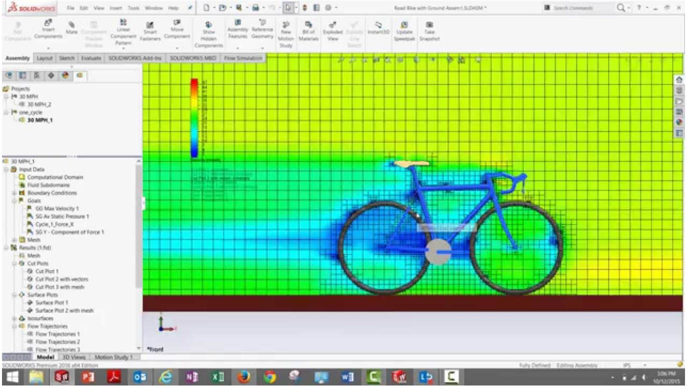
Let f represent *aerodynamic performance* and x represent *design parameters*, what is $\nabla f(x)$?



[1] https://www.youtube.com/watch?v=-mAHCq2dnKk&ab_channel=KapilGaitonde

Do we always have the gradient?

Let f represent *aerodynamic performance* and x represent *design parameters*, what is $\nabla f(x)$?



Idea: we *can* **evaluate** $f(x')$ for some design choice x'. [1]

[1] `https://www.youtube.com/watch?v=-mAHCq2dnKk&ab_channel=KapilGaitonde`

Zeroth order optimization

Obtain (approximate)

$$x^{\star} \in \operatorname*{argmin}_{x \in \mathcal{D}} f(x)$$

via function evaluations $f(x_0), f(x_1), \ldots, f(x_K)$ for some set of *selected* points $x_0, x_1, \ldots, x_K$.

[2] Conn, Scheinberg, and Vicente 2009.

[3] Nemirovsky and Yudin 1983; Flaxman, Kalai, and McMahan 2004; Nesterov and Spokoiny 2017.

Zeroth order optimization

Obtain (approximate)

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{D}} f(x)$$

via function evaluations $f(x_0), f(x_1), \ldots, f(x_K)$ for some set of *selected* points $x_0, x_1, \ldots, x_K$.

- *Model-based*: construct local model of f, optimize using that function[2].

[2] Conn, Scheinberg, and Vicente 2009.

[3] Nemirovsky and Yudin 1983; Flaxman, Kalai, and McMahan 2004; Nesterov and Spokoiny 2017.

Zeroth order optimization

Obtain (approximate)

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{D}} f(x)$$

via function evaluations $f(x_0), f(x_1), \ldots, f(x_K)$ for some set of *selected* points $x_0, x_1, \ldots, x_K$.

- *Model-based*: construct local model of f, optimize using that function[2].
- *Gradient-based*: approximate ∇f directly and apply gradient descent[3].

---

[2] Conn, Scheinberg, and Vicente 2009.

[3] Nemirovsky and Yudin 1983; Flaxman, Kalai, and McMahan 2004; Nesterov and Spokoiny 2017.

A first gradient-based approach

For any differentiable $f : \mathbb{R} \to \mathbb{R}$

$$\partial_x f(x) = \frac{f(x+\delta) - f(x)}{\delta} + \mathcal{O}(\delta).$$

[4] d'Aspremont 2008; Devolder, Glineur, and Nesterov 2014.

A first gradient-based approach

For any differentiable $f : \mathbb{R} \to \mathbb{R}$

$$\partial_x f(x) = \frac{f(x+\delta) - f(x)}{\delta} + \mathcal{O}(\delta).$$

Then, run *inexact* ($\delta > 0$) gradient descent

$$x_{k+1} = x_k - \mu_k \frac{f(x_k+\delta) - f(x_k)}{\delta}.$$

[4] d'Aspremont 2008; Devolder, Glineur, and Nesterov 2014.

