Adjusting the Centred Simplex Gradient to Compensate for Misaligned Sample Points

Yiwen Chen[1,2]

[1]School of Mathematical Sciences
Beihang University

[2]Department of Mathematics
University of British Columbia

July 20, 2022

Joint work with Dr. Warren Hare

Outline

Simplex Gradient, SG

Consider

$$f : \mathbb{R}^n \to \mathbb{R}$$

and a set

$$\mathbb{Y} = \{y_0, y_0 + d_1, ..., y_0 + d_n\}$$

poised for linear interpolation

The Simplex Gradient of f over $\mathbb{Y}$, denoted by $\nabla_S f(\mathbb{Y})$, is the gradient of the linear interpolation of f over $\mathbb{Y}$

An equivalent definition

Suppose $\mathbb{Y}$ is poised
Then

$$\nabla_S f(\mathbb{Y}) = L^{-\top} \delta_S^{f(\mathbb{Y})}$$

where

$$L = L(\mathbb{Y}) = [y_0 + d_1 - y_0 \cdots y_0 + d_n - y_0] = [d_1 \cdots d_n],$$

$$\delta_S^{f(\mathbb{Y})} = \begin{bmatrix} f(y_0 + d_1) - f(y_0) \\ \vdots \\ f(y_0 + d_n) - f(y_0) \end{bmatrix}$$

Centred Simplex Gradient, CSG

The reflection a poised set $\mathbb{Y}^+ = \{y_0, y_0 + d_1, \ldots, y_0 + d_n\}$ through y_0

- $\mathbb{Y}^- = \{y_0, y_0 - d_1, \ldots, y_0 - d_n\}$
- $\mathbb{Y}^-$ is also poised

The Centred Simplex Gradient of f over $\mathbb{Y} = \mathbb{Y}^+ \cup \mathbb{Y}^-$, denoted by $\nabla_{CS} f(\mathbb{Y})$, is given by

$$\nabla_{CS} f(\mathbb{Y}) = \frac{1}{2}\left(\nabla_S f(\mathbb{Y}^+) + \nabla_S f(\mathbb{Y}^-)\right)$$

Centred Simplex Gradient, CSG

The reflection a poised set $\mathbb{Y}^+ = \{y_0, y_0 + d_1, ..., y_0 + d_n\}$ through y_0

- $\mathbb{Y}^- = \{y_0, y_0 - d_1, ..., y_0 - d_n\}$
- $\mathbb{Y}^-$ is also poised

The Centred Simplex Gradient of f over $\mathbb{Y} = \mathbb{Y}^+ \cup \mathbb{Y}^-$, denoted by $\nabla_{CS} f(\mathbb{Y})$, is given by

$$\nabla_{CS} f(\mathbb{Y}) = \frac{1}{2}\left(\nabla_S f(\mathbb{Y}^+) + \nabla_S f(\mathbb{Y}^-)\right)$$

- Is equivalent to

$$\nabla_{CS} f(\mathbb{Y}) = \left(L^+ - L^-\right)^{-\top}\left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\mathbb{Y}^-)}\right)$$

where $L^+ = L(\mathbb{Y}^+), \quad L^- = L(\mathbb{Y}^-)$

Approximation accuracy

Let $\Delta = \overline{\text{diam}}(\mathbb{Y}^+) := \max_i \{\|y_i - y_0\|\}$

Then

- $\|\nabla f(y_0) - \nabla_S f(y_0)\| = \mathcal{O}(\Delta)$
- $\|\nabla f(y_0) - \nabla_{CS} f(y_0)\| = \mathcal{O}(\Delta^2)$

In this talk...

CSG requires

- A point of interest y_0
- A poised set $\mathbb{Y}^+$
- An exact reflection set $\mathbb{Y}^-$

What if the reflection set of $\mathbb{Y}^+$ is not exact?

