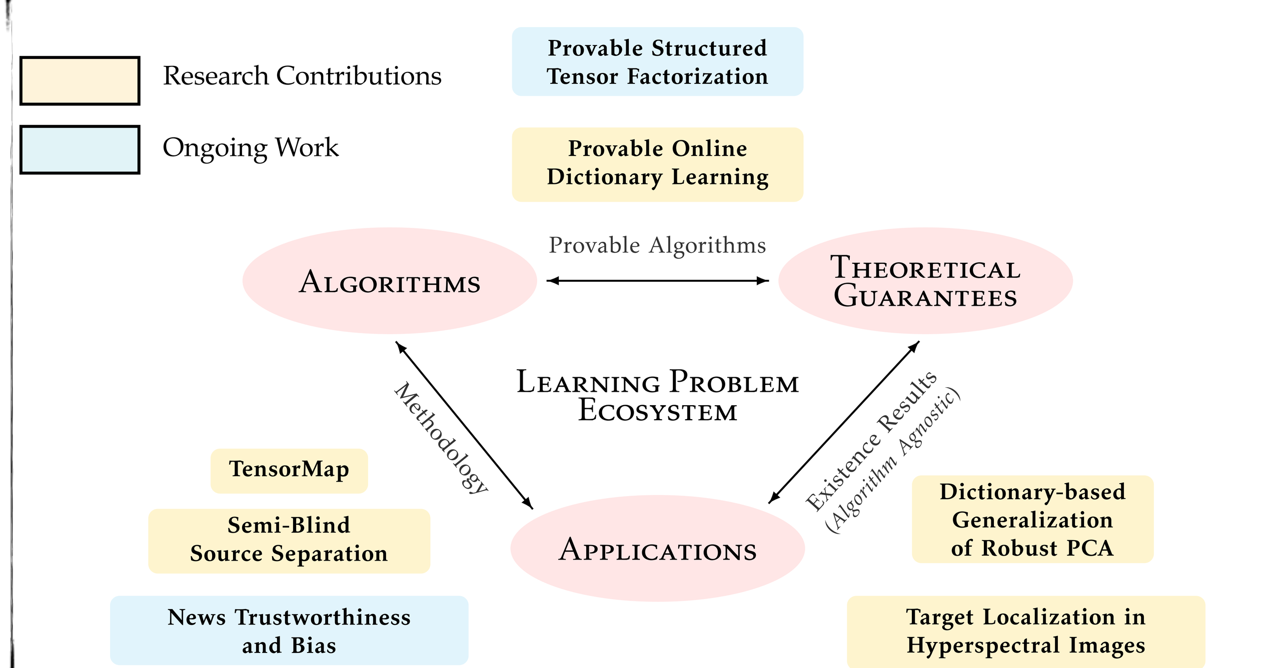
NOODL: Provable Online Dictionary Learning and Sparse Coding

Sirisha Rambhatla†, Xingguo Li‡, and Jarvis Haupt†

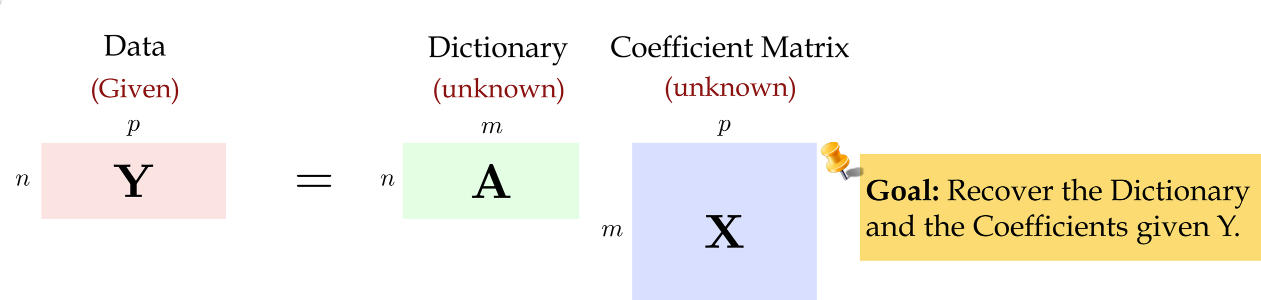
†Dept. of Electrical and Computer Engineering, University of Minnesota-Twin Cities, Minneapolis, MN. ‡Dept. of Computer Science, Princeton University, Princeton, NJ, USA