A first gradient-based approach

For any differentiable $f : \mathbb{R} \to \mathbb{R}$

$$\partial_x f(x) = \frac{f(x+\delta) - f(x)}{\delta} + \mathcal{O}(\delta).$$

Then, run *inexact* ($\delta > 0$) gradient descent

$$x_{k+1} = x_k - \mu_k \frac{f(x_k+\delta) - f(x_k)}{\delta}.$$

When does $f(x_k) \to f(x^\star)$?

[4] d'Aspremont 2008; Devolder, Glineur, and Nesterov 2014.

A first gradient-based approach

For any differentiable $f : \mathbb{R} \to \mathbb{R}$

$$\partial_x f(x) = \frac{f(x+\delta) - f(x)}{\delta} + \mathcal{O}(\delta).$$

Then, run *inexact* ($\delta > 0$) gradient descent

$$x_{k+1} = x_k - \mu_k \frac{f(x_k+\delta) - f(x_k)}{\delta}.$$

When does $f(x_k) \to f(x^\star)$? A bias prevails, $f(x_k) \to f(x^\star) + \mathcal{O}(\delta)$.[4]

[4] d'Aspremont 2008; Devolder, Glineur, and Nesterov 2014.

A first gradient-based approach

For any differentiable $f : \mathbb{R} \to \mathbb{R}$

$$\partial_x f(x) = \frac{f(x+\delta) - f(x)}{\delta} + \mathcal{O}(\delta).$$

Then, run *inexact* ($\delta > 0$) gradient descent

$$x_{k+1} = x_k - \mu_k \frac{f(x_k+\delta) - f(x_k)}{\delta}.$$

When does $f(x_k) \to f(x^\star)$? A bias prevails, $f(x_k) \to f(x^\star) + \mathcal{O}(\delta)$.[4]
Similarly, for $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) \approx \sum_{i=1}^{n} \frac{f(x+\delta e_i) - f(x)}{\delta} e_i.$$

[4] d'Aspremont 2008; Devolder, Glineur, and Nesterov 2014.

A second gradient-based approach

Idea, recall

$$\nabla f(x) \approx \sum_{i=1}^{n} \frac{f(x + \delta e_i) - f(x)}{\delta} e_i,$$

A second gradient-based approach

Idea, recall

$$\nabla f(x) \approx \sum_{i=1}^{n} \frac{f(x+\delta e_i)-f(x)}{\delta} e_i,$$

assume we find a *random variable* ξ such that

$$\nabla f(x) \approx \mathbb{E}_{\xi}\left[\frac{f(x+\delta\xi)-f(x)}{\delta}\xi\right], \quad \xi \sim \Xi.$$

Consider the **randomized** algorithm

$$x_{k+1} = x_k - \mu_k \frac{f(x_k+\delta\xi)-f(x_k)}{\delta}\xi, \quad \xi \sim \Xi.$$

A second gradient-based approach

Idea, recall

$$\nabla f(x) \approx \sum_{i=1}^{n} \frac{f(x+\delta e_i) - f(x)}{\delta} e_i,$$

assume we find a *random variable* ξ such that

$$\nabla f(x) \approx \mathbb{E}_{\xi}\left[\frac{f(x+\delta\xi) - f(x)}{\delta}\xi\right], \quad \xi \sim \Xi.$$

Consider the **randomized** algorithm

$$x_{k+1} = x_k - \mu_k \frac{f(x_k+\delta\xi) - f(x_k)}{\delta}\xi, \quad \xi \sim \Xi.$$

Performance criteria is weaker but cleaner $\mathbb{E}_{\xi}[f(x_k)] \to f(x^{\star})$.

Nemirovski and Yudin

Let $f : \mathbb{R}^n \to \mathbb{R}$, Nemirovski and Yudin[5] consider δ-*smoothing*

$$f_\delta(x) = \mathbb{E}_{v\sim\mathbb{B}^n}\left[f(x+\delta v)\right] = \frac{1}{\mathrm{vol}(\mathbb{B}^n)}\int_{\mathbb{B}^n} f(x+\delta v)dv, \tag{2a}$$

$$\nabla f_\delta(x) = \frac{n}{\delta}\mathbb{E}_{u\sim\mathbb{S}^{n-1}}\left[f(x+\delta u)u\right] = \frac{n}{\delta}\frac{1}{\mathrm{vol}(\mathbb{S}^{n-1})}\int_{\mathbb{S}^{n-1}} f(x+\delta u)\frac{u}{\|u\|_2}du. \tag{2b}$$

[5] Nemirovsky and Yudin 1983, credits usually given to Flaxman, Kalai, and McMahan 2004.

