SUPPLEMENTARY MATERIAL FOR NONLINEAR BLIND COMPRESSED SENSING UNDER SIGNAL-DEPENDENT NOISE

Rudrajit Das[1] *and Ajit Rajwade*[2]

[1]Dept. of Electrical Engineering, [2]Dept. of Computer Science & Engineering; IIT Bombay

I. MULTIPLICATIVE UPDATE RULE DERIVATION

In this section, we show the derivation of the multiplicative update rule discussed in Section 3 of the main paper. Recall that all square root and division operations are performed element-wise and the $\odot$ operator denotes the element-wise product of two vectors or matrices.

In order to estimate $\boldsymbol{\theta_i^{(t+1)}}$ (i.e. the sparse codes for the i^{th} signal, $1 \leq i \leq T$, at the $(t+1)^{\text{th}}$ iteration), given $\boldsymbol{\theta_i^{(t)}}$ and $\boldsymbol{A^{(t)}}$, we must minimize the following objective function:

$$J_\theta^{(i)} = \left\| \sqrt{\boldsymbol{y_i^{(1)}} + c\boldsymbol{1_{m_1}}} - \sqrt{\boldsymbol{\Phi_i^{(1)} A^{(t)} \theta_i} + c\boldsymbol{1_{m_1}}} \right\|^2 + \lambda \|\boldsymbol{\theta_i}\|_1 \tag{1}$$

$$\text{such that } \boldsymbol{\theta_i} \succeq \boldsymbol{0_k},$$

The gradient (technically sub-gradient as $\|\boldsymbol{\theta_i}\|_1$ is not differentiable at a zero value for any element of $\boldsymbol{\theta_i}$) of $J_\theta^{(i)}$ with respect to $\boldsymbol{\theta_i}$ evaluated at $\boldsymbol{\theta_i} = \boldsymbol{\theta_i^{(t)}}$ is:

$$\left.\frac{\partial J_\theta^{(i)}}{\partial \boldsymbol{\theta_i}}\right|_{\boldsymbol{\theta_i^{(t)}}} = (\boldsymbol{\Phi_i^{(1)} A^{(t)}})^T \left(\boldsymbol{1_m} - \frac{\sqrt{\boldsymbol{y_i^{(1)}} + c\boldsymbol{1_{m_1}}}}{\sqrt{\boldsymbol{\Phi_i^{(1)} A^{(t)} \theta_i^{(t)}} + c\boldsymbol{1_{m_1}}}} \right) + \lambda \mathbb{1}(\boldsymbol{\theta_i^{(t)}}) \tag{2}$$

As defined in the main paper, $c \triangleq \frac{3}{8}$ and $\mathbb{1}(z) = 1$ for $z > 0$ and $\mathbb{1}(z) = 0$ for $z = 0$. We use gradient descent to update $\boldsymbol{\theta_i^{(t)}}$ with step size $\boldsymbol{\eta_i^{(t)}}$ (represented as a vector due to different step-sizes for each component of $\boldsymbol{\theta_i^{(t)}}$) chosen as follows:

$$\boldsymbol{\eta_i^{(t)}} = \frac{\boldsymbol{\theta_i^{(t)}}}{(\boldsymbol{\Phi_i^{(1)} A^{(t)}})^T \boldsymbol{1_{m_1}}} \tag{3}$$

Using the step-size in (3), $\boldsymbol{\theta_i^{(t+1)}}$ becomes:

$$\boldsymbol{\theta_i^{(t+1)}} = \boldsymbol{\theta_i^{(t)}} - \boldsymbol{\eta_i^{(t)}} \odot \left.\frac{\partial J_\theta^{(i)}}{\partial \boldsymbol{\theta_i}}\right|_{\boldsymbol{\theta_i^{(t)}}}$$
$$= \max\left(\boldsymbol{0}, \left[\frac{\boldsymbol{\theta}}{(\boldsymbol{\Phi_i^{(1)} A^{(t)}})^T \boldsymbol{1_{m_1}}} \odot \left\{ (\boldsymbol{\Phi_i^{(1)} A^{(t)}})^T \left(\frac{\sqrt{\boldsymbol{y_i^{(1)}} + c\boldsymbol{1_{m_1}}}}{\sqrt{\boldsymbol{\Phi_i^{(1)} A^{(t)} \theta_i^{(t)}} + c\boldsymbol{1_{m_1}}}} \right) - \lambda \mathbb{1}(\boldsymbol{\theta_i^{(t)}}) \right\} \right] \right) \tag{4}$$

Observe that the $\boldsymbol{\theta_i^{(t+1)}}$ in (4) is the $\boldsymbol{\theta_i^{(\text{new})}}$ term in (5) of the main paper. Empirically, we observed in our experiments that this value of $\boldsymbol{\theta_i^{(t+1)}}$ obtained using the step-size in (3) of this document (henceforth referred to as $\boldsymbol{\theta_i^{(\text{new})}}$) does indeed decrease the objective function value. However, due to the intractability of the Hessian matrix, we cannot (at present) theoretically guarantee whether the step-size in (3) would indeed decrease the value of the objective function. For this, we propose reducing the magnitude of the step-size by multiplying it with a suitable scalar constant $0 < \beta_i^{(t)} \leq 1$ which ensures a decrease in the value of the objective function. More specifically:

$$\boldsymbol{\theta_i^{(t+1)}} = \boldsymbol{\theta_i^{(t)}} - \beta_i^{(t)} \boldsymbol{\eta_i^{(t)}} \odot \left.\frac{\partial J_\theta^{(i)}}{\partial \boldsymbol{\theta_i}}\right|_{\boldsymbol{\theta_i^{(t)}}}$$
$$= (1 - \beta_i^{(t)}) \boldsymbol{\theta_i^{(t)}} + \beta_i^{(t)} \boldsymbol{\theta_i^{(\text{new})}} \tag{5}$$

In the above equation, $\boldsymbol{\theta_i^{(\text{new})}}$ is as obtained in (4) of this paper. This completes the derivation of (4) of the main paper.