Overview: Dictionary learning models a given data sample (vector) as a sparse linear combination of a few columns of a matrix known as a *dictionary*. Here, the weights characterizing the sparse linear combination are known as *coefficients*. Since both the dictionary and the coefficients parameterizing the linear model are unknown, the associated optimization formulations are inherently non-convex. In this work, we develop NOODL — a Neurally plausible alternating Optimization-based Online Dictionary Learning algorithm which, to the best of our knowledge, is the first algorithm that provably recovers both factors of the dictionary learning model simultaneously, that too at a linear rate, under some relatively mild conditions.

0. Overview of Research Efforts



1. Dictionary Learning



Overview of Dictionary Learning Techniques

Popular but Limited or No Guarantees	Provable Algorithms
Regularized Least Squares-based (Olshausen et. al. '96, Lewicki et. al. '00, Lee et. al. '07, Mairal et. al. '09, Kreutz-Delgado et. al.'03) ***Greedy Methods:*** Method of Optimal Directions (Engan et. al. '99), and k-SVD (Aharon et. al. '06)	***Convex Relaxations*** ℓ_1 (Gribonval et. al. '10, Jenatton et. al. '12, Geng et. al. '14) ***Hard Constrained*** ℓ_0 (Agarwal et. al. '14, Arora et. al. '14 & '15)

2. Motivation

Error-in-Variables Model

Non-negligible error in dictionary → **Ability to estimate coefficients degrades** (Wainwright '09, Belloni '16) → **Limited Guarantees** (Donoho '06) → **No exact or support recovery guarantee!**

3. Our Contributions

An algorithm for online dictionary learning, which once initialized appropriately achieves:

Exact recovery of both factors | **Unbiased estimation** | **An Iterative Hard Thresholding (IHT)-based update for the coefficients** | **Linear convergence guarantees** | **Online algorithm**

Of independent interest:	"Neurally Plausible"	Scalable:	Easy to tune:
Analysis of IHT for exact recovery of sparse coefficients even when the dictionary is known.	Can be implemented in a neural architecture.	Amenable for large scale distributed implementation	Specific guidelines on choosing the parameters.

4. Main Result

Under certain conditions on initialization and incoherence, the current state-of-the-art guarantees the following with probability $(1-\delta_{[2]})$ for some small $\delta_{[2]}$.

Arora et. al. `15 [2] $\quad \mathbf{E}[\|\mathbf{A}^{(t)} - \mathbf{A}^*\|^2] \le (1-\omega)^t \|\mathbf{A}^{(0)} - \mathbf{A}^*\|^2 + \mathcal{O}(k/n)$ **Non-negligible Estimation Error in Dictionary!** *(Falls under Error-in-Variables Model)*

Under similar conditions on initialization and incoherence, our algorithm NOODL guarantees the following.

Main Result Under some assumptions on initialization and incoherence (shown below), when our algorithm NOODL is provided with $p = \tilde{\Omega}(mk^2)$ new samples at each iteration t generated according to the model described. Then for some $0 < \omega < 1/2$, the estimate $\mathbf{A}^{(t)}$ at (t)-th iteration satisfies

$$\|\mathbf{A}_i^{(t)} - \mathbf{A}_i^*\|^2 \le (1-\omega)^t \|\mathbf{A}_i^{(0)} - \mathbf{A}_i^*\|^2, \text{ for all } t = 1, 2, \ldots.$$

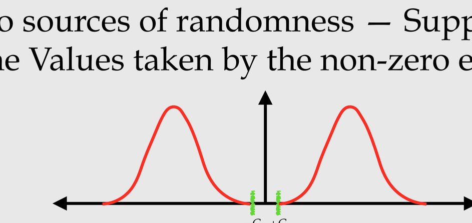
No Bias in Dictionary Estimation!