[6] Agarwal, Dekel, and Xiao 2010; Nesterov 2011.

Nemirovski and Yudin

Let $f : \mathbb{R}^n \to \mathbb{R}$, Nemirovski and Yudin[5] consider δ-*smoothing*

$$f_\delta(x) = \mathbb{E}_{v \sim \mathbb{B}^n}\left[f(x + \delta v)\right] = \frac{1}{\mathrm{vol}(\mathbb{B}^n)} \int_{\mathbb{B}^n} f(x + \delta v) dv, \tag{2a}$$

$$\nabla f_\delta(x) = \frac{n}{\delta} \mathbb{E}_{u \sim \mathbb{S}^{n-1}}\left[f(x + \delta u) u\right] = \frac{n}{\delta} \frac{1}{\mathrm{vol}(\mathbb{S}^{n-1})} \int_{\mathbb{S}^{n-1}} f(x + \delta u) \frac{u}{\|u\|_2} du. \tag{2b}$$

Natural *one-point* candidate to approximate ∇f

$$g_\delta(x) = \frac{n}{\delta} f(x + \delta u) u, \quad u \sim \mathbb{S}^{n-1}. \tag{3a}$$

---

[5] Nemirovsky and Yudin 1983, credits usually given to Flaxman, Kalai, and McMahan 2004.

[6] Agarwal, Dekel, and Xiao 2010; Nesterov 2011.

Nemirovski and Yudin

Let $f : \mathbb{R}^n \to \mathbb{R}$, Nemirovski and Yudin[5] consider δ-*smoothing*

$$f_\delta(x) = \mathbb{E}_{v\sim\mathbb{B}^n}\left[f(x+\delta v)\right] = \frac{1}{\mathrm{vol}(\mathbb{B}^n)}\int_{\mathbb{B}^n} f(x+\delta v)dv, \tag{2a}$$

$$\nabla f_\delta(x) = \frac{n}{\delta}\mathbb{E}_{u\sim\mathbb{S}^{n-1}}\left[f(x+\delta u)u\right] = \frac{n}{\delta}\frac{1}{\mathrm{vol}(\mathbb{S}^{n-1})}\int_{\mathbb{S}^{n-1}} f(x+\delta u)\frac{u}{\|u\|_2}du. \tag{2b}$$

Natural *one-point* candidate to approximate ∇f

$$g_\delta(x) = \frac{n}{\delta}f(x+\delta u)u, \quad u\sim\mathbb{S}^{n-1}. \tag{3a}$$

Observation[6]: give (3a) again the interpretation of a **directional derivative** and use the *multi-point* oracle

$$g'_\delta(x) = \frac{n}{\delta}\left(f(x+\delta u) - f(x)\right)u, \quad u\sim\mathbb{S}^{n-1}. \tag{3b}$$

[5] Nemirovsky and Yudin 1983, credits usually given to Flaxman, Kalai, and McMahan 2004.

[6] Agarwal, Dekel, and Xiao 2010; Nesterov 2011.

Nesterov and Spokoiny

For $f : \mathbb{R}^n \to \mathbb{R}$ (locally convex), *Gaussian* smoothing[7]

$$f_\gamma(x) = \frac{1}{\kappa} \int_{\mathbb{R}^n} f(x + \gamma u) e^{-\frac{1}{2}\|u\|_2^2} du \tag{4a}$$

$$\nabla f_\gamma(x) = \frac{1}{\kappa} \int_{\mathbb{R}^n} \frac{f(x + \gamma u) - f(x - \gamma u)}{2\gamma} e^{-\frac{1}{2}\|u\|_2^2} u du \tag{4b}$$

with $\|\nabla f - \nabla f_\gamma\| = \mathcal{O}(n\gamma^2)$.