Next, in order to estimate $\boldsymbol{A^{(t+1)}}$, given $\boldsymbol{\theta_i^{(t+1)}}$ and $\boldsymbol{A^{(t)}}$, we must minimize the following objective function:

$$J_{\boldsymbol{A}} = \sum_{i=1}^{T'} \left\| \sqrt{\boldsymbol{y_i^{(1)}} + c\boldsymbol{1_{m_1}}} - \sqrt{\boldsymbol{\Phi_i^{(1)} A \theta_i^{(t+1)}} + c\boldsymbol{1_{m_1}}} \right\|^2 \tag{6}$$

$$\text{such that } \boldsymbol{A} \succeq \boldsymbol{0_{n \times k}}.$$

The gradient of J_A with respect to $\boldsymbol{A}$ evaluated at $\boldsymbol{A} = \boldsymbol{A^{(t)}}$ is:

$$\left.\frac{\partial J_A}{\partial \boldsymbol{A}}\right|_{\boldsymbol{A^{(t)}}} = \sum_{i=1}^{T'} (\boldsymbol{\Phi_i^{(1)}})^T \gamma_i (\boldsymbol{\theta_i^{(t+1)}})^T \text{ where}$$

$$\boldsymbol{\gamma_i} = \left(\boldsymbol{1_{m_1}} - \frac{\sqrt{\boldsymbol{y_i^{(1)}} + c\boldsymbol{1_{m_1}}}}{\sqrt{\boldsymbol{\Phi_i^{(1)} A^{(t)} \theta_i^{(t+1)}} + c\boldsymbol{1_{m_1}}}} \right) \tag{7}$$

In this case, the step size $\boldsymbol{\eta^{(t)}}$ (this time a matrix) is chosen as follows:

$$\boldsymbol{\eta^{(t)}} = \frac{\boldsymbol{A^{(t)}}}{\sum_{i=1}^{T'} (\boldsymbol{\Phi_i^{(1)}})^T \boldsymbol{1_{m_1}} (\boldsymbol{\theta_i^{(t+1)}})^T} \tag{8}$$

Using the step-size in (8), $\boldsymbol{A^{(t+1)}}$ becomes:

$$\boldsymbol{A^{(t+1)}} = \boldsymbol{A^{(t)}} - \boldsymbol{\eta^{(t)}} \odot \frac{\partial J_A}{\partial \boldsymbol{A}}\bigg|_{\boldsymbol{A^{(t)}}}$$
$$= \left[\frac{\boldsymbol{A^{(t)}}}{\sum_{i=1}^{T'} (\boldsymbol{\Phi_i^{(1)}})^T \boldsymbol{1_{m_1}} (\boldsymbol{\theta_i^{(t+1)}})^T} \odot \right.$$
$$\left. \sum_{i=1}^{T'} (\boldsymbol{\Phi_i^{(1)}})^T \left(\frac{\sqrt{\boldsymbol{y_i^{(1)}} + c\boldsymbol{1_{m_1}}}}{\sqrt{\boldsymbol{\Phi_i^{(1)}} \boldsymbol{A^{(t)}} \boldsymbol{\theta_i^{(t+1)}} + c\boldsymbol{1_{m_1}}}} \right) (\boldsymbol{\theta_i^{(t+1)}})^T \right] \tag{9}$$

Observe that the $\boldsymbol{A^{(t+1)}}$ in (9) is the $\boldsymbol{A}^{(\text{new})}$ term in (7) of the main paper. Here also, empirically, we found that this value of $\boldsymbol{A^{(t+1)}}$ obtained using the step-size in (8) of this document (henceforth referred to as $\boldsymbol{A}^{(\text{new})}$) does indeed decrease the objective function value. However, once again, the Hessian is very complicated and we cannot (at present) theoretically guarantee whether the step-size in (8) would indeed decrease the value of the objective function. To ensure a decrease in the objective function value, we once again propose reducing the magnitude of the step-size by multiplying it with a suitable scalar constant $0 < \beta^{(t)} \leq 1$. More specifically:

$$\boldsymbol{A^{(t+1)}} = \boldsymbol{A^{(t)}} - \beta^{(t)} \boldsymbol{\eta^{(t)}} \odot \frac{\partial J_A}{\partial \boldsymbol{A}}\bigg|_{\boldsymbol{A^{(t)}}}$$
$$= (1 - \beta^{(t)}) \boldsymbol{A^{(t)}} + \beta^{(t)} \boldsymbol{A}^{(\text{new})} \tag{10}$$

In the above equation, $\boldsymbol{A}^{(\text{new})}$ is as obtained in (9) of this paper. This completes the derivation of (6) of the main paper and concludes the derivation of the multiplicative update rule.

II. Error Bound Derivation

In this section, we show the derivations of **Theorem 1** (as well as **Lemma 1**, **Lemma 2** and **Lemma 3**) and ε.

Firstly, we restate **Theorem 1** for ready reference of the reader.

Theorem 1. *Consider that the entries of all the sensing matrices $\{\boldsymbol{\Phi_i}\}_{i=1}^T$ are independently drawn from a $\{0,1\}$ Bernoulli distribution with probability of drawing 1 (and 0) being 0.5, denoted by $Bernoulli(0.5)$. Let $\boldsymbol{A_e}$ and $\{(\boldsymbol{\theta_i})_{\boldsymbol{e}}\}_{i=T'+1}^T$ be the estimates of $\boldsymbol{A}$ and $\{\boldsymbol{\theta_i}\}_{i=T'+1}^T$ respectively, upon running NLBCS. Suppose that the condition $\mathcal{C}$ in Eqn. (8) of the main paper holds. Also suppose that $\mathcal{V}$ is large enough so that $\widehat{E_v} \approx \mathbb{E}_{val}\left[\left\|\sqrt{\boldsymbol{y^{(2)}} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi^{(2)}} \boldsymbol{A_e} \boldsymbol{\theta_e} + c\boldsymbol{1_{m_2}}}\right\|^2\right]$ where the symbol E_{val} indicates that the expectation is over signals in $\mathcal{V}$ and including all possible $Bernoulli(0.5)$ sensing matrices. Then with probability $p(\delta, T - T')$, we have the following bound:*

$$\mathbb{E}_{val}\left[\|\boldsymbol{x} - \boldsymbol{x_e}\|^2\right] \leq \mathcal{O}\left(\frac{I\varepsilon}{1-\delta}\right)$$

where $\boldsymbol{x} = \boldsymbol{A\theta}$ and $\boldsymbol{x_e} = \boldsymbol{A_e \theta_e}$ are the actual and estimated (using the NLBCS Algorithm) values of the signal, I is the upper bound on the ℓ_1 norm of each signal as defined previously, δ is a parameter in $(0,1)$ and $p(\delta, T - T') \geq 1 - \mathcal{O}\Big(\exp\Big(-\widetilde{c}\delta^2 m_2 \sqrt{T - T'}\Big)\Big)$ with $\widetilde{c} > 0$ for large $(T - T')$ (i.e. the validation set is large enough).

