# <u>Gra</u>dient <u>C</u>ompressed S<u>e</u>nsing: A Query-Efficient Gradient Estimator for High-Dimensional Zeroth-Order Optimization

Ruizhong Qiu† Hanghang Tong† {rq5,htong}@illinois.edu

† UIUC ILLINOIS

ICML International Conference On Machine Learning

## HIGHLIGHTS

### ➢Query-efficient gradient estimator: GraCe
- Only $O(s \log\log(d/s))$ queries per step
- Still $O(1/T)$ convergence for nonconvex optimization

### ➢Relaxed sparsity assumption
- We assume **approximately** $s$-sparse gradients
- Weaker than *exact sparsity* and *compressibility*

### ➢Improvement of Indyk–Price–Woodruff (IPW)
- We reduce its constant by a factor of nearly **4300**
- Via our *dependent random partition* technique

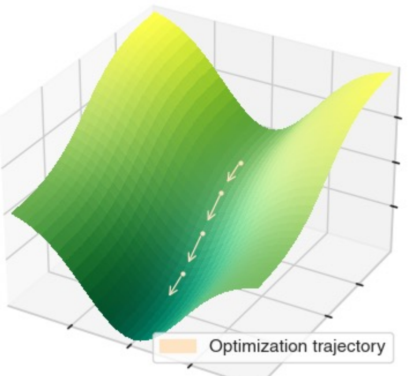
### ➢Strong empirical performance
- Evaluated under **10000**-dimensional functions
- Significantly outperforms 12 existing methods

## BACKGROUND

### ➢High-dim zeroth-order optimization (ZOO)

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$$

- High-dimensional space $\mathbb{R}^d$
- Possibly nonconvex function $f: \mathbb{R}^d \to \mathbb{R}$
- Only queries $f(\boldsymbol{x})$; no gradients $\nabla f(\boldsymbol{x})$
- Aim to use as few queries as possible




### ➢Our main assumption: $\rho$-approximate $s$-sparsity

$$\max_{I\subseteq[d]:\,|I|=s} \|\nabla_I f(\boldsymbol{x})\|_2^2 \ge \rho \|\nabla f(\boldsymbol{x})\|_2^2, \qquad \forall \boldsymbol{x}$$

- Approximation ratio $0 < \rho \le 1$; sparsity $1 \le s \le d$

### ➢Other standard assumptions:
- Lower boundedness: $f_* := \inf_{\boldsymbol{x}} f(\boldsymbol{x}) > -\infty$
- Lipschitz continuity: $|f(\boldsymbol{x}+\boldsymbol{u}) - f(\boldsymbol{x})| \le L_0\|\boldsymbol{u}\|_2$
- Lipschitz smoothness: $\|\nabla f(\boldsymbol{x}+\boldsymbol{u}) - \nabla f(\boldsymbol{x})\|_2 \le L_1\|\boldsymbol{u}\|_2$

## PROPOSED METHOD: GraCe

### ➢Core idea: Locate large-gradient dimensions
- Employ the **IPW** algorithm from compressed sensing
- **Adaptively encode** dimension information into queries

### ➢Motivating case: Approx. 1-sparse gradient
- Suppose that some $j \in [d]$ has sufficiently large $\frac{|\nabla_j f(\boldsymbol{x})|}{\|\nabla_{[d]\setminus j} f(\boldsymbol{x})\|_2}$
- Let $u_i' := \epsilon$ and $v_i' := \epsilon \cdot i$. Then $\frac{f(\boldsymbol{x}+\boldsymbol{v}')-f(\boldsymbol{x})}{f(\boldsymbol{x}+\boldsymbol{u}')-f(\boldsymbol{x})} \approx \frac{\epsilon\cdot j\cdot\nabla_j f(\boldsymbol{x})}{\epsilon\cdot\nabla_j f(\boldsymbol{x})} = j$
- For instance, $f(\boldsymbol{x}) := 0.1x_1 + x_2$ gives $\frac{f(\boldsymbol{0}+\boldsymbol{v}')-f(\boldsymbol{0})}{f(\boldsymbol{0}+\boldsymbol{u}')-f(\boldsymbol{0})} = \frac{2.1\epsilon}{1.1\epsilon} \approx 2$

### ➢General case: IPW + *dependent random partition*
- Divide dimensions into $O(s)$ groups $S$ of *fixed* size $O(d/s)$
- Each group has only 1 large-gradient dimension $j$ w.h.p.
- Use $O(\log\log|S|)$ **adaptive** queries to locate $j \in S$

## THEORETICAL GUARANTEES

### ➢Query complexity: $O(s \log\log(d/s))$
- Given any $\boldsymbol{x} \in \mathbb{R}^d$, any $\epsilon > 0$, and any $0 < \alpha < \rho$
- $O(s \log\log(d/s))$ queries can find a gradient estimate $\boldsymbol{g}$ s.t.

$$\|\boldsymbol{g}\|_2 \le \|\nabla f(\boldsymbol{x})\|_2 + O(\epsilon)$$
$$\mathbb{E}[\langle \nabla f(\boldsymbol{x}), \boldsymbol{g}\rangle | \boldsymbol{x}] \ge \alpha \|\nabla f(\boldsymbol{x})\|_2^2 - O(\epsilon)$$

### ➢Rate of convergence: $O(1/T)$ for nonconvex ZOO
- Given any initial point $\boldsymbol{x}_1 \in \mathbb{R}^d$, any step size $0 < \eta < \rho/L_1$, any $0 < \beta < 1$, and any $\Delta > 0$, under suitable hyperparams
- Suppose gradient descent + GraCe yields points $\boldsymbol{x}_2, \boldsymbol{x}_3, \ldots$
- With probability $\ge 1-\beta$, for all $T \ge 1$ simultaneously,