[7] Nesterov 2011; Nesterov and Spokoiny 2017.

Nesterov and Spokoiny

For $f : \mathbb{R}^n \to \mathbb{R}$ (locally convex), *Gaussian* smoothing[7]

$$f_\gamma(x) = \frac{1}{\kappa} \int_{\mathbb{R}^n} f(x + \gamma u) e^{-\frac{1}{2}\|u\|_2^2} du \tag{4a}$$

$$\nabla f_\gamma(x) = \frac{1}{\kappa} \int_{\mathbb{R}^n} \frac{f(x + \gamma u) - f(x - \gamma u)}{2\gamma} e^{-\frac{1}{2}\|u\|_2^2} u du \tag{4b}$$

with $\|\nabla f - \nabla f_\gamma\| = \mathcal{O}(n\gamma^2)$.

$$\textbf{Oracle}: \quad g_\gamma(x) = \frac{f(x + \gamma u) - f(x - \gamma u)}{2\gamma} u, \quad u \sim \mathcal{N}(0, I_n)$$

with $\mathbb{E}_{u \sim \mathcal{N}(0, I_n)} \left[\|g_\gamma(x)\|_2^2 \right] \leq \mathcal{O}(n^2\gamma^2 + n\|\nabla f(x)\|_2^2)$.

Algorithm: $x_{k+1} = x_k - \mu_k g_{\gamma_k}(x_k), \quad \mu_k = \mathcal{O}\left(\frac{1}{n \cdot L}\right)$.

Performance: for $\gamma_k \to 0$ and $\bar{x}_{K-1} := \frac{1}{K} \sum_{k=0}^{K-1} x_k$

$$\mathbb{E}[f(\bar{x}_{K-1})] - f(x^\star) \leq \mathcal{O}\left(\frac{n \cdot L \cdot \|x_0 - x^\star\|_2^2}{K}\right) = \mathcal{O}(n) \cdot \text{ gradient descent}$$

[7] Nesterov 2011; Nesterov and Spokoiny 2017.

Numerical considerations

All common oracles of the form

$$\textit{finite (forward) difference}\text{: } \frac{f(x+\delta, u) - f(x)}{\delta} u$$

$$\textit{central difference}\text{: } \frac{f(x+\delta u) - f(x-\delta u)}{2\delta} u,$$

Numerical considerations

All common oracles of the form

$$\textit{finite (forward) difference}: \ \frac{f(x+\delta, u) - f(x)}{\delta} u$$

$$\textit{central difference}: \ \frac{f(x+\delta u) - f(x-\delta u)}{2\delta} u,$$

with approximation errors $\mathcal{O}(\delta^p)$, $p \geq 1$, algorithms require $\delta_k = \mathcal{O}(\frac{1}{k})$.

Can we pick $\delta \downarrow 0$?

Numerical considerations

All common oracles of the form

$$\textit{finite (forward) difference}: \quad \frac{f(x+\delta, u) - f(x)}{\delta} u$$

$$\textit{central difference}: \quad \frac{f(x+\delta u) - f(x-\delta u)}{2\delta} u,$$

with approximation errors $\mathcal{O}(\delta^p)$, $p \geq 1$, algorithms require $\delta_k = \mathcal{O}(\frac{1}{k})$.

Can we pick $\delta \downarrow 0$?

For small δ, $f(x+\delta u) - f(x) \leq$ machine precision: *cancellation error*.

Numerical considerations

All common oracles of the form

$$\textit{finite (forward) difference}: \ \frac{f(x+\delta, u) - f(x)}{\delta} u$$

$$\textit{central difference}: \ \frac{f(x+\delta u) - f(x-\delta u)}{2\delta} u,$$

with approximation errors $\mathcal{O}(\delta^p)$, $p \geq 1$, algorithms require $\delta_k = \mathcal{O}(\frac{1}{k})$.

Can we pick $\delta \downarrow 0$?