Misaligned reflection set

Suppose the reflection set of $\mathbb{Y}^+$ is not exact

- $\widetilde{\mathbb{Y}}$ is provided instead, with $\widetilde{\mathbb{Y}} \approx \mathbb{Y}^-$

Misaligned reflection set

Suppose the reflection set of $\mathbb{Y}^+$ is not exact

- $\widetilde{\mathbb{Y}}$ is provided instead, with $\widetilde{\mathbb{Y}} \approx \mathbb{Y}^-$

Centred Simplex Gradient:

$$\nabla_{CS} f(\mathbb{Y}) = \left(L^+ - L^-\right)^{-\top} \left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\mathbb{Y}^-)}\right)$$

Misaligned reflection set

Suppose the reflection set of $\mathbb{Y}^+$ is not exact

- $\widetilde{\mathbb{Y}}$ is provided instead, with $\widetilde{\mathbb{Y}} \approx \mathbb{Y}^-$

Centred Simplex Gradient:

$$\nabla_{CS} f(\mathbb{Y}) = \left(L^+ - L^-\right)^{-\top} \left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\mathbb{Y}^-)}\right)$$

An ‘obvious’ approximate gradient would be

$$\nabla f(y_0) \approx \left(L^+ - \widetilde{L}\right)^{-\top} \left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right)$$

However...

Approximation accuracy

Let $\Delta = \overline{\mathrm{diam}}(\mathbb{Y}^+) := \max_i \{\|y_i - y_0\|\}$

Then

- $\|\nabla f(y_0) - \nabla_S f(y_0)\| = \mathcal{O}(\Delta)$
- $\|\nabla f(y_0) - \nabla_{CS} f(y_0)\| = \mathcal{O}(\Delta^2)$

Approximation accuracy

Let $\Delta = \overline{\mathrm{diam}}(\mathbb{Y}^+) := \max_i \{\|y_i - y_0\|\}$

Then

- $\|\nabla f(y_0) - \nabla_S f(y_0)\| = \mathcal{O}(\Delta)$
- $\|\nabla f(y_0) - \nabla_{CS} f(y_0)\| = \mathcal{O}(\Delta^2)$
- $\left\|\nabla f(y_0) - \left(L^+ - \widetilde{L}\right)^{-\top} \left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right)\right\| = \mathcal{O}(\Delta^2)$?

A simple example

Example. Let $f = x^2, y_0 = 0, \mathbb{Y}^+ = \{0, \Delta\}, \widetilde{\mathbb{Y}} = \{0, -0.9\Delta\}$
Then

$$\begin{aligned} L^+ &= \Delta \\ \widetilde{L} &= -0.9\Delta \\ \delta_S^{f(\mathbb{Y}^+)} &= f(\Delta) - f(0) = \Delta^2 \\ \delta_S^{f(\widetilde{\mathbb{Y}})} &= f(-0.9\Delta) - f(0) = 0.81\Delta^2 \end{aligned}$$

So

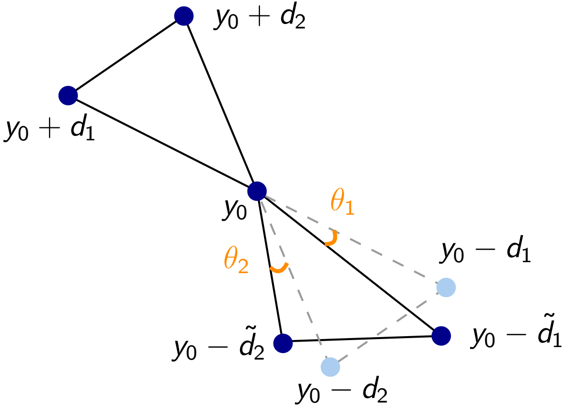
$$\left(L^+ - \widetilde{L}\right)^{-\top} \left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right) = 0.1\Delta$$

$$\left\| \nabla f(y_0) - \left(L^+ - \widetilde{L}\right)^{-\top} \left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right) \right\| = 0.1\Delta = \mathcal{O}(\Delta)$$

Outline

Structure of $\widetilde{\mathbb{Y}}$ relative to $\mathbb{Y}^-$

Let $\mathbb{Y}^+ = \{y_0, y_0 + d_1, \ldots, y_0 + d_n\}, \widetilde{\mathbb{Y}} = \{y_0, y_0 - \widetilde{d}_1, \ldots, y_0 - \widetilde{d}_n\}$



