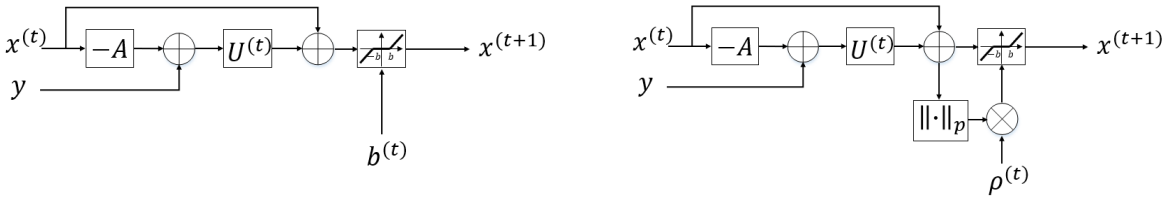
APPENDIX: LEARNED ISTA WITH ERROR-BASED THRESHOLDING FOR ADAPTIVE SPARSE CODING

Ziang Li[1] *Kailun Wu*[1] *Yiwen Guo* [2*] *Zhangshui Chang*[1*]

[1] Institute for Artificial Intelligence, Tsinghua University
[2] Independent Researcher



(a) LISTA (b) EBT-LISTA

Fig. 1: The t-th layer of LISTA and EBT-LISTA.

1. RELATED WORKS

Some prior work also aims at improving LISTA. For instance, [1] tried only to learn the step size in ISTA, while all other parameters in the network are not learned. [2] introduced the extragradient method in LISTA and proposed ELISTA, in which $x^{(t+\frac{1}{2})}$ is calculated to consider the curvature information. It designs multistage-thresholding functions to obtain effective sparse representation. [3] proposed hybrid ISTA and hybrid LISTA that incorporate free-form DNNs into ISTA and LISTA to improve the efficiency and flexibility without compromising the convergence rate. The most related work is ALISTA [4], in which analytic weights were obtained in a data-free manner. Benefiting from its training simplicity, ALISTA can be adopted to a scenario where robustness and adaptivity is required. We will show with experiments that our EBT can be combined with these methods to achieve further improvements.

[5] proposed Ada-LISTA to improve adaptivity to dictionary permutations and perturbations, which is quite different from our goal which is the adaptivity to possible variations in data (i.e., $x^\star$). Some other work introduced structured convolutional neural network to learn in image domains. For example, [6] proposed ISTA-net for image compressive sensing, and [7] proposed SCN for image super-resolution. These tasks are not the main interest of this paper.

2. EXPERIMENTS

2.1. Basic settings

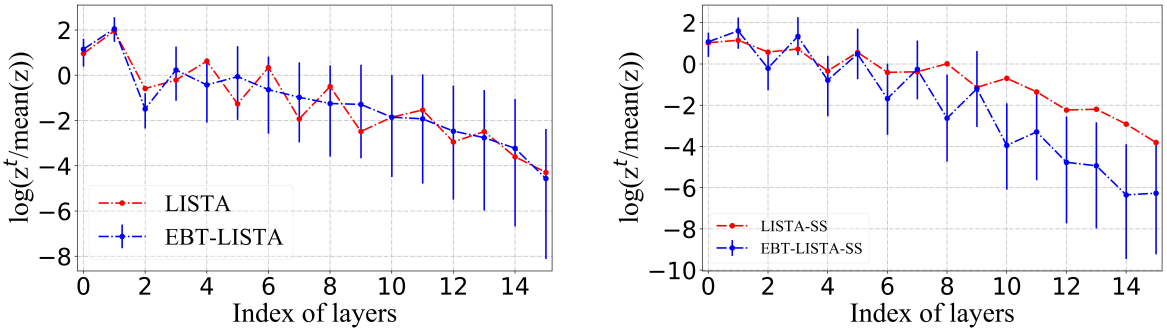
The network architectures and training strategies in our experiments mostly follow those of prior works [8, 9]. To be more specific, all the compared networks have $d = 16$ layers and the learnable parameters $W^{(t)}, U^{(t)}, \rho^{(t)}$ (or $b^{(t)}$ without EBT) are not shared among layers. The training batch size is 64, and we use the popular Adam optimizer [10] for training with its default hyper-parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Training is performed from layer to layer in a progressive way, i.e., if the validation loss of the current layer does not decrease for 4000 iterations, the next layer will then be trained. When training each layer, the learning rate is first initialized to 0.0005. It will first be decreased to 0.0001 and finally to 0.00001 if the validation loss does not decrease for 4000 iterations. Specifically, in the proposed methods, we impose the constraint between $W^{(t)}$ and $U^{(t)}$ and make sure it holds that $W^{(t)} = I - U^{(t)}A, \forall t$, i.e., the coupled constraints are introduced [8]. All experiments are performed on NVIDIA GeForce RTX 2080 Ti using TensorFlow [11].

In simulation experiments, we set $m = 250$, $n = 500$, and we generate the dictionary matrix A by using the standard Gaussian distribution. The indices of the non-zero entries in $x^\star$ are determined by a Bernoulli distribution letting its sparsity (i.e., the probability of any of its entries be zero) be p_b, while the magnitudes of the non-zero entries are sampled from the standard Gaussian distribution. The noise ε is sampled from a Gaussian distribution where the standard deviation is determined by the noise level. With $y = Ax^\star + \varepsilon$, we can randomly synthesize in-stream $x^\star$ and get a corresponding set of observations y for training, thus the number of training samples grows as the training proceeds. We also synthesize two sets for validation and test, respectively, each containing 1000 samples. The sparse coding performance of different models is evaluated by the normalized mean squared error (NMSE) in decibels (dB):

$$\text{NMSE}(x, x^\star) = 10 \log_{10} \left(\frac{\|x - x^\star\|_2^2}{\|x^\star\|_2^2} \right). \tag{1}$$

2.2. Simulation Results for Showing Disentanglement

We analyze the obtained threshold values in EBT-LISTA and EBT-LISTA-SS, i.e., $b^{(t)} = \rho^{(t)} \|U^{(t)}(Ax^{(t)} - y)\|_\phi$ (with $\phi = 2$), and compare them with the thresholds values obtained in LISTA and LISTA-SS. Note that the threshold values in our EBT-based models differ from sample to sample, we show the results in Figure 2. It can be seen that the learned thresholds in our EBT-based methods and the original LISTA and LISTA-SS are similar, which indicates that the introduced EBT mechanism does not modify the training dynamics of the original methods, and our EBT works by disentangling the reconstruction error and learnable parameters.