For small δ, $f(x+\delta u) - f(x) \leq$ machine precision: *cancellation error*.

Ignored (?) in most optimization literature.

Beautiful insight from complex analysis

As pioneered by[8], let $f : \mathbb{R} \to \mathbb{R}$ be real-analytic ($f \in C^\omega(\mathbb{R})$) and consider

$$f(x + i\delta) = f(x) + \partial_x f(x) i\delta - \frac{1}{2}\partial_x^2 f(x)\delta^2 - \frac{1}{6}\partial_x^3 f(x) i\delta^3 + O(\delta^4), \quad i^2 = -1.$$

[8] Lyness and Moler 1967; Squire and Trapp 1998; Martins, Sturdza, and Alonso 2003; Abreu et al. 2018.

[9] A value of $\delta = 10^{-100}$ (!) is successfully used in National Physical Laboratory software Cox and Harris 2004, Page 44.

Beautiful insight from complex analysis

As pioneered by[8], let $f : \mathbb{R} \to \mathbb{R}$ be real-analytic ($f \in C^\omega(\mathbb{R})$) and consider

$$f(x+i\delta) = f(x) + \partial_x f(x) i\delta - \frac{1}{2}\partial_x^2 f(x)\delta^2 - \frac{1}{6}\partial_x^3 f(x) i\delta^3 + O(\delta^4), \quad i^2 = -1.$$

such that (for $z \in \mathbb{C}$, $z = \Re(z) + \Im(z)$):

$$\Im\left(f(x+i\delta)\right) = \partial_x f(x)\delta - \frac{1}{6}\partial_x^3 f(x)\delta^3 + O(\delta^5)$$

[8] Lyness and Moler 1967; Squire and Trapp 1998; Martins, Sturdza, and Alonso 2003; Abreu et al. 2018.

[9] A value of $\delta = 10^{-100}$ (!) is successfully used in National Physical Laboratory software Cox and Harris 2004, Page 44.

Beautiful insight from complex analysis

As pioneered by[8], let $f : \mathbb{R} \to \mathbb{R}$ be real-analytic ($f \in C^\omega(\mathbb{R})$) and consider

$$f(x+i\delta) = f(x) + \partial_x f(x) i\delta - \frac{1}{2}\partial_x^2 f(x)\delta^2 - \frac{1}{6}\partial_x^3 f(x) i\delta^3 + O(\delta^4), \quad i^2 = -1.$$

such that (for $z \in \mathbb{C}$, $z = \Re(z) + \Im(z)$):

$$\Im\left(f(x+i\delta)\right) = \partial_x f(x)\delta - \frac{1}{6}\partial_x^3 f(x)\delta^3 + O(\delta^5)$$

and thus

$$\partial_x f(x) = \frac{\Im\Big(f(x+i\delta)\Big)}{\delta} + O(\delta^2), \quad f(x) = \Re(f(x+i\delta)) + O(\delta^2).$$

[8] Lyness and Moler 1967; Squire and Trapp 1998; Martins, Sturdza, and Alonso 2003; Abreu et al. 2018.

[9] A value of $\delta = 10^{-100}$ (!) is successfully used in National Physical Laboratory software Cox and Harris 2004, Page 44.

Beautiful insight from complex analysis

As pioneered by[8], let $f : \mathbb{R} \to \mathbb{R}$ be real-analytic ($f \in C^\omega(\mathbb{R})$) and consider

$$f(x+i\delta) = f(x) + \partial_x f(x) i\delta - \frac{1}{2}\partial_x^2 f(x)\delta^2 - \frac{1}{6}\partial_x^3 f(x) i\delta^3 + O(\delta^4), \quad i^2 = -1.$$

such that (for $z \in \mathbb{C}$, $z = \Re(z) + \Im(z)$):

$$\Im\left(f(x+i\delta)\right) = \partial_x f(x)\delta - \frac{1}{6}\partial_x^3 f(x)\delta^3 + O(\delta^5)$$

and thus

$$\partial_x f(x) = \frac{\Im\big(f(x+i\delta)\big)}{\delta} + O(\delta^2), \quad f(x) = \Re(f(x+i\delta)) + O(\delta^2).$$

Cancellation errors are impossible[9].