Furthermore, given $R = \log(n)$, with probability at least $(1-\delta_{\text{alg}})$ for some small constant δ_{alg}, the coefficient estimate $\hat{\mathbf{x}}_i^{(t)}$ at t-th iteration has the correct signed-support and satisfies

$$(\hat{\mathbf{x}}_i^{(t)} - \mathbf{x}_i^*)^2 = \mathcal{O}(k(1-\omega)^{t/2}\|\mathbf{A}_i^{(0)} - \mathbf{A}_i^*\|), \text{ for all } i \in (\mathbf{x}^*).$$

Simultaneous Coefficient Recovery!

Assumptions

Incoherence of Dictionary	Properties of Coefficients	Good Dictionary Initialization	Sparsity	Parameter Choice
Columns of $\mathbf{A}^*$are sufficiently "spread-out"	Two sources of randomness — Support and the Values taken by the non-zero entries 	$\|\mathbf{A}^{(0)} - \mathbf{A}^*\| \le 2\|\mathbf{A}^*\|$ $\|\mathbf{A}_i^{(0)} - \mathbf{A}_i^*\| \le \epsilon_0$	$k = \mathcal{O}(\sqrt{n})$	$\eta_A = \Theta(m/k)$ $\eta_x < c_1(\epsilon_t, \mu, n, k) < 1$ $\tau = c_2(k)$

5. NOODL: Neurally Plausible alternating Optimization-based Online Dictionary Learning

Algorithm 1: NOODL: Neurally plausible alternating Optimization-based Online Dictionary Learning.

Input: Fresh data (samples) $\mathbf{y}_{(j)} \in \mathbb{R}^n$ for $j = 1 \ldots p$ per iteration t, η_A chosen according to our assumptions, $\mathrm{T} = \Omega(\log(1/\epsilon_{\mathrm{T}}))$, and $\mathrm{R} = \Omega(\log(1/\delta_{\mathrm{R}}))$

Initialize: Estimate $\mathbf{A}^{(0)}$, which is $(\epsilon_0, 2)$-near to $\mathbf{A}^*$ for $\epsilon_0 = \mathcal{O}^*(1/\log(n))$

for $t = 0$ ***to*** $\mathrm{T}-1$ **do**

- **Predict: (Estimate Coefficients)** — **Form Coefficient Estimate**
- **for** $j = 1$ ***to*** p **do**
  - **Initialize:** $\mathbf{x}_{(j)}^{(0)} = \mathcal{T}_{C/2}(\mathbf{A}^{(t)\top}\mathbf{y}_{(j)})$
  - **for** $r = 0$ ***to*** $\mathrm{R}-1$ **do** — *Algorithmic Novelty*
    - Choose $\eta_x^{(r)}$ and $\tau^{(r)}$ as per assumptions
    - Update: $\mathbf{x}_{(j)}^{(r+1)} = \mathcal{T}_{\tau^{(r)}}(\mathbf{x}_{(j)}^{(r)} - \eta_x^{(r)}\, \mathbf{A}^{(t)\top}(\mathbf{A}^{(t)}\mathbf{x}_{(j)}^{(r)} - \mathbf{y}_{(j)}))$
  - **end** — Iterative Hard Thresholding
- **end**
- $\hat{\mathbf{x}}_{(j)} := \mathbf{x}_{(j)}^{(\mathrm{R})}$ for each $j \in [p]$
- **Learn: (Update Dictionary)** — **Update Dictionary**
- Form empirical gradient estimate : $\hat{\mathbf{g}}^{(t)} = \frac{1}{p}\sum_{j=1}^{p}(\mathbf{A}^{(t)}\hat{\mathbf{x}}_{(j)} - \mathbf{y}_{(j)})\mathrm{sign}(\hat{\mathbf{x}}_{(j)})^\top$
- Take a gradient step: $\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} - \eta_A\, \hat{\mathbf{g}}^{(t)}$
- Normalize: $\mathbf{A}_i^{(t+1)} = \mathbf{A}_i^{(t+1)}/\|\mathbf{A}_i^{(t+1)}\|$ for all $i \in [m]$. — Approximate Gradient Descent

