Structure-aware and Large-scale ZORO

Zeroth Order Online Meeting

Presenter: Yuchen Lou
Supervised by Hanqin Cai, Daniel Mckenzie, Wotao Yin

July 2020

Table of Contents

Contents

Motivation

Aim: Improve efficiency and storage of ZORO.

Trick: Improve CoSaMP

- Special structures of gradients $\Longrightarrow$ structure-based CoSaMP $\Longrightarrow$ reduce function queries $\Longrightarrow$ increase efficiency
- Special structures of sensing matrix $\Longrightarrow$ reduce information needed for building sensing matrix $\Longrightarrow$ save storage

Contents

Structure-based compressed sensing

Gradient estimation $\iff$ Sparse recovery problem (solved by CoSaMP)

$$\hat{g} = \arg\min_{v \in \mathbb{R}^n} ||Zv - y||_2^2 \quad s.t.\, ||v||_0 \leq s$$

where $Z \in \mathbb{R}^{m \times n}$ is the sensing matrix whose $i-$th row is $\frac{1}{\sqrt{m}} z_i^T$ and $\{z_i\}_{i=1}^m$ are Rademacher random vectors.

Structure-based compressed sensing

Gradient estimation $\Longleftrightarrow$ Sparse recovery problem (solved by CoSaMP)

$$\hat{g} = \arg\min_{v \in \mathbb{R}^n} ||Zv - y||_2^2 \quad s.t. \, ||v||_0 \leq s$$

where $Z \in \mathbb{R}^{m \times n}$ is the sensing matrix whose $i-$th row is $\frac{1}{\sqrt{m}} z_i^T$ and $\{z_i\}_{i=1}^m$ are Rademacher random vectors.

- In normal compressed sensing, $m = \mathcal{O}(s \log(n/s))$ to satisfy Restricted Isometry Property (RIP) with high probability.
- According to [Baraniuk et al., 2010], with specific structures of the signals, $m = \mathcal{O}(s)$.
- Two structures considered: binary tree structure, block structure.

Structure-based CoSaMP

Notations:

$\mathcal{M}$: the space of a certain structure (e.g. $\mathcal{T}$ stands for the tree structure).

$\mathcal{M}_s$: the space of $\mathcal{M}$ structure $s-$sparse signals.

$\mathbb{M}(x,s)$: the algorithms that obtains the best approximation of x in $\mathcal{M}_s$, i.e. $\mathbb{M}(x,s) = \arg\min_{\bar{x}\in\mathcal{M}_s} ||x-\bar{x}||_2$.

$\mathbb{M}_2(x,s)$: the algorithms that obtains the best approximation of x in $\mathcal{M}_s \bigoplus \mathcal{M}_s$, i.e. $\mathbb{M}(x,s) = \arg\min_{\bar{x}\in\mathcal{M}_s\bigoplus\mathcal{M}_s} ||x-\bar{x}||_2$.

Structure-based CoSaMP

Notations:

$\mathcal{M}$: the space of a certain structure (e.g. $\mathcal{T}$ stands for the tree structure).
$\mathcal{M}_s$: the space of $\mathcal{M}$ structure $s-$sparse signals.
$\mathbb{M}(x,s)$: the algorithms that obtains the best approximation of x in $\mathcal{M}_s$, i.e. $\mathbb{M}(x,s) = \arg\min_{\bar{x}\in\mathcal{M}_s} ||x-\bar{x}||_2$.
$\mathbb{M}_2(x,s)$: the algorithms that obtains the best approximation of x in $\mathcal{M}_s \bigoplus \mathcal{M}_s$, i.e. $\mathbb{M}(x,s) = \arg\min_{\bar{x}\in\mathcal{M}_s\bigoplus\mathcal{M}_s} ||x-\bar{x}||_2$.

Structure-based CoSaMP:
Implement a given structure $\mathcal{M}$ to the usual CoSaMP. Solving:

$$\arg\min_{v} ||Zv-y||_2 \;\; s.t. \; v\in\mathcal{M}_s$$

Structure-based CoSaMP

Algorithm 1: Model-based CoSaMP [Baraniuk et al., 2010]

Inputs: CS matrix Φ, measurements y, structured sparse approximation algorithm $\mathbb{M}$

Outputs: $s-$sparse approximation $\hat{x}$.

