Learning a Compressed Sensing Measurement Matrix via Gradient Unrolling

Shanshan Wu[1], **Alex Dimakis**[1], Sujay Sanghavi[1], Felix Yu[2], Dan Holtmann-Rice[2], Dmitry Storcheus[2], Afshin Rostamizadeh[2], Sanjiv Kumar[2]

[1]*University of Texas at Austin,* [2]*Google Research*

Wed Jun 12th 06:30 -- 09:00 PM @ Pacific Ballroom #189

Learning a Compressed Sensing Measurement Matrix via Gradient Unrolling

Shanshan Wu[1], **Alex Dimakis**[1], Sujay Sanghavi[1], Felix Yu[2], Dan Holtmann-Rice[2], Dmitry Storcheus[2], Afshin Rostamizadeh[2], Sanjiv Kumar[2]

[1]*University of Texas at Austin,* [2]*Google Research*

12th 06:30 -- 09:00 PM @ Pacific Ballroom #189

Motivation

- Goal: Create good representations for sparse data

Motivation

- Goal: Create good representations for sparse data
  - Amazon employee dataset: $d = 15k$, nnz = 9
  - RCV1 text dataset: $d = 47k$, nnz = 76
  - Wiki multi-label dataset: $d = 31k$, nnz = 19

One-hot encoded
categorical data
+Text parts

  - eXtreme Multi-label Learning (XML).

(Multiple labels per item, from a very large class of labels)

Motivation

- Goal: Create good representations for sparse data
  - Amazon employee dataset: $d = 15k$, nnz = 9
  - RCV1 text dataset: $d = 47k$, nnz = 76
  - Wiki multi-label dataset: $d = 31k$, nnz = 19

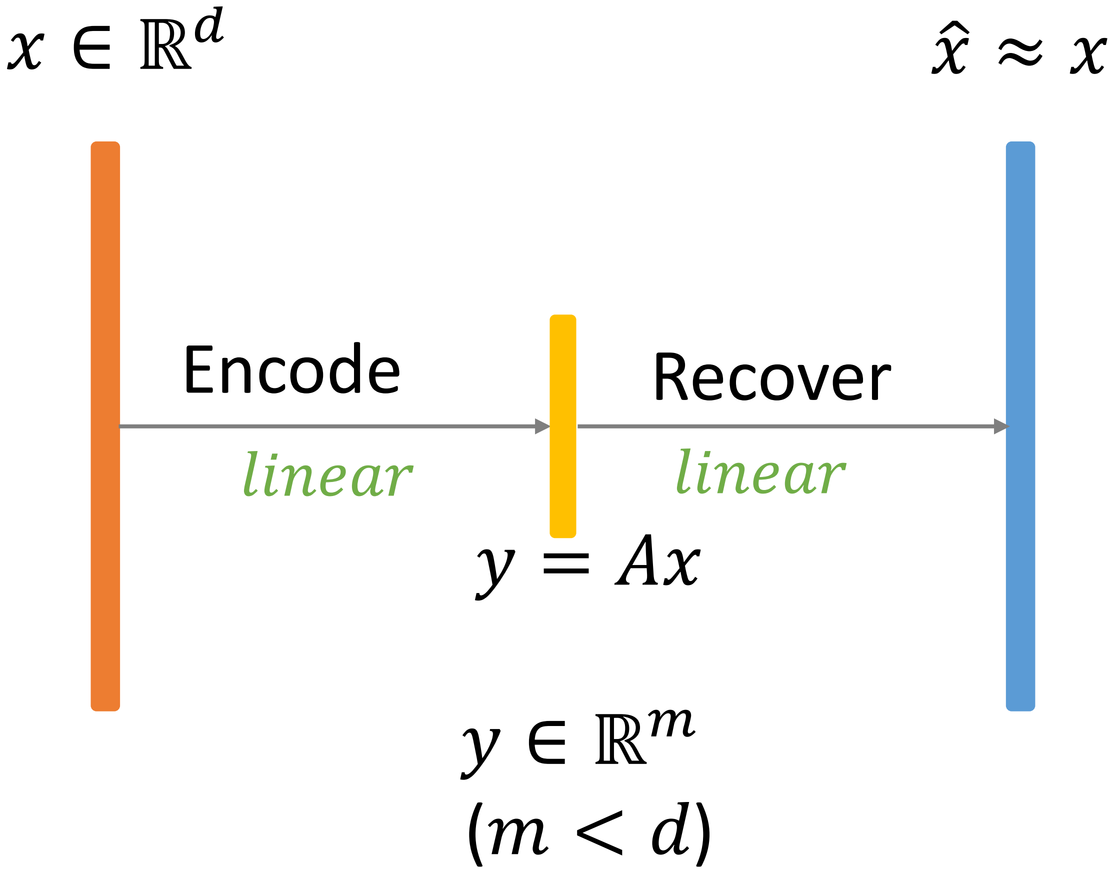
  One-hot encoded categorical data +Text parts

  - eXtreme Multi-label Learning (XML).

