Link to paper:

https://goo.gl/rB3eBN

Minimax Reconstruction in (Convolutional) Sparse Dictionary Learning

Shashank Singh, Barnabás Póczos, and Jian Ma

Carnegie Mellon University

Carnegie Mellon University

Introduction

- Sparse dictionary learning (a.k.a. sparse coding) is widely used to denoise data
- *Convolutional* Sparse Dictionary Learning (CSDL) is popular for data with translation-invariant features (e.g., images, sound, movies, genomics, etc.) [1]
  - Translation invariant dictionary $\Rightarrow$ smaller dictionary and greater sparsity
- We study minimax denoising of convolutionally sparse data

Contributions

1. First bounds on minimax reconstruction/denoising risk of CSDL.
2. Most work in compressed sensing assumes mutually independent noise; we show this often-unrealistic assumption is not necessary for CSDL denoising.
3. Prior theory for sparse dictionary learning makes strong assumptions to ensure identifiability of dictionary (e.g. incoherence or restricted isometry properties). We show that, unlike dictionary recovery, sparse dictionary denoising *requires no assumptions whatsoever on dictionary.*

Notation

- **Multi-convolution:** For two matrices $R \in \mathbb{R}^{(N-n+1)\times K}$ and $D \in \mathbb{R}^{n \times K}$ with equal numbers of columns, we define multi-convolution $\otimes$ by

$$R \otimes D = \sum_{k=1}^{K} R_k * D_k \in \mathbb{R}^N,$$

where $*$ denotes the usual convolution operator.

- **Matrix Norms:** For $A \in \mathbb{R}^{n\times m}$ and $p \in [0,\infty]$, we write

$$\|A\|_{p,q} := \left(\sum_{j=1}^{m} \left(\sum_{i=1}^{n} |A_{i,j}|^p \right)^{q/p} \right)^{1/q}.$$

- **Problem Domain:** For $N, K, n \in \mathbb{N}$ and $\lambda \geq 0$, we write

$$\mathcal{S}_\lambda := \left\{ (R, D) \in \mathbb{R}^{(N-n+1)\times K} \times \mathbb{R}^{n\times K} : \|D\|_{2,\infty} \leq 1, \|R\|_{1,1} \leq \lambda \right\}.$$

Modeling Assumptions

- Notation:
  - Single Observation $Y \in \mathbb{R}^N$
  - True signal $X \in \mathbb{R}^N$
  - Noise $\epsilon \in \mathbb{R}^N$
  - Dictionary $D \in \mathbb{R}^{n\times K}$
  - Encoding $R \in \mathbb{R}^{(N-n+1)\times K}$

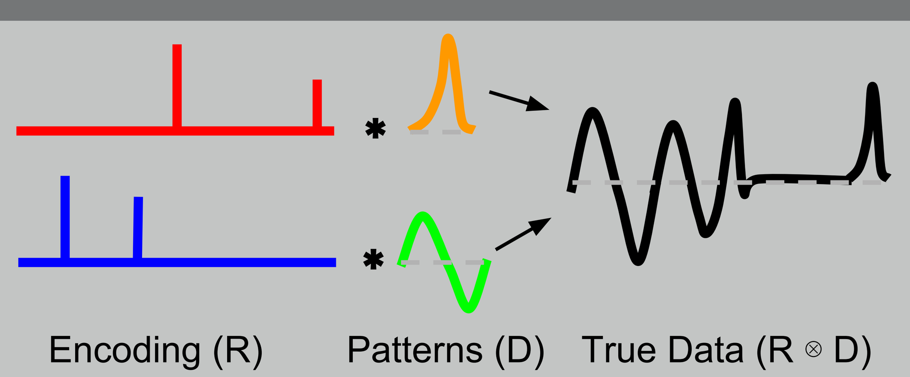


Figure: Temporal Linear Generative Model.

- **Temporal Linear Generative Model (TLGM)** [2]:

$$Y = X + \epsilon, \quad \text{where } X = R \otimes D, \quad \text{for some } (R, D) \in \mathcal{S}_\lambda.$$

- We consider several possible noise assumptions:
  1. ϵ is called **componentwise σ^2-sub-Gaussian** ($\epsilon \in \textbf{CSG}(\sigma^2)$) if
  $$\max_{i\in\{1,\dots,N\}} \mathbb{E}\left[e^{t\epsilon_i}\right] \leq e^{t^2\sigma^2/2}, \quad \text{for all } t \in \mathbb{R}.$$
  2. ϵ is called **jointly σ^2-sub-Gaussian** ($\epsilon \in \textbf{JSG}(\sigma^2)$) if
  $$\mathbb{E}\left[e^{\langle t,\epsilon\rangle}\right] \leq e^{\|t\|_2^2\sigma^2/2}, \quad \text{for all } t \in \mathbb{R}^N.$$
  Note: In general, $\textbf{JSG}(\sigma^2) \subseteq \textbf{CSG}(\sigma^2)$ and $\textbf{CSG}(\sigma^2) \subseteq \textbf{JSG}(N\sigma^2)$. If the entries of ϵ are mutually independent, then $\textbf{CSG}(\sigma^2) \subseteq \textbf{JSG}(\sigma^2)$.
  3. Paper also has bounds under weaker bounded-moment assumptions.