(a) LISTA (b) LISTA-SS

Fig. 2: Thresholds obtained from different methods across layers.

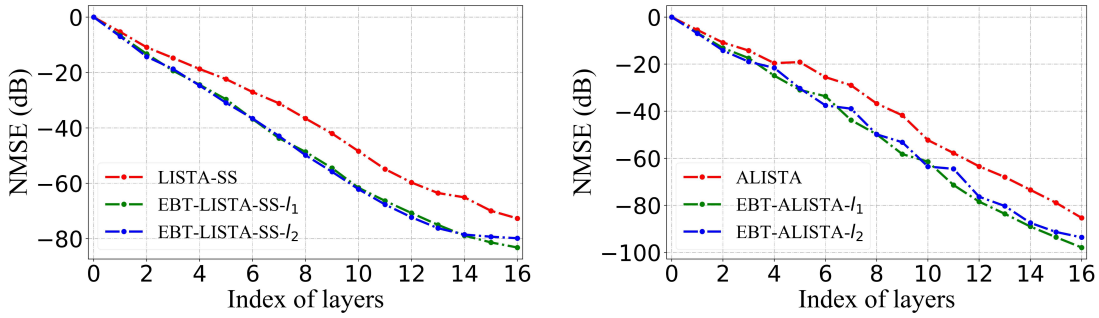
.

2.3. Validation of Theorem 2

Figure 3a shows how the NMSE of EBT-LISTA-SS varies across layers. In addition to the l_1 norm (i.e., $b^{(t)} = \rho^{(t)} \|U^{(t)}(Ax^{(t)} - y)\|_1$) concerned in the theorem, we also test EBT-LISTA-SS with the l_2 norm (i.e., by letting $b^{(t)} = \rho^{(t)} \|U^{(t)}(Ax^{(t)} - y)\|_2$). It can be seen that, with both the l_1 and l_2 norms, EBT-LISTA-SS leads to consistently faster convergence than LISTA-SS. Also, it is clear that there exist two convergence phases for EBT-LISTA-SS and LISTA-SS, and the later phase is indeed faster than the earlier phase. With faster convergence, EBT-LISTA-SS finally achieves superior performance. The experiment is performed in the noiseless case with $p_b = 0.95$. Similar observations can be made on other variants of ALISTA (e.g., ALISTA [4], see Figure 3b).

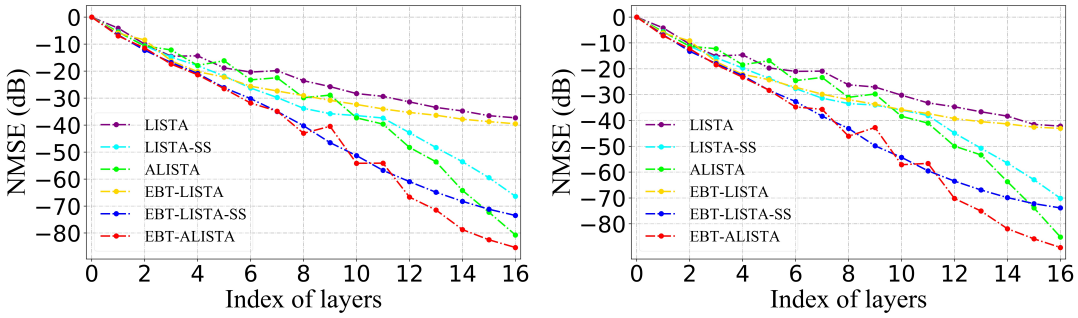
2.4. Random sparsity

We here consider the scenario where the sparsity of data follows a certain distribution. We test with two distributions of sparsity (i.e., p_b): $p_b \sim U(0.9, 1)$ (uniform distribution) and $p_b \sim N(0.95, 0.025)$ with an constraint of $p \in [0.9, 1]$ (truncated Gaussian distribution). The comparison results in such settings are shown in Figure 4. Our EBT leads to huge advantages in all the models and settings, indicating that the conclusion in Theorem 1 can be extended to broader distributions of data sparsity.



(a) LISTA-SS (b) ALISTA

Fig. 3: Validation of Theorem 2: there exist two convergence phases and our EBT accelerates the convergence of LISTA-SS, in particular in the first phase.

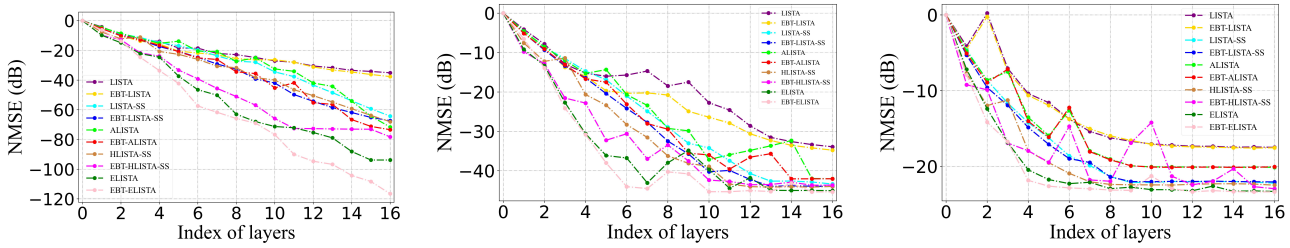


(a) Uniform distribution (b) Truncated Gaussian distribution

Fig. 4: NMSE of different sparse coding methods when the sparsity of the data follows a certain distribution.

