OPTIMIZATION GUARANTEES FOR ISTA AND ADMM BASED UNFOLDED NETWORKS

Wei Pu[⋆] *Yonina C. Eldar*[†] *Miguel R.D. Rodrigues*[⋆]

[⋆] Department of Electronic and Electrical Engineering, University College London, UK
[†] Weizmann Institute of Science, Rehovot, Israel.

ABSTRACT

Recently, unfolding techniques have been widely utilized to solve the inverse problems in various applications. In this paper, we study optimization guarantees for two popular unfolded networks, i.e., unfolded networks derived from iterative soft thresholding algorithms (ISTA) and derived from Alternating Direction Method of Multipliers (ADMM). Our guarantees – leveraging the Polyak-Lojasiewicz* (PL*) condition – state that the training (empirical) loss decreases to zero with the increase in the number of gradient descent epochs provided that the number of training samples is less than some threshold that depends on various quantities underlying the desired information processing task. Our guarantees also show that this threshold is larger for unfolded ISTA in comparison to unfolded ADMM, suggesting that there are certain regimes of number of training samples where the training error of unfolded ADMM does not converge to zero whereas the training error of unfolded ISTA does. A number of numerical results are provided backing up our theoretical findings.

***Index Terms*—** Algorithm unfolding, optimization guarantee, Polyak-Lojasiewicz* (PL*) condition

1. INTRODUCTION

To deal with inverse problems [1–3], there have been mainly three classes of approaches: 1) model-based, 2) data-driven and more recently 3) model-aware data-driven approaches. Algorithm unfolding or unrolling [4–8] is a popular model-aware data-driven approach for inverse problems. Algorithm unfolding techniques connect model-based iterative algorithms such as iterative soft thresholding algorithms (ISTA) to neural network architectures, wherein a diagram representation of one iteration step reveals its resemblance to a single network layer [5].

Generally speaking, for both purely data-driven and unfolding approaches, the more training data we have, the better the performance on unseen (testing) data. However, focusing on the training dataset, the authors in [9, 10] suggested an opposite view from the perspective of optimization. They argued that when using gradient-based methods to train deep neural networks, it is more difficult to make the training loss converge to zero with more training samples. Specifically, leveraging the Polyak-Lojasiewicz* (PL*) condition, they have established an optimization guarantee stating that the training error of a deep neural network optimized using gradient descent can only converge to zero provided that the number of training samples is smaller than a threshold.

In this paper, motivated by [9, 10], we offer optimization guarantees for two popular unfolded networks, i.e., unfolded ISTA [4] and unfolded Alternating Direction Method of Multipliers (ADMM) [6]. We show that the training losses both for ISTA and ADMM based networks converge to zero provided that number of training samples is below a certain threshold depending on various parameters associated with the problem. We also show this threshold is lower for ADMM based networks compared to ISTA based networks, which implies that there are certain regimes where an ADMM network training error does not converge to zero whereas the ISTA network training error does.

Note that our work differs from [11, 12] as the authors in [11, 12] studied the convergence of the error of the output after several layers if the weights in unfolded ISTA are well-learnt, while we focus on the convergence of training loss. Our work also differs from [9, 10] as we study the optimization guarantees of the unfolded networks in inverse problems, analyze the relationship between optimization guarantees and various parameters associated with the inverse problem, and compare the guarantees of different unfolded networks.

The remainder of the paper is organized as follows: Section 2 formulates the optimization problem of training the unfolded networks. Section 3 describes optimization guarantees of unfolded ISTA and ADMM. Sections 4 presents experimental results to support the theoretical findings. Finally, in section 5, we draw conclusions.

2. PROBLEM FORMULATION

Consider a linear inverse problem given by:

$$y = Ax + e, \tag{1}$$

where $y \in \mathbb{R}^n$ is the observation vector, $x \in \mathbb{R}^m$ is the target vector to be recovered, $A \in \mathbb{R}^{n\times m}$ describes the forward model with $m > n$ and $e \in \mathbb{R}^n$ is the noise. The goal is to reconstruct x from y.