end

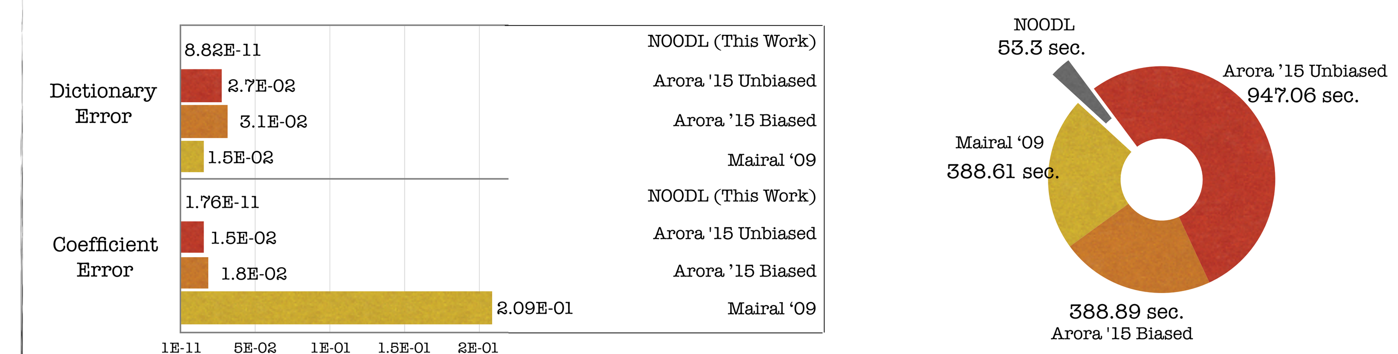
Output: The dictionary $\mathbf{A}^{(\mathrm{T})}$ and coefficient estimates $\hat{\mathbf{x}}_{(j)}$ for $j \in [1, p]$ at each iterate t.

Definitions

$\mathbf{y}_{(j)} \in \mathbb{R}^n$	$\mathbf{A}^{(0)} \in \mathbb{R}^{n\times m}$	R	η_x	T	η_A
j-th data sample/ vector	Initial dictionary estimate $\|\mathbf{A}_i^{(0)} - \mathbf{A}_i^*\| \le \epsilon_0$ where $\epsilon_0 = \mathcal{O}^*(1/\log(n))$ and $\|\mathbf{A}^{(0)} - \mathbf{A}^*\| \le 2\|\mathbf{A}^*\|$	Number of inner loop i.e. coefficient update iterations $\mathrm{R} = \Omega(\log(1/\delta_{\mathrm{R}}))$ where $(1-\eta_x)^{\mathrm{R}} \le \delta_{\mathrm{R}}$	Learning rate for coefficient update	Number of outer loop i.e. dictionary update iterations $\mathrm{T} = \Omega(\log(1/\epsilon_{\mathrm{T}}))$ where $\|\mathbf{A}_i^{(\mathrm{T})} - \mathbf{A}^*\| \le \epsilon_{\mathrm{T}}$	Learning rate for dictionary update
$\mathbf{x}_{(j)}^{(0)} \in \mathbb{R}^m$			τ		$\hat{\mathbf{g}}^{(t)}$
Sparse coefficient vector with at most k non-zeros			Threshold for hard thresholding step		Empirical gradient matrix at t-th iterate