Next, we first state **Lemma 1** and then prove it.

Lemma 1. *Under the settings of* ***Theorem 1****, we have:*

$$\mathbb{E}_{val}\left[\left\|\sqrt{\boldsymbol{\Phi^{(2)}x} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi^{(2)}x_e} + c\boldsymbol{1_{m_2}}}\right\|^2\right] \leq \frac{m_2 \varepsilon}{4}$$

Proof: Since we have a large number of validation examples, i.e. $(T - T') \to \infty$, we have:

$$\widehat{E_v} = \sum_{i=T'+1}^{T} \frac{\left\|\sqrt{\boldsymbol{y_i^{(2)}} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi_i^{(2)}} \boldsymbol{A_e} (\boldsymbol{\theta_i})_{\boldsymbol{e}} + c\boldsymbol{1_{m_2}}}\right\|^2}{(T - T')}$$
$$\approx \mathbb{E}_{val}\left[\left\|\sqrt{\boldsymbol{y^{(2)}} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi^{(2)}} \boldsymbol{A_e \theta_e} + c\boldsymbol{1_{m_2}}}\right\|^2\right] \tag{11}$$

The above expectation is over the validation set. Now:

$$\mathbb{E}_{val}\left[\left\|\sqrt{\boldsymbol{y^{(2)}} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi^{(2)}} \boldsymbol{A_e \theta_e} + c\boldsymbol{1_{m_2}}}\right\|^2\right]$$
$$= \mathbb{E}_{val}\left[\|\boldsymbol{d^{(2)}} + \boldsymbol{z^{(2)}}\|^2\right] \tag{12}$$

where $\boldsymbol{d^{(2)}} = \sqrt{\boldsymbol{\Phi^{(2)} A\theta} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi^{(2)} A_e \theta_e} + c\boldsymbol{1_{m_2}}}$ and $\boldsymbol{z^{(2)}}$ is the m_2-dimensional noise term, each entry of which is approximately distributed as $\mathcal{N}(0, 1/4)$ (as per the Anscombe transform). Then (12) further simplifies as:

$$\mathbb{E}_{val}\left[\|\boldsymbol{d^{(2)}} + \boldsymbol{z^{(2)}}\|^2\right] =$$
$$\mathbb{E}_{val}\left[\|\boldsymbol{d^{(2)}}\|^2\right] + \mathbb{E}_{val}\left[\|\boldsymbol{z^{(2)}}\|^2\right] + 2\,\mathbb{E}_{val}\left[\langle \boldsymbol{d^{(2)}}, \boldsymbol{z^{(2)}} \rangle\right] \tag{13}$$

In (13), clearly $\mathbb{E}_{val}\left[\|\boldsymbol{z^{(2)}}\|^2\right] = m_2/4$. Now recall that $\boldsymbol{A_e}$ was computed over the training set and so it is independent of the noise for the validation set. Also, $\boldsymbol{\theta_e}$ was computed (for both the training as well as the validation set) over $\boldsymbol{\phi^{(1)}}$ and hence it depends only on $\boldsymbol{z^{(1)}}$, i.e. the first m_1 entries of the overall m-dimensional noise vector $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z^{(1)}} \\ \boldsymbol{z^{(2)}} \end{bmatrix}$. Thus, $\boldsymbol{d^{(2)}}$ which depends on $\boldsymbol{A_e}$ and $\boldsymbol{\theta_e}$ is independent of $\boldsymbol{z^{(2)}}$. Hence, we have $\mathbb{E}_{val}\left[\langle \boldsymbol{d^{(2)}}, \boldsymbol{z^{(2)}} \rangle\right] = 0$. **This is the rationale for splitting the observations and sensing matrices into two disjoint parts.** Hence, we have:

$$\mathbb{E}_{val}\left[\|\boldsymbol{d^{(2)}} + \boldsymbol{z^{(2)}}\|^2\right] = \mathbb{E}_{val}\left[\|\boldsymbol{d^{(2)}}\|^2\right] + \frac{m_2}{4} + 0$$
$$\implies \mathbb{E}_{val}\left[\|\boldsymbol{d^{(2)}}\|^2\right] \approx \widehat{E_v} - \frac{m_2}{4} \leq \frac{m_2 \varepsilon}{4}. \tag{14}$$

The last inequality in (14) follows from (8) of the main paper. Recall that $\boldsymbol{x} = \boldsymbol{A\theta}$ and $\boldsymbol{x_e} = \boldsymbol{A_e\theta_e}$. Thus, $\boldsymbol{d^{(2)}}$ becomes:

$$\boldsymbol{d^{(2)}} = \sqrt{\boldsymbol{\Phi^{(2)}x} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi^{(2)}x_e} + c\boldsymbol{1_{m_2}}}.$$

Using the above, (14) can be written as:

$$\mathbb{E}_{val}\left[\left\|\sqrt{\boldsymbol{\Phi^{(2)}x} + c\boldsymbol{1_{m_2}}} - \sqrt{\boldsymbol{\Phi^{(2)}x_e} + c\boldsymbol{1_{m_2}}}\right\|^2\right] \leq \frac{m_2\varepsilon}{4}. \tag{15}$$

This finishes the proof of **Lemma 1** of the main paper.

Now, we first state **Lemma 2** and then prove it.

