BEYOND INDEPENDENT MEASUREMENTS: GENERAL COMPRESSED SENSING WITH GNN APPLICATION

A. Naderi & Y. Plan

NeurIPS 2021 Workshop on Deep Learning and Inverse Problems

COMPRESSED SENSING 101

$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{w}$

- $\mathbf{M} \in \mathbb{R}^{l \times n}, \; l < n$
- $\mathbf{x} \in \{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\|_0 \leq s\} \subset \mathbb{R}^n$
- $\mathbf{w}$ is unknown noise

Theorem [Candès et al. '05]

RIP $\implies$ BPDN exactly recovers $\mathbf{x}$

ACCURACY

EFFICIENCY

Theorem [Baraniuk et al. '08]

$\mathbf{M}$ (sub-)gaussian with $l \gtrsim s \log\frac{n}{s} \implies$ RIP w.h.p.

- Basis Pursuit Denoising (aka LASSO)

$\hat{\mathbf{x}} = \operatorname{argmin} \|\mathbf{z}\|_1$ s.t. $\|\mathbf{y} - \mathbf{M}\mathbf{z}\|_2 \leq \sigma$

[Chen et al. '94, Tibshirani '96]

BEYOND SPARSITY

- For a general signal structure, is an efficient and accurate reconstruction possible?
    - What measurement matrices work well?
        - Partial answer: random (sub-)gaussian matrices
    - How many measurements are needed?
        - It depends how LARGE the structure set is!

Definition (Gaussian complexity)

For $S \subset \mathbb{R}^n$, define $w(S) := \mathbb{E} \sup_{\mathbf{x} \in S} |\langle \mathbf{x}, \mathbf{g} \rangle|$,
where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

LET'S START WITH A SIMPLE MODEL…

- $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, $\mathbf{x} \in T \subset \mathbb{R}^n$, $q \coloneqq w^2\big((T - T) \cap \mathbb{S}^{n-1}\big)$
- $\mathbf{A} \in \mathbb{R}^{m \times n}$ has independent standard normal rows $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
- $\hat{\mathbf{x}} \in \underset{\mathbf{z} \in T}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2$
- If $m > \mathcal{O}(q)$, w.h.p.,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \frac{\sqrt{q}}{m} \|\mathbf{w}\|_2$$

[Gordon '88, Rudelson et al. '08, Stojnic '09, Chandrasekaran et al. '12]

LET'S ALLOW OPTIMIZATION ERROR…

- $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, $\mathbf{x} \in T \subset \mathbb{R}^n$, $q \coloneqq w^2\big((T - T) \cap \mathbb{S}^{n-1}\big)$
- $\mathbf{A} \in \mathbb{R}^{m \times n}$ has independent standard normal rows $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
- $\hat{\mathbf{x}} \in T$ satisfies $\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_2^2 \leq \min_{\mathbf{z} \in T} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \epsilon^2$
- If $m > \mathcal{O}(q)$, w.h.p.,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \frac{\sqrt{q}}{m} \|\mathbf{w}\|_2 + \frac{\epsilon}{\sqrt{m}}$$

LET'S ALLOW MODEL MISMATCH TOO…

- $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, $\mathbf{x}$ is close to $T \subset \mathbb{R}^n$, $q \coloneqq w^2\big((T - T) \cap \mathbb{S}^{n-1}\big)$
- $\mathbf{A} \in \mathbb{R}^{m \times n}$ has independent standard normal rows $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$
- $\hat{\mathbf{x}} \in T$ satisfies $\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_2^2 \leq \min_{\mathbf{z} \in T} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \epsilon^2$
- If $m > \mathcal{O}(q)$, w.h.p.,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \frac{\sqrt{q}}{m} \|\mathbf{w}\|_2 + \frac{\epsilon}{\sqrt{m}} + \mathrm{dist}(\mathbf{x}, T)$$

WHAT IF MEASUREMENTS ARE SUB-GAUSSIAN?

- $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, $\mathbf{x}$ is close to $T \subset \mathbb{R}^n$, $q \coloneqq w^2\big((T-T) \cap \mathbb{S}^{n-1}\big)$
- $\mathbf{A} \in \mathbb{R}^{m \times n}$ has independent, mean-zero, isotropic, sub-gaussian rows with $\|\mathbf{a}_i\|_{\psi_2} \leq K$
- $\hat{\mathbf{x}} \in T$ satisfies $\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_2^2 \leq \min_{\mathbf{z} \in T} \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_2^2 + \epsilon^2$
- If $m > \mathcal{O}(K^2 \log K\, q)$, w.h.p.,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \frac{K\sqrt{q}}{m} \|\mathbf{w}\|_2 + \frac{\epsilon}{\sqrt{m}} + K \operatorname{dist}(\mathbf{x}, T)$$

[Mendelson et al. ‘07, Tropp ‘14, Liaw et al. ’16, Jeong et al. ‘20]

TIME TO MIX THE MEASUREMENTS…

- $\mathbf{y} = \mathbf{BAx} + \mathbf{w}$, $\mathbf{x}$ is close to $T \subset \mathbb{R}^n$, $q := w^2\big((T - T) \cap \mathbb{S}^{n-1}\big)$
- $\mathbf{A} \in \mathbb{R}^{m\times n}$ has independent, mean-zero, isotropic, sub-gaussian rows with $\|\mathbf{a}_i\|_{\psi_2} \leq K$
- $\mathbf{B} \in \mathbb{R}^{l\times m}$ is arbitrary
- $\hat{\mathbf{x}} \in T$ satisfies $\|\mathbf{y} - \mathbf{BA}\hat{\mathbf{x}}\|_2^2 \leq \min_{\mathbf{z}\in T}\|\mathbf{y} - \mathbf{BAz}\|_2^2 + \epsilon^2$
- If $\mathrm{sr}(\mathbf{B}) > \mathcal{O}(K^2 \log K\, q)$, w.h.p.,