Initialize $\hat{x}_0 = 0$, $d = y$, $i = 0$

while halting criterion false **do**

1. $i \leftarrow i + 1$
2. $e \leftarrow \Phi^T d$
3. $\Omega \leftarrow \mathrm{supp}(\mathbb{M}_2(e, s))$ ($\Omega \leftarrow \mathrm{supp}(e_{2s})$ in normal CoSaMP)
4. $T \leftarrow \Omega \cup \mathrm{supp}(\hat{x}_{i-1})$
5. $b|_T \leftarrow {\Phi_T}^\dagger y, b|_{T^C}$
6. $\hat{x}_i \leftarrow \mathbb{M}(b, s)$ ($\hat{x}_i \leftarrow b_s$ in normal CoSaMP)
7. $d \leftarrow y - \Phi \hat{x}_i$

end while

return $\hat{x} \leftarrow \hat{x}_i$

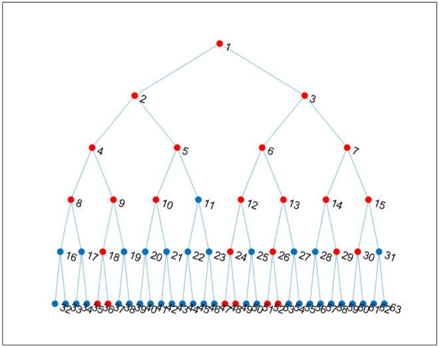
Tree Structure

connected sub-tree: a set of indices Ω forms a connected sub-tree if, whenever an index $i \in \Omega$, then its parent is also in Ω.

Definition: Tree Structure

Define the set of s−tree sparse signals as

$$\mathcal{T}_s = \{x|_{\Omega^C} = 0,\ |\Omega| = s,\ \Omega \text{ forms a connected subtree}\}$$



Tree Structure CoSaMP

Seek an algorithm $\mathbb{T}(x, s)$ to solve the approximation repeatedly in CoSaMP:

$$x_s^{\mathcal{T}} = \arg\min_{\bar{x} \in \mathcal{T}_s} ||x - \bar{x}||_2 \tag{1}$$

Algorithms exist:

- condensing sort and select algorithm (CSSA) [Baraniuk and Jones, 1994].
- exact tree projection algorithm [Cartis and Thompson, 2013].
- fast approximation algorithm [Hegde et al., 2014].

Tree Structure CoSaMP

Seek an algorithm $\mathbb{T}(x,s)$ to solve the approximation repeatedly in CoSaMP:

$$x_s^{\mathcal{T}} = \arg\min_{\bar{x}\in\mathcal{T}_s} ||x-\bar{x}||_2 \tag{1}$$

Algorithms exist:

- condensing sort and select algorithm (CSSA) [Baraniuk and Jones, 1994].
- exact tree projection algorithm [Cartis and Thompson, 2013].
- fast approximation algorithm [Hegde et al., 2014].

(1) is equivalent to solve

$$\arg\max_{\Omega} ||x_\Omega||_2 \;\; s.t. \, x_\Omega \in \mathcal{T}_s$$

In [Hegde et al., 2014], relax on the sparsity to solve

$$\arg\max_{\Omega} ||x_\Omega||_2^2 - \lambda|\Omega| = \arg\max_{\Omega} \sum_{i\in\Omega} (|x_i|^2 - \lambda)$$

where Ω is a support of tree structure (without sparsity constraint). The algorithm returns a tree support Ω which is $s \le |\Omega| \le cs$, for some chosen c.

Tree Approximation Algorithm [Hegde et al., 2014]

Algorithm 2: Tree Sparse Approximation

Begin

function TreeApprox(x, s, c, δ)

$x_{\max} \leftarrow \max_{i\in[n]} |x_i|,\ x_{\min} \leftarrow \min_{i\in[n], x_i>0} |x_i|$

$\lambda_l \leftarrow x_{\max}^2,\ \lambda_r \leftarrow 0,\ \epsilon \leftarrow \frac{\delta x_{\min}^2}{s}$

while $\lambda_l - \lambda_r > \epsilon$ **do**

$\lambda_m \leftarrow \frac{\lambda_l + \lambda_r}{2}$

$\Omega \leftarrow$ FindTree(x, λ_m)

if $s \le |\Omega| \le cs$ **then**

return Ω

else if $|\Omega| < s$ **then**

$\lambda_l \leftarrow \lambda_m$

else

$\lambda_r \leftarrow \lambda_m$

return $\Omega \leftarrow$FindTree(x, λ_l)

function FindTree(x, λ)