Lemma 2. *Given that **Lemma 1** holds, we have:*

$$\mathbb{E}_{val}\left[\|\boldsymbol{\Phi^{(2)}}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] \leq \mathcal{O}(Im_2\varepsilon).$$

Proof: Here we try to upper bound the expected value of the squared norm of $\boldsymbol{\Phi^{(2)}}(\boldsymbol{x} - \boldsymbol{x_e})$ over the validation set, using (15) as follows:

$$\begin{aligned}
&\mathbb{E}_{val}\left[\|\boldsymbol{\Phi^{(2)}}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] = \\
&\quad \mathbb{E}_{val}\left[\left\|\left(\boldsymbol{\Phi^{(2)}x} + c\boldsymbol{1}_{m_2}\right) - \left(\boldsymbol{\Phi^{(2)}x_e} + c\boldsymbol{1}_{m_2}\right)\right\|^2\right] \\
&\qquad \leq B\,\mathbb{E}_{val}\left[\|\boldsymbol{d^{(2)}}\|^2\right] \text{ where} \\
&B \triangleq \max_{1\leq j\leq m_2}\left(\sqrt{\boldsymbol{\Phi^{(2)}x} + c\boldsymbol{1}_{m_2}} + \sqrt{\boldsymbol{\Phi^{(2)}x_e} + c\boldsymbol{1}_{m_2}}\right)_j^2; \\
&\qquad \boldsymbol{v}_j \text{ denotes the } j^{th} \text{ element of the vector } \boldsymbol{v}.
\end{aligned} \tag{16}$$

In order to show how we can obtain the inequality in (16), consider two d dimensional non-negative vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Then:

$$\|\boldsymbol{u} - \boldsymbol{v}\|^2 = \sum_{r=1}^{d}(\boldsymbol{u}_r - \boldsymbol{v}_r)^2 = \sum_{r=1}^{d}(\boldsymbol{u}_r - \boldsymbol{v}_r)^2$$

$$= \sum_{r=1}^{d}(\sqrt{\boldsymbol{u}_r} - \sqrt{\boldsymbol{v}_r})^2(\sqrt{\boldsymbol{u}_r} + \sqrt{\boldsymbol{v}_r})^2 \leq \sum_{r=1}^{d}(\sqrt{\boldsymbol{u}_r} - \sqrt{\boldsymbol{v}_r})^2 M,$$

where $M \triangleq \max_{1\leq r\leq d}(\sqrt{\boldsymbol{u}_r} + \sqrt{\boldsymbol{v}_r})^2$. Thus, we get:

$$\|\boldsymbol{u} - \boldsymbol{v}\|^2 \leq M\|\sqrt{\boldsymbol{u}} - \sqrt{\boldsymbol{v}}\|^2.$$

This completes the proof of the inequality in (16).

Now, noting that Φ is a $Bernoulli(0.5)$ random matrix, it is trivial to see that $B = \mathcal{O}(I)$, where I is the maximum ℓ_1 norm of a signal (as specified in the main paper). Using this as well (15) and (16), we get:

$$\mathbb{E}_{val}\left[\|\boldsymbol{\Phi^{(2)}}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] \leq \frac{m_2\varepsilon}{4}\mathcal{O}(I) = \mathcal{O}(Im_2\varepsilon). \tag{17}$$

This finishes the proof of **Lemma 2** of the main paper.

Finally, we state the complete version of **Lemma 3** (recall that in the main paper, we had stated only an abridged version of it) and then prove it, while also precisely quantifying $p(\delta, T - T^{'})$. Before that let $T_{\mathcal{V}} = T - T^{'}$.

Lemma 3. *Let $\delta \in (0, 1)$. Then with probability $p(\delta, T_{\mathcal{V}})$, the following bound holds:*

$$\mathbb{E}_{val}\left[\|(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] \leq \frac{\mathbb{E}_{val}\left[\|\boldsymbol{\Phi^{(2)}}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right]}{m_2(1-\delta)},$$

where $p(\delta, T_{\mathcal{V}}) = 1 - 2\exp(-c_1c_2\delta\min(c_2\delta\Gamma_2, \Gamma_\infty))$, c_1 and c_2 are positive constants, and

$$\Gamma_2 = m_2\Big(\sum_{i=T'+1}^{T}\|\boldsymbol{x_i} - (\boldsymbol{x_i})_{\boldsymbol{e}}\|^2\Big)^2\Big/\Big(\sum_{i=T'+1}^{T}\|\boldsymbol{x_i} - (\boldsymbol{x_i})_{\boldsymbol{e}}\|^4\Big)$$

$$\Gamma_\infty = m_2\Big(\sum_{i=T'+1}^{T}\|\boldsymbol{x_i} - (\boldsymbol{x_i})_{\boldsymbol{e}}\|^2\Big)\Big/\Big(\max_{T'<i\leq T}\|\boldsymbol{x_i} - (\boldsymbol{x_i})_{\boldsymbol{e}}\|^2\Big).$$

Further, when $T_{\mathcal{V}}$ is large enough, we have $p(\delta, T_{\mathcal{V}}) \geq 1 - \mathcal{O}\Big(\exp\big(-\widetilde{c}\delta^2 m_2\sqrt{T_{\mathcal{V}}}\big)\Big)$ with $\widetilde{c} > 0$.

Note that in the main paper, we had not introduced the variable $T_{\mathcal{V}}$, due to which the abridged version of **Lemma 3** in the main paper is stated in terms of $(T - T^{'})$ instead of $T_{\mathcal{V}}$.

Proof: We express $\boldsymbol{\Phi^{(2)}} = \frac{1}{2}(\boldsymbol{\widetilde{\Phi}^{(2)}} + \boldsymbol{1}_{m\times n})$ where the entries of $\boldsymbol{\widetilde{\Phi}^{(2)}}$ are such that it contains 1 (and correspondingly -1) wherever $\boldsymbol{\Phi^{(2)}}$ contains a 1 (and correspondingly 0) and $\boldsymbol{1}_{m\times n}$ is a $m \times n$ matrix consisting of only 1's. Thus, we have:

$$\begin{aligned}
&\mathbb{E}_{val}\left[\|\boldsymbol{\Phi^{(2)}}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] = \\
&\quad \mathbb{E}_{val}\left[\|\widetilde{\boldsymbol{\Phi}}^{(2)}(\boldsymbol{x} - \boldsymbol{x_e}) + \boldsymbol{1}_{m\times n}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] = \\
&\mathbb{E}_{val}\left[\|\widetilde{\boldsymbol{\Phi}}^{(2)}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] + \mathbb{E}_{val}\left[\|\boldsymbol{1}_{m\times n}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right] \\
&\qquad + \mathbb{E}_{val}\left[(\boldsymbol{x} - \boldsymbol{x_e})^T\boldsymbol{1}_{m\times n}^T\widetilde{\boldsymbol{\Phi}}^{(2)}(\boldsymbol{x} - \boldsymbol{x_e})\right]
\end{aligned} \tag{18}$$

