Zeroth-Order Regularized Optimization (ZORO)

William de Vazelhes

Overleaf

2021

1. Introduction on ZO: Table of Contents

1. Introduction on ZO: Zeroth-order optimization

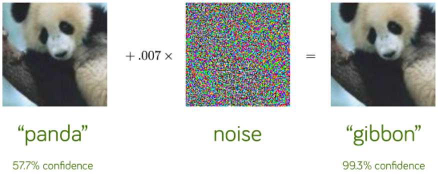
Gradient descent:

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

What if we don't know ∇f, but only f ?

1. Introduction on ZO: Applications

- Black-Box Adversarial attacks



from:https://arxiv.org/abs/1412.6572

- Reinforcement learning

from:https://www.cs.toronto.edu/ vmnih/docs/dqn.pdf

1. Introduction on ZO: Idea to approximate the gradient

Idea: approximate the gradient using using finite differences coordinate-wise, e.g. with $u_i = (0, .., \overset{\overset{i}{\downarrow}}{1}, ..., 0)$.

$$(\hat{\nabla} f(x))_i = \frac{f(x + \mu u_i) - f(x)}{\mu} \tag{1}$$

1. Introduction on ZO: Two points approximation

(from Liu et al. (2020), and https://scholar.harvard.edu/files/yujietang/files/slides_2019_zero-order_opt_tutorial.pdf)

$$\hat{\nabla} f(x) = d\frac{f(\boldsymbol{x}+\mu\boldsymbol{u}) - f(x)}{\mu}\boldsymbol{u} \quad \text{with} \quad \boldsymbol{u} \sim \mathrm{Uni}(\mathbb{S}_{n-1})$$

Unbiased, w.r.t a smoothed version of f:

$$\mathbb{E}_{u\sim p}\left[\hat{\nabla} f(x)\right] = \nabla f_\mu(x)$$

with

$$f_\mu(x) \triangleq \mathbb{E}_{u\sim \mathrm{Uni}(\mathbb{B}_n)}\left[\hat{\nabla} f(\boldsymbol{x}+\mu\boldsymbol{u})\right]$$

1. Introduction on ZO: Noisy observations

Generally, in ZO, we don't observe f directly but a noisy version of f:

$$E_f(x, \xi)$$

noise: ξ
noise can be additive or not, bounded variance/magnitude, zero mean or not
Examples:

- physical simulation, reinforcement learning (noisy rewards)
- bi-level optimization: inner problem is solved inexactly

1. Introduction on ZO: Applied to many settings, with many techniques

- Many settings:
  - Stochastic/Deterministic
  - Convex or not, Smooth, Strongly Convex...
- Many techniques:
  - Variance reduction
  - Frank-Wolfe, Proximal, Coordinate descent...

1. Introduction on ZO: Comparison of ZO algorithms

https://arxiv.org/pdf/2106.02958.pdf#page=5

2. Curse of dimensionality for ZO Curse of dimensionality for ZO

- Number of operations: We see in the table above that there is often a $\mathcal{O}(d)$ factor $\rightarrow$ impractical for large d

2. Curse of dimensionality for ZO Curse of dimensionality for ZO

- Number of operations: We see in the table above that there is often a $\mathcal{O}(d)$ factor $\rightarrow$ impractical for large d
- Can we do better ?

2. Curse of dimensionality for ZO Curse of dimensionality for ZO

- Number of operations: We see in the table above that there is often a $\mathcal{O}(d)$ factor $\rightarrow$ impractical for large d
- Can we do better ?
- Not without assumptions: Jamieson et al. (2012)

2. Curse of dimensionality for ZO An impossibility result

Jamieson et al. (2012) "Query Complexity of Derivative-Free Optimization"

- $\mathcal{F}_{\tau,L,\mathcal{B}}$: class of **all** the τ strongly convex functions, L-Lipschitz, defined on convex set $\mathcal{B} \subset \mathbb{R}^d$, with noisy observations: $E_f(x) = f(x) + w$, $\mathbb{E}[w] = 0$, $\mathbb{E}[w^2] = \sigma^2$:
- If $d \geq 8$ and sufficiently large T:

$$\inf_{\hat{x}^T} \sup_{f \in \mathcal{F}_{\tau,L,\mathcal{B}}} \mathbb{E}[f(\hat{x}^T) - f(x_f^*)] \geq c \left(\frac{d\sigma^2}{T} \right)^{\frac{1}{2}}$$

