Supplementary material for Memory-efficient model based deep learning with convergence and robustness guarantees

Aniket Pramanik, *Student Member, IEEE,* M. Bridget Zimmerman, Mathews Jacob, *Fellow, IEEE,*

I. DEQ: FORWARD AND BACKWARD PROPAGATION

This section provides more details on the implementation of the DEQ based MOL algorithm, briefly described in V.A. During inference and training, we use the forward iterations $\mathbf{x}_{n+1} = \mathcal{T}_{\rm MOL}(\mathbf{x}_n) + \mathbf{z}$ until convergence as shown in the pseudo-code in Algorithm 1. The expanded equation is,

$$\mathbf{x}_{n+1} = \underbrace{(\mathbf{I}+\alpha\lambda\mathbf{A}^H\mathbf{A})^{-1}\Big((1-\alpha)\mathbf{x}_n + \alpha\mathcal{H}_\theta(\mathbf{x}_n)\Big)}_{\mathcal{T}_{\rm MOL}(\mathbf{x}_n)} + \underbrace{(\mathbf{I}+\alpha\lambda\mathbf{A}^H\mathbf{A})^{-1}\Big(\alpha\lambda\mathbf{A}^H\mathbf{b}\Big)}_{\mathbf{z}} \tag{1}$$

$$= \mathcal{T}_{\rm MOL}(\mathbf{x}_n) + \mathbf{z} \tag{2}$$

from equation (14) in the main paper. We terminate the algorithm at the n^{th} iteration, when $\mathbf{x}_n$ satisfies,

$$e_n = \frac{\|\mathbf{x}_n - \mathbf{x}_{n-1}\|_2}{\|\mathbf{x}_{n-1}\|_2} < \kappa. \tag{3}$$

We set $\kappa = 1 \times 10^{-4}$ for the experiments. We denote the fixed point of the algorithm as $\mathbf{x}^*(\mathbf{b})$, such that $\mathbf{x}^*(\mathbf{b}) \approx \mathcal{T}_{\rm MOL}\left(\mathbf{x}^*(\mathbf{b})\right) + \mathbf{z}$.

DEQ schemes [1] rely on fixed point iterations for backpropagating the gradients. Using chain rule, the gradient of the loss $\mathcal{C}$ with respect to the CNN parameters θ is computed as $\nabla_\theta\mathcal{C} = (\nabla_\theta\mathbf{x}^*)^T(\nabla_{\mathbf{x}^*}\mathcal{C})$. Using fixed point relation $\mathbf{x}^* = \mathcal{T}_{\rm MOL}(\mathbf{x}^*) + \mathbf{z}$, one obtains $\nabla_\theta\mathbf{x}^* = (\mathcal{I} - \nabla_{\mathbf{x}}\mathcal{T}_{\rm MOL}(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^*})^{-1}\left(\nabla_\theta\mathcal{T}_{\rm MOL}(\mathbf{x}^*)\right)$ which translates to

$$\nabla_\theta\mathcal{C} = (\nabla_\theta\mathcal{T}_{\rm MOL}(\mathbf{x}^*))^\intercal \underbrace{(\mathcal{I} - \nabla_{\mathbf{x}}\mathcal{T}_{\rm MOL}(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^*})^{-\intercal}\left(\nabla_{\mathbf{x}^*}\mathcal{C}\right)}_{\mathbf{q}} \tag{4}$$

Here, $\mathbf{q}$ is solved using fixed point iterations [1], [2] as

$$\mathbf{q} = (\nabla_{\mathbf{x}}\mathcal{T}_{\rm MOL}(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^*})^\intercal\ \mathbf{q} + \nabla_{\mathbf{x}^*}\mathcal{C} \tag{5}$$

with initialization as $\mathbf{q}_0 = 0$. This iteration is evaluated until convergence, using the similar termination conditions as in (3). The computed $\mathbf{q}$ is substituted in (4) to obtain the gradient $\nabla_\theta\mathcal{C} = (\nabla_\theta\mathcal{T}_{\rm MOL}(\mathbf{x}^*))^\intercal\ \mathbf{q}$. The pseudo-code for the backpropagation steps are shown in Algorithm 2.

Algorithm 1 : Forward propagation in MOL: input $\mathbf{b}$

```
$\mathbf{x}_0 = \lambda_0(\mathbf{I}+\lambda_0\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{b}$
$\mathbf{z} = (\mathbf{I}+\alpha\lambda\mathbf{A}^H\mathbf{A})^{-1}\Big(\alpha\lambda\mathbf{A}^H\mathbf{b}\Big)$
$e = \infty$
while $e > \kappa$ do
    $\mathbf{x}_{\rm old} = \mathbf{x}$
    $\mathbf{x} = \mathcal{T}_{\rm MOL}(\mathbf{x}) + \mathbf{z}$ from (1)
    $e = \|\mathbf{x}_{\rm old} - \mathbf{x}\|_2^2 / \|\mathbf{x}_{\rm old}\|_2^2$
end while
Return $\mathbf{x}$
```