Let us now analyze the last expectation (third term) obtained in the final step of (18). As mentioned before, we are taking expectation over all possible signals (drawn from the same distribution as that of the training signals) as well as sensing matrices (whose entries are i.i.d $Bernoulli(0.5)$ random variables) for the validation set. Let us first take expectation with respect to the (modified) sensing matrix $\widetilde{\boldsymbol{\Phi}}^{(2)}$ whose entries are ± 1 with 0.5 probability each. As a result the value of $\mathbb{E}_{\boldsymbol{\widetilde{\Phi}^{(2)}}}\left[\boldsymbol{1}_{m\times n}^T\widetilde{\boldsymbol{\Phi}}^{(2)}\right] = \boldsymbol{0_{n\times n}}$ (i.e. the $n \times n$ zero matrix). Therefore, the third expectation term turns out to be 0, i.e. $\mathbb{E}_{val}\left[(\boldsymbol{x} - \boldsymbol{x_e})^T\boldsymbol{1}_{m\times n}^T\widetilde{\boldsymbol{\Phi}}^{(2)}(\boldsymbol{x} - \boldsymbol{x_e})\right] = 0$. Also the second expectation term (i.e. $\mathbb{E}_{val}\left[\|\boldsymbol{1}_{m\times n}(\boldsymbol{x} - \boldsymbol{x_e})\|^2\right]$)

is non-negative. Using these two facts in (18) and (17), we get:

$$\mathbb{E}_{val}\left[\|\widetilde{\mathbf{\Phi}}^{(2)}(\boldsymbol{x}-\boldsymbol{x_e})\|^2\right] \leq \mathbb{E}_{val}\left[\|\mathbf{\Phi^{(2)}}(\boldsymbol{x}-\boldsymbol{x_e})\|^2\right] \leq \mathcal{O}(Im_2\varepsilon). \quad (19)$$

Let us consider all the true and estimated signals belonging to the validation set, which are $\{\boldsymbol{x}_i\}_{i=T'+1}^{T}$ and $\{(\boldsymbol{x}_i)_e\}_{i=T'+1}^{T}$. For ease of notation, let $T'' = T' + 1$ and $\boldsymbol{\Delta x}_i = \boldsymbol{x_i} - (\boldsymbol{x_i})_{\boldsymbol{e}}$. Since the binary random variable taking values of ± 1 with 0.5 probability each is a sub-Gaussian random variable with zero mean and unit variance, using Theorem 1 of [1], we have:

$$\mathbb{P}\left(\left|\sum_{i=T''}^{T}\frac{\|\widetilde{\mathbf{\Phi}}_i^{(2)}\boldsymbol{\Delta x_i}\|^2}{m_2 T_{\mathcal{V}}} - \sum_{i=T''}^{T}\frac{\|\boldsymbol{\Delta x_i}\|^2}{T_{\mathcal{V}}}\right| > \delta\sum_{i=T''}^{T}\frac{\|\boldsymbol{\Delta x_i}\|^2}{T_{\mathcal{V}}}\right)$$
$$\leq 2\exp\left(-c_1\min\left(\frac{c_2^2\delta^2}{\|\phi\|_{\psi_2}^4}\Gamma_2, \frac{c_2\delta}{\|\phi\|_{\psi_2}^2}\Gamma_\infty\right)\right). \quad (20)$$

In (20):

$$\Gamma_2 = m_2\Big(\sum_{i=T''}^{T}\|\boldsymbol{\Delta x_i}\|^2\Big)^2\Big/\Big(\sum_{i=T''}^{T}\|\boldsymbol{\Delta x_i}\|^4\Big)$$

$$\Gamma_\infty = m_2\Big(\sum_{i=T''}^{T}\|\boldsymbol{\Delta x_i}\|^2\Big)\Big/\Big(\max_{T'<i\leq T}\|\boldsymbol{\Delta x_i}\|^2\Big)$$

$\|\phi\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}\Big(\mathbb{E}\left[|\phi|^p\right]\Big)^{1/p}$ (as defined in [1])

We get $\|\phi\|_{\psi_2} = 1$ for binary ± 1 random variables.

And finally, c_1 and c_2 are positive constants.

Now since $T_{\mathcal{V}}$ is large (as mentioned earlier), (20) in this document is approximately:

$$\mathbb{P}\left(\left|\frac{1}{m_2}\mathbb{E}_{val}\left[\|\widetilde{\mathbf{\Phi}}^{\mathbf{(2)}}\boldsymbol{\Delta x}\|^2\right] - \mathbb{E}_{val}\left[\|\boldsymbol{\Delta x}\|^2\right]\right| > \delta\,\mathbb{E}_{val}\left[\|\boldsymbol{\Delta x}\|^2\right]\right)$$
$$\leq 2\exp(-c_1c_2\delta\min(c_2\delta\Gamma_2, \Gamma_\infty)) \quad (21)$$

Thus from (21), we have:

$$\mathbb{P}\Big(m_2(1-\delta)\,\mathbb{E}_{val}\left[\|\boldsymbol{\Delta x}\|^2\right] < \mathbb{E}_{val}\left[\|\widetilde{\mathbf{\Phi}}^{\mathbf{(2)}}\boldsymbol{\Delta x}\|^2\right] < m_2(1+\delta)\,\mathbb{E}_{val}\left[\|\boldsymbol{\Delta x}\|^2\right]\Big) \geq$$
$$1 - 2\exp(-c_1c_2\delta\min(c_2\delta\Gamma_2, \Gamma_\infty)) = p(\delta, T_{\mathcal{V}}). \quad (22)$$

We want $p(\delta, T_{\mathcal{V}})$ to be close to 1 for large values of $T_{\mathcal{V}}$ for δ not too small. For this, we need to estimate the order of Γ_2 and Γ_∞. We provide estimates for Γ_2 and Γ_∞ in the regime of large $T_{\mathcal{V}}$, where the quantities in the numerators and denominators of Γ_2 and Γ_∞ can be approximated by (and will be close to) their respective expected values.