Tree Approximation Algorithm [Hegde et al., 2014]

```
    CalculateBest(1, x, λ)
    return Ω ← FindSupport(1)
function CalculateBest(i, x, λ)
    b_i ← |x_i|^2 − λ
    for j ∈ children(i) do
        CalculateBest(j, x, λ)
        b_i ← b_i + b_j
    b_i ← max(0, b_i)
function FindSupport(i)
    if b_i = 0 then
        Ω_i ← {}
    else
        Ω_i ← {i}
        for j ∈ children(i) do
            Ω_i ← Ω_i ∪ FindSupport(j)
return Ω_i
```

RIP on Tree Structure CoSaMP

According to [Baraniuk et al., 2010], a subgaussian random matrix has the $\mathscr{T}_s$–RIP property with constant $\delta_{\mathscr{T}_s}$ and probability $1 - e^{-t}$ if the number of measurements obeys

$$m \geq \frac{2}{c\delta^2_{\mathscr{T}_s}}(s \ln \frac{48}{\delta_{\mathscr{T}_s}} + \ln \frac{512}{se^2} + t)$$

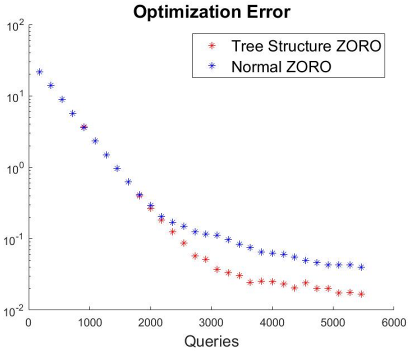
which is $\mathscr{O}(s)$.

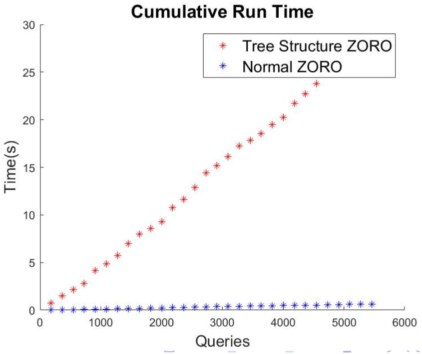
Numerical Results on QP

$$\min\ f(x) = x^T Q x$$

where Q is diagonal with tree structure diagonal entries.

$$n = 2^{10} - 1 = 1023, s = 26, m = 7s = 182$$





Block Structure

Definition: Block Sparsity

Define the set of S−block sparse signals as

$$\mathscr{B}_S = \{x = [x_1^T \ \cdots \ x_N^T]^T \in \mathbb{R}^{JN} \text{ such that}$$

$$x_n \neq 0 \text{ for } n \notin \Omega, \Omega \subseteq \{1, 2, ..., N\}, |\Omega| = S\}$$

Example

$$x = \begin{bmatrix} 0 & 0 & | & 1 & 1 & | & 0 & 0 & | & 0 & 0 & | & 1 & 1 \end{bmatrix}$$

$x \in \mathbb{R}^{2\times 5}$, $J = 2$, $N = 5$, $S = 2$, total sparsity $s = 4$.

Block Structure CoSaMP

Block-structure-based algorithm $\mathbb{B}(x, S)$:

$$x_s^{\mathscr{B}} = \arg\min_{\bar{x} \in \mathscr{B}_S} ||x - \bar{x}||_2$$

Block Structure CoSaMP

Block-structure-based algorithm $\mathbb{B}(x, S)$:

$$x_s^{\mathscr{B}} = \arg\min_{\bar{x} \in \mathscr{B}_S} ||x - \bar{x}||_2$$

Solution: *block-wise hard thresholding*: let ρ be the S−th largest l_2-norm among the blocks of x. Then the solution $x_S^{\mathscr{B}} = [x_{S,1}^{\mathscr{B}}, ..., x_{S,N}^{\mathscr{B}}]$ satisfies

$$x_{S,n}^{\mathscr{B}} = \begin{cases} x_n & ||x_n||_2 \geq \rho \\ 0 & ||x_n||_2 < \rho \end{cases}$$