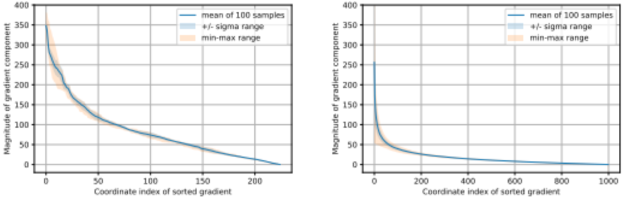
c depends on the oracle and function class parameters + geometry of $\mathcal{B}$, but is independent of T and n.
$\implies$ cannot optimize in less than $O(d/\epsilon^2)$ iterations

$$\text{(because} \quad \epsilon \geq C\frac{\sqrt{d}}{\sqrt{T}} \implies T \geq O(\frac{d}{\epsilon^2}))$$

2. Curse of dimensionality for ZO Doing better: restricting the class of functions

- Let's make some assumptions
- ZORO Cai et al. (2020) (Zeroth Order Regularized Optimization Method): assumes the gradients are either:
  - s sparse: $(\forall x \in \mathbb{R}^d : ||\nabla f(x)||_0 \leq s)$
  - or: compressible: $|\nabla f(x)|_{(i)} \leq i^{-1/p} ||\nabla f(x)||_2, p \in (0, 1)$
- Also: $||\nabla^2 f(x)||_1 \leq H$, L smooth, noisy oracle $E_f = f(x) + \xi$, $|\xi| < \sigma$, f coercive, ∇f coercive
- conv. rate in $\mathcal{O}(s \log(d))$!

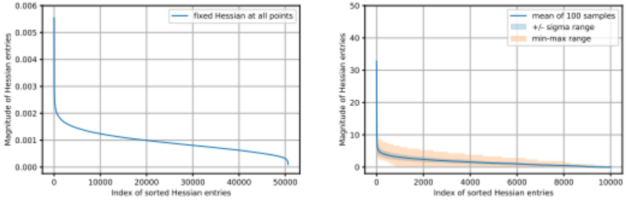
3. ZORO: Are those assumptions verified ?



(a) Asset management

(b) Imagenet adversarial attack

Figure 1: Sorted gradient components at 100 random points in real-world optimization problems. Such decays indicate the gradients are compressible.



(a) Asset management

(b) Imagenet adversarial attack

Figure 2: Sorted Hessian entries at 100 random points in real-world optimization problems. Such decays indicate the Hessians are weakly sparse. Note that the asset management problem has fixed Hessian at all points, thus there is no variance.

3. ZORO: ZORO

Idea: use compressed sensing to estimate $\nabla f(x)$:
For μ small enough (Taylor expansion), with direction z_i (assume unit norm here):

$$\frac{f(x+\mu z_i) - f(x)}{\mu} \approx z_i^T \nabla f(x)$$

so, for n directions ($n << d$):

$$y \triangleq \begin{bmatrix} \frac{f(x+\mu z_1)-f(x)}{\mu} \\ \frac{f(x+\mu z_2)-f(x)}{\mu} \\ \vdots \\ \frac{f(x+\mu z_n)-f(x)}{\mu} \end{bmatrix} \approx Z^T \nabla f(x)$$

3. ZORO: ZORO

Idea: use compressed sensing to estimate $\nabla f(x)$:
For μ small enough (Taylor expansion), with direction z_i (assume unit norm here):

$$\frac{f(x+\mu z_i)-f(x)}{\mu} \approx z_i^T \nabla f(x)$$

so, for n directions ($n << d$):

$$y \triangleq \begin{bmatrix} \frac{f(x+\mu z_1)-f(x)}{\mu} \\ \frac{f(x+\mu z_2)-f(x)}{\mu} \\ \vdots \\ \frac{f(x+\mu z_n)-f(x)}{\mu} \end{bmatrix} \approx Z^T \nabla f(x)$$

Inverse problem: We observe $y = Z^T \nabla f(x) \rightarrow$ What is a good estimate of $\nabla f(x)$?

3. ZORO: ZORO

Idea: use compressed sensing to estimate $\nabla f(x)$:
For μ small enough (Taylor expansion), with direction z_i (assume unit norm here):

$$\frac{f(x+\mu z_i)-f(x)}{\mu} \approx z_i^T \nabla f(x)$$

so, for n directions ($n << d$):

$$y \triangleq \begin{bmatrix} \frac{f(x+\mu z_1)-f(x)}{\mu} \\ \frac{f(x+\mu z_2)-f(x)}{\mu} \\ \vdots \\ \frac{f(x+\mu z_n)-f(x)}{\mu} \end{bmatrix} \approx Z^T \nabla f(x)$$