Let $\mathbb{E}\left[\|\boldsymbol{\Delta x}\|^2\right] = e$ and $\mathbb{E}\left[\|\boldsymbol{\Delta x}\|^4\right] = \alpha\,\mathbb{E}\left[\|\boldsymbol{\Delta x}\|^2\right]^2 = \alpha e^2$ where $\alpha \geq 1$ is a positive constant (since for a random variable X, we have $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$). Further, we have:

$$\Gamma_2 = m_2\frac{\Big(\sum_{i=T''}^{T}\|\boldsymbol{\Delta x_i}\|^4\Big) + \Big(\sum_{i,j=T''}^{T}\|\boldsymbol{\Delta x_i}\|^2\|\boldsymbol{\Delta x_j}\|^2\Big)}{\Big(\sum_{i=T''}^{T}\|\boldsymbol{\Delta x_i}\|^4\Big)}$$
$$\approx m_2\frac{T_{\mathcal{V}}\alpha e^2 + T_{\mathcal{V}}(T_{\mathcal{V}}-1)e^2}{T_{\mathcal{V}}\alpha e^2} \sim \mathcal{O}(m_2T_{\mathcal{V}}).$$

The above approximate equality is obtained by replacing all the summations by their respective expected values which should hold approximately for large $T_{\mathcal{V}}$.

Now, let the PDF and CDF of $\|\boldsymbol{\Delta x}\|^2$ be denoted by $f(.)$ and $F(.)$, respectively. Then the CDF of $\max_{T'<i\leq T}\|\boldsymbol{\Delta x_i}\|^2$ is simply given by $F(.)^{T_{\mathcal{V}}}$. Thus:

$$\mathbb{E}\left[\max_{T'<i\leq T}\|\boldsymbol{\Delta x_i}\|^2\right] = \int_{t=0}^{\infty}T_{\mathcal{V}}[F(t)]^{(T_{\mathcal{V}}-1)}f(t)tdt$$
$$\leq T_{\mathcal{V}}\Big(\int_{t=0}^{\infty}[F(t)]^{(2T_{\mathcal{V}}-2)}f(t)dt\Big)^{0.5}\Big(\int_{t=0}^{\infty}t^2f(t)dt\Big)^{0.5}.$$

The last step follows from the Cauchy-Schwarz inequality for integrals. It is easy to see that in the obtained inequality, the first integral (inside square root) using the substitution $u = F(t)$, becomes:

$$\int_0^1 u^{(2T_{\mathcal{V}}-2)}du = 1/(2T_{\mathcal{V}}-1).$$

Also, the second integral (also inside square root) is equal to $\mathbb{E}\left[\|\boldsymbol{\Delta x}\|^4\right] = \alpha e^2$. So, we have:

$$\mathbb{E}\left[\max_{T'<i\leq T}\|\boldsymbol{\Delta x_i}\|^2\right] \leq T_{\mathcal{V}}\frac{1}{\sqrt{2T_{\mathcal{V}}-1}}\sqrt{\mathbb{E}\left[\|\boldsymbol{\Delta x}\|^4\right]}$$
$$= \frac{T_{\mathcal{V}}\sqrt{\alpha}e}{\sqrt{2T_{\mathcal{V}}-1}}.$$

Using the above estimate, we have:

$$\Gamma_\infty \gtrsim m_2(T_{\mathcal{V}}e)/\Big(\frac{T_{\mathcal{V}}\sqrt{\alpha}e}{\sqrt{2T_{\mathcal{V}}-1}}\Big) \sim \mathcal{O}(m_2\sqrt{T_{\mathcal{V}}}).$$

Therefore, $\min(c_2\delta\Gamma_2, \Gamma_\infty)) \geq \mathcal{O}(\delta m_2\sqrt{T_{\mathcal{V}}})$. So with probability greater than $\Big\{1 - \mathcal{O}(\exp(-\widetilde{c}\delta^2m_2\sqrt{T_{\mathcal{V}}}))\Big\}$, the following holds:

$$\mathbb{E}_{val}\left[\|\boldsymbol{\Delta x}\|^2\right] < \frac{\mathbb{E}_{val}\left[\|\widetilde{\mathbf{\Phi}}^{\mathbf{(2)}}\boldsymbol{\Delta x}\|^2\right]}{m_2(1-\delta)}$$

or

$$\mathbb{E}_{val}\left[\|(\boldsymbol{x}-\boldsymbol{x_e})\|^2\right] \leq \frac{\mathbb{E}_{val}\left[\|\widetilde{\mathbf{\Phi}}^{\mathbf{(2)}}(\boldsymbol{x}-\boldsymbol{x_e})\|^2\right]}{m_2(1-\delta)}$$
$$\leq \frac{\mathbb{E}_{val}\left[\|\mathbf{\Phi^{(2)}}(\boldsymbol{x}-\boldsymbol{x_e})\|^2\right]}{m_2(1-\delta)}. \quad (23)$$

This finishes the proof of **Lemma 3**.

Finally, combining **Lemma 1**, **Lemma 2** and **Lemma 3**, we get:

$$\mathbb{E}_{val}\left[\|(\boldsymbol{x}-\boldsymbol{x_e})\|^2\right] \leq \mathcal{O}\left(\frac{I\epsilon}{1-\delta}\right). \tag{24}$$

This completes the proof of **Theorem 1**.

Next, we first state and then prove **Corollary 1**.