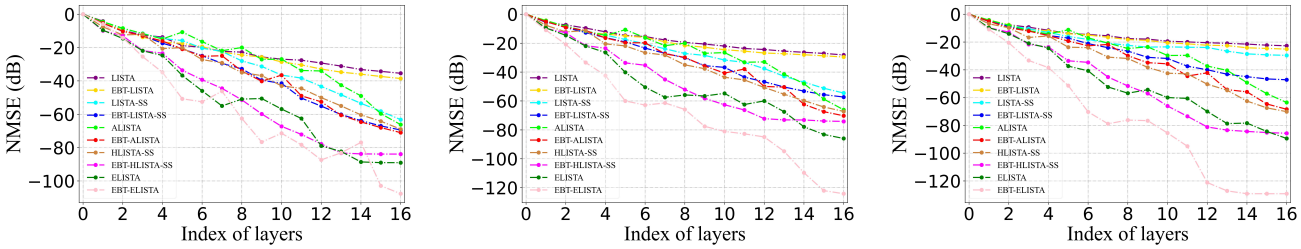
2.5. Additional Comparison with competitors

More experiment results than Figure 2 in the main paper are given here. The performances of different networks under different noise levels (with $p_b = 0.9$) are shown in Figure 5. It can be seen that when combined with LISTA and its variants, our EBT achieves better or similar performance. The figures show that the performance of our EBT is more promising in the noiseless or low noise cases (i.e., SNR=∞ and SNR=40dB), while in very noisy scenarios it provides little help. Figure 6 shows the results of our methods with EBT (i.e. EBT-LISTA, EBT-LISTA-SS, EBT-ALISTA, EBT-ELISTA, and EBT-HLISTA-SS) and other competitors under different condition number (with $p_b = 0.9$). From Figure 6, we can find that our EBT leads to better performance as shown in the results in the main paper.



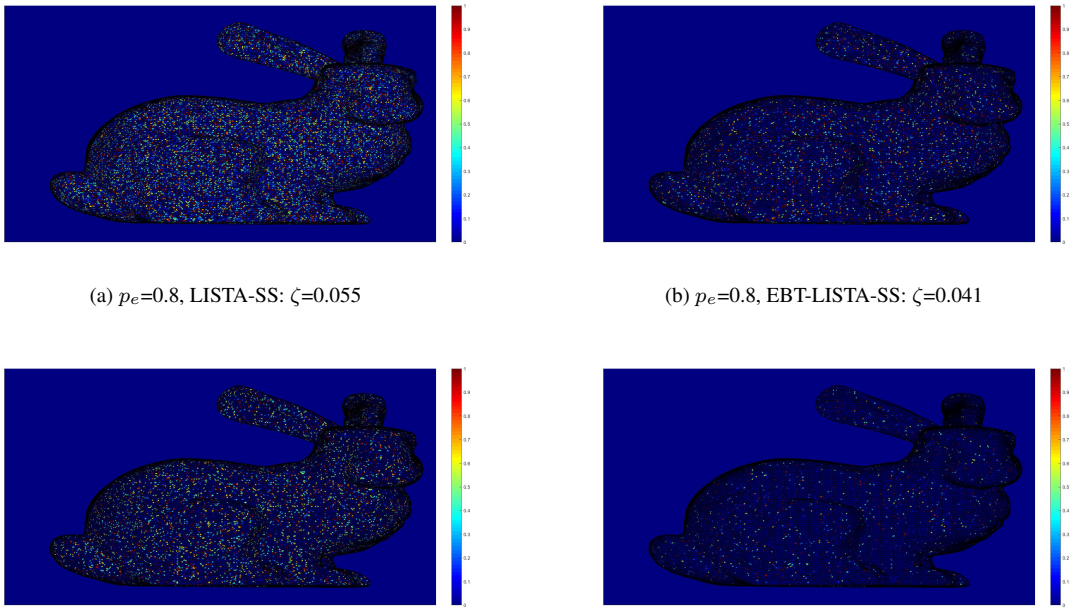
(a) SNR=∞ (b) SNR=40dB (c) SNR=20dB

Fig. 5: NMSE of different sparse coding methods under different noise levels with $p_b = 0.9$. It can be seen that our EBT performs favorably well under SNR=∞ and SNR=40dB.



(a) condition number=3 (b) condition number=30 (c) condition number=100

Fig. 6: NMSE of different sparse coding methods under different condition numbers with $p_b = 0.9$.



(a) p_e=0.8, LISTA-SS: ζ=0.055 (b) p_e=0.8, EBT-LISTA-SS: ζ=0.041

(c) p_e=0.9, LISTA-SS: ζ=0.0067 (d) p_e=0.9, EBT-LISTA-SS: ζ=0.0026

Fig. 7: Reconstruction 3D error maps of different methods in different settings. ζ here is the mean estimation error in degree. Note that the maximal error is 0.1 and 0.03 in theory when $p_e = 0.8$ and $p_e = 0.9$, respectively. It can be seen that with EBT-LISTA-SS considerably outperforms LISTA-SS in this task.

2.6. Photometric Stereo Analysis

Here we give a detailed discussion on photometric stereo analysis. To be specific, the task solves the problem of estimating the normal direction of a Lambertian surface, given q observations under different light directions. It can be formulated as

$$o = \rho L v + e, \tag{2}$$

where $o \in \mathbb{R}^q$ is the observation, $v \in \mathbb{R}^3$ represents the normal direction of the Lambertian surface to be estimated, $L \in \mathbb{R}^{q\times 3}$ represents the normalized light directions, e is the noise vector, and ρ is the diffuse albedo scalar. Although the normal vector v is unconstrained in Eq. (2), the noise vector e is found to be generally sparse [12, 13]. Therefore, we may estimate the noise e first. We introduce the orthogonal complement of L, denoted by $L^\dagger$, to rewrite Eq. (2) as

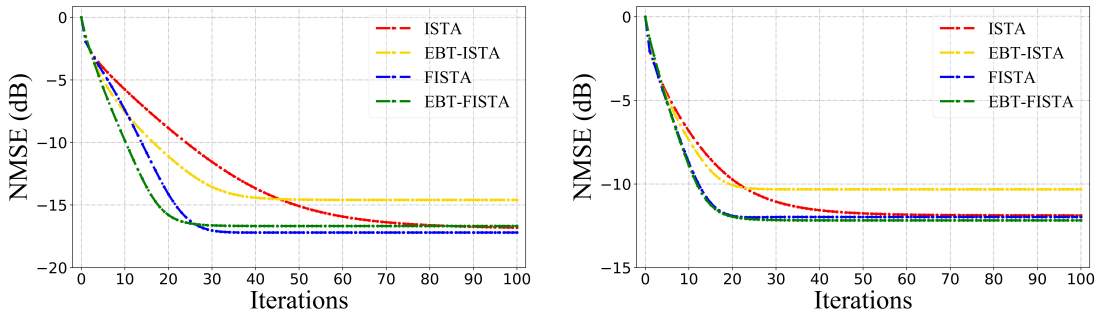
$$L^\dagger o = \rho L^\dagger L v + L^\dagger e = L^\dagger e. \tag{3}$$