  (Multiple labels per item, from a very large class of labels)
- Unlike image/video data, there is no notion of spatial/time locality. No CNN
- Reduce the dimensionality via a linear sketching/embedding

Want: Beyond sparsity, learn additional structure

Representing vectors in low-dimension



$x \in \mathbb{R}^d$

$\hat{x} \approx x$

Encode

linear

Recover

linear

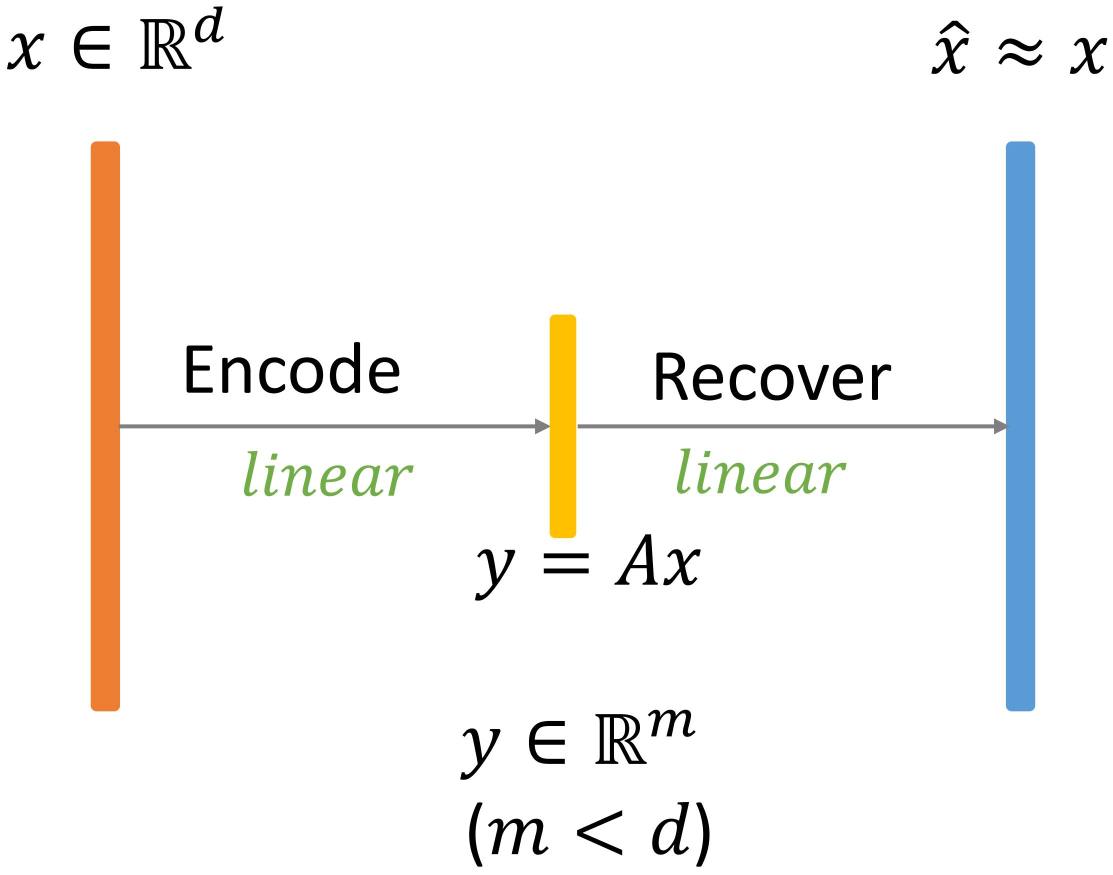
$y = Ax$

$y \in \mathbb{R}^m$

$(m < d)$