Structure of $\widetilde{\mathbb{Y}}$ relative to $\mathbb{Y}^-$

For all $i \in \{1, ..., n\}$,

- The Stretching Parameter k_i is given by

$$k_i = \frac{\left\|\widetilde{d}_i\right\|}{\|d_i\|}$$

Structure of $\widetilde{\mathbb{Y}}$ relative to $\mathbb{Y}^-$

For all $i \in \{1, ..., n\}$,

- The Stretching Parameter k_i is given by

$$k_i = \frac{\left\|\widetilde{d}_i\right\|}{\|d_i\|}$$

- The Rotation Angle θ_i is the angle between d_i and $\widetilde{d}_i$, given by

$$\theta_i = \cos^{-1}\left(\frac{d_i^\top \widetilde{d}_i}{\|d_i\| \left\|\widetilde{d}_i\right\|}\right) \in [0, \pi]$$

Adapted Centred Simplex Gradient, ACSG

The Adapted Centered Simplex Gradient of f over $\mathbb{Y} = \mathbb{Y}^+ \cup \widetilde{\mathbb{Y}}$, denoted by $\nabla_{ACS} f(\mathbb{Y})$, is given by

$$\nabla_{ACS} f(\mathbb{Y}) = \left(L^+ D - \widetilde{L}\right)^{-\top} \left(D\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right)$$

where

$$D = \begin{bmatrix} k_1^2 & & \\ & \ddots & \\ & & k_n^2 \end{bmatrix}$$

Adapted Centred Simplex Gradient, ACSG

The Adapted Centered Simplex Gradient of f over $\mathbb{Y} = \mathbb{Y}^+ \cup \widetilde{\mathbb{Y}}$, denoted by $\nabla_{ACS} f(\mathbb{Y})$, is given by

$$\nabla_{ACS} f(\mathbb{Y}) = \left(L^+ D - \widetilde{L}\right)^{-\top} \left(D \delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right)$$

where

$$D = \begin{bmatrix} k_1^2 & & \\ & \ddots & \\ & & k_n^2 \end{bmatrix}$$

Note:

- When all $\theta_i = 0$ and $k_i = 1$, $\nabla_{ACS} f(\mathbb{Y}) = \nabla_{CS} f(\mathbb{Y})$

Outline

Error bound

Theorem. $f \in \mathcal{C}^{2+}$ on $B_{\bar{\Delta}}(y_0)$ with constant C, $\overline{\mathrm{diam}}(\mathbb{Y}^+), \overline{\mathrm{diam}}(\widetilde{\mathbb{Y}}) \leq \bar{\Delta}$
Then

$$\|\nabla f(y_0) - \nabla_{ACS} f(\mathbb{Y})\| \leq \frac{K}{2} \max\left\{k_i^2\right\} \sqrt{n} \left\| \left(\widehat{L}^+ D - \widehat{\widetilde{L}}\right)^{-1} \right\| \max\left\{\theta_i\right\} \Delta + \frac{C}{6} \max\left\{k_i^2 \left(1 + k_i\right)\right\} \sqrt{n} \left\| \left(\widehat{L}^+ D - \widehat{\widetilde{L}}\right)^{-1} \right\| \Delta^2$$

where

$$\Delta = \overline{\mathrm{diam}}(\mathbb{Y}^+), \quad \widehat{L}^+ = \frac{1}{\Delta} L^+, \quad \widehat{\widetilde{L}} = \frac{1}{\Delta} \widetilde{L}$$

Error bound

Theorem. $f \in \mathcal{C}^{2+}$ on $B_{\bar{\Delta}}(y_0)$ with constant C, $\overline{\mathrm{diam}}(\mathbb{Y}^+), \overline{\mathrm{diam}}(\widetilde{\mathbb{Y}}) \leq \bar{\Delta}$
Then

$$\begin{aligned}\|\nabla f(y_0) - \nabla_{ACS} f(\mathbb{Y})\| \leq & \kappa_\theta \max\{\theta_i\}\Delta + \kappa_\Delta \Delta^2 \\ = & \mathcal{O}(\Theta\Delta + \Delta^2)\end{aligned}$$

I.e., ACSG has $\mathcal{O}(\Theta\Delta + \Delta^2)$ accuracy, where $\Theta = \max\{\theta_i\}$

Proof overview

Part 1: Suppose all $\theta_i = 0$, by Taylor expansion

$$f(y_0 + d_i) = f(y_0) + \nabla f(y_0)^\top d_i + \frac{1}{2} d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \tag{1}$$

$$f(y_0 - \widetilde{d}_i) = f(y_0) - k_i \nabla f(y_0)^\top d_i + \frac{1}{2} k_i^2 d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \tag{2}$$

