Descriptions of Projects for Summer 2020 CSST

HanQin Cai, Yuchen Lou and Daniel McKenzie

2020

Abstract

Let's outline the scope and goals of the various projects we are pursuing this summer. At the moment, I'm focusing on the projects that have not yet been clearly defined. So, I won't go into too much detail on Tree and Block Sparse ZORO.

1 Structured Sparsity Extensions to ZORO

- As discussed, we want to extend ZORO to be able to deal with gradients that exhibit structured sparsity. The two cases we have examined are tree-sparse and block-sparse.

- We note that in both cases, the structure has to be known *a priori*. Clearly, this is unrealistic in some applications. Thus, it is of interest to see whether the tree structure can be *learned*. Specifically, this would mean the following:
  - For the first T iterations, use regular ZORO. Compute the gradient estimators $\hat{\mathbf{g}}_k$ and let $S_k = \text{supp}(\hat{\mathbf{g}}_k)$.
  - Assume that there exists an underlying tree $\mathcal{T}$, such that all gradients for this problem are $\mathcal{T}$-sparse. For now lets assume that all gradients should be s-tree-sparse (*i.e.* have s non-zero entries that form a subtree of $\mathcal{T}$).
  - Observe that each S_k is the set of vertices of an s-vertex subtree of $\mathcal{T}$. Interestingly, observe that S_k alone does not tell us any parent/child relationships.
  - Let's write $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the vertices and $\mathcal{E}$ represents the edges. Lets assume that T is large enough that $\cup_{k=1}^{T} S_k = \mathcal{V}$.
  - Here is the interesting question: Can we infer $\mathcal{E}$ from $\{S_1, \ldots, S_T\}$? This seems hard, but one could exploit the tree structure. For example if S_{k_1} and S_{k_2} differ by one index, call it i, then we can infer that i cannot be a parent to any of the vertices in S_{k_1} or S_{k_2}.

Suppose that this problem can be solved, and now let T be the minimum number such that the problem is solvable with $S_1, \ldots, S_T$. Then one can make an adaptive algorithm that runs regular ZORO for T iterations, then reconstructs $\mathcal{T}$, and then switches to tree-sparse ZORO for the remainder of the optimization process.

- Possible applications. We (i.e. HanQin and Daniel) have some expertise in adversarial attacks on classification algorithms. This is quite a trendy application of zeroth-order optimization, so I think it could be a nice way to apply structure-aware ZORO. Three possible ways to do this (and don't worry if this doesn't all make sense to you at this stage):
  - Attacks on decision trees. This paper: [CLC+18] details a way to do this. They use a generic zeroth-order algorithm to do the attack, but it can probably be sped up by using a tree-sparsity-aware zeroth-order algorithm. The downside is that one would need to know the tree, which would mean it's only partially black-box.
  - Attacks that are sparse in the wavelet domain on Convolutional Neural Networks for image classification.
  - Block-sparse attacks on Convolutional Neural Networks for image classification.This paper: [XLZ+18] has some details on this, but I haven't read it closely.

2 Exploiting Block Sparsity for High Dimensional ZORO

- Suppose that we are solving $\text{minimize}_{x \in \mathbb{R}^d} f(x)$ where d is extremely large.
- ZORO generates $z_i \in \mathbb{R}^d$ with $z_{i,j} = \pm 1$ and then uses finite differences to get $y_i \approx z_i^\top \mathbf{g}_k$.
- As we've discussed, the z_i are then stacked into a matrix $Z \in \mathbb{R}^{m \times d}$. By construction, this matrix is now full (*i.e.* not sparse. When d is large this can require too much memory, and operations with this matrix might be too slow.
- So, my idea is the following:
  - Divide $\{1, \ldots, d\}$ into D blocks of size d/D. Lets label the blocks $\mathbf{b}_1, \ldots, \mathbf{b}_D$. Note that $\mathbf{b}_a = \{a * (d/D) + 1, \ldots, (a+1) * (d/D)\}$.
  - Now, only generate z_i that are *supported on only one block*. That is, $\text{supp}(z_i) = \mathbf{b}_{a_i}$ for some a_i. There are several ways one could do it. Given i one could choose the a_i uniformly at random. Alternatively, at the k-th iteration of the gradient descent part of ZORO, one could fix a block $\mathbf{b}_{a_k}$, and then choose all the z_i at this iteration to have

support $\mathbf{b}_{a_k}$. This turns ZORO into a sort of block coordinate descent algorithm, which could be very interesting to analyze.

