Zeroth-Order Hard-Thresholding: Gradient Error vs. Expansivity

William de Vazelhes [1], Hualin Zhang [2], Huimin Wu [2], Xiao-Tong Yuan [2], Bin Gu [1]

[1]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
[2]Nanjing University of Information Science and Technology (NUIST)

Abstract

Problem: $\min_{\boldsymbol{x}\in\mathbb{R}^d} \{f(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}} f(\boldsymbol{x},\boldsymbol{\xi})\}$, s.t. $\|\boldsymbol{x}\|_0 \leq k$

We optimize a function under **hard sparsity constraints** (ℓ_0), with only access to **functions evaluation (ZO)**. We reveal a conflict between the ZO gradient error and the expansivity of hard-thresholding, which results into a **minimum number of random directions** q necessary for our convergence result to hold. We show that the query complexity (QC) is **dimension independent** (if f is smooth), or **weakly dimension dependent** (if f is RSS). We confirm the **efficiency of our algorithm experimentally**.

Related Works

- **StoIHT** [4] Stochastic Hard-Thresholding algorithm (first order (FO))
- **RSPGF** [2] Proximal ZO with ℓ_1 penalty
- **ZSCG** [1] Frank-Wolfe ZO with ℓ_1 ball constraint
- **ZORO** [3] Proximal ZO algorithm with ℓ_1 penalty, retrieving ∇f by CoSaMP

Type	Name	Assumptions	#IZO(=QC)/#IFO
FO/ℓ_0	StoIHT [4]	RSS, RSC	$\mathcal{O}(\kappa \log(\frac{1}{\varepsilon}))$
ZO/ℓ_1	RSPGF [2]	smooth	$\mathcal{O}(\frac{d}{\varepsilon^2})$
ZO/ℓ_1	ZSCG [1]	convex, smooth	$\mathcal{O}(\frac{d}{\varepsilon^2})$
ZO/ℓ_1	ZORO [3]	∇f s-sparse, $\nabla^2 f$ weakly-sparse, f smooth & RSC$_{\text{other}}$	$\mathcal{O}(s \log(d) \log(\frac{1}{\varepsilon}))$
ZO/ℓ_0	**SZOHT**	RSS, RSC	$\mathcal{O}((\boldsymbol{k}+\frac{\boldsymbol{d}}{\boldsymbol{s_2}})\kappa^2 \log(\frac{1}{\varepsilon}))$
ZO/ℓ_0	**SZOHT**	smooth, RSC	$\mathcal{O}(\boldsymbol{k}\kappa^2 \log(\frac{1}{\varepsilon}))$

Assumptions

- (ν_s, s)-RSC: $\forall (\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^d$ s.t. $\|\boldsymbol{x}-\boldsymbol{y}\|_0 \leq s$: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{\nu_s}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2$
- (L_s, s)-RSS: $\forall (\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^d$ s.t. $\|\boldsymbol{x}-\boldsymbol{y}\|_0 \leq s$: $\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}) - \nabla f_{\boldsymbol{\xi}}(\boldsymbol{y})\| \leq L_s\|\boldsymbol{x}-\boldsymbol{y}\|$
- $\sigma^2 := \mathbb{E}_{\boldsymbol{\xi}}[\|\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x}^*)\|_\infty^2]$ is finite

The SZOHT algorithm

Initialization: *learning rate* : η, *max. iter.*: T, *size of support*: s_2, *num. of random directions*: q, *num. of coordinates kept*: $k = \mathcal{O}(\kappa^4 k^*)$, *init.*: $\boldsymbol{x}_0$ *with* $\|\boldsymbol{x}_0\|_0 \leq k^*$ *(e.g.* $\boldsymbol{x}_0 = \boldsymbol{0}$*)*.
Output: $\boldsymbol{x}_T$.
for $t = 1, ..., T$ **do**
  Sample $\boldsymbol{\xi}$ (for instance sample a train sample)
  for $i = 1, ..., q$ **do**
    Sample a random support $S \sim \mathcal{U}(\binom{[d]}{s_2})$
    Sample a random direction $\boldsymbol{u}_i$ from the unit sphere supported on S:
    $\boldsymbol{u}_i \sim \mathcal{U}\left(\mathcal{S}^d_S\right)$
    Compute $\hat{\nabla} f_{\boldsymbol{\xi}}(\boldsymbol{x}_{t-1}; \boldsymbol{u}_i) = \frac{d}{\mu}\left(f_{\boldsymbol{\xi}}(\boldsymbol{x}_{t-1}+\mu\boldsymbol{u}_i) - f_{\boldsymbol{\xi}}(\boldsymbol{x}_{t-1})\right)\boldsymbol{u}_i$
  end
  Compute $\hat{\nabla} f_{\boldsymbol{\xi}}(\boldsymbol{x}_{t-1}) = \frac{1}{q}\sum_{j=1}^q \hat{\nabla} f_{\boldsymbol{\xi}}(\boldsymbol{x}_{t-1}; \boldsymbol{u}_j)$ **# ZO grad.**
  Compute $\boldsymbol{x}_t = \boldsymbol{\Phi_k}(\boldsymbol{x}_{t-1} - \eta\hat{\nabla} f_{\boldsymbol{\xi}}(\boldsymbol{x}_{t-1}))$ **# Hard-Thresholding**
  (Φ_k : keeps only the top-k entries (sets others to 0))
end