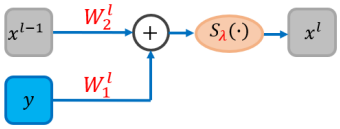

Fig. 1. Structure of the l-th layer of unfolded ISTA.

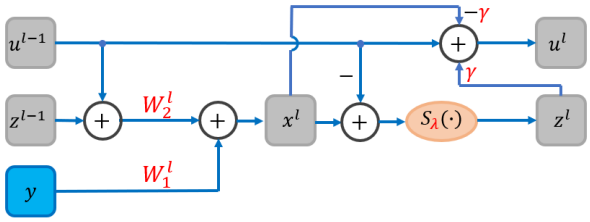

Fig. 2. Structure of the l-th layer of unfolded ADMM network.

This class of problems can often be solved by assuming that the vector of interest is sparse. This involves posing an optimization problem such as least absolute shrinkage and selection operator (Lasso) [13] that can in turn be solved using well-known methods such as ISTA [14] or ADMM [15].

Alternatively, these problems can also be solved using algorithm unfolding or unrolling techniques whereby solvers such as ISTA or ADMM are mapped onto a neural network architecture, whose parameters can then be further tuned using gradient descent or some other variant based on the availability of a series of examples $(x_p, y_p), p = 1, ..., P$.

In particular, for a L-layer unfolded ISTA network, the output of the l-th layer as shown in Fig. 1 is

$$x^l = \mathcal{S}_\lambda(W_1^l y + W_2^l x^{l-1}) \tag{2}$$

where $l = 1, 2, \cdots, L$ denotes the layer number, W_1^l and W_2^l are the weights at the l-th layer, and $\mathcal{S}_\lambda(\cdot)$ is the soft-thresholding function defined as

$$\mathcal{S}_\lambda\{x\} = \text{sign}(x)\max(|x| - \lambda, 0). \tag{3}$$

In (2), W_1^l, W_2^l and λ are learnable parameters.

In turn, for a L-layer unfolded ADMM network, the output of the l-th layer as shown in Fig. 2 is

$$\begin{aligned} x^l &= W_1^l y + W_2^l(z^{l-1} + u^{l-1}) \\ z^l &= \mathcal{S}_\lambda(x^l - u^{l-1}) \\ u^l &= u^{l-1} - \gamma(x^l - z^l) \end{aligned} \tag{4}$$

where $\gamma \geq 0$ is the step size. In an unfolded ADMM network, W_1^l, W_2^l, λ and γ are learnable parameters.

Given a training dataset $(x_p, y_p), p = 1, \ldots, P$, the parameters of these networks are then learnt by optimizing the squared loss function given by

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{p=1}^{P} \|x_p^L(y_p) - x_p\|^2, \tag{5}$$

where x_p^L denotes network output given network input y_p, x_p is the corresponding ground truth, $\mathbf{w}$ denotes the learnable parameters $\mathbf{w} = \text{vec}(\mathbf{W})$, and [1]

$$\mathbf{W} = \begin{bmatrix} W_1^{(1)} & W_2^{(1)} \\ \vdots & \vdots \\ W_1^{(L)} & W_2^{(L)} \end{bmatrix}. \tag{6}$$

We focus on gradient descent procedures to optimize the loss function in (5) whereby the parameters estimate at epoch t depends on its estimate at epoch $t-1$ as follows:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \nabla \mathcal{L}(\mathbf{w}_{t-1}), \tag{7}$$

where η is the step size, t denotes the epoch number and $\mathbf{w}_0$ is the initialized parameters.

Our goal is to understand whether the training loss in (5) converges to zero as the number of gradient descent epochs increases to infinity, both for unfolded ISTA and ADMM networks.

3. OPTIMIZATION GUARANTEES

Our guarantees are based on the PL* condition. In particular, it has been shown in [10] that gradient descent converges for neural networks that fulfill this condition. It has also been shown in [10] that the PL* condition is satisfied provided that the number of training samples is smaller than a constant related to the spectral norm of the Hessian of a deep neural network.