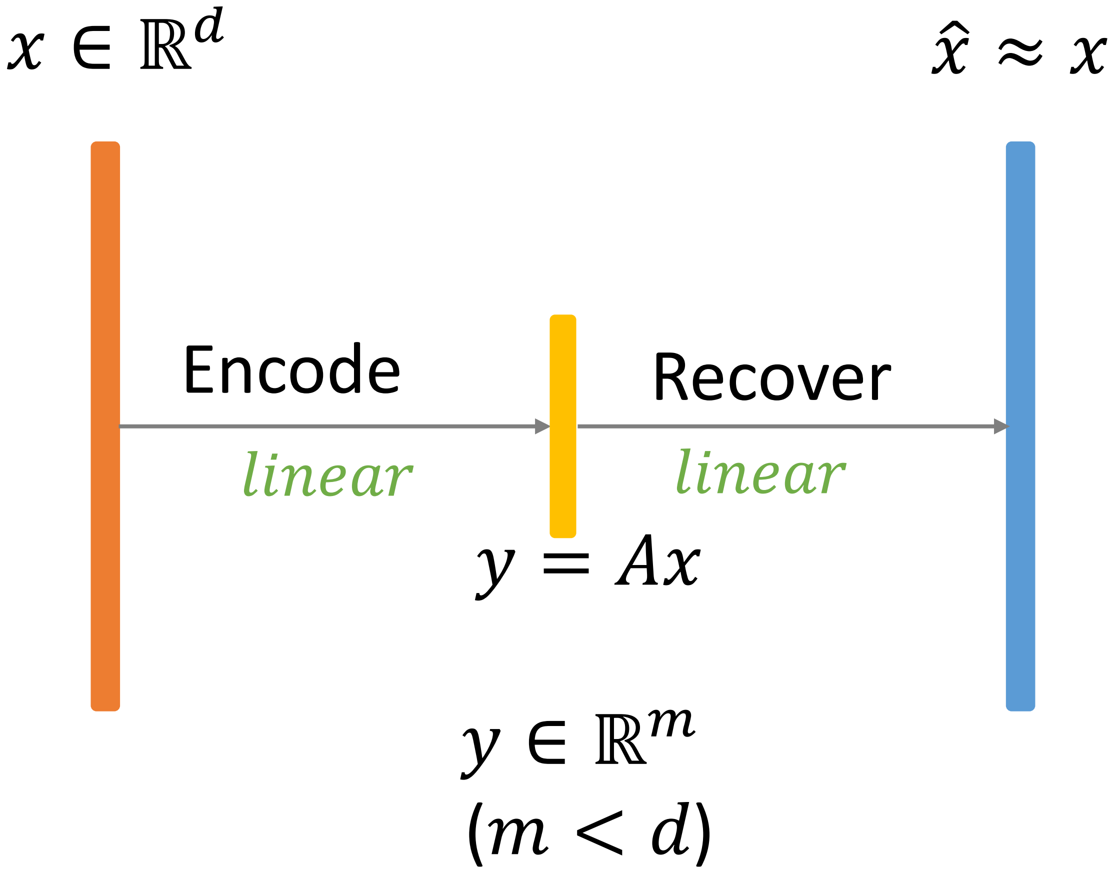
- $A \in \mathbb{R}^{m \times d}$ Measurement matrix
- If we ask: Linear compression,
- And Linear recovery
- Best learned measurement/reconstruction matrices for l2 norm?

Representing vectors in low-dimension



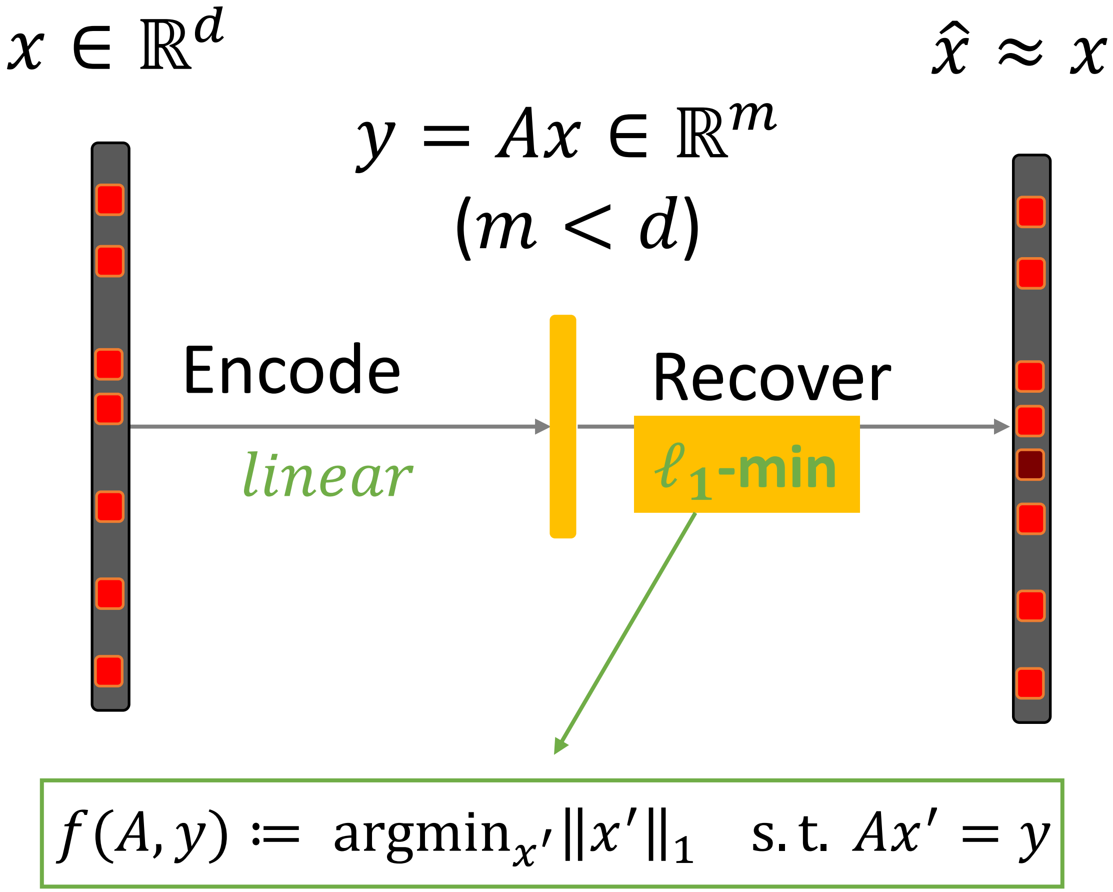
- $A \in \mathbb{R}^{m \times d}$ Measurement matrix
- If we ask Linear compression,
- And Linear recovery
- Best learned measurement/reconstruction matrices for l2 norm?
- **PCA**