Proof overview

Part 1: Suppose all $\theta_i = 0$, by Taylor expansion

$$f(y_0 + d_i) = f(y_0) + \nabla f(y_0)^\top d_i + \frac{1}{2} d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \quad (1)$$

$$f(y_0 - \widetilde{d}_i) = f(y_0) - k_i \nabla f(y_0)^\top d_i + \frac{1}{2} k_i^2 d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \quad (2)$$

Applying k_i^2(1)-(2), we have

$$\begin{aligned} & k_i^2 \left(f(y_0 + d_i) - f(y_0) \right) - \left(f(y_0 - \widetilde{d}_i) - f(y_0) \right) \\ = & \nabla f(y_0)^\top \left(k_i^2 d_i + k_i d_i \right) + \mathcal{O}(\Delta^3) \end{aligned} \quad (3)$$

Proof overview

Part 1: Suppose all $\theta_i = 0$, by Taylor expansion

$$f(y_0 + d_i) = f(y_0) + \nabla f(y_0)^\top d_i + \frac{1}{2} d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \quad (1)$$

$$f(y_0 - \widetilde{d}_i) = f(y_0) - k_i \nabla f(y_0)^\top d_i + \frac{1}{2} k_i^2 d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \quad (2)$$

Applying k_i^2(1)-(2), we have

$$\begin{aligned} & k_i^2 \left(f(y_0 + d_i) - f(y_0) \right) - \left(f(y_0 - \widetilde{d}_i) - f(y_0) \right) \\ = & \nabla f(y_0)^\top \left(k_i^2 d_i + k_i d_i \right) + \mathcal{O}(\Delta^3) \end{aligned} \quad (3)$$

Using techniques similar to Simplex Gradient analysis, we obtain

$$\|\nabla f(y_0) - \nabla_{ACS} f(\mathbb{Y})\| \leq \kappa_\Delta \Delta^2$$

Part 2: Suppose all $k_i = 1$, by Taylor expansion

$$f(y_0 + d_i) = f(y_0) + \nabla f(y_0)^\top d_i + \frac{1}{2} d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \quad (4)$$

$$f(y_0 - \widetilde{d}_i) = f(y_0) - \nabla f(y_0)^\top A_{\theta_i} d_i + \frac{1}{2} d_i^\top A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i} d_i + \mathcal{O}(\Delta^3) \quad (5)$$

Part 2: Suppose all $k_i = 1$, by Taylor expansion

$$f(y_0 + d_i) = f(y_0) + \nabla f(y_0)^\top d_i + \frac{1}{2} d_i^\top \nabla^2 f(y_0) d_i + \mathcal{O}(\Delta^3) \tag{4}$$

$$f(y_0 - \widetilde{d}_i) = f(y_0) - \nabla f(y_0)^\top A_{\theta_i} d_i + \frac{1}{2} d_i^\top A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i} d_i + \mathcal{O}(\Delta^3) \tag{5}$$

Applying (4)-(5), we have

$$\begin{aligned} & f(y_0 + d_i) - f(y_0 - \widetilde{d}_i) \\ =& \nabla f(y_0)^\top (d_i + A_{\theta_i} d_i) + \frac{1}{2} d_i^\top \left(\nabla^2 f(y_0) - A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i} \right) d_i + \mathcal{O}(\Delta^3) \end{aligned} \tag{6}$$

Separate

$$\nabla^2 f(y_0) - A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i}$$

into two symmetric matrices S_1 and S_2, so

$$\begin{aligned}\left|\frac{1}{2} d_i^\top \left(\nabla^2 f(y_0) - A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i}\right) d_i\right| &\leq \left|\frac{1}{2} d_i^\top S_1 d_i\right| + \left|\frac{1}{2} d_i^\top S_2 d_i\right| \\ &\leq \frac{1}{2} \left(\max\left\{|\lambda_{S_1}|\right\} + \max\left\{|\lambda_{S_2}|\right\}\right) \|d_i\|^2 \\ &\leq \kappa \max\left\{\theta_i\right\} \Delta^2\end{aligned}$$