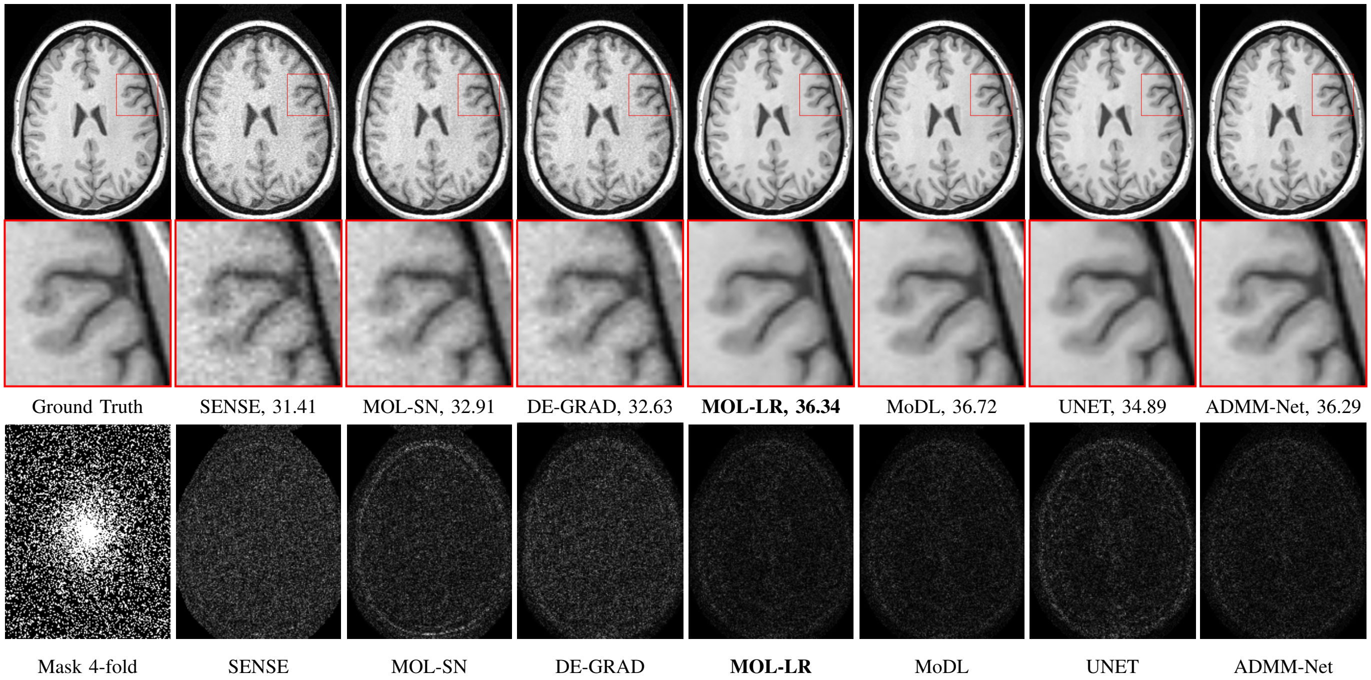
Algorithm 2 : Backpropagation in MOL: input $\nabla_{\mathbf{x}^*}\mathcal{C}$

```
$\mathbf{q}_0 = 0$
$e = \infty$
while $e > \kappa$ do
    $\mathbf{q}_{\rm old} = \mathbf{q}$
    $\mathbf{q} = (\nabla_{\mathbf{x}}\mathcal{T}_{\rm MOL}(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^*})^\intercal\ \mathbf{q} + \nabla_{\mathbf{x}^*}\mathcal{C}$ in (5)
    $e = \|\mathbf{q}_{\rm old} - \mathbf{q}\|_2^2 / \|\mathbf{q}_{\rm old}\|_2^2$
end while
Return $\nabla_\theta\mathcal{C} = (\nabla_\theta\mathcal{T}_{\rm MOL}(\mathbf{x}^*))^\intercal\ \mathbf{q}$
```

II. PERFORMANCE COMPARISON ON PARALLEL MRI FROM CALGARY BRAIN DATA

Qualitative analysis of the reconstructions from different algorithms on four-fold accelerated brain MRI data is shown in Fig. 1. Proposed MOL-LR performs at par with ten-iterations of the unrolled MoDL and ADMM-Net. A slight drop in PSNR for MOL-LR is due to Lipschitz constraint on the CNN block. MOL-SN uses spectral normalization of each layer of the CNN which gives stricter bounds on its Lipschitz and thus provides lower performance. The drop in performance of DE-GRAD is also for similar reasons as in MOL-SN. UNET has a lower performance compared to unrolled algorithms MoDL and ADMM-Net. The error images show high amplitude of errors for SENSE, MOL-SN, DE-GRAD and UNET. Proposed MOL-LR and unrolled MoDL, ADMM-Net show fewer errors in the error images and that too with much lower amplitude. These comparisons show that the Lipschitz regularization strategy offers better performance than spectral normalization. Similar trends are also visible in the quantitative analysis of these methods over 20 subjects as reported in Table I.