Representing vectors in low-dimension



- $A \in \mathbb{R}^{m \times d}$ Measurement matrix
- If we ask Linear compression,
- And Linear recovery
- Best learned measurement/reconstruction matrices for l2 norm?
- **PCA**
- **But if x is sparse we can do better**

Compressed Sensing (Donoho; Candès et al.; …)



$$f(A, y) \coloneqq \operatorname{argmin}_{x'} \|x'\|_1 \quad \text{s.t. } Ax' = y$$

- $A \in \mathbb{R}^{m \times d}$ Measurement matrix
- If we ask Linear compression,
- Recovery by convex opt
  - ℓ_1-min, Lasso,…
- **Near-perfect recovery for sparse vectors.**
- **Provably for Gaussian random A.**

Comp dès et al.; ...)

$x \in \mathbb{R}$

easurement matrix

r compression,

vex opt

,...

ecovery for sparse

r Gaussian random A.

1. If our vectors are
sparse +additional unknown structure
(e.g. one-hot encoded features, text+features, XML, etc)

2. Can we **LEARN** a measurement matrix A

3. Make it work well for convex opt decoder

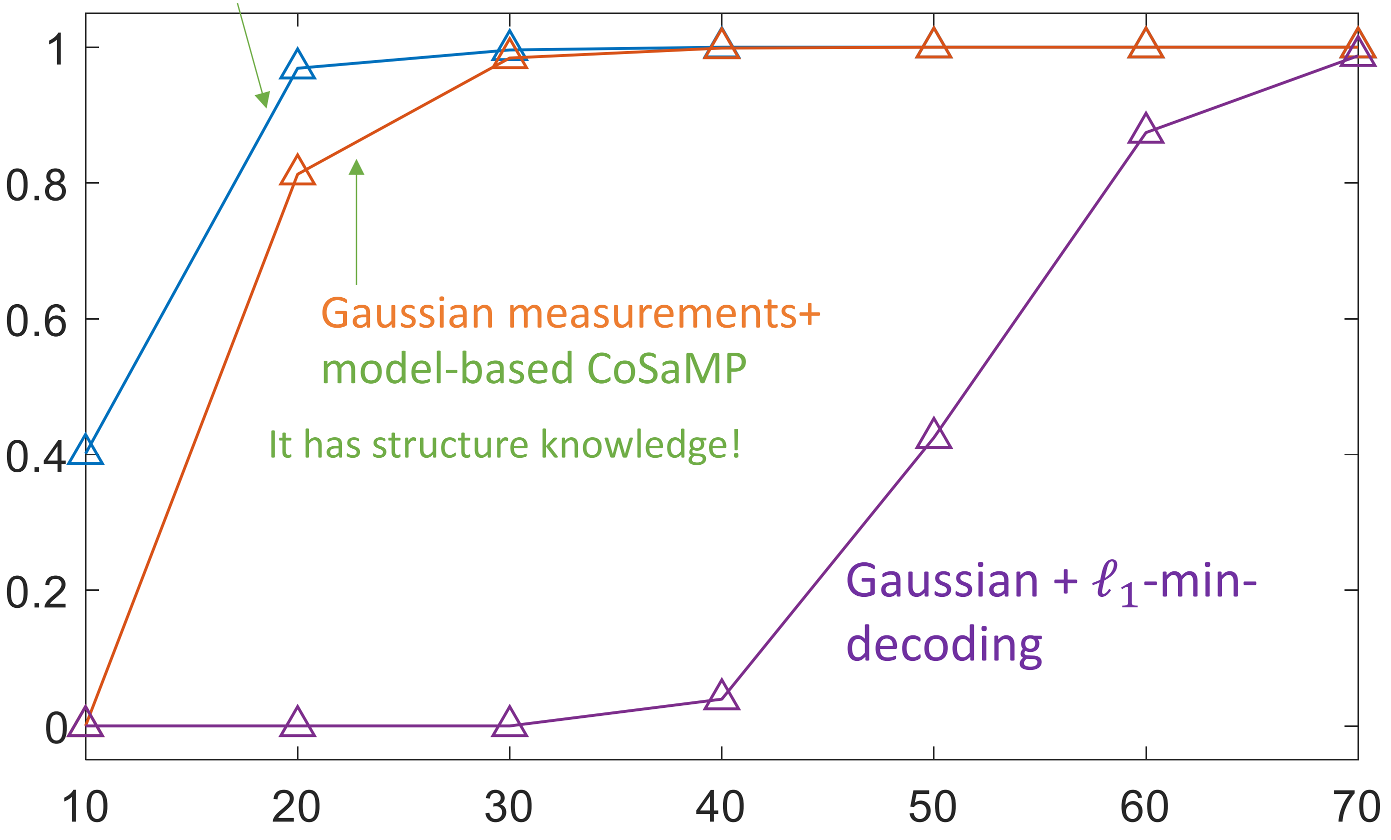
Comparisons of the recovery performance

Learned measurements + ℓ_1-min decoding

[our method]