– That's where this paper: [CA20] comes in. Hopefully we can use their analysis to establish when the matrix Z has the RIP, and hence use the analysis in the ZORO paper to bound $\|\mathbf{g}_k - \hat{\mathbf{g}}_k\|$.

3 Block Coordinate descent for ZORO

This section generalizes the previous one. Suppose again that we are solving $\text{minimize}_{x\in\mathbb{R}^d}$ where d is extremely large. Let's assume that we have a block-diagonal sensing matrix:

$$Z = \begin{bmatrix} Z_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Z_C \end{bmatrix} \in \mathbb{R}^{m\times d}$$

where $Z_c \in \mathbb{R}^{m_c \times d_c}$.

1. The blocks can either be randomly assigned or can correspond to user-specified parts of the problem. Let's assume that they are equally sized, so that $m_c = m/C$ and $d_c = d/c$.

2. Observe that the problem $y = Z\mathbf{g}$ decouples into problems $y_c = Z_c\mathbf{g}_c$ for $c = 1, \dots, C$.

3. Let us assume that $\|\mathbf{g}_c\|_0 \leq 1.1s/C$, where $s = \|\mathbf{g}\|_0$. We can guarantee this with high probability, if the blocks are random. Alternatively we can make this a requirement that the user-specified blocks must satisfy.

4. In the random block case, one should have $s = \alpha d$ where $\alpha \in (0, 1)$ and C a fixed constant, independent of d. Then:

$$\mathbb{P}\left[\|\mathbf{g}_c\|_0 \leq f(\alpha, d)s/C\right] \leq 1 - \frac{1}{d} \text{ with } f(\alpha, d) \to 0 \text{ as } d \to \infty$$

   from Hoeffding's inequality or the Chernoff Bounds.

5. Note that each Z_c will satisfy the $1.1s/C$ RIP with high probability, as long as m/C is sufficiently large. Thus CoSaMP will succesfully solve each problem:

$$\hat{\mathbf{g}}^{(c)} = \text{argmin}_{\mathbf{g}'\in\mathbb{R}^{d_c}} \left\{ \|y_c - Z_c\mathbf{g}'\|_2 \text{ subject to: } \|\mathbf{g}'\|_0 \leq s_c \right\} \qquad (1)$$

   Note further that one could do this in parallel.

6. More interestingly, one could solve (1) for only one c. This would lead to an *inexact block coordinate descent scheme*: $x_{k+1} = x_k - \alpha\hat{\mathbf{g}}^{(c)}$, where we are abusing notation slightly by considering $\hat{\mathbf{g}}^{(c)}$ to be both a vector in $\mathbb{R}^{d_c}$ and a vector in $\mathbb{R}^n$ padded with zeros.

7. For $x_{k+1} = x_k - \alpha \hat{\mathbf{g}}^{(c)}$, we must have $\alpha \leq \frac{1}{L^{(c)}}$ where $\|\nabla_c f(x) - \nabla_c f(y)\| \leq L^{(c)}\|x - y\|$. The point is $L = \min_c L^{(c)}$, so coordinate descent can, in principle, take larger steps (and therefore converge faster).

8. The paper [TRG16] studies this exact problem, and provides rates of convergence.

4 Variance Reduction for ZORO

Yuchen's note sent on Friday 12th June summarizes this part quite nicely. The only thing I would add is that I think we can further exploit the structure of the error terms, $e_i = z_i^\top \nabla^2 f(x) z_i$. In particular, one can use the Hanson-Wright inequality (see pg. 139 of this book: [Ver18]) to get bounds of the form:

$$\mathbb{P}\left[\left|z_i^\top \nabla^2 f(x) z_i - \mathbb{E}\left[z_i^\top \nabla^2 f(x) z_i\right]\right| \geq t\right] \leq e^{-ct}$$

Of course one would then also need to bound the difference between the true mean, $\mathbb{E}\left[z_i^\top \nabla^2 f(x) z_i\right]$, and the sample mean:

$$\bar{e} = \frac{1}{m}\sum_i e_i = \frac{1}{m}\sum_i z_i^\top \nabla^2 f(x) z_i$$

so getting sharper bounds than those given by Popoviciu's inequality might not be so straightforward.

References

[CA20] Il Yong Chun and Ben Adcock. Uniform recovery from subgaussian multi-sensor measurements. *Applied and Computational Harmonic Analysis*, 48(2):731–765, 2020.

[CLC+18] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.

[TRG16] Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact coordinate descent: complexity and preconditioning. *Journal of Optimization Theory and Applications*, 170(1):144–176, 2016.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[XLZ+18] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. *arXiv preprint arXiv:1808.01664*, 2018.