Separate

$$\nabla^2 f(y_0) - A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i}$$

into two symmetric matrices S_1 and S_2, so

$$\begin{aligned}\left|\frac{1}{2} d_i^\top \left(\nabla^2 f(y_0) - A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i}\right) d_i\right| \leq & \left|\frac{1}{2} d_i^\top S_1 d_i\right| + \left|\frac{1}{2} d_i^\top S_2 d_i\right| \\ \leq & \frac{1}{2}\left(\max\left\{|\lambda_{S_1}|\right\} + \max\left\{|\lambda_{S_2}|\right\}\right)\|d_i\|^2 \\ \leq & \kappa \max\left\{\theta_i\right\} \Delta^2\end{aligned}$$

Using techniques similar to Simplex Gradient analysis, we obtain

$$\|\nabla f(y_0) - \nabla_{ACS} f(\mathbb{Y})\| \leq \kappa_\theta \max\left\{\theta_i\right\} \Delta + \kappa_\Delta \Delta^2$$

Separate

$$\nabla^2 f(y_0) - A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i}$$

into two symmetric matrices S_1 and S_2, so

$$\begin{aligned}\left|\frac{1}{2} d_i^\top \left(\nabla^2 f(y_0) - A_{\theta_i}^\top \nabla^2 f(y_0) A_{\theta_i}\right) d_i\right| \leq & \left|\frac{1}{2} d_i^\top S_1 d_i\right| + \left|\frac{1}{2} d_i^\top S_2 d_i\right| \\ \leq & \frac{1}{2}\left(\max\{|\lambda_{S_1}|\} + \max\{|\lambda_{S_2}|\}\right)\|d_i\|^2 \\ \leq & \kappa \max\{\theta_i\}\Delta^2\end{aligned}$$

Using techniques similar to Simplex Gradient analysis, we obtain

$$\|\nabla f(y_0) - \nabla_{ACS} f(\mathbb{Y})\| \leq \kappa_\theta \max\{\theta_i\}\Delta + \kappa_\Delta \Delta^2$$

Part 3: Combine Part 1 and Part 2

Outline

Comparison to the direct generalization

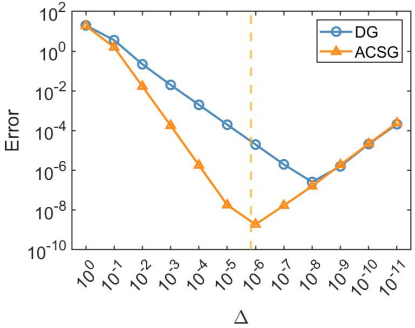
Two formulae:

- $\nabla f(y_0) \approx \left(L^+ - \widetilde{L}\right)^{-\top} \left(\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right)$ (Direct Generalization)
- $\nabla f(y_0) \approx \left(L^+ D - \widetilde{L}\right)^{-\top} \left(D\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right)$ (ACSG)

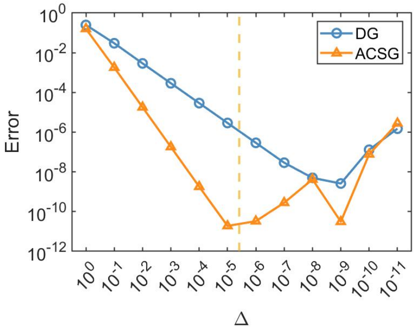
Test problems:

- $f(x) = \sum_{i=1}^{n} [\sin(ix_i) + \cos(ix_i)]$ at $y_0 = \mathrm{e}_1$
- $f(x) = e^{-\|x\|^2}$ at $y_0 = \mathrm{e}_1$