Corollary 1. *Suppose on running the NLBCS Algorithm under the same settings as described in **Theorem 1**, (8) in the main paper holds with ε given by (9) in the main paper, then the following bound holds with probability $p(\delta, T-T^{'})$:*

$$\mathbb{E}_{val}\left[\frac{\|\boldsymbol{x}-\boldsymbol{x_e}\|}{\|\boldsymbol{x}\|}\right] \leq \mathcal{O}\left(\sqrt{\frac{\zeta n}{I\sqrt{m_2}(1-\delta)}}\right).$$

Proof: We have:

$$\mathbb{E}_{val}\left[\frac{\|\boldsymbol{x}-\boldsymbol{x_e}\|}{\|\boldsymbol{x}\|}\right] \leq \sqrt{\mathbb{E}_{val}\left[\frac{\|\boldsymbol{x}-\boldsymbol{x_e}\|^2}{\|\boldsymbol{x}\|^2}\right]}$$

$$\leq \sqrt{\mathbb{E}_{val}\left[\frac{n\|\boldsymbol{x}-\boldsymbol{x_e}\|^2}{\kappa I^2}\right]} = \sqrt{\frac{n}{\kappa I^2}\,\mathbb{E}_{val}\left[\|\boldsymbol{x}-\boldsymbol{x_e}\|^2\right]}.$$

The first step follows from the fact that $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$ for any random variable X. The second step follows from the fact that $n\|\boldsymbol{x}\|^2 \geq \|\boldsymbol{x}\|_1^2 = \kappa I^2$ where $0<\kappa<1$ (since I is the maximum signal intensity), which is obtained using the Cauchy-Schwarz inequality. Now putting $\epsilon = 2\zeta/\sqrt{m_2}$ from (9) of the main paper into **Theorem 1** and then substituting it in the above inequality, we get:

$$\mathbb{E}_{val}\left[\frac{\|\boldsymbol{x}-\boldsymbol{x_e}\|}{\|\boldsymbol{x}\|}\right] \leq \mathcal{O}\left(\sqrt{\frac{n}{\kappa I^2}\frac{I}{(1-\delta)}\frac{2\zeta}{\sqrt{m_2}}}\right)$$

$$= \mathcal{O}\left(\sqrt{\frac{\zeta n}{I\sqrt{m_2}(1-\delta)}}\right).$$

This finishes the proof of **Corollary 1**.

III. CHOOSING ε

Recall that the condition to terminate our algorithm, i.e. (8) of the main paper, is:

$$m_2/4 \leq \widehat{E_v} \leq m_2(1+\varepsilon)/4.$$

We also provide a value of ε in (9) of the main paper which is:

$$\varepsilon = 2\zeta/\sqrt{m_2}.$$

In this section, we show how to obtain the above value of ε. We choose ε such that:

$$\mathbb{P}\left(\widehat{E_v} > \frac{m_2(1+\varepsilon)}{4}\right) < p.$$

or

$$\mathbb{P}\left(\widehat{E_v^S} > \frac{m_2 T_{\mathcal{V}}(1+\varepsilon)}{4}\right) < p \text{ where} \tag{25}$$

$$\widehat{E_v^S} = \sum_{i=T'+1}^{T}\left\|\sqrt{\boldsymbol{y_i^{(2)}} + c\boldsymbol{1}_{m_2}} - \sqrt{\boldsymbol{\Phi_i^{(2)}}\boldsymbol{A_e}(\boldsymbol{\theta_i})_{\boldsymbol{e}} + c\boldsymbol{1}_{m_2}}\right\|^2$$

$$= \sum_{i=T'+1}^{T}\sum_{j=1}^{m_2}\left(\sqrt{\boldsymbol{y_i^{(2)}} + c\boldsymbol{1}_{m_2}} - \sqrt{\boldsymbol{\Phi_i^{(2)}}\boldsymbol{A_e}(\boldsymbol{\theta_i})_{\boldsymbol{e}} + c\boldsymbol{1}_{m_2}}\right)_j^2.$$

In the above equation p is a probability value of our choice and we would ideally want to choose a small value of p. We set $p = \exp\left(-\zeta^2 T_{\mathcal{V}}\right)$ where ζ is a non-zero constant of our choice.

Now notice that $\widehat{E_v^S}$ is just the sum of the squares of $m_2 T_{\mathcal{V}}$ i.i.d Gaussian random variables with mean 0 and variance 1/4. They are identically distributed because of the Anscombe transform. The independence holds because every component of the measurement vector (i.e. $(\boldsymbol{y_i})_j$) made for a particular signal (i.e. $\boldsymbol{x_i}$) is corrupted with noise independent of the other components and the same logic also applies to two different signals (i.e. $\boldsymbol{x_i}$ and $\boldsymbol{x_{i'}}$). Hence, $4\widehat{E_v^S}$ is a chi-squared random variable with $m_2 T_{\mathcal{V}}$ degrees of freedom.

Let $\widetilde{E_v^S} = 4\widehat{E_v^S}$ and $M_{val} = m_2 T_{\mathcal{V}}$. Thus (25) reduces to:

$$\mathbb{P}\left(\widetilde{E_v^S} > M_{val}(1+\varepsilon)\right) < p = \exp\left(-\zeta^2 T_{\mathcal{V}}\right). \tag{26}$$

The bound given in [2] (Lemma 1 page 1325) states that if Z is a chi-squared random variable with u degrees of freedom then:

$$\mathbb{P}(Z > u + 2\sqrt{uz} + 2z) < \exp(-z).$$

For our case, $u = M_{val} = m_2 T_{\mathcal{V}}$. Thus, we have:

$$\frac{u}{2} + \left(\sqrt{\frac{u}{2}} + \sqrt{2z}\right)^2 = u(1+\varepsilon).$$

$$z = \frac{u}{2}(1+\varepsilon-\sqrt{1+2\varepsilon}) \approx \frac{u\varepsilon^2}{4} \text{ for } \varepsilon \to 0.$$

The above approximate equality is obtained using the binomial expansion of $\sqrt{1+2\varepsilon}$ upto the second order term and neglecting all other higher order terms for small ε.
But we also want: $p = \exp\left(-\zeta^2 T_{\mathcal{V}}\right) = \exp(-z)$ or $z = \zeta^2 T_{\mathcal{V}}$. This gives us:

$$\frac{u\varepsilon^2}{4} = \zeta^2 T_{\mathcal{V}} \implies \varepsilon = 2\frac{\zeta}{\sqrt{m_2}}. \tag{27}$$

This proves (9) of the main paper and completes the discussion on how to obtain ε.

IV. ADDITIONAL EXPERIMENTS

In this section, we present the results of some experiments which show that learning a data-dependent dictionary using our algorithm is better than just using an off the shelf dictionary (and optimizing only over the sparse codes). We used non-negative matrix factorization (NMF) to obtain a decent off the shelf matrix.

We took nine 512×512 gray scale images. The 9 images along with their names (used in this section) are shown in Fig. 3. Similar to the experiment in the main paper, we divided each image into $64 \times 64 = 4096 = T$ non-overlapping (to maintain independence of the signals) patches of size 8×8 each, followed by reshaping of each patch to form a $64 (= n) \times 1$ vector. Thus, we had T signals of dimension $n \times 1$, per image. The size of the dictionary used in all our experiments was $k = 16$. We present the results for two different cases.