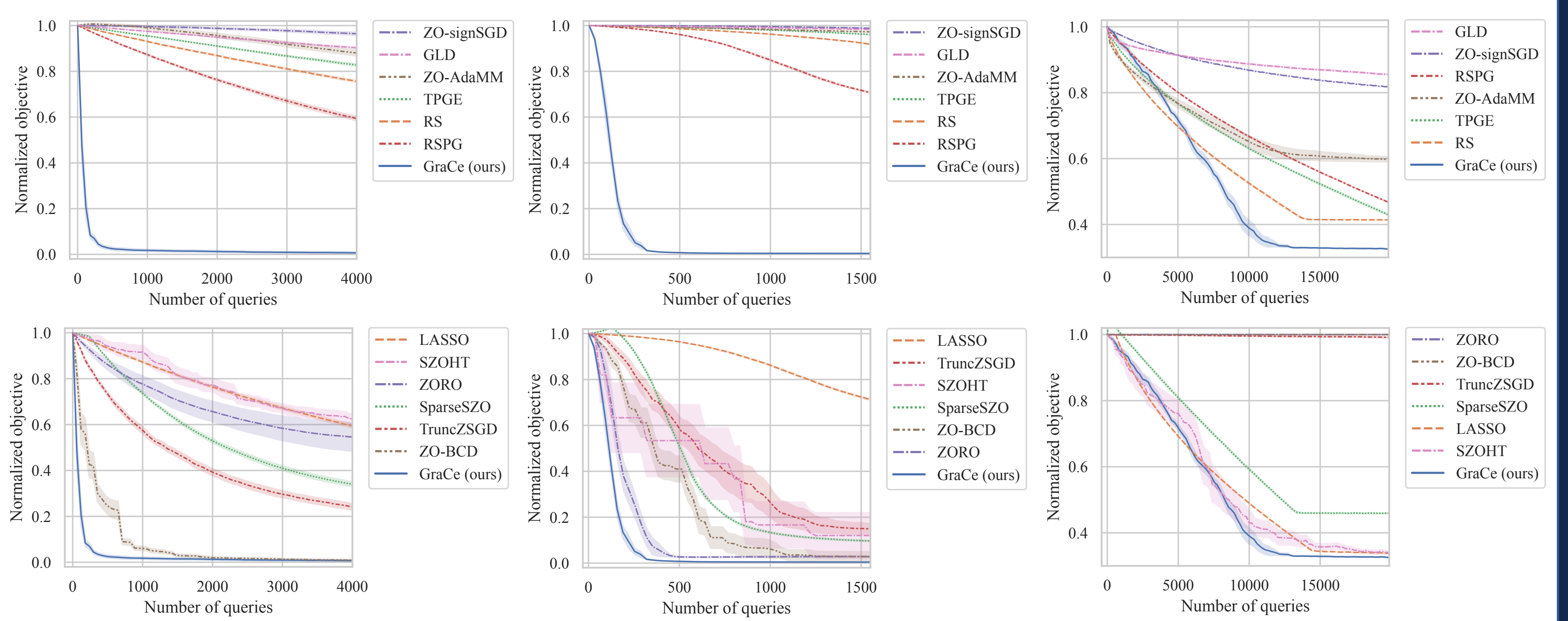
$$\min_{t=1,\ldots,T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \le \frac{\dfrac{1 + \dfrac{2(1-L_1\eta)}{L_1\eta\beta}}{\eta - \dfrac{L_1\eta^2}{2}}(f(\boldsymbol{x}_1) - f_*) + \Delta}{T}$$

## COMPARISON w/ BASELINES

### ➢Theoretical comparison:

| Type | Method | Queries per step | Rate of convergence |
|---|---|---|---|
| Full Gradient | RS (Ghadimi & Lan, 2012) | $O(1)$ | $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|_2^2] \le O(\frac{\sqrt{d}}{\sqrt{T}} + \frac{d}{T})$ |
| | TPGE (Duchi et al., 2015) | $O(1)$ | $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|_2^2] \le O(\frac{\sqrt{d}}{\sqrt{T}})$ |
| | RSPG (Ghadimi et al., 2016) | $O(q)$ | $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|_2^2] \le O(\frac{d}{q} + \frac{d^2}{qT})$ |
| | ZO-signSGD (Liu et al., 2019) | $O(bq)$ | $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|_2] \le O(\frac{\sqrt{d}\sqrt{q+d}}{\sqrt{bq}} + \frac{\sqrt{d}}{\sqrt{T}})$ |
| | ZO-AdaMM (Chen et al., 2019) | $O(1)$ | $\mathbb{E}[\|\nabla f(\boldsymbol{x}_\tau)\|_2^2] \le O(\frac{d}{\sqrt{T}} + \frac{d^2}{T})$ |
| Sparse Gradient | ZORO (Cai et al., 2022) | $O(s \log \frac{d}{s})$ | $\|\nabla f(\boldsymbol{x}_\tau)\|_2^2 \le O(\frac{1}{T})$ w.h.p. |
| | **GraCe** (ours) | $O(s \log\log \frac{d}{s})$ | $\|\nabla f(\boldsymbol{x}_\tau)\|_2^2 \le O(\frac{1}{T})$ w.h.p. |

### ➢Empirical comparison:
- Evaluated under **10000**-dimensional functions
- 2 synthetic functions (left) and 1 real-world function (right)




### ➢Conclusions:
- GraCe consistently outperforms 12 existing ZOO methods
- GraCe achieves **fastest** convergence with **fewest** queries

## ACKNOWLEDGEMENTS

NSF · USDA NIFA · U.S. Department of Homeland Security · Air Force Office of Scientific Research · C3.ai Digital Transformation Institute