$\theta = 0, k = 0.75$

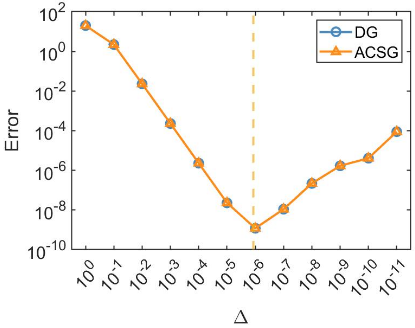


$$f(x) = \sum_{i=1}^{n} [\sin(ix_i) + \cos(ix_i)]$$

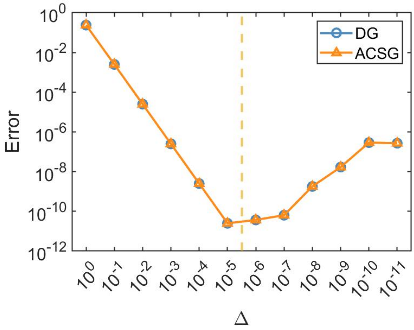


$$f(x) = e^{-\|x\|^2}$$

$\theta = 0, k = 1$

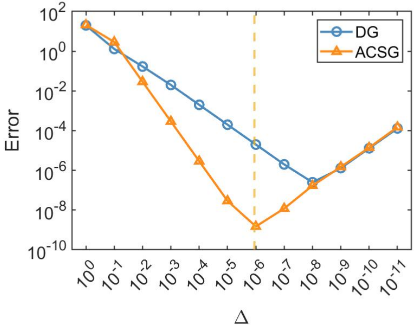


$$f(x) = \sum_{i=1}^{n} [\sin(ix_i) + \cos(ix_i)]$$

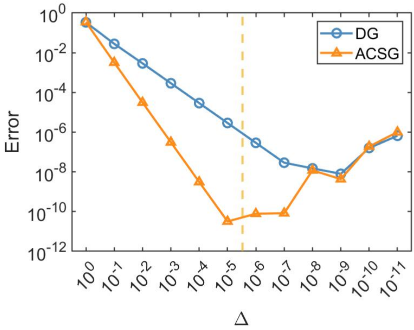


$$f(x) = e^{-\|x\|^2}$$

$\theta = 0, k = 1.25$



$$f(x) = \sum_{i=1}^{n} [\sin(ix_i) + \cos(ix_i)]$$



$$f(x) = e^{-\|x\|^2}$$

Other experiments

Comparison to the direct generalization:

- Fix $\theta = 0.1$, repeat previous experiments
- Result: Similar pattern as $\theta = 0$

Other experiments

Comparison to the direct generalization:

- Fix $\theta = 0.1$, repeat previous experiments
- Result: Similar pattern as $\theta = 0$

Relation of error to θ:

- Fix $k \approx 1$, consider $\theta \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-10}\}$, shrink Δ
- Result: As $\theta \to 0$, error goes from $\mathcal{O}(\Delta)$ to $\mathcal{O}(\Delta^2)$

Recall: ACSG has $\mathcal{O}(\Theta\Delta + \Delta^2)$ accuracy, where $\Theta = \max\{\theta_i\}$

Other experiments

Relation of error to k:

- Fix $\theta = 0$, consider $k \in \{1, 10^1, 10^2, 10^4, 10^{16}, 10^{64}, 10^{128}\}$, shrink Δ
- Result: As $k \to \infty$, ACSG goes from CSG to SG

Proof:

$$
\begin{aligned}
&\lim_{k\to\infty} \left(L^+ D - \widetilde{L}\right)^{-\top} \left(D\delta_S^{f(\mathbb{Y}^+)} - \delta_S^{f(\widetilde{\mathbb{Y}})}\right) \\
=&\lim_{k\to\infty} \left(L^+\right)^{-\top} D^{-\top} D\delta_S^{f(\mathbb{Y}^+)} \\
=&\left(L^+\right)^{-\top} \delta_S^{f(\mathbb{Y}^+)} \\
=&\nabla_S f(\mathbb{Y}^+)
\end{aligned}
$$

Outline

Conclusions

Summary

- We generalized CSG and developed ACSG
- ACSG has $\mathcal{O}(\Theta\Delta + \Delta^2)$ accuracy, where $\Theta = \max\{\theta_i\}$
- When $\mathbb{Y}$ is not perfectly symmetric, ACSG outperforms CSG
- ACSG could be used to reduce functions calls in DFO algorithms

Next steps

- Develop algorithms to find the best pair of $\mathbb{Y}^+$ and $\widetilde{\mathbb{Y}}$ efficiently
- Explore properties of underdetermined and overdetermined ACSG

Thank you

- Chen, Hare. *Adapting the Centred Simplex Gradient to Compensate for Misaligned Sample Points.* Preprint available by emailing `yiwchen@student.ubc.ca` or `warren.hare@ubc.ca`