Block Structure CoSaMP

Block-structure-based algorithm $\mathbb{B}(x, S)$:

$$x_s^{\mathscr{B}} = \arg\min_{\bar{x}\in\mathscr{B}_S} ||x - \bar{x}||_2$$

Solution: *block-wise hard thresholding*: let ρ be the S−th largest l_2-norm among the blocks of x. Then the solution $x_S^{\mathscr{B}} = [x_{S,1}^{\mathscr{B}}, ..., x_{S,N}^{\mathscr{B}}]$ satisfies

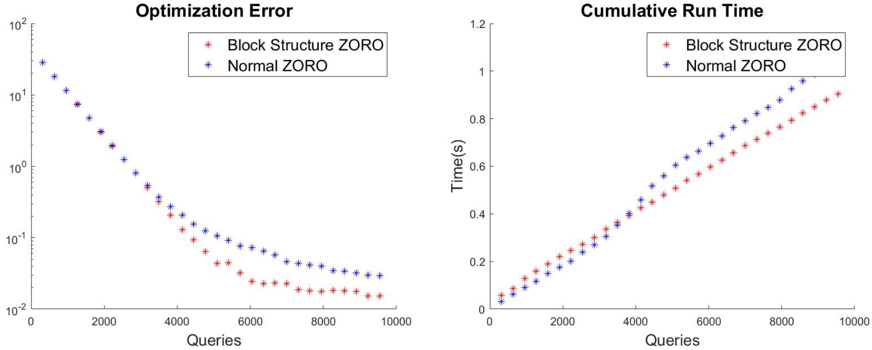
$$x_{S,n}^{\mathscr{B}} = \begin{cases} x_n & ||x_n||_2 \geq \rho \\ 0 & ||x_n||_2 < \rho \end{cases}$$

RIP [Baraniuk et al., 2010]
The number of measurements for robust recovery:
$m = \mathscr{O}(JS + S\log(N/S))$, which is a substantial improvement over $m = \mathscr{O}(JS\log(N/S))$

Numerical Results on QP

Similar tests on noisy QP $f(x) = x^T Qx$

$$J = 50, N = 20, S = 1, s = 50, m = 6\text{ceil}(SJ + S\log(N/S)) = 318$$



Quick Summary and Future Works

- By exploiting specific structures, m is reduced.
- In ZORO, estimation uses $E_f(x + \delta z_i)$ for $i \in [m]$.
- Further study on learning the underlying structure with first few ZORO iterations.
- Still Looking for applications of non-trivial structured gradient.
- Possible future directions: learning to optimize, hyperparameter optimization...

Contents

Motivation and Ideas

- Storage problem for large-scale ZORO (e.g. adversarial attack on 1024×1024 image).
- Consider special forms of sensing matrix Z.
- Two cases: *Random Block Diagonal Matrix* and *Circulant Matrix*.
- Trade-off: to satisfy the RIP, the order of measurements will increase.

Random Block Diagonal Matrix

Random block diagonal matrix takes the form:

$$\Phi = \begin{pmatrix} \Phi_1 & 0 & \cdots & 0 \\ 0 & \Phi_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Phi_J \end{pmatrix} \in \mathbb{C}^{m \times n}$$

where J is the number of blocks, and ϕ_i are formed by random (sub-)Gaussian vectors.

Random Block Diagonal Matrix

Random block diagonal matrix takes the form:

$$\Phi = \begin{pmatrix} \Phi_1 & 0 & \cdots & 0 \\ 0 & \Phi_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Phi_J \end{pmatrix} \in \mathbb{C}^{m \times n}$$

where J is the number of blocks, and ϕ_i are formed by random (sub-)Gaussian vectors.

- *Distinct Block Diagonal Matrix (DBD)*: Φ_i are distinct and independently drawn.
- *Repeated Block Diagonal Matrix (RBD)*: Φ_i are identical.

Measurements for RIP

For DBD case, according to [Eftekhari et al., 2015] and [Koep et al., 2019], the number of measurements needed to satisfy RIP with high probability is

$$m \gtrsim \delta^{-2} \tilde{\mu}^2(U) \cdot s \cdot \log^2(s) \log^2(n)$$

where U is the orthogonal basis, $\tilde{\mu}(U) := \min\{\sqrt{J}, \mu(U)\}$, $\mu(U) = \sqrt{n} \max_{p,q \in [n]} |U(p,q)|$, s is the sparsity and n is the ambient dimension.

