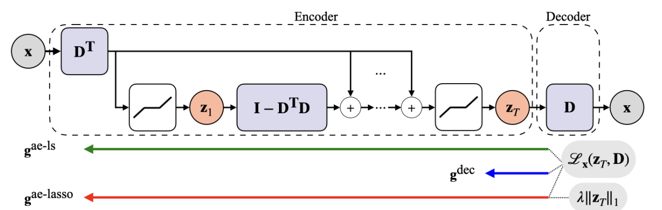
Interpretable Unrolled Dictionary Learning Networks

Bahareh Tolooshams and Demba Ba
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

I. Introduction

The dictionary learning problem, representing data $\boldsymbol{x} \in \mathbb{R}^m$ as a combination of a few atoms from a dictionary $\boldsymbol{D} \in \mathbb{R}^{m\times p}$, has long stood as a popular method for learning representations in statistics and signal processing [1, 2, 3, 4]. The most popular dictionary learning algorithm alternates between sparse coding and dictionary update steps. Sparse coding has been utilized to construct neural architectures through recurrent sparsifying encoders [5], initiating a growing literature on constructing interpretable unrolled networks [6, 7]. We offer the theoretical analysis of unrolled sparse coding. We address the following challenge; the vanilla unrolled sparse coding computes a biased code estimate; this results in a biased estimate of the backward gradient. We reduce this bias and demonstrate unrolled interpretability.

Given $\boldsymbol{x}$ and $\boldsymbol{D}$, the problem of recovering the sparse coefficients $\boldsymbol{z} \in \mathbb{R}^p$ is referred to as sparse coding, and can be solved through the lasso [8] $\ell_{\boldsymbol{x}}(\boldsymbol{D}) \coloneqq \min_{\boldsymbol{z}\in\mathbb{R}^p} \mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}, \boldsymbol{D}) + h(\boldsymbol{z})$ where $\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}, \boldsymbol{D}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{z}\|_2^2$, and $h(\boldsymbol{z}) = \lambda\|\boldsymbol{z}\|_1$. The problem aims to recover a dictionary $\boldsymbol{D}^*$ that generates the data, i.e., $\boldsymbol{x} = \boldsymbol{D}^*\boldsymbol{z}^*$ where $\boldsymbol{z}^*$ is sparse. Prior to unrolled networks, gradient-based dictionary learning relied on analytic gradients. With unrolled networks, backpropagation gained attention for parameter estimation [9, 10, 11][1].



1: Unrolled dictionary learning.

Unrolled dictionary learning is constructed as following: sparse coding is converted into an encoder by unfolding T iterations of ISTA $(\boldsymbol{z}_{t+1} = \Phi(\boldsymbol{z}_t, \boldsymbol{D}) = \mathcal{P}_{\alpha\lambda}(\boldsymbol{z}_t - \alpha\nabla_1\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}_t, \boldsymbol{D})))$ with $\mathcal{P}_b(v) \triangleq \operatorname{sign}(v)\max(|v| - b, 0)$ [12, 13]. The decoder is $\hat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{z}_T$. We recover $\boldsymbol{D}^*$ by backpropagated gradient (i.e., $\boldsymbol{D}^{(l+1)} = \boldsymbol{D}^{(l)} - \eta\boldsymbol{g}_T^{(l)}$) (See Fig. 1). Backpropagation through the decoder results in the analytic gradient $\boldsymbol{g}_T^{\text{dec}} \triangleq \nabla_2\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}_T, \boldsymbol{D})$. The gradients $\boldsymbol{g}_T^{\text{ae-lasso}} \triangleq \nabla_2\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}_T, \boldsymbol{D}) + \frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{D}}(\nabla_1\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}_T, \boldsymbol{D}) + \partial h(\boldsymbol{z}_T))$ and $\boldsymbol{g}_T^{\text{ae-ls}} \triangleq \nabla_2\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}_T, \boldsymbol{D}) + \frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{D}}\nabla_1\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}_T, \boldsymbol{D})$ are computed by backpropagation through the autoencoder using the lasso and least-squares objectives. We show how using $\boldsymbol{g}_T^{\text{ae-ls}}$ is a better gradient estimator to recover $\boldsymbol{D}^*$. The desired direction is $\boldsymbol{g}^* \triangleq \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}}[\nabla_2\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{z}^*, \boldsymbol{D})]$.