[8] Lyness and Moler 1967; Squire and Trapp 1998; Martins, Sturdza, and Alonso 2003; Abreu et al. 2018.

[9] A value of $\delta = 10^{-100}$ (!) is successfully used in National Physical Laboratory software Cox and Harris 2004, Page 44.

Example

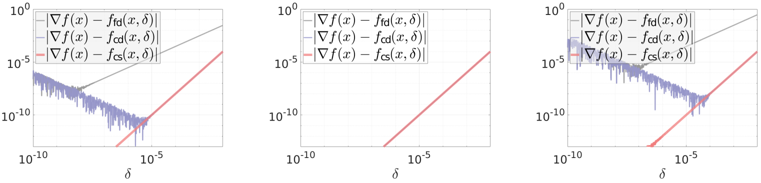
For $f(x) = x^3$, approximate $\nabla f(x)$ at $x \in \{-1, 0, 10\}$ using

$$\text{(forward difference)} \quad f_{\mathsf{fd}}(x, \delta) = \frac{f(x+\delta) - f(x)}{\delta}, \tag{5a}$$

$$\text{(central difference)} \quad f_{\mathsf{cd}}(x, \delta) = \frac{f(x+\delta) - f(x-\delta)}{2\delta}, \tag{5b}$$

$$\text{(complex step)} \quad f_{\mathsf{cs}}(x, \delta) = \frac{\Im\left(f(x+i\delta)\right)}{\delta} \tag{5c}$$

and compare the error for $\delta \downarrow 0$.



(a) $x = -1$ (b) $x = 0$ (c) $x = 10$

Complex-step oracle[11]

Let $f \in C^\omega(\mathcal{D})$, then

$$f_\delta(x) = \mathbb{E}_{v \sim \mathbb{B}^n}\left[\Re\left(f(x + i\delta v)\right)\right]$$
$$\nabla f_\delta(x) = \tfrac{n}{\delta} \cdot \mathbb{E}_{u \sim \mathbb{S}^{n-1}}\left[\Im\left(f(x + i\delta u)\right) u\right]$$

with $\|\nabla f_\delta - \nabla f\|_2 \leq \mathcal{O}(n\delta^2)$.

[10] The paper provides similar results for strong-convex and non-convex functions.

[11] Jongeneel, Yue, and Kuhn 2021.

Complex-step oracle[11]

Let $f \in C^{\omega}(\mathcal{D})$, then

$$f_{\delta}(x) = \mathbb{E}_{v \sim \mathbb{B}^n}\left[\Re\left(f(x + i\delta v)\right)\right]$$
$$\nabla f_{\delta}(x) = \tfrac{n}{\delta} \cdot \mathbb{E}_{u \sim \mathbb{S}^{n-1}}\left[\Im\left(f(x + i\delta u)\right) u\right]$$

with $\|\nabla f_{\delta} - \nabla f\|_2 \leq \mathcal{O}(n\delta^2)$.

Oracle: $g_{\delta}(x) = \tfrac{n}{\delta}\Im\left(f(x + i\delta u)\right) u, \quad u \sim \mathbb{S}^{n-1}.$

with $\mathbb{E}_{u \sim \mathbb{S}^{n-1}}\left[\|g_{\delta}(x)\|_2^2\right] \leq \mathcal{O}(n^2\delta^2 + n\|\nabla f(x)\|_2^2)$.

Algorithm: $x_{k+1} = x_k - \mu_k g_{\delta_k}(x_k), \quad \mu_k = \mathcal{O}\left(\frac{1}{n \cdot L}\right)$

Performance: for f *convex* $\delta_{k-1} = \mathcal{O}(\frac{1}{k})$ and $\bar{x}_{K-1} := \frac{1}{K}\sum_{k=0}^{K-1} x_k$

[10] The paper provides similar results for strong-convex and non-convex functions.