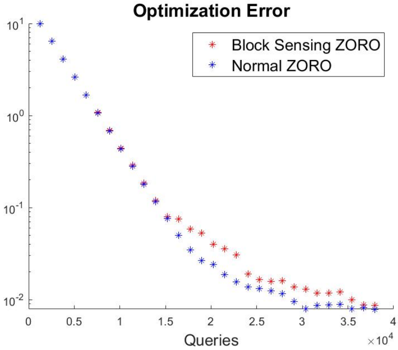
Remarks

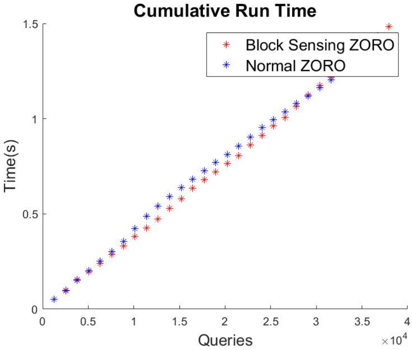
- $1 \leq \tilde{\mu}(U) \leq \sqrt{J}$. The result is dependent of the basis (or dictionary). For Fourier basis, $m = \mathcal{O}(s \log^2(n))$; and for canonical basis, $m = \mathcal{O}(Js \log^2(n))$.
- Similar results with the same order hold for RBD case, by changing coherence $\tilde{\mu}(U)$ to a block-wise coherence $\gamma(U)$.

Numerical results

Simple tests on QP $f(x) = x^T Qx$.

$$n = 1000, s = 10, m = 5\text{ceil}(s\log(s)^2\log(n)^2/10) = 1265$$





Circulant Matrix

Circulant matrix takes the form:

$$C = \begin{pmatrix} t_n & t_{n-1} & \cdots & \cdots & t_1 \\ t_1 & t_n & t_{n-1} & \ddots & \vdots \\ t_2 & t_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_n & t_{n-1} \\ t_{n-1} & \cdots & t_2 & t_1 & t_n \end{pmatrix}$$

- Only **one row/column** is needed to store.
- This row/column is generated randomly as a subgaussian vector.

Computation efficiency

Let $C \in \mathbb{R}^{n \times n}$ be the circulant matrix and c be its first column. Let F_n be the discrete Fourier transform, then we have

$$C = \frac{1}{n} F_n^{-1} \text{diag}(F_n c) F_n$$

Moreover, to compute the product Cx in compressed sensing for some x, we can apply the convolution:

$$Cx = c * x = F_n^{-1}[F_n(c) \cdot F_n(x)]$$

By *fft* & *ifft*, complexity $\mathcal{O}(n^2) \to \mathcal{O}(n \log(n))$.

Computation efficiency

Let $C \in \mathbb{R}^{n\times n}$ be the circulant matrix and c be its first column. Let F_n be the discrete Fourier transform, then we have

$$C = \frac{1}{n} F_n^{-1} \text{diag}(F_n c) F_n$$

Moreover, to compute the product Cx in compressed sensing for some x, we can apply the convolution:

$$Cx = c * x = F_n^{-1}[F_n(c) \cdot F_n(x)]$$

By *fft* & *ifft*, complexity $\mathcal{O}(n^2) \to \mathcal{O}(n \log(n))$.

Dependency issue: from [Yin et al., 2010], for sparse signals under DCT, the recovery is not robust.
Possible solution: instead of randomizing c, randomize $d := F_n c$ with $|d_i| = 1$ to establish C.

Measurements for RIP

According to [Huang et al., 2018], for the vector c being a random vector with zero mean and $\mathbb{E}(c_i^2) = 1$ and $|c_i| \leq a$ for some constant $a \geq 1$, then the required number of measurements is

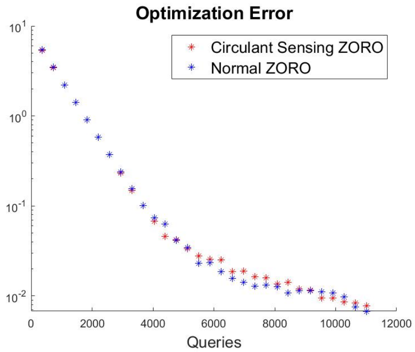
$$m \gtrsim \log^2(\frac{1}{\delta})\delta^{-2} s \log^2(s/\delta) \log(n)$$

with the requirement $s \lesssim \frac{n}{\log^4(n)}$ to satisfy RIP with high probability.
$m = \mathcal{O}(s \log^2(s) \log(n))$

Numerical Tests

Simple tests on QP $f(x) = x^T Qx$.

$$n = 1000, s = 10, m = \text{ceil}(s \log(s)^2 \log(n)) = 367$$



Contents

ZORO Application: Adversarial Attack

Problem considered:

$$\min_{\delta} f(x+\delta)+\lambda||\delta||_0$$

where we consider to minimize the attack loss $f(x+\delta)$ and l_0 distortion.