Definition (stable rank)

$$\mathrm{sr}(\mathbf{B}) := \frac{\|\mathbf{B}\|_F^2}{\|\mathbf{B}\|^2} = \frac{\sum_i \sigma_i^2}{\max_i \sigma_i^2}$$

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \frac{K\sqrt{q}}{\|\mathbf{B}\|_F\sqrt{\mathrm{sr}(\mathbf{B})}}\|\mathbf{w}\|_2 + \frac{\epsilon}{\|\mathbf{B}\|_F} + K\frac{\sqrt{l}}{\sqrt{\mathrm{sr}(\mathbf{B})}}\mathrm{dist}(\mathbf{x}, T)$$

[Jeong et al. '20]

OUR MODEL IN SUMMARY

$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{w}$

- $\mathbf{M} = \mathbf{B}\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{l \times m}$ deterministic, $\mathbf{A} \in \mathbb{R}^{m \times n}$ sub-gaussian
- $\mathbf{x} \in T \subset \mathbb{R}^n$ arbitrary cone (or, $\mathrm{dist}(\mathbf{x}, T)$ small)
- $\mathbf{w}$ is a fixed/random noise

- Empirical Risk Minimization (probably non-convex)

  $\cancel{\hat{\mathbf{x}} \in \mathrm{argmin}_{\mathbf{z} \in T} \|\mathbf{y} - \mathbf{M}\mathbf{z}\|_2}$

  $\hat{\mathbf{x}}$ satisfies $\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}\|_2^2 \leq \min_{\mathbf{z} \in T} \|\mathbf{y} - \mathbf{M}\mathbf{z}\|_2^2 + \epsilon^2$

GENERAL COMPRESSED SENSING

Theorem [Naderi & Plan '21]

Given the last slide settings, w.h.p,

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2 \lesssim \frac{K\sqrt{q}}{\|\mathbf{B}\|_F\sqrt{\mathrm{sr}(\mathbf{B})}}\|\mathbf{w}\|_2 + \frac{\epsilon}{\|\mathbf{B}\|_F} + K\frac{\sqrt{l}}{\sqrt{\mathrm{sr}(\mathbf{B})}}\mathrm{dist}(\mathbf{x},T)$$

- Denoising for fixed and random noise
- Dependence among rows and columns allowed
- Suitable for GNN setting
    - No known convex relaxation (convex theories [Chandrasekaran et al. '12, Tropp '15] fail)
    - No precise solution for the non-convex program
    - Signals don't necessarily belong to the range

A QUICK ASIDE: HOW LARGE IS THE RANGE OF A GNN?

Proposition [Naderi & Plan '21]

Let $G: \mathbb{R}^k \to \mathbb{R}^n$ be a d-layer fully-connected feedforward neural network with ReLU activation function, i.e. $G(\mathbf{z}) = \sigma\left(\mathbf{A}_d \sigma\left(\mathbf{A}_{d-1}\sigma(\dots \mathbf{A}_2 \sigma(\mathbf{A}_1 \mathbf{z}) \dots)\right)\right)$, where $\mathbf{A}_i \in \mathbb{R}^{p_i \times p_{i-1}}$, $p_0 = k$, $p_d = n$, and $\sigma(\cdot) = \max(0, \cdot)$ applies entrywise. Let $T = \operatorname{ran}(G)$. Then

$$w(T \cap \mathbb{S}^{n-1}) \leq w\left((T - T) \cap \mathbb{S}^{n-1}\right) \lesssim \sqrt{kd \log\left(\frac{p'}{k}\right)},$$

where $p' = \left(\prod_{i=1}^{d} p_i\right)^{1/d}$.

- $\operatorname{ran}(G)$ is a cone since $\mathbf{x} = G(\mathbf{z})$ implies $\lambda\mathbf{x} = G(\lambda\mathbf{z})$ for $\lambda \geq 0$
- Subspace counting argument similar to [Bora et al. '17]

THEORY APPLIED TO NEURAL NETS

Corollary [Naderi & Plan '21]

Let $G: \mathbb{R}^k \to \mathbb{R}^n$, $G(\mathbf{z}) = \sigma\left(\mathbf{A}_d \sigma\left(\mathbf{A}_{d-1} \sigma(\dots \mathbf{A}_2 \sigma(\mathbf{A}_1 \mathbf{z}) \dots)\right)\right)$ be a trained GNN. Then given $m = \mathcal{O}\left(kd \log \frac{p'}{k}\right)$ sub-gaussian noisy measurements, an approximate empirical risk minimizer robustly recovers signals close to $\mathrm{ran}(G)$ with high probability, i.e.

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \lesssim \frac{\sqrt{kd \log p'/k}}{m} \|\mathbf{w}\|_2 + \frac{\epsilon}{\sqrt{m}} + \mathrm{dist}(\mathbf{x}, \mathrm{ran}(G))$$

- [Bora et al. '17] requires $m = \mathcal{O}(kd \log n)$ measurements for a similar bound
- We also exhibit denoising phenomenon
- Dependence on the model parameters is highlighted (some are provably optimal)
- Dependent measurements are allowed

THE UNIVERSITY OF BRITISH COLUMBIA