[11] Jongeneel, Yue, and Kuhn 2021.

Complex-step oracle[11]

Let $f \in C^{\omega}(\mathcal{D})$, then

$$f_{\delta}(x) = \mathbb{E}_{v \sim \mathbb{B}^n}\left[\Re\left(f(x + i\delta v)\right)\right]$$
$$\nabla f_{\delta}(x) = \tfrac{n}{\delta} \cdot \mathbb{E}_{u \sim \mathbb{S}^{n-1}}\left[\Im\left(f(x + i\delta u)\right) u\right]$$

with $\|\nabla f_{\delta} - \nabla f\|_2 \leq \mathcal{O}(n\delta^2)$.

Oracle: $g_{\delta}(x) = \tfrac{n}{\delta}\Im\left(f(x + i\delta u)\right) u, \quad u \sim \mathbb{S}^{n-1}.$

with $\mathbb{E}_{u \sim \mathbb{S}^{n-1}}\left[\|g_{\delta}(x)\|_2^2\right] \leq \mathcal{O}(n^2\delta^2 + n\|\nabla f(x)\|_2^2)$.

Algorithm: $x_{k+1} = x_k - \mu_k g_{\delta_k}(x_k), \quad \mu_k = \mathcal{O}\left(\frac{1}{n \cdot L}\right)$

Performance: for f *convex* $\delta_{k-1} = \mathcal{O}(\frac{1}{k})$ and $\bar{x}_{K-1} := \frac{1}{K}\sum_{k=0}^{K-1} x_k$

$$\mathbb{E}[f(\bar{x}_{K-1})] - f(x^{\star}) \leq \mathcal{O}\left(\frac{n \cdot L \cdot \|x_0 - x^{\star}\|_2^2}{K}\right) = \mathcal{O}(n) \cdot \text{ gradient descent}^{10}.$$

[10] The paper provides similar results for strong-convex and non-convex functions.

[11] Jongeneel, Yue, and Kuhn 2021.

Example: worst function in the world

Consider the test function from Nesterov 2003, Section 2.1.2

$$f_n(x) = L\left(\frac{1}{2}\left[(x^{(1)})^2 + \sum_{i=1}^{n-1}(x^{(i+1)} - x^{(i)})^2 + (x^{(n)})^2\right] - x^{(1)}\right) \qquad (7)$$

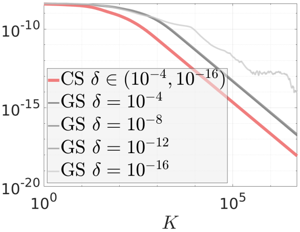
for $x_0 = 0$, $L = 10^{-8}$, $L_1(f) = 4L$ and $(x^\star)^{(i)} = 1 - i/(n+1)$ with $x^{(i)}$.

Example: worst function in the world

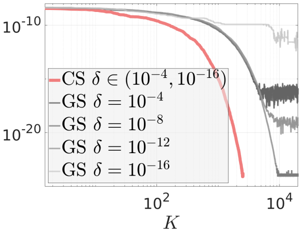
Consider the test function from Nesterov 2003, Section 2.1.2

$$f_n(x) = L\left(\frac{1}{2}\left[(x^{(1)})^2 + \sum_{i=1}^{n-1}(x^{(i+1)} - x^{(i)})^2 + (x^{(n)})^2\right] - x^{(1)}\right) \qquad (7)$$

for $x_0 = 0$, $L = 10^{-8}$, $L_1(f) = 4L$ and $(x^\star)^{(i)} = 1 - i/(n+1)$ with $x^{(i)}$.



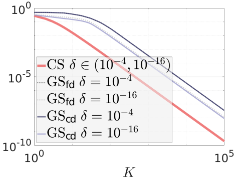
(ia) Suboptimality gap $f(\bar{x}_K) - f^\star$ for the test function (7).



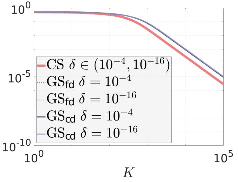
(ib) Suboptimality gap $f(x_K) - f^\star$ for the test function (7).