We attacked Inception-V3 model [Szegedy et al., 2016] on ImageNet, on a randomly selected **subspace** of 2000 variables, and compared ZORO with ZO-AdaMM, ZO-SGD, and ZO-SCD [Chen et al., 2019]. The results from [Cai et al., 2020] are:

Table 2: Attack success rate (ASR), average final ℓ_0 distortion (as a percentage of the total pixels), average final ℓ_2 distortion, and average iteration of first successful attack for different attack methods.

METHODS	ASR	ℓ_0 DIST	ℓ_2 DIST	ITER
ZO-SCD	78 %	0.89%	57.5	240
ZO-SGD	78%	100%	37.9	159
ZO-AdaMM	81%	100%	28.2	172
ZORO	**90%**	**0.73%**	**21.1**	**59**

Median Filter

We also apply a median filter to attempt mitigating the adversarial attack. The results are:

Table 1: Recovery success rate (RSR), original image distortion rate, and total prediction accuracy reduction (TPAR) for different median filter sizes.

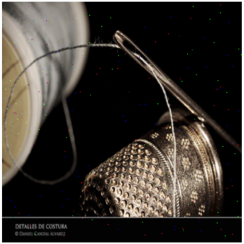
MED. FILTER	RSR	DIST RATE	TPAR
size = 2	86 %	8%	21%
size = 3	**92** %	7%	**14**%
size = 4	76 %	14%	34%
size = 5	69 %	29%	53%

Adversarial Attack Examples

(a) True label: "corn" → Mislabeled: "ear, spike, capitulum"

(b) True label: "plastic bag" → Mislabeled: "shower cap"

(c) True label: "water ouzel, dipper" → Mislabeled: "otter"



(d) True label: "thimble" → Mislabeled: "measuring cup"

References I

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. (2010).
Model-based compressive sensing.
IEEE Transactions on information theory, 56(4):1982–2001.

Baraniuk, R. G. and Jones, D. L. (1994).
A signal-dependent time-frequency representation: Fast algorithm for optimal kernel design.
IEEE Transactions on Signal Processing, 42(1):134–146.

Cai, H., Mckenzie, D., Yin, W., and Zhang, Z. (2020).
Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling.
arXiv preprint arXiv:2003.13001.

Cartis, C. and Thompson, A. (2013).
An exact tree projection algorithm for wavelets.
IEEE Signal Processing Letters, 20(11):1026–1029.

Chen, X., Liu, S., Xu, K., Li, X., Lin, X., Hong, M., and Cox, D. (2019).
Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization.
In *Advances in Neural Information Processing Systems*, pages 7204–7215.

Eftekhari, A., Yap, H. L., Rozell, C. J., and Wakin, M. B. (2015).
The restricted isometry property for random block diagonal matrices.
Applied and Computational Harmonic Analysis, 38(1):1–31.

Hegde, C., Indyk, P., and Schmidt, L. (2014).
A fast approximation algorithm for tree-sparse recovery.
In *2014 IEEE International Symposium on Information Theory*, pages 1842–1846. IEEE.

References II

Huang, M., Pang, Y., and Xu, Z. (2018).
Improved bounds for the rip of subsampled circulant matrices.
arXiv preprint arXiv:1808.07333.

Koep, N., Behboodi, A., and Mathar, R. (2019).
The restricted isometry property of block diagonal matrices for group-sparse signal recovery.
arXiv preprint arXiv:1901.06214.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016).
Rethinking the inception architecture for computer vision.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Yin, W., Morgan, S., Yang, J., and Zhang, Y. (2010).
Practical compressive sensing with toeplitz and circulant matrices.
In *Visual Communications and Image Processing 2010*, volume 7744, page 77440K. International Society for Optics and Photonics.

Thank You!