Gradient Error

Proposition 1: For a support $F \subset [d]$ of size s, q random directions, and random supports of size s_2, with $f_{\boldsymbol{\xi}}$ (L_{s_2}, s_2)-RSS, with $\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x})$ the hard thresholding of $\nabla f_{\boldsymbol{\xi}}(\boldsymbol{x})$ on F (that is, we set all coordinates not in F to 0), we have:

$$\mathbb{E}\|\hat{\nabla}_F f_{\boldsymbol{\xi}}(\boldsymbol{x}) - \nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x})\|^2 \leq \epsilon_{err}\|\nabla_F f_{\boldsymbol{\xi}}(\boldsymbol{x})\|^2 + C_2\|\nabla_{F^c} f_{\boldsymbol{\xi}}(\boldsymbol{x})\|^2 + C_3\mu^2$$

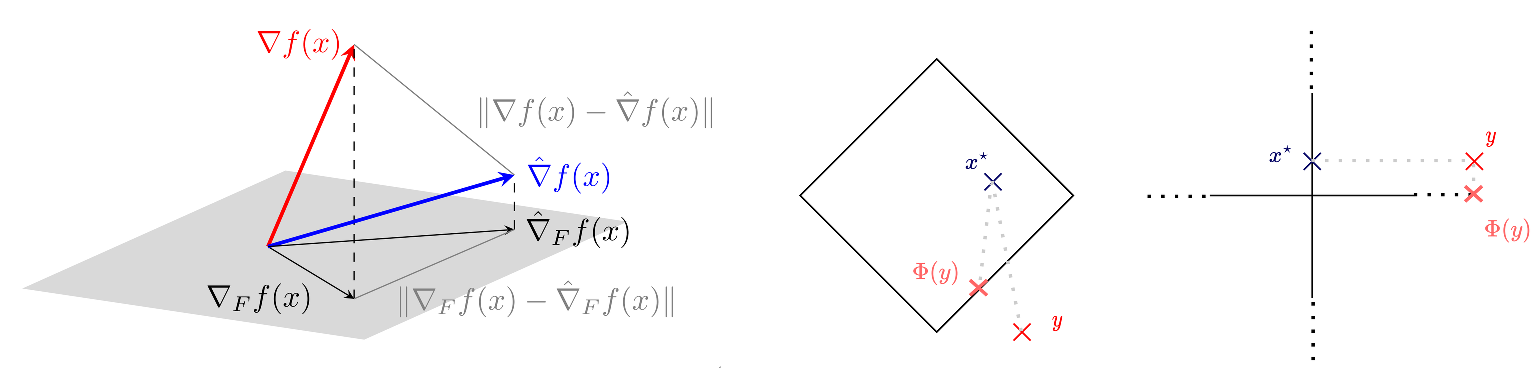
with $\epsilon_{err} = \mathcal{O}\left(1 + \frac{s+d/s_2}{q}\right)$, $C_2 = \mathcal{O}\left(\frac{s}{q}\right)$, $C_3 = \mathcal{O}\left(L_{s_2}^2\left(\frac{ss_2}{q}(d+ss_2)+sd\right)\right)$

Expansivity

Projection on the ℓ_0 ball ($\mathcal{B}_{\ell_0,k}$) is not non-expansive:

$$\forall \boldsymbol{y} \in \mathbb{R}^d, \boldsymbol{x}^* \in \mathcal{B}_{\ell_0,k^*} : \|\Phi_k(\boldsymbol{y}) - \boldsymbol{x}^*\| \leq \gamma\|\boldsymbol{y}-\boldsymbol{x}^*\| \quad \text{with} \quad \gamma > 1$$

$$\gamma = \sqrt{1 + \left(k^*/k + \sqrt{(4+k^*/k)\,k^*/k}\right)/2} \quad \text{(from [5])}$$



(a) Gradient Error (b) Expansivity

Convergence Analysis

Convergence rate: (Theorem 1)