CSDL Estimator

$$\widehat{X}_\lambda = \widehat{R}_\lambda \otimes \widehat{D}_\lambda \quad \text{where} \quad \left(\widehat{R}_\lambda, \widehat{D}_\lambda\right) := \operatorname*{argmin}_{(R,D)\in\mathcal{S}_\lambda} \|Y - R \otimes D\|_2^2.$$

- Computable by alternating (between R and D) projected gradient descent
- Paper also has similar results for $\|R\|_{1,1}$-penalized version

References

[1] Vardan Papyan, Jeremias Sulam, and Michael Elad. Working locally thinking globally-part I: Theoretical guarantees for convolutional sparse coding. *arXiv preprint arXiv:1607.02005*, 2016.

[2] Bruno A. Olshausen. Probabilistic Models of the Brain. page 257, 2002.

[3] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 407–410. IEEE, 2015.

Theoretical Results: Upper Bounds

- **Lemma 1 (Oracle Inequality):** If $Y = X + \epsilon$, then

$$\|X - \widehat{X}_\lambda\|_2^2 \leq \inf_{(R,D)\in\mathcal{S}_\lambda} \underbrace{\|X - R\otimes D\|_2^2}_{\text{model misspecification}} + \underbrace{2\langle \epsilon, \widehat{X}_\lambda - R \otimes D\rangle}_{\text{statistical error}}.$$

$\Rightarrow \widehat{X}_\lambda$ robust to violation of TLGM assumption.

- **Theorem 2:** Under TLGM with $\epsilon \in \textbf{CSG}(\sigma^2)$,

$$\frac{1}{N}\mathbb{E}\left[\left\|\widehat{X}_\lambda - X\right\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}.$$

- **Theorem 3:** Under TLGM with $\epsilon \in \textbf{JSG}(\sigma^2)$,

$$\frac{1}{N}\mathbb{E}\left[\left\|\widehat{X}_\lambda - X\right\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2\log(2(N-n+1))}}{N}.$$

- **Note:** Under TLGM, we always have $\frac{1}{N}\mathbb{E}\left[\left\|\widehat{X}_0 - X\right\|_2^2\right] \leq \frac{\lambda^2}{N} \Rightarrow$ Under extreme sparsity/noise ($\lambda \ll \sigma\sqrt{n\log N}$), trivial estimate $\widehat{X} = 0$ is better.

Theoretical Results: Lower Bounds

- **Minimax Error:** For $\lambda \in [0,\infty]$, $N > n \in \mathbb{N}$, and a class $\mathcal{E}$ of $\mathbb{R}^N$-valued random variables,

$$M(\lambda, N, n, \mathcal{E}) := \inf_{\widehat{X}:\mathbb{R}^N\to\mathbb{R}^N} \sup_{(R,D)\in\mathcal{S}_\lambda, \epsilon\in\mathcal{E}} \frac{1}{N}\mathbb{E}\left[\left\|\widehat{X}(Y) - X\right\|_2^2\right].$$

- **Theorem 4 (Componentwise Sub-Gaussian Noise):**

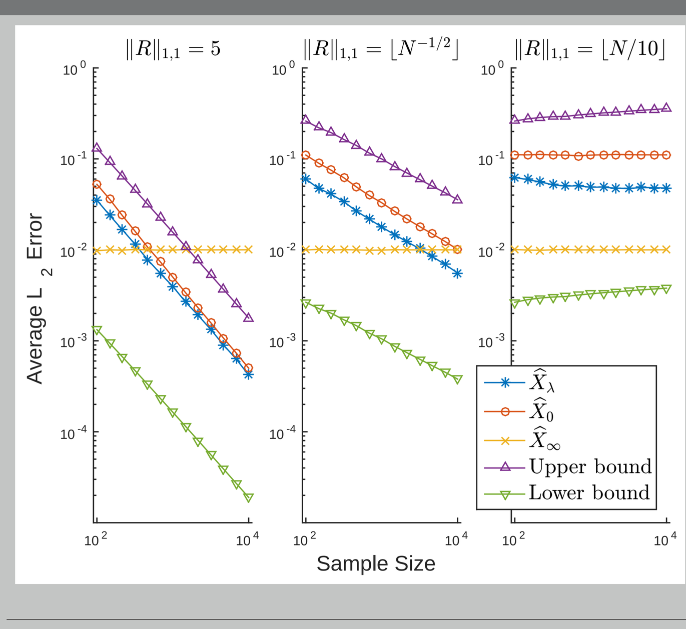
$$M(\lambda, N, n, \textbf{CSG}(\sigma^2)) \geq \frac{\lambda}{8N}\min\left\{\lambda, \sigma\sqrt{n\log(N-n+1)}\right\}.$$