In line with the work in [10], we provide optimization guarantees of the unfolded ISTA and ADMM by the following two steps: 1) we provide a bound on the spectral norm of the Hessian matrix of the unfolded networks and 2) we provide a threshold on the number of training samples below which gradient descent is guaranteed to converge. [2]

3.1. Convergence guarantees of unfolded network

We now offer our main results. For any learning parameter set $\mathbf{w}$ in parameter space $\{\mathcal{S} : \|\mathbf{w} - \mathbf{w}_0\|_F \leq R\}$, where $R > 0$ is a constant independent on the size of the network, i.e., m (dimension of target vector x), n (dimension of observation vector y), and L (number of network layers), we have

$$\|\mathbf{H}(\mathbf{w})\| \leq c_H, \tag{8}$$

[1] Note that in this paper, the activation functions in both unfolded ISTA and ADMM networks are assumed to be twice differentiable activation function like tanh or sigmoid instead of soft-thresholding, and γ in (4) is assumed to be a constant rather than learnable parameter in the unfolded ADMM network.

[2] To simplify our analysis, we also make some assumptions on the initialization of the unfolded networks, which are described in [16].

where $\|\mathbf{H}(\mathbf{w})\|$ is the spectral norm of the Hessian matrix of unfolded ISTA or ADMM, and c_H is a constant depending on the size of the unfolded network. Note that c_H differs for unfolded ISTA and ADMM.

Theorem 1: For the unfolded ISTA or ADMM, if the number of training samples P satisfies

$$P \leq \left(\frac{c}{c_H}\right)^2, \tag{9}$$

where c is a constant independent on the size of the unfolded network and c_H is the threshold of the spectral norm of the Hessian matrix of unfolded ISTA or ADMM depending on the type of network, i.e., ADMM or ISTA, then we can state that:

- **Existence of a solution:** There exists a solution $\mathbf{w}^* \in \mathcal{S}$ such that $x_p^L = x_p$ for $p = 1, 2, \cdots, P$.
- **Convergence:** If we use gradient descent method to train the unfolded network, and the step size η is smaller than a certain value, which is of the order of $O(1/\mathcal{L}(\mathbf{w}_0))$ and independent on the size of the unfolded network, then the training loss behaves as follows:

$$\mathcal{L}(\mathbf{w}_t) \leq (1 - \eta\mu)^t \mathcal{L}(\mathbf{w}_0), \tag{10}$$

  where μ is a constant independent of the network size.

The proof of Theorem 1 is omitted here due to space limitations, but it can be shown that whereas c_H differs for unfolded ISTA and ADMM, we have $c_H = O(\sqrt{m})$, $c_H = O(1/\sqrt{n})$ and $c_H = O((c_1)^L)$ both for unfolded ISTA and ADMM, where $c_1 > 0$ is a constant. We conclude the following.

Firstly, the number of training samples should be smaller than some threshold in order for the training loss to decrease to zero as the number of gradient descent epochs increases to infinity. This threshold decreases with the increase in dimensionality of the target vector, m. We attribute this to the fact that a higher target vector dimensionality is associated with a higher task complexity, resulting in a more complex optimization problem.

Secondly, this threshold increases with the increase in dimensionality of the measurement vector, n. We in turn attribute this to the fact that a higher measurement vector dimensionality is associated with a lower task complexity, resulting in a simpler optimization problem. The size of the network is also additionally larger, allowing one to deal with larger training sets.

3.2. Convergence guarantees comparison

In order to compare unfolded ISTA to unfolded ADMM networks, it can be shown that

$$c_{H,\text{ista}} \leq c_{H,\text{admm}}, \tag{11}$$

where $c_{H,\text{ista}}$ and $c_{H,\text{admm}}$ denote the bounds of Hessian spectral norms of unfolded ISTA and ADMM, respectively. We also note equality holds provided that $\gamma = 0$ and $u^0 = 0$.

This indicates that the threshold on the number of training samples guaranteeting gradient descent convergence of the unfolded ISTA is larger than that of the unfolded ADMM, implying that the unfolded ISTA is capable of dealing with a larger amount of training samples than unfolded ADMM in terms of the convergence of training loss. We explain this by contrasting the structures of unfolded ISTA and ADMM networks shown in Figs. 1 and 2. We think that whereas the more complex structure associated with ADMM networks in relation to ISTA based ones – exhibiting more shortcuts and skip connections – can help with generalization (as documented in [8]), it also results nonetheless in a more complex optimization procedure, that manifests itself in a more stringent requirement in number of training samples for gradient descent to converge.