II. Main Results

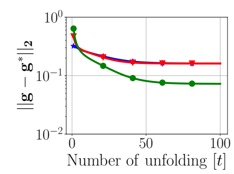
Assumptions: The code $\boldsymbol{z}^*$ is at most s-sparse with the support $S^* = \operatorname{supp}(\boldsymbol{z}^*)$. Given the support, $\boldsymbol{z}_{S^*}^*$ is i.i.d, $\mathbb{E}[\boldsymbol{z}_{(j)}^* \mid j \in S^*] = 0$ and $\mathbb{E}[\boldsymbol{z}_{(S)}^*\boldsymbol{z}_{(S^*)}^{*\text{T}} \mid S^*] = \boldsymbol{I}$. $\boldsymbol{D}^*$ is μ-incoherent with $\mu = \mathcal{O}(\log(m))$. $\|\boldsymbol{D}_j^*\|_2 = 1$ and $\|\boldsymbol{D}^*\|_2 = \mathcal{O}(\sqrt{p/m})$, and $p = \mathcal{O}(m)$. $\forall j$ $\|\boldsymbol{D}_j^{(0)} - \boldsymbol{D}_j^*\|_2 \le \delta$ and $\|\boldsymbol{D}^{(0)} - \boldsymbol{D}^*\|_2 \le 2\|\boldsymbol{D}^*\|_2$. $\boldsymbol{D}^{(l)}$ is μ_l-incoherent and $\|\boldsymbol{D}_j^{(l)} - \boldsymbol{D}_j^*\|_2 \le \delta_l$ with $\delta_l = \mathcal{O}^*(1/\log p)$.

Results: Thm. II.1 provides an upper bound (as a function of dictionary error, amount of unrolling, and sparse regularizer) on the error between the true code $\boldsymbol{z}^*$ and the sparse latent code $\boldsymbol{z}_t$. The λ term in the upper bound shows that the code error when we strictly perform ℓ_1-norm based sparse coding does not go to zero.

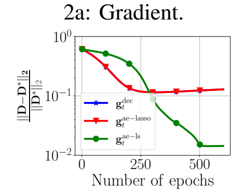
Theorem II.1. *If $s = \mathcal{O}^*(\sqrt{m}/\mu\log m)$, and the regularizer and step size are $\lambda_t^{(l)} = \lambda = \frac{\mu_l}{\sqrt{m}}\|\boldsymbol{z}^* - \boldsymbol{z}_0\|_1 + a_\gamma = \boldsymbol{\Omega}(\frac{s\log m}{\sqrt{m}})$ and $\alpha^{(l)} \le 1 - \frac{2\lambda_t^{(l)} - (1 - \frac{\delta_l^2}{2})C_{\min}}{\lambda_{t-1}^{(l)}}$, then with high probability, $|\boldsymbol{z}_{t,(j)}^{(l)} - \boldsymbol{z}_{(j)}^*| \le \mathcal{O}(\sqrt{s\|\boldsymbol{D}_j^{(l)} - \boldsymbol{D}_j^*\|_2} + e_{t,j}^{(l)} + \lambda)$ where $e_{t,j}^{(l)} \to 0$ as $t \to \infty$.*

Given the forward pass, Thm. II.2 shows that training unrolled dictionary learning network using gradient descent with $\boldsymbol{g}_T^{\text{dec}}$ results in a contractive mapping, hence recovery of the dictionary up to a $\boldsymbol{D}^*$ neighborhood mainly characterized by the regularization parameter λ.

Theorem II.2. *If $s = \mathcal{O}(\sqrt{m})$, $\eta = \mathcal{O}(\frac{p}{s(1-\delta_l^2/2)})$, and the regularizer λ and α are set according to above, then with high probability $\|\boldsymbol{D}_j^{(l+1)} - \boldsymbol{D}_j^*\|_2^2 \le (1-\psi)\|\boldsymbol{D}_j^{(l)} - \boldsymbol{D}_j^*\|_2^2 + \epsilon_\lambda^{(l)}$ where $\epsilon_\lambda^{(l)} \coloneqq \eta\frac{2p}{s(1-\langle \boldsymbol{D}_j^* - \boldsymbol{D}_j^{(l)}, \boldsymbol{D}_j^*\rangle)}\lambda^2$.*



2a: Gradient.



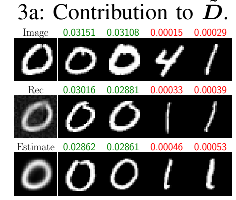
2b: Dictionary.

To reduce this bias in the dictionary update, we propose to use backpropagation using the reconstruction loss $\boldsymbol{g}_\infty^{\text{ae-ls}}$ instead of the analytic gradient $\boldsymbol{g}_\infty^{\text{dec}}$. Theorem II.3 compares the gradients for appropriately large T; it shows that $\boldsymbol{g}_T^{\text{ae-ls}}$ is a better estimator of the desired direction $\boldsymbol{g}^*$.