Fraction of exactly recovered points

Exact recovery:
$\|x - \hat{x}\|_2 \leq 10^{-10}$



Gaussian measurements+ model-based CoSaMP

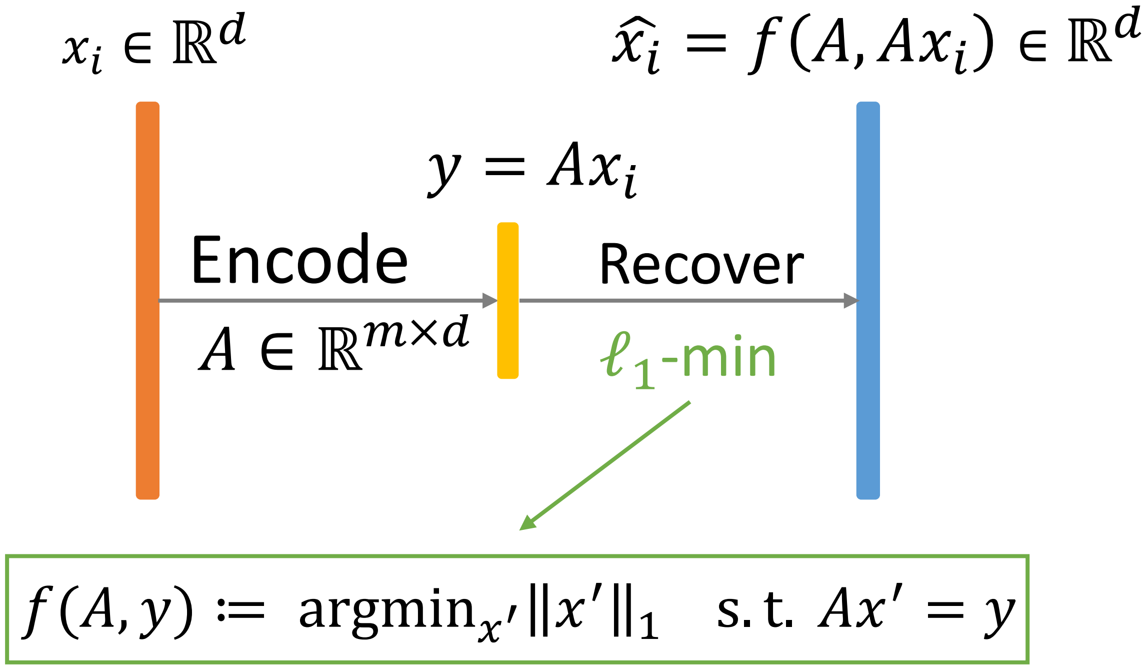
It has structure knowledge!

Gaussian + ℓ_1-min-decoding

Number of measurements (m)