Aniket Pramanik and Mathews Jacob are from the Department of Electrical and Computer Engineering at the University of Iowa, Iowa City, IA, 52242, USA (e-mail: aniket-pramanik@uiowa.edu; mathews-jacob@uiowa.edu). M. Bridget Zimmerman is from the Department of Biostatistics at the University of Iowa, Iowa City, IA, 52242, USA (e-mail: bridget-zimmerman@uiowa.edu). This work is supported by grants NIH R01 AG067078 and R01 EB031169.



Ground Truth | SENSE, 31.41 | MOL-SN, 32.91 | DE-GRAD, 32.63 | **MOL-LR, 36.34** | MoDL, 36.72 | UNET, 34.89 | ADMM-Net, 36.29

Mask 4-fold | SENSE | MOL-SN | DE-GRAD | **MOL-LR** | MoDL | UNET | ADMM-Net

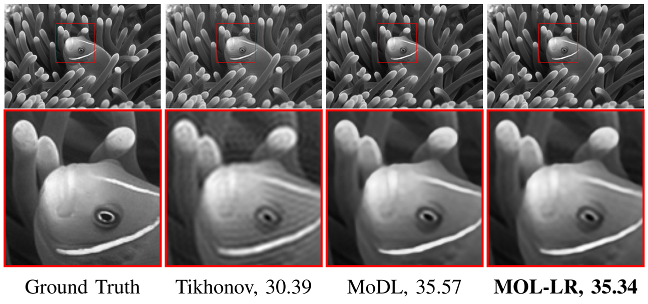
Fig. 1. Reconstruction results of 4x accelerated multi-channel brain data. PSNR (dB) values are reported for each case. The image in the first row and column was undersampled using a Cartesian 2D non-uniform variable-density mask as shown in the first column, second row. The top row shows reconstructions (magnitude images), while the bottom row shows corresponding error images. We note that the quality of the **MOL-LR** reconstructions is comparable to unrolled methods MoDL and ADMM-Net. MOL-SN and DE-GRAD show significantly lower performance due to spectral normalization of weights, resulting in stricter bounds on the Lipschitz constant of its CNN.

	Four-fold Brain MRI		
Methods	PSNR	SSIM	Run-time
SENSE	31.13 ± 1.67	0.982 ± 0.025	**0.06s**
MOL-SN	33.05 ± 1.13	0.986 ± 0.014	0.29s
DE-GRAD	32.84 ± 1.21	0.985 ± 0.016	0.24s
MOL-LR	36.56 ± 0.82	0.992 ± 0.008	0.35s
MoDL	**36.89 ± 0.75**	**0.993 ± 0.007**	0.15s
MoDL-LR	34.63 ± 1.02	0.990 ± 0.011	0.15s
ADMM-Net	36.77 ± 0.76	0.993 ± 0.008	0.16s
ADMM-Net-LR	34.51 ± 1.03	0.990 ± 0.011	0.16s
UNET	34.69 ± 1.03	0.988 ± 0.011	0.07s
UNET-LR	33.42 ± 1.15	0.986 ± 0.014	0.07s

TABLE I
QUATITATIVE COMPARISONS ON 2D DATASETS WITH 4-FOLD UNDERSAMPLING USING CARTESIAN 2D NON-UNIFORM VARIABLE DENSITY MASK. PSNR IN DB, SSIM AND MEAN RUN-TIME PER SLICE IN SECONDS ARE REPORTED. THE PSNR AND SSIM VALUES ARE IN MEAN ± STANDARD DEVIATION FORMAT.

III. Illustration in Image Super-resolution Setting

The proposed theory and algorithms are broadly applicable to general linear inverse problems. We now show the preliminary feasibility of the proposed MOL-LR approach in three-fold super-resolution image reconstruction in Fig. 2. MOL-LR is compared against MoDL and Tikhonov regularized reconstructions, respectively. Similar to the parallel MRI setting, MoDL and MOL-LR offer similar performance while outperforming the Tikhonov regularized reconstruction in terms of PSNR. There is visible aliasing and blurring in Tikhonov, which is reduced to an appreciable extent in MoDL and MOL-LR. All the models have been trained and tested on the Berkeley Segmentation Dataset and Benchmark [3]. The dataset consists of 300 images, out of which 200 are split for training and the remaining ones for testing. While the preliminary experiments in this context are encouraging, more experiments are needed to compare MOL to state-of-the-art super-resolution methods. We plan to pursue this in the future.



Ground Truth | Tikhonov, 30.39 | MoDL, 35.57 | **MOL-LR, 35.34**

Fig. 2. Results of 3x super-resolution reconstruction of natural grayscale images. PSNR (dB) values are reported for each case. The proposed MOL-LR is compared against MoDL and Tikhonov regularized reconstructions. The Tikhonov reconstruction shows blurring and Gibbs ringing, which is reduced significantly in MOL-LR and MoDL, which show similar performance in terms of PSNR.

References

[1] D. Gilton, G. Ongie, and R. Willett, "Deep equilibrium architectures for inverse problems in imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1123–1133, 2021.

[2] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.