Inverse problem: We observe $y = Z^T \nabla f(x) \rightarrow$ What is a good estimate of $\nabla f(x)$?
Answer:

$$\hat{\nabla} f(x) = \arg \min_{g/||g||_0 \leq s} ||Z^T g - y||_2$$

3. ZORO: Algorithm

Algorithm 1 ZORO

Input: x_0: initial point; s: gradient sparsity level; α: step size; δ: query radius, K: number of iterations.
$m \leftarrow b_1 s \log(d/s)$ where b_1 is as in Theorem2.2 of ZORO paper. Typically, $b_1 \approx 1$ is appropriate
$z_1, \ldots, z_m \leftarrow$ i.i.d. Rademacher random vectors
for $k = 0$ **to** K **do**
    for $i = 1$ **to** m **do**
        $y_i \leftarrow (f(x + \delta z_i) - f(x))/\delta$
    end for
    $\mathrm{y} \leftarrow \frac{1}{\sqrt{m}}[y_1, \ldots, y_m]^\top$
    $Z \leftarrow \frac{1}{\sqrt{m}}[z_1, \ldots, z_m]^\top$
    $\hat{\mathrm{g}}_k \approx \arg\min_{\|\mathrm{g}\|_0 \le s} \|Z\mathrm{g} - \mathrm{y}\|_2$ by CoSaMP
    $x_{k+1} \leftarrow x_k - \alpha \hat{\mathrm{g}}_k$
end for
Output: x_K: minimizer of the function.

3. ZORO: Query complexity, compressible gradients

- f **convex**
  stepsize: $\alpha = 1/L$, s large enough s.t. $b_4 s^{1/2-1/p} \leq 0.35$
  if $\epsilon > b_3 R\sqrt{2\sigma H/(1-8\psi^2)}$
  ZORO finds ϵ-optimal solution in:

  $$O\left(s\log(d)\frac{1}{\epsilon}\right)$$

  with probability:

  $$1 - 2(s/d)^{b_2 s}$$

- f **restricted ν-SC**
  ZORO finds ϵ-optimal solution in:

  $$O\left(s\log(d)\log(\frac{1}{\epsilon})\right)$$

  with probability:

  $$1 - 2(s/d)^{b_2 s}$$

4. ZO-BCD: ZO-BCD, exactly sparse gradients

- Separate the features into J blocks
- At each iteration:
  - 1) select a block i
  - 2) approximate the gradient along this block using sparse recovery (like ZORO)
- Do a gradient step

4. ZO-BCD: Even further improvements
ZO-BCD-RC

Additional techniques:

- Make randomized blocks $\rightarrow$ to divide and conquer the sparsity
- Reuse the same Rademacher vectors for each block at each iteration
- Don't sample d/J Rademacher, take random columns from a circulant matrix created by one vector

$$\mathcal{C}(v) = \begin{pmatrix} v_1 & v_2 & \cdots & v_{d/J} \\ v_{d/J} & v_1 & \cdots & v_{d/J-1} \\ \vdots & \ddots & \ddots & \vdots \\ v_2 & \cdots & v_{d/J} & v_1 \end{pmatrix}. \tag{2}$$

4. ZO-BCD: Convergence rate

stepsize: $\alpha = 1/L$, query radius $\delta = 2\sqrt{\sigma/H}$, assume $4\rho^{4n} + \frac{16\tau^2\sigma H}{c_1 L_{max}} < 1$, $s > s_{\text{exact}}$, choose number of CoSaMP n and ϵ such that:
$\frac{c_1}{2}\left(2\rho^{2n} + \sqrt{\rho^n + \frac{16\tau^2\sigma H}{c_1\zeta L_{max}}}\right) < \epsilon < f(x_0 - f^*)$

- With probability at least $1 - \zeta - \tilde{\mathcal{O}}\left(\frac{J^2}{\epsilon}\exp(\frac{-0.01 s_{\text{exact}}}{3J})\right)$, ZO-BCD-R finds an ϵ-solution in $\tilde{\mathcal{O}}(s/\epsilon)$ queries, using $\tilde{\mathcal{O}}(sd/J^2)$ FLOPS per iteration and $\tilde{\mathcal{O}}(sd/J)$ total memory.
- With probability at least $1 - \tilde{\mathcal{O}}\left(\frac{J^2}{\epsilon}\exp(\frac{-0.01 s_{\text{exact}}}{3J})\right) - (d/J)^{\log(d/J)\log^2(4.4s/J)}$, ZO-BCD-RC finds an ϵ-solution in $\tilde{\mathcal{O}}(s/\epsilon)$ queries, requiring $\tilde{\mathcal{O}}(d/J)$ FLOPS per iteration and $\mathcal{O}(d/J)$ total memory.