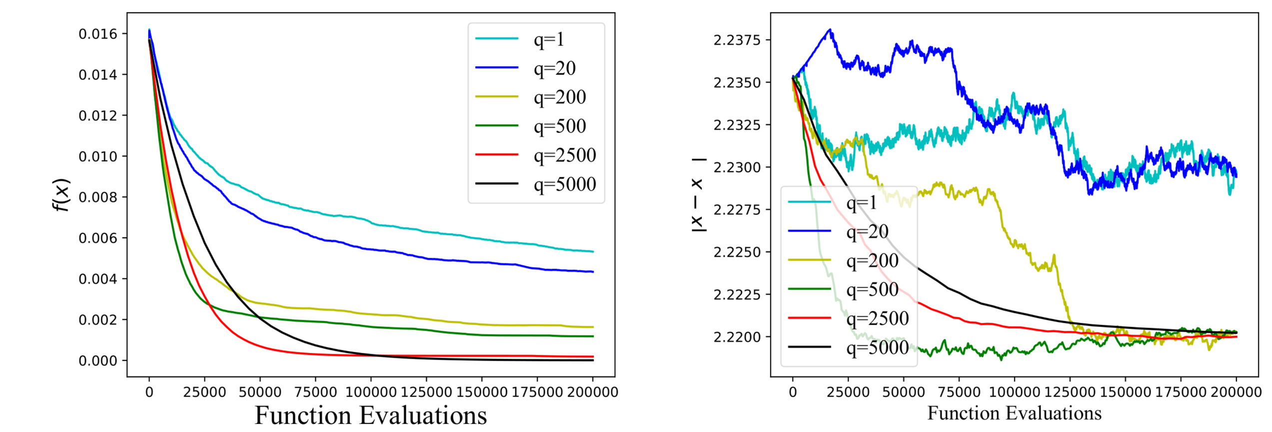
$$\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}^*\| \leq (\rho\gamma)^t\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| + (\cdot)\sigma + (\cdot)\mu$$

$$\text{with} \quad \eta = \frac{\nu}{(4\epsilon_{err}+1)L^2}, \rho = 1 - \frac{\nu^2}{(4\epsilon_{err}+1)L^2}$$

We want $\rho\gamma < 1$ for convergence, so we need q **large enough**.

- **Necessary condition on** q (Remark 4): $q \geq 4\kappa^2\sqrt{\frac{k^* d}{s_2}}$
- **Sufficient condition on** q (Corollaries 1 & 2):
  With $k \geq (86\kappa^2 - 12\kappa^2)k^*$, $\kappa := \frac{L_{s'}}{\nu_s}$, and $s' := \max(s_2, s)$:
  - if f is s'-RSS: take $q \geq 2s + 6\frac{d}{s_2}$, to get **QC:** $\mathcal{O}(\kappa^2(\boldsymbol{k}+\frac{\boldsymbol{d}}{\boldsymbol{s_2}})\log(\frac{1}{\varepsilon}))$
    $\implies$ **Weakly dimension dependent**
    (e.g. if $s_2 = d/m, m \in [\![1, d]\!]$)
  - if f is smooth & $s_2 = d$: take $q \geq 2(s+2)$, to get **QC:** $\mathcal{O}(\kappa^2\boldsymbol{k}\log(\frac{1}{\varepsilon}))$
    $\implies$ **Dimension independent**

Sensitivity Analysis



(c) $f(\boldsymbol{x})$ (d) $\|\boldsymbol{x}-\boldsymbol{x}^*\|$

Evolution of $f(\boldsymbol{x})$ and $\|\boldsymbol{x}-\boldsymbol{x}^*\|$, on a **toy quadratic problem**, for several values of q. If q is too small, the **gradient error is too large** and even if we decrease the learning rate accordingly, we **cannot make enough progress to counterbalance expansivity**, and don't converge anymore.