On the basis of the above equation, the estimation of e is basically a noiseless sparse coding problem, where $L^\dagger$ is the dictionary matrix A, e is the sparse code $x^\star$ to be estimated in the reformulated problem, and $L^\dagger o$ is the observation y. Once we have gotten a reasonable estimation of e, we can further obtain v by using the equation $v = L^\dagger(o - e)$.

Other than Table 1 in main paper, we here build the reconstruction 3D error maps for LISTA-SS and EBT-LISTA-SS when $p_e = 0.8$ and $p_e = 0.9$, as shown in Figure 7. Note that in the picture, brighter means larger estimation error. The results from the figure also show that EBT-LISTA-SS outperforms LISTA-SS in photometric stereo analysis.

2.7. EBT mechanism on (F)ISTA

We further test the proposed EBT mechanism on the standard ISTA and FISTA. We set the regularization coefficients λ in the Lasso problem (and ISTA algorithm) as 0.1 and 0.2, respectively. With EBT incorporated, we use $\lambda\|Ax^{(t)} - y\|_1/\gamma$ as the threshold at the t-th layer, and we use λ/γ for (F)ISTA. From the experiment results shown in Figure 8, we can find that our EBT mechanism leads to faster convergence, however, the final performance is not satisfactory. This can possibly be ascribed to the convergence speed, considering that LISTA has linear convergence in theory while ISTA and FISTA converge in a sub-linear manner.



(a) λ=0.1 (b) λ=0.2

Fig. 8: NMSE of different sparse coding methods where different regularization coefficients λ are considered.

3. THEORETICAL ANALYSIS

Here, we provide theoretical analyses of the theorems shown in the main paper. Before delve deep into the proof, we first give some important notations. We use $\mathrm{supp}(x)$ to represent the support set of vector x. We denote $\mathcal{S}$ as the support set of $x^\star$, and $|\mathcal{S}|$ is the number of the elements in the set $\mathcal{S}$. We use x_i denotes the i-th element of a vector x, and A_{ij} denotes the element of matrix A placed on i-th raw and j-th column.

3.1. Proof of Theorem 1

Remind that the our EBT-LISTA is formulated as $x^{(0)} = 0$, and, for $t = 0, \ldots, d$,

$$\begin{aligned} x^{(t+1)} &= \mathrm{sh}_{b^{(t)}}((I - U^{(t)}A)x^{(t)} + U^{(t)}y), \\ b^{(t)} &= \rho^{(t)}\|U^{(t)}(Ax^{(t)} - y)\|_\phi. \end{aligned} \tag{4}$$

Theorem 1. (Convergence of EBT-LISTA) For EBT-LISTA formulated in Eq. (4) where $\phi = 1$, $x^\star$ is sampled from $\gamma(B, s)$. If s is small such that $\mu(A)(2s-1) < 1$, $U^{(t)} \in \mathcal{W}(A)$ and $\rho^{(t)} = \frac{\mu(A)}{1-\mu(A)s}$, the estimation $x^{(t)}$ at the t-th layer satisfies

$$\|x^{(t)} - x^\star\|_2 \leq q_0 \exp(c_1 t),$$

where $q_0 < sB$ and $c_1 < c_0$ hold with the probability of $1 - \mu(A)s$.

Proof. The proving process of our Theorem 1 is similar to Theorem 2 in [8].
We first prove the no "false positive" property, i.e., $\forall t \geq 0$, we have $\mathrm{supp}(x^{(t)}) \subset \mathcal{S}$. We use the Mathematical Induction

to assume $\text{supp}(x^{(t)}) \subset \mathcal{S}$ and consider $x^{(t+1)}$. For $i \notin \mathcal{S}$, i.e., $(x^\star)_i = 0$. If $(x^{(t+1)})_i \neq 0$, note that $y = Ax^\star$, there is

$$\begin{aligned}
b^{(t)} &< |(x^{(t+1)})_i| \\
&< |[(I - U^{(t)}A)x^{(t)} + U^{(t)}Ax^\star]_i| \\
&= |[(I - U^{(t)}A)(x^{(t)} - x^\star) + x^\star]_i| \\
&\leq |[(I - U^{(t)}A)(x^{(t)} - x^\star)]_i| + |(x^\star)_i| \\
&= |[(I - U^{(t)}A)(x^{(t)} - x^\star)]_i| \\
&= |\sum_j (I - U^{(t)}A)_{ij}(x^{(t)} - x^\star)_j| \\
&\leq \sum_j |(I - U^{(t)}A)_{ij}(x^{(t)} - x^\star)_j| \\
&\leq \sum_j \mu(A)|(x^{(t)} - x^\star)_j| \\
&\leq \mu(A)\|x^{(t)} - x^\star\|_1.
\end{aligned} \tag{5}$$

From above derivation, we can also conclude that

$$|[(I - U^{(t)}A)(x^{(t)} - x^\star)]_i| \leq \mu(A)\|x^{(t)} - x^\star\|_1.$$

Since $\text{supp}(x^{(t)}) \subset \mathcal{S}$, we have $|\text{supp}(x^{(t)} - x^\star)| \leq |\mathcal{S}|$, there is

$$\|(I - U^{(t)}A)(x^{(t)} - x^\star)\|_1 \leq |\mathcal{S}|\mu(A)\|x^{(t)} - x^\star\|_1.$$