Theorem II.3. *$\boldsymbol{g}_T^{\text{ae-lasso}}$ is equivalent to $\boldsymbol{g}_T^{\text{dec}}$ as $T \to \infty$ (Fig. 2a). $\|\boldsymbol{g}_T^{\text{ae-lasso}} - \boldsymbol{g}^*\|_2 \le \mathcal{O}(\|\boldsymbol{D} - \boldsymbol{D}^*\|_2 + \delta^* + C_{\text{lasso}})$ and $\|\boldsymbol{g}_T^{\text{ae-ls}} - \boldsymbol{g}^*\|_2 \le \mathcal{O}(\|\boldsymbol{D} - \boldsymbol{D}^*\|_2 + \delta^*)$, where δ^* is proportional to the biased code estimate, and $C_{\text{lasso}} \coloneqq \mathcal{O}(\lambda\sqrt{s})$. Hence, $\boldsymbol{g}_T^{\text{ae-ls}}$ is a better estimator of $\boldsymbol{g}^*$ (Fig. 2a), and $\boldsymbol{D}^*$ neighbourhood at which $\boldsymbol{g}_T^{\text{ae-ls}}$ is guaranteed to converge to is smaller than of the $\boldsymbol{g}_T^{\text{ae-lasso}}$ and $\boldsymbol{g}_T^{\text{ae-dec}}$ (Fig. 2b).*

Interpretability: We build a mathematical relation between the network weights, training data, and test reconstruction. Thm. II.4 characterizes stationary points of the trained network and proves that the dictionary interpolates the training data, i.e., $\tilde{\boldsymbol{D}}_j = \boldsymbol{X}(\boldsymbol{G}^{-1}\boldsymbol{w}_j) = \sum_{k=1}^n(\boldsymbol{G}^{-1}\boldsymbol{w}_j)_k\boldsymbol{x}^k$ (Fig. 3a, green for high and red for low contribution). We write the reconstruction of a new example $\boldsymbol{x}^j$ as a linear combination of all the training examples, i.e., $\hat{\boldsymbol{x}}^j = \tilde{\boldsymbol{D}}\hat{\boldsymbol{z}}^j = \sum_{k=1}^n\boldsymbol{\beta}_k^j\boldsymbol{x}^k$ where $\boldsymbol{\beta}_k^j = \sum_{a=1}^n\boldsymbol{G}_{ka}^{-1}\langle\tilde{\boldsymbol{z}}^a, \hat{\boldsymbol{z}}^j\rangle$ (Fig. 3b, images with high contribution (green) are similar to the input image, and those with low (red) are different).

3a: Contribution to $\tilde{\boldsymbol{D}}$.



3b: Contribution for $\hat{\boldsymbol{x}}^j$.

Theorem II.4. *Consider $\min_{\boldsymbol{Z},\boldsymbol{D}}\ \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{Z}\|_F^2 + \lambda\|\boldsymbol{Z}\|_1 + {}^{\omega}\!/_{2}\|\boldsymbol{D}\|_F^2$, where $\boldsymbol{X} = [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n] \in \mathbb{R}^{m\times n}$ and $\boldsymbol{Z} = [\boldsymbol{z}^1, \ldots, \boldsymbol{z}^n] \in \mathbb{R}^{p\times n}$. Let $\tilde{\boldsymbol{Z}}$ be the given converged codes, then network stationary points follows $\tilde{\boldsymbol{D}} = \boldsymbol{X}\boldsymbol{G}^{-1}\tilde{\boldsymbol{Z}}^T$, where we denote $\boldsymbol{G} \coloneqq (\tilde{\boldsymbol{Z}}^T\tilde{\boldsymbol{Z}} + \omega\boldsymbol{I})$.*

[1]This work is presented as a talk at the Conference on the Mathematical Theory of Deep Neural Networks, 2022.

REFERENCES

[1] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[2] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[3] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, “Proximal methods for hierarchical sparse coding,” *The Journal of Machine Learning Research*, vol. 12, pp. 2297–2334, 2011.

[4] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach, “Supervised dictionary learning,” in *Proceedings of Advances in Neural Information Processing Systems*, vol. 21, 2009, pp. 1–8.

[5] K. Gregor and Y. LeCun, “Learning fast approximations of sparse coding,” in *Proceedings of international conference on international conference on machine learning*, 2010, pp. 399–406.

[6] J. R. Hershey, J. L. Roux, and F. Weninger, “Deep unfolding: Model-based inspiration of novel deep architectures,” *arXiv:1409.2574*, pp. 1–27, 2014.

[7] V. Monga, Y. Li, and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing,” *arXiv:1912.10557*, pp. 1–27, 2019.

[8] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[9] B. Tolooshams, S. Dey, and D. Ba, “Deep residual autoencoders for expectation maximization-inspired dictionary learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2415–2429, 2021.

[10] B. Tolooshams, A. Song, S. Temereanca, and D. Ba, “Convolutional dictionary learning based auto-encoders for natural exponential-family distributions,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 9493–9503.

[11] B. Malézieux, T. Moreau, and M. Kowalski, “Understanding approximate and unrolled dictionary learning for pattern recovery,” in *Proceedings of International Conference on Learning Representations*, 2022.

[12] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[13] T. Blumensath and M. E. Davies, “Iterative thresholding for sparse approximations,” *Journal of Fourier analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.