6. Experimental Results

Recovery and Timing Performance

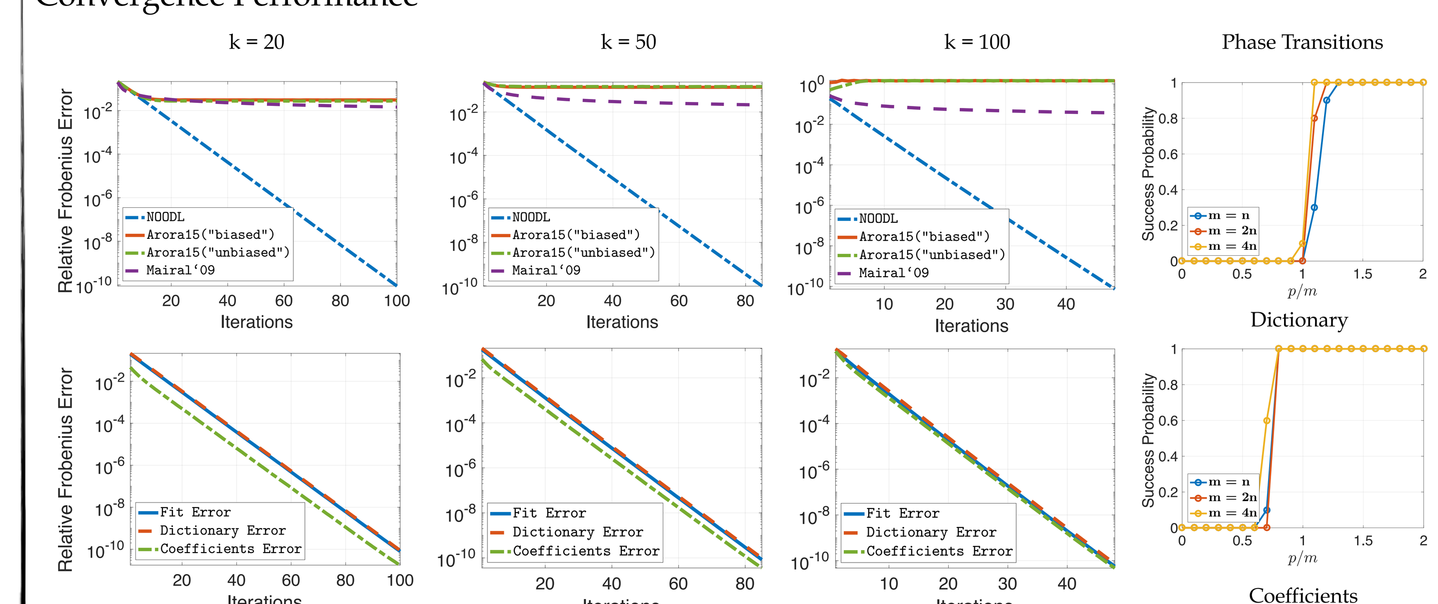


Recovery Performance | Time Per Iteration

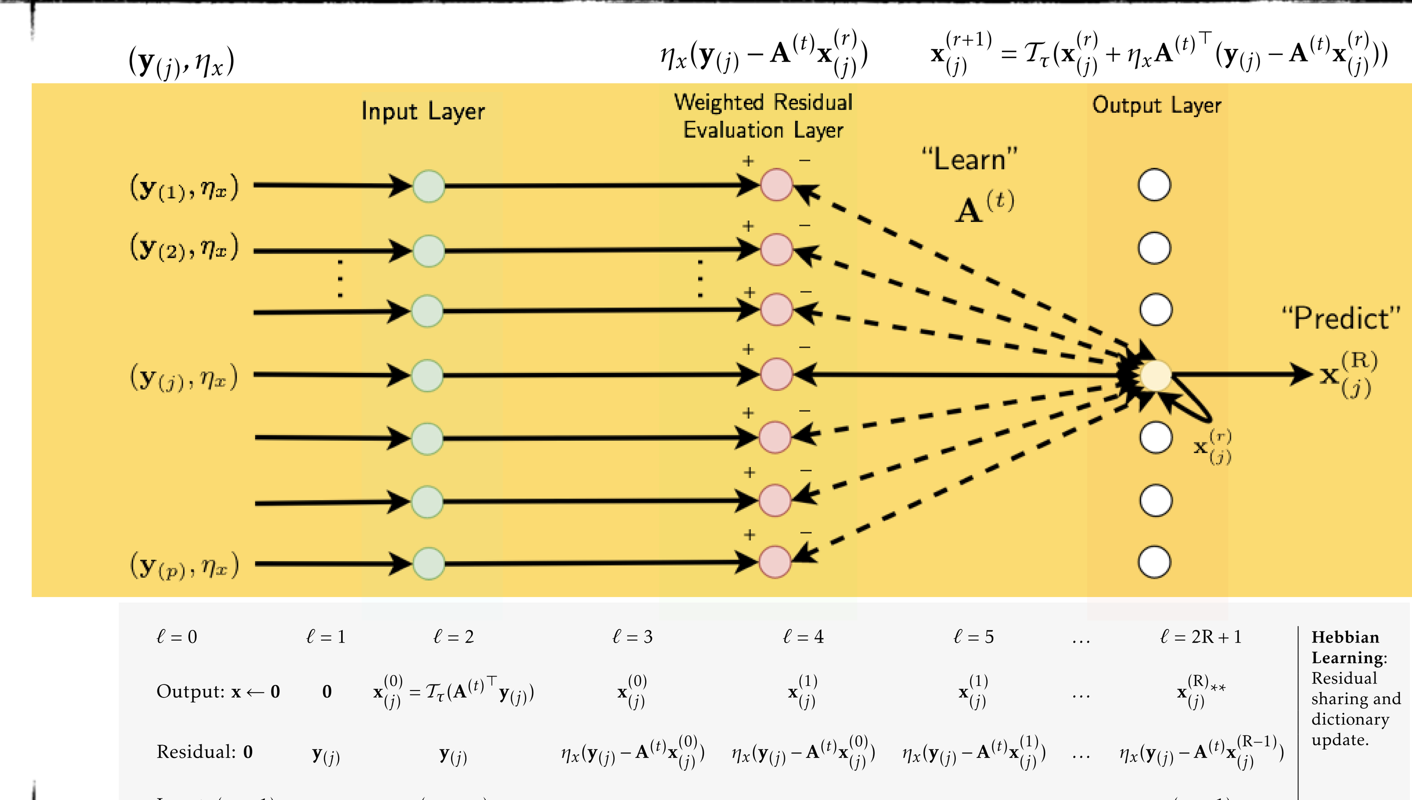
Here, n = 1000, m = 1500, and k = 20. For the techniques presented in [2], we scan across 50 values of the regularization parameter for sparse approximation for coefficient estimation after learning the dictionary, while for Mairal '09 [3] , we scanning across 10 values of the regularization parameters at each step. See [1] for additional results.

Take-away: NOODL is significantly more accurate and faster than the current state-of-the-art techniques, while also providing guarantees on recovery of both factors!

Convergence Performance



7. A Prototype Neural Implementation



8. Selected References

[1] S. Rambhatla, X. Li, and J. Haupt. NOODL: Provable Online Learning for Dictionary Learning and Sparse Coding. To appear in the proceedings of the *International Conference on Learning Representations (ICLR)*, 2019.

[2] S. Arora, R. Ge, T. Ma and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. *In Conference on Learning Theory (COLT)*, 2015.

[3] J. Mairal, F. Bach, J. Ponce and G. Sapiro. Online dictionary learning for sparse coding. In Proceedings of the *International Conference on Machine Learning (ICML)*. 2009.

9. Acknowledgements

- The authors graciously acknowledge support from DARPA Young Faculty Award, Grant No. N66001-14-1-4047.
- Also, the authors would like to thank University Imaging Centers at the University of Minnesota for their services.

Sirisha Rambhatla, email: rambh002@umn.edu, web: sirisharambhatla.com