4. EXPERIMENTS

We now perform various experiments to support the intuitions in the previous section:

- We show how the training loss behaves as a function of the number of training samples. See Fig. 3.
- We show how the training loss behaves as a function of the number of training epochs, with the number of training sample satisfying the optimization guarantee both for the ISTA based network and the ADMM one. See Fig. 4.
- We also show how the training loss behaves as a function of the number of training epochs, with the number of training sample satisfying the optimization guarantee for the ISTA based network but not the ADMM one. See Fig. 5.

Our experimental set-up involves the generation of synthetic data, where $x \in \mathbb{R}^{200}$ is a vector with sparsity equal to four with the non-zero elements randomly chosen within the interval $(0, 1]$, and $e \in \mathbb{R}^{25}$ is a Gaussian vector with zero mean and variance 0.03. The measurement matrix $A \in \mathbb{R}^{25\times 200}$ is Gaussian with $\|A\|_F = 10$, and $y \in \mathbb{R}^{25}$ is generated via the model in (1). We generate P different sample pairs (x, y) where the matrix A remains fixed. We use sigmoid function as activation function in each layer for both unfolded ISTA and ADMM networks, and set $L = 3$. Additionally, in unfolded ADMM, $\gamma = 1$. During training, we use gradient descent with step size $\eta = 10^{-6}$.

We carried out the experiments over a number of trials associated with different number of training samples P. Fig. 3 depicts the evolution of the average MSE of the training samples as a function of the sample number. Here, average

MSE is calculated by

$$\frac{1}{RP}\sum_{p=1}^{P}\sum_{r=1}^{R}\|x_{p,r}(y_{p,r})-\hat{x}_{p,r}\| \tag{12}$$

where $y_{p,r}$ is the input of the p-th sample on the r-th trial, $x_{p,r}$ is the output of the p-th sample on the r-th trial, and $\hat{x}_{p,r}$ is the corresponding ground truth. We set the number of trials R to be equal to 100 in our experiments. During training the network, the stop condition is that the changing rate of loss, i.e., $(\mathcal{L}(\mathbf{w}_t)-\mathcal{L}(\mathbf{w}_{t-1})/\mathcal{L}(\mathbf{w}_t)$, is smaller than a pre-determined threshold 10^{-3}. In line with Theorem 1, it could be observed that for both unfolded ISTA and ADMM networks, the average MSE increases rapidly once the number of training samples exceeds some threshold. Additionally, the threshold is approximately 2000 for unfolded ADMM and 3000 for unfolded ISTA. The threshold of unfolded ISTA is larger than that of unfolded ADMM.

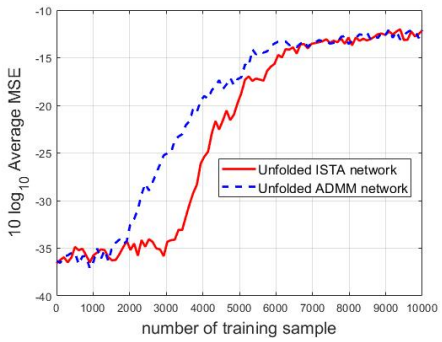

Fig. 3. Average MSE versus sample number.

Fig. 4 depicts the evolution of training loss as a function of the number of epochs with $P = 1000$ training samples. Here $P = 1000$ is smaller than the thresholds of both unfolded ISTA and ADMM networks, and therefore, the optimization guarantees of both unfolded ISTA and ADMM networks in Theorem 1 are satisfied. It can be seen that the losses for both unfolded ISTA and ADMM networks converge rapidly to zero with an approximate exponential linear convergence rate.

Fig. 5 depicts the evolution of training loss as a function of the number of epochs with $P = 2500$ training samples. Note that $P = 2500$ is smaller than the threshold of unfolded ISTA, and larger than the threshold of unfolded ADMM. The training loss of unfolded ISTA converges with an approximate exponential linear convergence rate. However, the training loss of unfolded ADMM decreases in the first 1200 epochs while it does not appear to change significantly in the last 800 epochs.