Applications

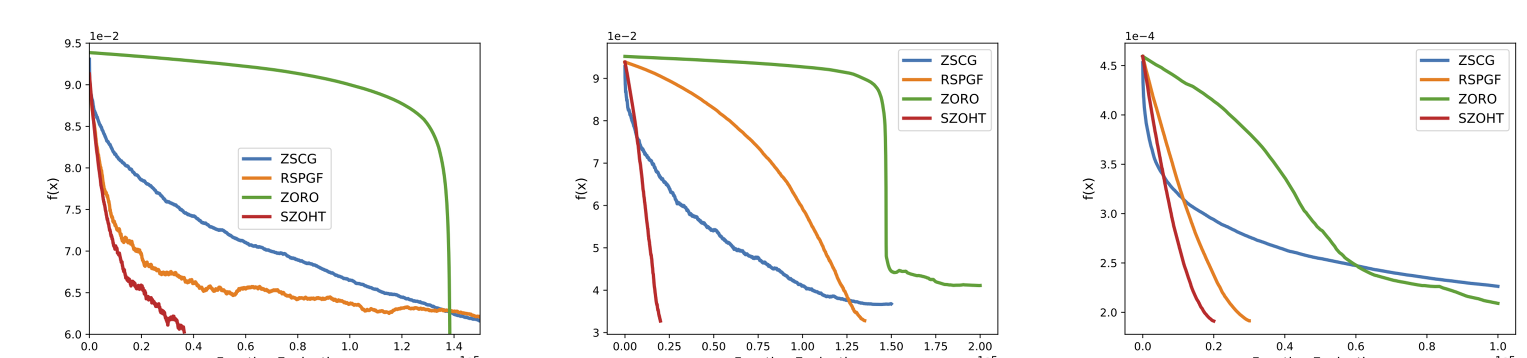
- Asset management [6], (a), (b), (c) :

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{\boldsymbol{x}^\top \boldsymbol{C}\boldsymbol{x}}{2\left(\sum_{i=1}^d \boldsymbol{x}_i\right)^2} + \lambda\left(\min\left\{\frac{\sum_{i=1}^d \boldsymbol{m}_i\boldsymbol{x}_i}{\sum_{i=1}^d \boldsymbol{x}_i} - r, 0\right\}\right)^2 \quad \text{s.t.} \quad \|\boldsymbol{x}\|_0 \leq k$$

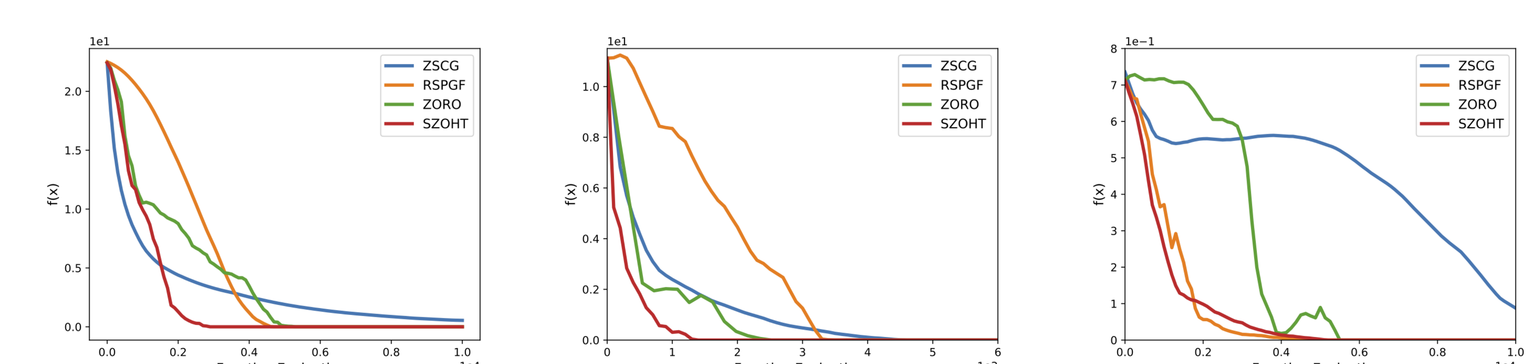
- Few pixels adv. attacks [7], (d), (e), (f) :

$$\min_{\boldsymbol{\delta}} f(\boldsymbol{x}+\boldsymbol{\delta}) \text{ s.t. } \|\boldsymbol{\delta}\|_0 \leq k$$

- Comparison with ZSCG [1], RSPGF [2], ZORO [3]: **improved QC**



(e) port3 (f) port4 (g) port5



(h) MNIST (i) CIFAR (j) Imagenet

Figure 1. $f(\boldsymbol{x})$ vs. # queries

References

[1] K. Balasubramanian and S. Ghadimi, "Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[2] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Mathematical Programming*, vol. 155, no. 1, pp. 267–305, 2016.
[3] H. Cai, D. McKenzie, W. Yin, and Z. Zhang, "Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling," *SIAM Journal on Optimization*, vol. 32, no. 2, pp. 687–714, 2022.
[4] N. Nguyen, D. Needell, and T. Woolf, "Linear convergence of stochastic iterative greedy algorithms with sparse constraints," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 6869–6895, 2017.
[5] J. Shen and P. Li, "A tight bound of hard thresholding," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 7650–7691, 2017.
[6] T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha, "Heuristics for cardinality constrained portfolio optimisation," *Computers & Operations Research*, vol. 27, no. 13, pp. 1271–1302, 2000.
[7] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.