Expt. 1: Here, we considered all the images except Lena (i.e. 8 images). We performed NMF on the $8T$ (since there are T signals per image) signals with size of the basis set to $n \times k$ (recall that $k = 16$). We then used the obtained NMF basis as our off the shelf dictionary (call it $\boldsymbol{A}_{NMF}$) to recover the Lena image given compressive measurements of each of its 8×8 patches, corrupted with Poisson noise (i.e. our measurement and noise model). In one case (call this **Case 1**), we kept the dictionary fixed equal to $\boldsymbol{A}_{NMF}$ and only optimized over the sparse codes by minimizing (1) of this paper (while keeping $\boldsymbol{A}^{(t)} = \boldsymbol{A}_{NMF}\ \forall\ t, i$). In another case (call this **Case 2**), we applied our algorithm using $\boldsymbol{A}_{NMF}$ as the initialization for $\boldsymbol{A}$ (i.e. $\boldsymbol{A}^{(0)} = \boldsymbol{A}_{NMF}$), and evolved both $\boldsymbol{A}$ as well as the sparse codes. In both the cases, we worked with $m = \{16, 24, 32, 40\}$. For $m = 16$ and 24, we chose $m_1 = m - 4$ and $m_2 = 4$, whereas for $m = 32$ and 40, we chose $m_1 = m - 5$ and $m_2 = 5$. We set the value of $\lambda = 0.125 m_1$ for both cases.

Table I shows the $RRMSE$ values for both cases in **Expt. 1**. Observe that the $RRMSE$ values in **Case 1** are significantly higher than those in **Case 2** for $m = 16, 24$. For $m = 32, 40$, the $RRMSE$ values in the two cases are much closer (although they are still lower in **Case 2**).

RRMSE-	m=16	m=24	m=32	m=40
Case 1	0.2383	0.1678	0.1368	0.1201
Case 2	0.1295	0.1156	0.1085	0.1015

Table I. $RRMSE$ values in **Expt. 1**.

Fig. 1 shows the reconstructed Lena images in both the cases with $m = 16$. Visually, the reconstructed image in **Case 1** is quite noisy. The reconstructed image in **Case 2** is much cleaner although it has artifacts from other images.

(a) **Case 1** (b) **Case 2**

Fig. 1. Recovered Lena images in **Expt. 1** with $m = 16$.

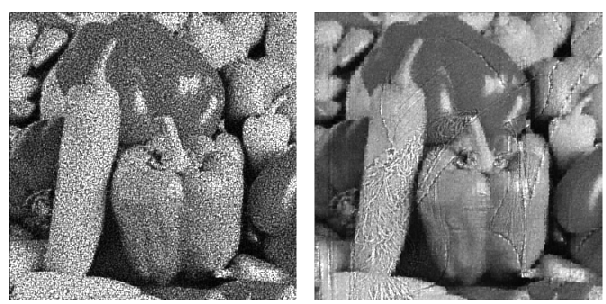
Expt. 2: Here, everything was the same as in **Expt. 1** except that we replaced Lena in **Expt. 1** with Peppers here.

Table II shows the $RRMSE$ values for both cases in **Expt. 2**. Here, the $RRMSE$ values in **Case 1** are much higher than those in **Case 2** for $m = 16, 24, 32$. For $m = 40$, the $RRMSE$ values in the two cases are much closer (although they are still lower in **Case 2**).

RRMSE-	m=16	m=24	m=32	m=40
Case 1	0.3320	0.2240	0.1822	0.1418
Case 2	0.1714	0.1515	0.1414	0.1325

Table II. $RRMSE$ values in **Expt. 2**.

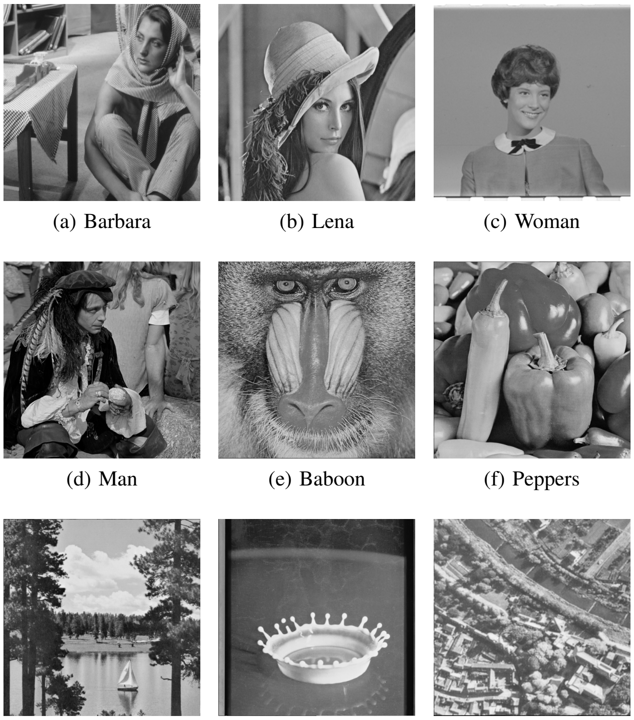
Fig. 2 shows the reconstructed Peppers images in both the cases with $m = 16$. Even here, the reconstructed image in **Case 1** is quite noisy while **Case 2**'s is much cleaner although with artifacts from other images.



(a) **Case 1** (b) **Case 2**

Fig. 2. Recovered Peppers images in **Expt. 2** with $m = 16$.

From the two experiments, we conclude that our algorithm initialized with the separately learnt NMF basis performs better than just using the same NMF basis as a fixed dictionary, especially when the number of measurements is small.



(a) Barbara (b) Lena (c) Woman

(d) Man (e) Baboon (f) Peppers

(g) Sailboat (h) Splash (i) Aerial

Fig. 3. The 9 images used in our experiments.

V. REFERENCES

[1] M. Aghagolzadeh and H. Radha, “Joint estimation of dictionary and image from compressive samples,” *IEEE Trans. Computational Imaging*, 2017.

[2] B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection,” *Annals of Statistics*, pp. 1302–1338, 2000.