Since $\|U^{(t)}A(x^{(t)} - x^\star)\|_1 = \|(x^{(t)} - x^\star) - (I - U^{(t)}A)(x^{(t)} - x^\star)\|_1$, we have

$$(1 - |\mathcal{S}|\mu(A))\|x^{(t)} - x^\star\|_1 \leq \|U^{(t)}A(x^{(t)} - x^\star)\|_1 \leq (1 + |\mathcal{S}|\mu(A))\|x^{(t)} - x^\star\|_1. \tag{6}$$

As $b^{(t)} = \rho^{(t)}\|U^{(t)}(Ax^{(t)} - y)\|_1$ and $\rho^{(t)} = \frac{\mu(A)}{1-\mu(A)s} \geq \frac{\mu(A)}{1-\mu(A)|\mathcal{S}|}$, there is

$$b^{(t)} = \rho^{(t)}\|U^{(t)}A(x^{(t)} - x^\star)\|_1 \geq \mu(A)\|x^{(t)} - x^\star\|_1. \tag{7}$$

Eq. (5) and (7) are conflicted, which means $(x^{(t+1)})_i = 0$ if $(x^\star)_i = 0$, i.e. $\text{supp}(x^{(t+1)}) \subset \mathcal{S}$. Note that $x^{(0)} = 0 \subset \mathcal{S}$, due to the Mathematical Induction, the no "false positive" property has been proved.

Due to the inequality $x - b \leq \text{sh}_b(x) \leq x + b$, we have $|\text{sh}_b(x) - \zeta| \leq |x - \zeta| + |b|$, $\forall \zeta$. Thus, when we consider the absolute value of the i-th element of $x^{(t+1)} - x^\star$, from Eq. (4), we have

$$\begin{aligned}
|(x^{(t+1)} - x^\star)_i| &= |(\text{sh}_{b^{(t)}}((I - U^{(t)}A)x^{(t)} + U^{(t)}y) - x^\star)_i| \\
&\leq |(((I - U^{(t)}A)x^{(t)} + U^{(t)}y) - x^\star)_i| + |b^{(t)}| \\
&= |(((I - U^{(t)}A)x^{(t)} + U^{(t)}Ax^\star) - x^\star)_i| + |b^{(t)}| \\
&\leq |((I - U^{(t)}A)(x^{(t)} - x^\star))_i| + |b^{(t)}|.
\end{aligned} \tag{8}$$

Since $\text{supp}(x^{(t+1)}) \subset \mathcal{S}$, we have $\|x^{(t+1)} - x^\star\|_1 = \sum_{i\in\mathcal{S}} |(x^{(t+1)} - x^\star)_i|$. There we have

$$\begin{aligned}
\|x^{(t+1)} - x^\star\|_1 &\leq \sum_{i\in\mathcal{S}} (|((I - U^{(t)}A)(x^{(t)} - x^\star))_i| + |b^{(t)}|) \\
&= \sum_{i\in\mathcal{S}} (|\sum_{j\in\mathcal{S}\setminus\{i\}} (I - U^{(t)}A)_{ij}(x^{(t)} - x^\star)_j| + |b^{(t)}|) \\
&\leq \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}\setminus\{i\}} |(I - U^{(t)}A)_{ij}(x^{(t)} - x^\star)_j| + |\mathcal{S}||b^{(t)}| \\
&\leq (|\mathcal{S}| - 1)\mu(A)\|x^{(t)} - x^\star\|_1 + |\mathcal{S}|\rho^{(t)}\|U^{(t)}A(x^{(t)} - x^\star)\|_1 \\
&\leq (|\mathcal{S}| - 1)\mu(A)\|x^{(t)} - x^\star\|_1 + \frac{\mu(A)|\mathcal{S}|}{1 - \mu(A)s}\|U^{(t)}A(x^{(t)} - x^\star)\|_1. \\
&\leq (|\mathcal{S}| + |\mathcal{S}|\frac{1 + \mu(A)s}{1 - \mu(A)s} - 1)\mu(A)\|(x^{(t)} - x^\star)\|_1.
\end{aligned} \tag{9}$$

The final step holds because $|\mathcal{S}| \leq s$ and Eq.(6). The l_2 error bound of t-th output of EBT-LISTA can be calculated as

$$\begin{aligned}\|x^{(t)} - x^\star\|_2 &\leq \|x^{(t)} - x^\star\|_1 \\ &\leq ((|\mathcal{S}| + |\mathcal{S}|\frac{1+\mu(A)s}{1-\mu(A)s} - 1)\mu(A))^t \|(x^{(0)} - x^\star)\|_1 \\ &\leq q_0 \exp(c_1 t),\end{aligned} \tag{10}$$

where $q_0 = \|x^\star\|_1$, and $c_1 = \log((|\mathcal{S}| + |\mathcal{S}|\frac{1+\mu(A)s}{1-\mu(A)s} - 1)\mu(A))$. Compare c_1 with c_0, we have

$$\exp(c_0) - \exp(c_1) = 2\mu(A)(s - \frac{|\mathcal{S}|}{1-\mu(A)s}) > 0 \tag{11}$$

hold when $|\mathcal{S}| < s(1 - \mu(A)s)$. Under this circumstance, we have

$$q_0 = \|x^\star\|_1 \leq |\mathcal{S}|B < s(1-\mu(A)s)B \leq sB. \tag{12}$$

Note that $x^\star$ is sampled from $\gamma(B, s)$, Eq. (11) and (12) hold with the probability with of $1 - \eta$, where

$$\begin{aligned}\eta &= \frac{s - |\mathcal{S}|}{s} \\ &= \mu(A)s.\end{aligned} \tag{13}$$

3.2. Proof of Lemma 2

Remind that the update rule of LISTA with support selection is formulated as $x^{(0)} = 0$, and, for $t = 0, \ldots, d$,

$$x^{(t+1)} = \mathrm{shp}_{(b^{(t)}, p^{(t)})}((I - U^{(t)}A)x^{(t)} + U^{(t)}y), \tag{14}$$

Lemma 2. (Convergence of LISTA with support selection) For LISTA with support selection formulated in Eq. (14), $x^\star$ is sampled from $\gamma(B, s)$. If s is small such that $\mu(A)(2s-1) < 1$, $U^{(t)} \in \mathcal{W}(A)$ and $b^{(t)} = \mu(A)\sup_{x^\star} \|x^{(t)} - x^\star\|_1$, there actually exist two convergence phases.