5. Other approaches: Results on sparsity

- Wang et al. (2018): one of the first works, that achieved $log(d)$ dependence on the dimension. Stronger assumption: s-sparsity of the gradient
- Liu and Yang (2021): less assumptions (only approximate sparsity of the solution), worse bounds, but still logarithmic in d

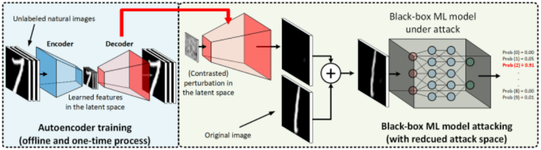
5. Other approaches: Comparison

Table: Comparison of query complexity results for techniques using sparsity (taken from Liu and Yang (2021)): $D_0 := \|\boldsymbol{x}^1 - \boldsymbol{x}^*\|^2$ Although $D_0 \sim \mathcal{O}(d)$ in general, it can be $\mathcal{O}(s)$ when $\boldsymbol{x}^*$ has only s-many nonzero components and the initial solution is chosen to be sparse (e.g., the initial solution can be the all-zero vector).

Algorithms	Complexity	Assumption
Wang et al. (2018)	$\mathcal{O}\left(\frac{s(\ln d)^3}{\epsilon^3}\right)$	s-sparse gradient Bounded 1-norm of gradient Bounded 1-norm of Hessian Additive randomness Function sparsity $\|\boldsymbol{x}^*\|_1 \le R$
Cai et al. (2020) (ZORO)	$\mathcal{O}\left(s \cdot \ln d \cdot \ln\left(\frac{1}{\epsilon}\right)\right)$	Compressible gradient Bounded 1-norm of Hessian Restricted strong convexity Additive randomness Coercivity
Balasubramanian and Ghadimi (2018)	$\mathcal{O}\left(\left(\frac{D_0 s^2}{\epsilon} + \frac{D_0 s}{\epsilon^2}\right)(\ln d)^2\right)$ $= \mathcal{O}\left(\left(\frac{s^3}{\epsilon} + \frac{s^2}{\epsilon^2}\right)(\ln d)^2\right)$	s-sparse gradient $\boldsymbol{x}^*$ is s-sparse
Liu and Yang (2021)	$\mathcal{O}\left(\frac{(D_0+R)^3 \ln d}{\epsilon^3}\right)$	$\|\boldsymbol{x}^*\|_1 \le R$
Liu and Yang (2021)	$\mathcal{O}\left(\frac{(s+D_0+R)^2 \ln d}{\epsilon^2}\right)$ $= \mathcal{O}\left(\frac{(s+R)^2 \ln d}{\epsilon^2}\right)$	$\|\boldsymbol{x}^*\|_1 \le R$ $\boldsymbol{x}^*$ is s-sparse Strong convexity

5. Other approaches: Others

- finite-sum results, e.g.: Liu et al. (2018)
- Holder continuous gradient Shibaev et al. (2021): rate depend on the exponent in the Holder continuous gradient
- Golovin et al. (2019)
  $f(x) = g(Ax) \implies \nabla f(x) = A^T \nabla g(Ax)$: gradient always spanned by a few colums (if $k << d$). Random descent method that uses that geometry: result in $O(k \log(d))$.
- Optimization on a manifold Li et al. (2021): data is embedded in a Riemannian manifold (of dim. k): result in $O(k)$
- AutoZOOM Tu et al. (2019): Use of autoencoders to encode the data in a low dimensional space: good empirical results but no convergence rate



5. Other approaches: Promising directions

- distributed setting: the distributed setting is a natural setting for ZO: Zhang et al. (2021)
- Use some other assumptions on f or ∇f and/or improve existing results when having the useful assumptions

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3459–3468, 2018.

HanQin Cai, Daniel Mckenzie, Wotao Yin, and Zhenliang Zhang. Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. 2020.

Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi Zhang. Gradientless descent: High-dimensional zeroth-order optimization. In *International Conference on Learning Representations*, 2019.

Kevin G Jamieson, Robert D Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. *arXiv preprint arXiv:1209.2434*, 2012.

Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Stochastic zeroth-order riemannian derivative estimation and optimization. 2021.

Hongcheng Liu and Yu Yang. A dimension-insensitive algorithm for stochastic zeroth-order optimization. *arXiv preprint arXiv:2104.11283*, 2021.

Liu Liu, Minhao Cheng, Cho-Jui Hsieh, and Dacheng Tao. Stochastic zeroth-order optimization via variance reduction method. *arXiv e-prints*, pages arXiv–1805, 2018.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Zeroth-order methods for noisy hölder-gradient functions. *Optimization Letters*, pages 1–21, 2021.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 33, pages 742–749, 2019.

Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1356–1365. PMLR, 2018.

Qingsong Zhang, Bin Gu, Zhiyuan Dang, Cheng Deng, and Heng Huang. Desirable companion for vertical federated learning: New zeroth-order gradient based algorithm. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2598–2607, 2021.