Learning a measurement matrix

- Training data: n sparse vectors $x_1, x_2, \dots, x_n \in \mathbb{R}^d$

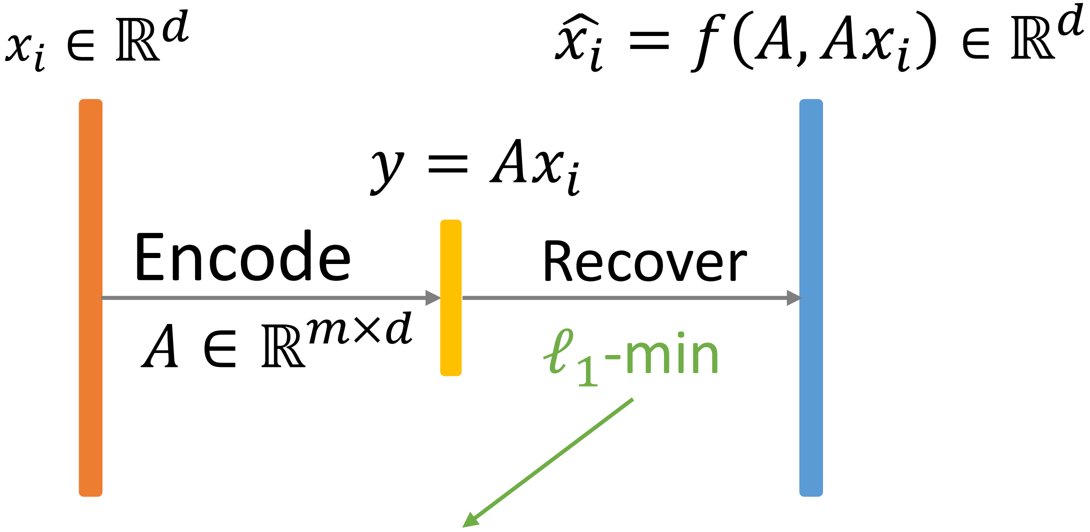


$x_i \in \mathbb{R}^d$
$y = Ax_i$
$\widehat{x_i} = f(A, Ax_i) \in \mathbb{R}^d$
Encode
$A \in \mathbb{R}^{m \times d}$
Recover
ℓ_1-min

$$f(A, y) := \operatorname{argmin}_{x'} \|x'\|_1 \quad \text{s.t. } Ax' = y$$

Learning a measurement matrix

- Training data: n sparse vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$



$x_i \in \mathbb{R}^d$

$\widehat{x_i} = f(A, Ax_i) \in \mathbb{R}^d$

$y = Ax_i$

Encode

$A \in \mathbb{R}^{m\times d}$

Recover

ℓ_1-min

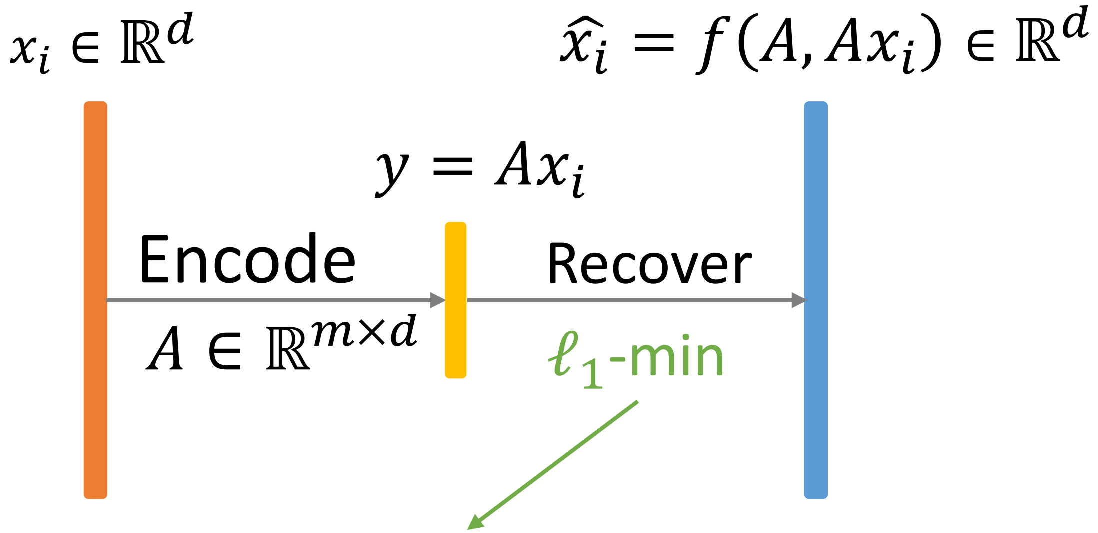
$$f(A, y) \coloneqq \operatorname{argmin}_{x'} \|x'\|_1 \quad \text{s.t. } Ax' = y$$

Objective function:

$$\min_{A \in \mathbb{R}^{m\times d}} \quad \sum_{i=1}^{n} \|x_i - f(A, Ax_i)\|_2^2$$

Learning a measurement matrix

- Training data: n sparse vectors $x_1, x_2, \dots, x_n \in \mathbb{R}^d$



$x_i \in \mathbb{R}^d$

$\hat{x}_i = f(A, Ax_i) \in \mathbb{R}^d$

$y = Ax_i$

Encode

$A \in \mathbb{R}^{m \times d}$

Recover

ℓ_1-min

$$f(A, y) := \operatorname{argmin}_{x'} \|x'\|_1 \quad \text{s.t. } Ax' = y$$

Objective function:

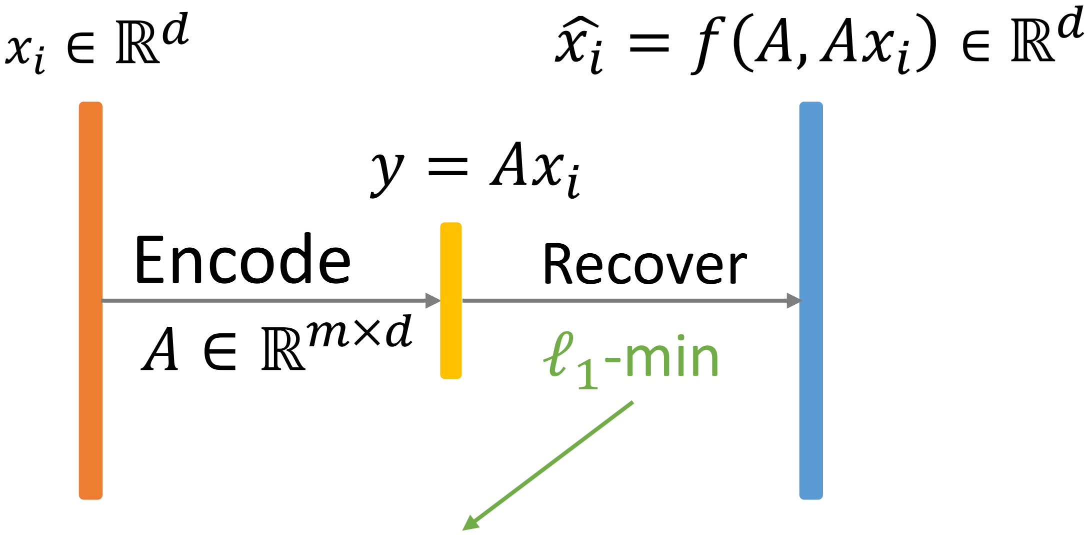
$$\min_{A \in \mathbb{R}^{m \times d}} \sum_{i=1}^{n} \|x_i - f(A, Ax_i)\|_2^2$$

Problem:

How to compute gradient w.r.t. A?

Learning a measurement matrix

- Training data: n sparse vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$



$x_i \in \mathbb{R}^d$

$\hat{x}_i = f(A, Ax_i) \in \mathbb{R}^d$

$y = Ax_i$

Encode

$A \in \mathbb{R}^{m \times d}$

Recover

ℓ_1-min

$$f(A, y) \coloneqq \operatorname{argmin}_{x'} \|x'\|_1 \quad \text{s.t. } Ax' = y$$

Objective function:

$$\min_{A \in \mathbb{R}^{m \times d}} \sum_{i=1}^{n} \|x_i - f(A, Ax_i)\|_2^2$$

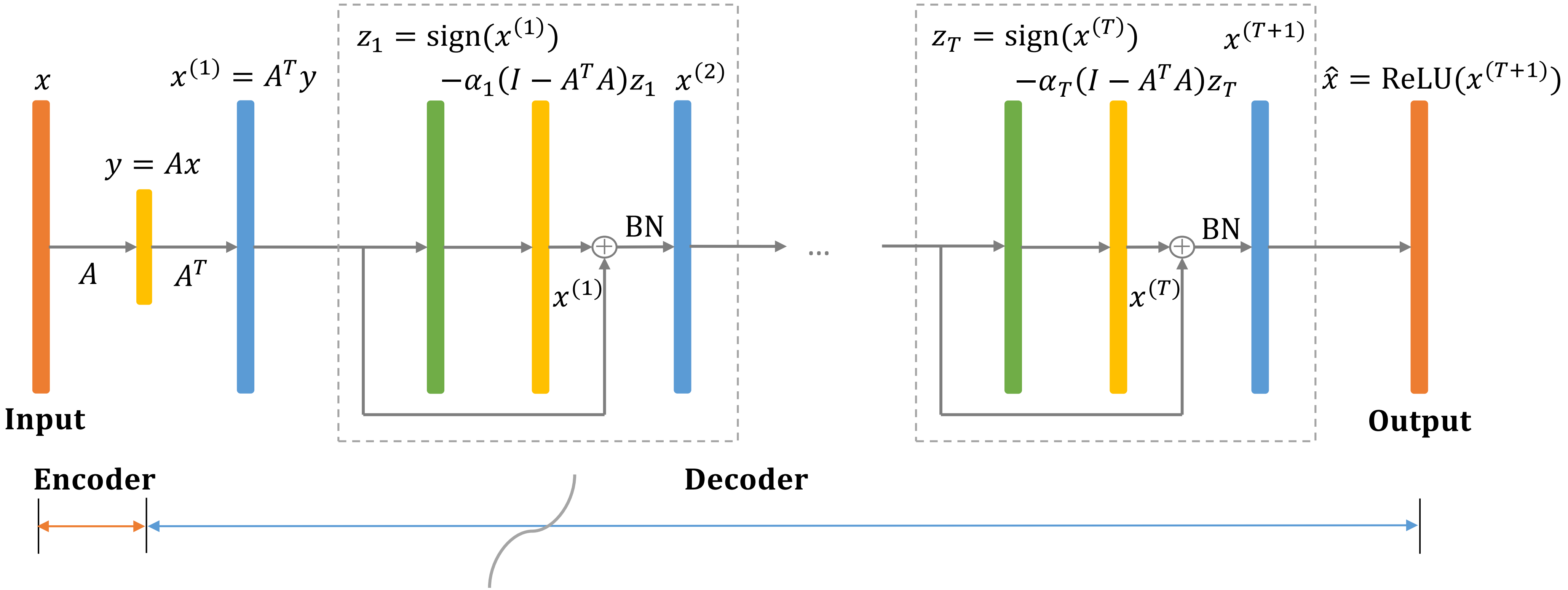
Problem:

How to compute gradient w.r.t. A?