In the first phase, i.e., $t \leq t_0$, the t-th layer estimation $x^{(t)}$ satisfies

$$\|x^{(t)} - x^\star\|_2 \leq sB \exp(c_2 t),$$

where $c_2 \leq \log((2s-1)\mu(A))$. In the second phase, i.e., $t > t_0$, the estimation $x^{(t)}$ satisfies

$$\|x^{(t)} - x^\star\|_2 \leq C\|x^{(t-1)} - x^\star\|_2,$$

where $C \leq s\mu(A)$.

Proof. We want to stress that the proving process of Lemma 2 is inspired by Theorem 3 in [8]. First, we have

$$\begin{aligned}(x^{(t+1)} - x^\star)_i =& \mathrm{shp}_{(b^{(t)}, p^{(t)})}((I - U^{(t)}A)x^{(t)} + U^{(t)}y)_i - x_i^\star \\ =& \begin{cases}\mathrm{sh}_{b^{(t)}}((I - U^{(t)}A)x^{(t)} + U^{(t)}y)_i - x_i^\star, & i \in S_{p^{(t+1)}}, \\ ((I - U^{(t)}A)x^{(t)} + U^{(t)}y)_i - x_i^\star, & i \notin S_{p^{(t+1)}}.\end{cases}\end{aligned} \tag{15}$$

Where $S_{p^{(t+1)}}$ is the set of the index of the largest $p\%$ elements (in absolute value) in vector $x^{(t+1)}$. We let $g_i^{(t)} = 0$, when $i \in S_{p^{(t)}}$, and $g_i^{(t)} = 1$ otherwise. Similar to Eq. (8), there is

$$|(x^{(t+1)} - x^\star)_i| \leq |(I - U^{(t)}A)(x^{(t)} - x^\star)_i| + |b^{(t)} \odot g_i^{(t+1)}|. \tag{16}$$

Since $b^{(t)} = \mu(A)\sup_{x^\star}\|x^{(t)} - x^\star\|_1$, same as the standard LISTA, LISTA with support selection is also "no false positive" [8], i.e., $\text{supp}(x^{(t+1)}) \subset \mathcal{S}$. Therefore, we have $\|x^{(t+1)} - x^\star\|_1 = \sum_{i\in\mathcal{S}}(x^{(t+1)} - x^\star)_i$. Similar to Eq.(9), we have

$$\begin{aligned}\|x^{(t+1)} - x^\star\|_1 \leq& \sum_{i\in\mathcal{S}}(|(I - U^{(t)}A)(x^{(t)} - x^\star)_i| + |b^{(t)} \odot g_i^{(t+1)}|)\\ =&\sum_{i\in\mathcal{S}}(|\sum_{j\in\mathcal{S}\setminus\{i\}}(I - U^{(t)}A)_{ij}(x^{(t)} - x^\star)_j| + |b^{(t)} \odot g_i^{(t+1)}|)\\ \leq&\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}\setminus\{i\}}|(I - U^{(t)}A)_{ij}(x^{(t)} - x^\star)_j| + \sum_{i\in\mathcal{S}}|b^{(t)} \odot g_i^{(t+1)}|\\ \leq&(|\mathcal{S}| - 1)\mu(A)\|x^{(t)} - x^\star\|_1 + \sum_{i\in\mathcal{S}}|b^{(t)} \odot g_i^{(t+1)}|.\end{aligned} \tag{17}$$

We let S_{t+1} denote the number of non-zero entries in $x^{(t+1)}$. Also, P_{t+1} denotes the number of the largest $p^{(t+1)}\%$ elements (in absolute value) in $x^{(t+1)}$. Therefore the number of zero entries in $g(x^{(t+1)})$ is $\min(S_{t+1}, P_{t+1})$. Then Eq.(17) can be calculated as

$$\begin{aligned}\|x^{(t+1)} - x^\star\|_1 \leq&(|\mathcal{S}| - 1)\mu(A)\|x^{(t)} - x^\star\|_1 + \sum_{i\in\mathcal{S}}|b^{(t)} \odot g(x_i^{(t+1)})|\\ \leq&(|\mathcal{S}| - 1)\mu(A)\|x^{(t)} - x^\star\|_1 + (|\mathcal{S}| - \min(S_{t+1}, P_{t+1}))|b^{(t)}|\\ =&(|\mathcal{S}| - 1)\mu(A)\|x^{(t)} - x^\star\|_1 + (|\mathcal{S}| - \min(S_{t+1}, P_{t+1}))\mu(A)\sup_{x^\star}\|x^{(t)} - x^\star\|_1.\end{aligned} \tag{18}$$

We now take the supremum of Eq.(18), there is

$$\sup_{x^\star}\|x^{(t+1)} - x^\star\|_1 \leq (2s - 1 - \min(S_{t+1}, P_{t+1}))\mu(A)\sup_{x^\star}\|x^{(t)} - x^\star\|_1. \tag{19}$$

Note that $\|x^\star\|_1 \leq sB$. Assume $k = \arg\min_t(S_t, P_t)$, the l_2 upper bound of t-th output can be calculated as

$$\begin{aligned}\|x^{(t)} - x^\star\|_2 \leq \|x^{(t)} - x^\star\|_1 \leq& \sup_{x^\star}\|x^{(t)} - x^\star\|_1\\ \leq& \left(\prod_{i=1}^{t}(2s - 1 - \min(S_i, P_i))\mu(A)\right)\sup_{x^\star}\|x^{(0)} - x^\star\|_1\\ \leq& ((2s - 1 - \min(S_k, P_k))\mu(A))^t sB\\ \leq& sB\exp(c_2 t),\end{aligned} \tag{20}$$

where $c_2 = \log((2s - 1 - \min(S_k, P_k))\mu(A))$. Apparently, we have $c_2 \leq c_0 = \log((2s - 1)\mu(A))$.