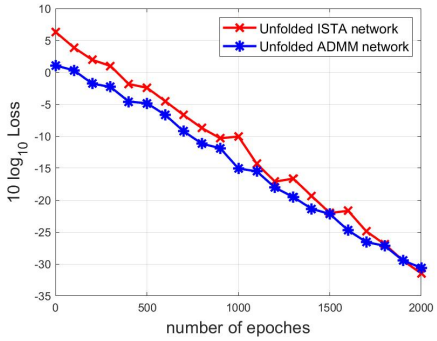

Fig. 4. Loss versus number of epochs with 1000 training samples.

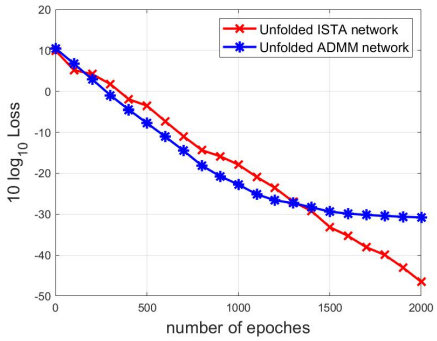

Fig. 5. Loss versus number of epochs with 2500 training samples.

5. CONCLUSION

Unfolding techniques have achieved significant success in solving inverse problems. In this paper, we study optimization guarantees for popular unfolded networks. We demonstrate that to guarantee that the training loss of unfolded ISTA or ADMM converges to zero using gradient descent techniques, the number of training samples should be smaller than a threshold that depends on various parameters associated with the underlying task. We also demonstrate that this threshold for unfolded ISTA is larger than that of the unfolded ADMM indicating there are certain regimes where the training loss of unfolded ADMM does not converge to zero whereas the training loss of unfolded ISTA does.

This perhaps surprising result contrasts with other results indicating that ADMM networks generalize better than ISTA based ones given access to sufficiently large training sets [8]. This motivates studying further the interplay between optimization and generalization of unfolded networks.

6. REFERENCES

[1] E. J. Candes, J. Romberg, and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] Y. C. Eldar and G. Kutyniok, "Compressed Sensing: Theory and Applications", Cambridge University Press, May 2012.

[4] K. Gregor and Y. LeCun, Learning fast approximations of sparse coding, in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, pp. 399C406, 2010.

[5] V. Monga, Y. Li, Y. C. Eldar. Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing. *IEEE Signal Process. Magazine*, vol. 38, no. 2, pp. 18–44, 2021.

[6] Y. Yang, J. Sun, H. LI, and Z. Xu, ADMM-CSNet: A Deep Learning Approach for Image Compressive Sensing, *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1C1, to appear, 2019.

[7] Y. Li, M. Tofighi, J. Geng, V. Monga and Y. C. Eldar, "Efficient and Interpretable Deep Blind Image Deblurring Via Algorithm Unrolling," *IEEE Trans. Comp. Imag.*, vol. 6, pp. 666-681, 2020.

[8] Y. Chen and T. Pock, Trainable Nonlinear Reaction Diffusion: A Flexible Framework for Fast and Effective Image Restoration, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, 2017.

[9] C. Liu, L. Zhu, M. Belkin, "On the linearity of large non-linear models: when and why the tangent kernel is constant why the tangent kernel is constant", https://arxiv.org/abs/2010.01092

[10] C. Liu, L. Zhu, M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks", https://arxiv.org/pdf/2003.00307.pdf

[11] Liu J, Chen X, Wang Z, et al, "ALISTA: Analytic Weights are as good as learned weights in LISTA," in *International Conference on Learning Representations (ICLR)*, 2019.

[12] Chen X , Liu J , Wang Z , et al, "Theoretical Linear Convergence of Unfolded ISTA and its Practical Weights and Thresholds," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Canada.

[13] Tibshirani, and J. Ryan, "The Lasso Problem and Uniqueness," *Electronic Journal of Statistics* vol. 7, no. 1, pp. 1456-1490, 2013.

[14] B. Amir, and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. , no. 1, pp. 183-202, 2009.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1-122, 2011.

[16] Supplementary materials of "Optimization Guarantees for Unfolded Networks".