Key idea:

Replace $f(A, y)$ by a few steps of projected subgradient

ℓ_1-AE: a novel autoencoder architecture

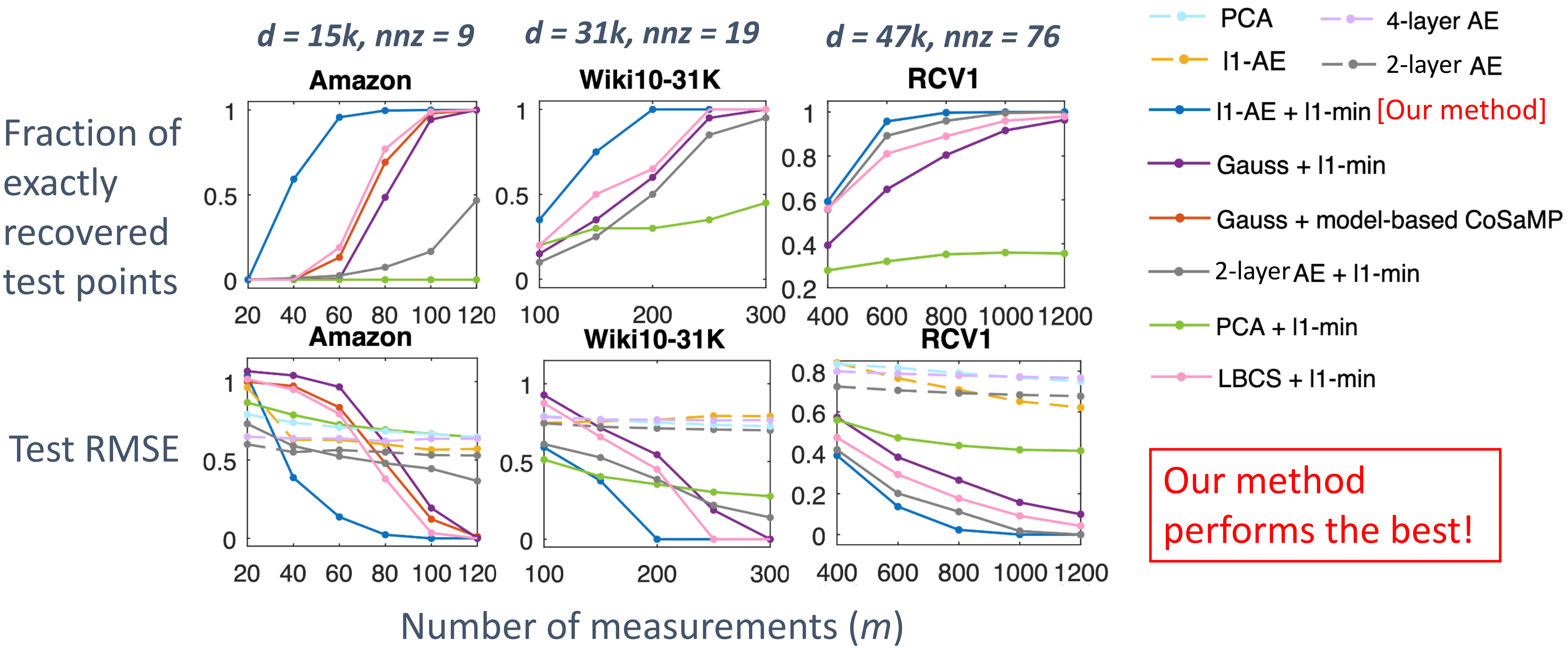


x
$x^{(1)} = A^T y$
$y = Ax$
A
A^T
Input
$z_1 = \text{sign}(x^{(1)})$
$-\alpha_1(I - A^T A)z_1$
$x^{(2)}$
BN
$x^{(1)}$
...
$z_T = \text{sign}(x^{(T)})$
$x^{(T+1)}$
$-\alpha_T(I - A^T A)z_T$
BN
$x^{(T)}$
$\hat{x} = \text{ReLU}(x^{(T+1)})$
Output
Encoder
Decoder

One step of projected subgradient

$$x^{(t+1)} = x^{(t)} - \alpha_t(I - A^T A)\text{sign}(x^{(t)})$$

Real sparse datasets



Our method performs the best!

Summary

- **Key idea:** We learn a compressed sensing measurement matrix by unrolling the projected subgradient of ℓ_1-min decoder
- Implemented as an autoencoder ℓ_1-AE
- Compared 12 algorithms over 6 datasets (3 synthetic and 3 real)
- Our method created perfect reconstruction with 1.1-3X fewer measurements compared to previous state-of-the-art methods
- Applied to Extreme multilabel classification, our method outperforms

SLEEC (Bhatia et al., 2015)

Wed Jun 12th 06:30 -- 09:00 PM @ Pacific Ballroom #189