From Eq.(20), we have $\|x^{(t)} - x^\star\|_1 \leq sB\exp(c_2 t)$, which means l_1 error bound can approaches to 0. Thus, there exists a t^*, when $t > t^*$, $\|x^{(t)} - x^\star\|_1 \leq min_{i\in\mathcal{S}}(x^\star)_i$. Note that $|x_i^{(t)} - (x^\star)_i| \leq \|x^{(t)} - x^\star\|_1$. If $i \in \mathcal{S}$, i.e., $(x^\star)_i \neq 0$, there exists $x_i^{(t)} \neq 0$, which means $\mathcal{S} \subset \text{supp}(x^{(t)})$. Recall the "no false positive" property, i.e., $\text{supp}(x^{(t)}) \subset \mathcal{S}$, we can conclude that $\text{supp}(x^{(t)}) = \mathcal{S}$. Since P_t increases layerwise and P_{max} is set as the upper bound of $|\mathcal{S}|$, there exists t' statisfies $P_{t'} > s$, we let $t_0 = \max(t^*, t')$, if $t > t_0$, there exists $P_t \geq |\mathcal{S}|$ and $\text{supp}(x^{(t)}) = \mathcal{S}$. Under this circumstance, if $i \in \mathcal{S}$, we have $x_i^{(t)} \neq 0$ and $i \in S_{p_t}$, which means every element in $\mathcal{S}$ will be selected as support. There we have

$$\begin{aligned}x_i^{(t+1)} - (x^\star)_i =&\text{shp}_{(b^{(t)},p^{(t)})}(((I - U^{(t)}A)x^{(t)} + U^{(t)}Ax^\star)_i) - (x^\star)_i\\ =&((I - U^{(t)}A)x^{(t)} + U^{(t)}Ax^\star)_i - (x^\star)_i\\ =&((I - U^{(t)}A)(x^{(t)} - x^\star))_i.\end{aligned} \tag{21}$$

We let $x_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ denote the vector that keeps the elements with indices of x in $\mathcal{S}$ and removes the others. Similarly, we let $M(\mathcal{S}, \mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the submatrix of matrix M which keeps the row and column if the index belongs to $\mathcal{S}$. Then, we

have

$$\begin{aligned}\|x^{(t+1)}-x^\star\|_2 &= \|(x^{(t+1)}-x^\star)_{\mathcal{S}}\|_2\\ &= \|((I-U^{(t)}A)(x^{(t)}-x^\star))_{\mathcal{S}}\|_2\\ &= \|(I-U^{(t)}A)(\mathcal{S},\mathcal{S})(x^{(t)}-x^\star)_{\mathcal{S}}\|_2\\ &\le \|(I-U^{(t)}A)(\mathcal{S},\mathcal{S})\|_2\|(x^{(t)}-x^\star)_{\mathcal{S}}\|_2\\ &= C\|(x^{(t)}-x^\star)\|_2,\end{aligned} \tag{22}$$

where $C=\|(I-U^{(t)}A)(\mathcal{S},\mathcal{S})\|_2$. Further we have $C\le\|(I-U^{(t)}A)(\mathcal{S},\mathcal{S})\|_F\le\sqrt{|\mathcal{S}|^2\mu(A)^2}\le s\mu(A)$.

3.3. Proof of Theorem 2

Remind that our EBT-LISTA with support selection can be formulated as $x^{(0)}=0$ and and for $t=0,\dots,d$,

$$\begin{aligned}x^{(t+1)} &= \mathrm{shp}_{(b^{(t)},p^{(t)})}((I-U^{(t)}A)x^{(t)}+U^{(t)}y),\\ b^{(t)} &= \rho^{(t)}\|U^{(t)}(Ax^{(t)}-y)\|_\phi.\end{aligned} \tag{23}$$

Theorem 2. (Convergence of EBT-LISTA with support selection) For EBT-LISTA with support selection and $\phi=1$, $x^\star$ is sampled from $\gamma(B,s)$. If s is small such that $\mu(A)(2s-1)<1$, $U^{(t)}\in\mathcal{W}(A)$, and $\rho^{(t)}=\frac{\mu(A)}{1-\mu(A)s}$, there exist two convergence phases.

In the first phase, i.e., $t\le t_1$, the t-th layer estimation $x^{(t)}$ satisfies

$$\|x^{(t)}-x^\star\|_2\le q_1\exp(c_3t),$$

where $c_3<c_2$, $q_1<sB$ and $t_1<t_0$ hold with a probability of $1-\mu(A)s$. In the second phase, i.e., $t>t_1$, the estimation $x^{(t)}$ satisfies

$$\|x^{(t)}-x^\star\|_2\le C\|x^{(t-1)}-x^\star\|_2,$$

where $C\le s\mu(A)$.

Proof. From Eq. (23), similar to Eq. (15) and (16), we have

$$|(x^{(t+1)}-x^\star)_i|\le|(I-U^{(t)}A)(x^{(t)}-x^\star)_i|+|b^{(t)}\odot g_i^{(t+1)}|, \tag{24}$$

where $b^{(t)}=\rho^{(t)}\|U^{(t)}(Ax^{(t)}-y)\|_1=\frac{\mu(A)}{1-\mu(A)s}\|U^{(t)}A(x^{(t)}-x^\star)\|_1$. Same as the origin EBT-LISTA, EBT-LISTA with support selection is also "no false positive" and Eq. (6) hold either. Therefore $\|U^{(t)}A(x^{(t)}-x^\star)\|_1\le(1+|\mathcal{S}|\mu(A))\|x^{(t)}-x^\star\|_1\le(1+s\mu(A))\|x^{(t)}-x^\star\|_1$. Similar to Eq. (17) and (18), there is