- **Theorem 5 (Jointly Sub-Gaussian Noise):**

$$M(\lambda, N, n, \textbf{JSG}(\sigma^2)) \geq \frac{\lambda}{8N}\min\left\{\lambda, \sigma\sqrt{\log(N-n+1)}\right\}.$$

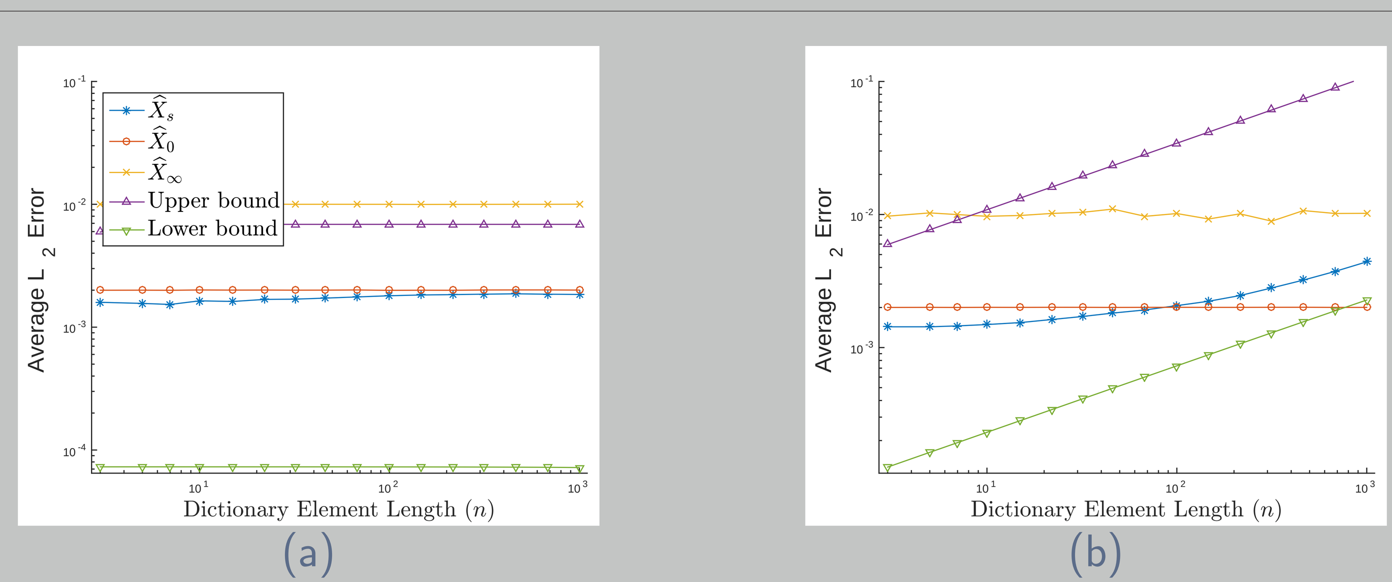
- **Note:** Lower bounds hold even when D is known in advance; estimating X is about as hard as estimating R (convolutional sparse recovery).

Simulation Results



Experiment 1: Average $\mathcal{L}_2$-error as a function of signal length N, with sparsity scaling as:

1. $\|R\|_{1,1} = 5$ (Panel 1)
2. $\|R\|_{1,1} = \left\lfloor \sqrt{N} \right\rfloor$ (Panel 2)
3. $\|R\|_{1,1} = \lfloor N/10 \rfloor$ (Panel 3)



Experiment 2: Average $\mathcal{L}_2$-error as a function of dictionary element length n, when entries of noise ϵ are (a) IID and (b) perfectly correlated.

Conclusions

- (Convolutional) sparse dictionary denoising is essentially **assumption-free**.
- For fixed n, CSDL is worst-case consistent (in reconstruction risk) if and only if

$$\frac{\lambda\sigma\sqrt{\log(N)}}{N} \to 0.$$

- When noise is independent, error is independent of dictionary atom length n.
- Similar results hold for classical dictionary learning (replace $n \to d$, $N \to \frac{N}{d}$).
- Alternating minimization appears minimax-rate optimal, consistent with recent results suggesting that local optima in SDL are often global optima [3]

Acknowledgements

This material is based upon work supported by the NSF Graduate Research Fellowship DGE-1252522 (to S.S.). The work was also partly supported by NSF IIS-1563887, Darpa D3M program, and AFRL (to B.P.), and NSF IIS-1717205 and NIH HG007352 (to J.M.).

http://www.andrew.cmu.edu/user/sss1/

sss1@andrew.cmu.edu