Figure: The single-point Complex-smoothing (CS) compared to the multi-point Gaussian smoothing (GS) method from Nesterov and Spokoiny 2017.

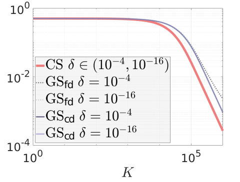
Example: strong convexity $f(x) = \frac{1}{2}\|x\|_2^2$



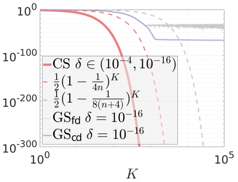
(a) $f(\bar{x}_K) - f^\star$, $n = 10^0$.



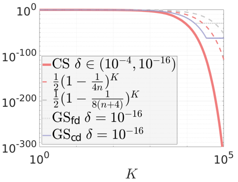
(b) $f(\bar{x}_K) - f^\star$, $n = 10^2$.



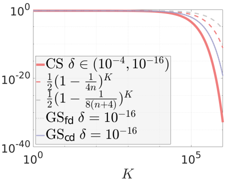
(c) $f(\bar{x}_K) - f^\star$, $n = 10^4$.



(d) $f(x_K) - f^\star$, $n = 10^0$.



(e) $f(x_K) - f^\star$, $n = 10^2$.



(f) $f(x_K) - f^\star$, $n = 10^4$.

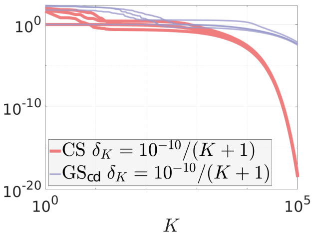
Figure: The single-point Complex-smoothing (CS) compared to the multi-point Gaussian smoothing (GS) method from Nesterov and Spokoiny 2017, Equation (55).

Example: non-convex optimization

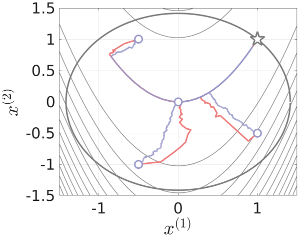
Consider a *Rosenbrock* optimization problem

$$\underset{x \in \sqrt{2}\mathbb{B}^2}{\text{minimize}} \quad (1 - x^{(1)})^2 + 100\left((x^{(2)} - (x^{(1)})^2\right)^2. \tag{8}$$

with $x^\star = (1, 1)$.



(a) Suboptimality gap $f(x_K) - f^\star$ for (8).



(b) Paths taken corresponding to Figure 4a.

Figure: The single-point Complex-smoothing (CS) method versus Gaussian-smoothing Nesterov and Spokoiny 2017.

The End

Many open problems remain. For more, see

(a) Arkadi Semenovich Nemirovsky and David Borisovich Yudin (1983). "Problem complexity and method efficiency in optimization.". In:

(b) Boris Teodorovich Polyak and Aleksandr Borisovich Tsybakov (1990). "Optimal order of accuracy of search algorithms in stochastic optimization". In: *Problemy Peredachi Informatsii* 26.2, pp. 45–53

(c) Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan (2004). "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *CoRR*

(d) John C Duchi et al. (2015). "Optimal rates for zero-order convex optimization: The power of two function evaluations". In: *IEEE Transactions on Information Theory* 61.5, pp. 2788–2806

(e) Francis Bach and Vianney Perchet (2016). "Highly-smooth zero-th order online optimization". In: *Conference on Learning Theory*, pp. 1–27

(f) Yurii Nesterov and Vladimir Spokoiny (2017). "Random gradient-free minimization of convex functions". In: *Foundations of Computational Mathematics* 17.2, pp. 527–566

(g) **Wouter Jongeneel, Man-Chung Yue, and Daniel Kuhn (2021). "Small Errors in Random Zeroth Order Optimization are Imaginary". In: arXiv:** 2103.05478

Thank you! contact: `wjongeneel.nl` or `wouter.jongeneel@epfl.ch`