$$\begin{aligned}\|x^{(t+1)}-x^\star\|_1 =& \sum_{i\in\mathcal{S}}|(x^{(t+1)}-x^\star)_i|\\ \le& \sum_{i\in\mathcal{S}}(|(I-U^{(t)}A)(x^{(t)}-x^\star)_i|+|b^{(t)}\odot g(x_i^{(t+1)})|)\\ =& \sum_{i\in\mathcal{S}}(|\sum_{j\in\mathcal{S}\setminus\{i\}}(I-U^{(t)}A)_{ij}(x^{(t)}-x^\star)_j|+|b^{(t)}\odot g_i^{(t+1)}\\ \le& \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}\setminus\{i\}}|(I-U^{(t)}A)_{ij}(x^{(t)}-x^\star)_j|+\sum_{i\in\mathcal{S}}|b^{(t)}\odot g_i^{(t+1)}|\\ \le& (|\mathcal{S}|-1)\mu(A)\|x^{(t)}-x^\star\|_1+(|\mathcal{S}|-\min(S_{t+1},P_{t+1}))|b^{(t)}|\\ \le& (|\mathcal{S}|-1)\mu(A)\|x^{(t)}-x^\star\|_1+(|\mathcal{S}|-\min(S_{t+1},P_{t+1}))\mu(A)\frac{1+\mu(A)s}{1-\mu(A)s}\|(x^{(t)}-x^\star)\|_1\\ \le& (\frac{2}{1-\mu(A)s}|\mathcal{S}|-\frac{1+\mu(A)s}{1-\mu(A)s}\min(S_{t+1},P_{t+1})-1)\mu(A)\|(x^{(t)}-x^\star)\|_1.\end{aligned} \tag{25}$$

Similar to Eq. (20), the l_2 error bound can be calculated as

$$\begin{aligned}\|x^{(t)}-x^\star\|_2 \leq & \|x^{(t)}-x^\star\|_1 \\ \leq & \prod_{i=1}^{t}\left[(\frac{2}{1-\mu(A)s}|\mathcal{S}|-\frac{1+\mu(A)s}{1-\mu(A)s}\min(S_i,P_i)-1)\mu(A)\right]\|(x^{(0)}-x^\star)\|_1 \\ \leq & \left[(\frac{2}{1-\mu(A)s}|\mathcal{S}|-\frac{1+\mu(A)s}{1-\mu(A)s}\min(S_k,P_k)-1)\mu(A)\right]^t\|x^\star\|_1 \\ \leq & q_1\exp(c_3 t),\end{aligned} \tag{26}$$

where we have $q_1=\|x^\star\|_1$, and $c_3=\log((\frac{2}{1-\mu(A)s}|\mathcal{S}|-\frac{1+\mu(A)s}{1-\mu(A)s}\min(S_k,P_k)-1)\mu(A))$, and .

Compare c_3 with c_2, we have

$$\begin{aligned}\exp(c_2)-\exp(c_3) &= 2\mu(A)(s-\frac{|\mathcal{S}|}{1-\mu(A)s}+\frac{2\mu(A)s}{1-\mu(A)s}\min(S_k,P_k)) \\ &\geq 2\mu(A)(s-\frac{|\mathcal{S}|}{1-\mu(A)s})>0\end{aligned} \tag{27}$$

hold when $|\mathcal{S}|<s(1-\mu(A)s)$. Under this circumstance, we have

$$q_1=\|x^\star\|_1\leq|\mathcal{S}B<s(1-\mu(A)s)B\leq sB. \tag{28}$$

Note that $x^\star$ is sampled from $\gamma(B,s)$, Eq. (27) and (28) hold with the probability with of $1-\eta$, where

$$\begin{aligned}\eta &= \frac{s-|\mathcal{S}|}{s} \\ &= \mu(A)s.\end{aligned} \tag{29}$$

Similar to LISTA with support selection, there exists a t^{**}, when $t>t^{**}$, $\|x^{(t)}-x^\star\|_1\leq min_{i\in\mathcal{S}}(x^\star)_i$. Therefore, $\mathrm{supp}(x^{(t)})=\mathcal{S}$. Recall that $c_3<c_2$ holds with the probability of $1-\eta$, we have $t^{**}<t^*$ with the probability of $1-\eta$. When we use the same settings for $p^{(t)}$ as LISTA-SS, then we have same t' satisfying $P_{t'}>s$. Let $t_1=\max(t^{**},t')$, we have $t_1\leq t_0$ with the probability of $1-\eta$. When $t>t_1$, we have $x_i^{(t)}\neq 0$ and $i\in S_{p_t}$. Same as Eq. (21) and (22), there is

$$\|x^{(t+1)}-x^\star\|_2\leq C\|(x^{(t)}-x^\star)\|_2, \tag{30}$$

where $C=\|(I-U^{(t)}A)(\mathcal{S},\mathcal{S})\|_2\leq\|(I-U^{(t)}A)(\mathcal{S},\mathcal{S})\|_F\leq\sqrt{|\mathcal{S}|^2\mu(A)^2}\leq s\mu(A)$.

4. REFERENCES

[1] Pierre Ablin, Thomas Moreau, Mathurin Massias, and Alexandre Gramfort, "Learning step sizes for unfolded sparse coding," in *Advances in Neural Information Processing Systems*, 2019, pp. 13100–13110.

[2] Yangyang Li, Lin Kong, Fanhua Shang, Yuanyuan Liu, Hongying Liu, and Zhouchen Lin, "Learned extragradient ista with interpretable residual structures for sparse coding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 8501–8509.

[3] Ziyang Zheng, Wenrui Dai, Duoduo Xue, Chenglin Li, Junni Zou, and Hongkai Xiong, "Hybrid ista: unfolding ista with convergence guarantees using free-form deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] Jialin Liu, Xiaohan Chen, Zhangyang Wang, and Wotao Yin, "Alista: Analytic weights are as good as learned weights in lista," in *International Conference on Learning Representations (ICLR)*, 2019.

[5] Aviad Aberdam, Alona Golts, and Michael Elad, "Ada-lista: Learned solvers adaptive to varying models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] Jian Zhang and Bernard Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1828–1837.

[7] Ding Liu, Zhaowen Wang, Bihan Wen, Jianchao Yang, Wei Han, and Thomas S Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3194–3207, 2016.

[8] Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin, "Theoretical linear convergence of unfolded ista and its practical weights and thresholds," in *Advances in Neural Information Processing Systems*, 2018, pp. 9061–9071.

[9] Kailun Wu, Yiwen Guo, Ziang Li, and Changshui Zhang, "Sparse coding with gated learned ista," in *Proceedings of the International Conference on Learning Representations*, 2020.

[10] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: a system for large-scale machine learning.," in *Osdi*. Savannah, GA, USA, 2016, vol. 16, pp. 265–283.

[12] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 703–717.

[13] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa, "Robust photometric stereo using sparse regression," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 318–325.