A VARIATIONAL BAYESIAN INFERENCE-INSPIRED UNROLLED DEEP NETWORK FOR COMPRESSED SENSING

Menghong Cai[1], Jun Fang[1], Huiping Duan[1], Xiaoyu Li[1],Hongbin Li[2]

[1]University of Electronic Science and Technology of China
[2]Stevens Institute of Technology
JunFang@uestc.edu.cn Hongbin.Li@stevens.edu

ABSTRACT

Deep learning (DL)-based algorithms were recently developed for compressed sensing. Specifically, a class of deep unfolding-based methods have gained much attention due to their ability of integrating the power of learning with the algorithmic structure. Nevertheless, most of these deep unfolding-based methods still require a large number of learnable parameters. In this paper, we develop an unrolled deep network for compressed sensing within an inverse-free variational Bayesian framework. Compared with existing networks, the proposed unrolled deep network has substantially fewer learnable parameters. The proposed network can therefore achieve better recovery performance with fewer training samples. This makes it useful for scenarios where training samples are costly to be acquired.

Index Terms— Compressed sensing, variational Bayesian inference, unrolled deep networks.

1. INTRODUCTION

Compressed sensing is a new paradigm for signal sampling and reconstruction and its core problem is to recover a sparse signal from a significantly smaller number of linear measurements. Over the past years, a lot of efforts has been made on sparse signal recovery. A plethora of algorithms such as the orthogonal matching pursuit (OMP) [1], the convex relaxation-based basis pursuit (BP) method [2], the iterative reweighted algorithms [3], and Bayesian inference-based methods [4, 5] were proposed for sparse signal recovery.

More recently, due to the enormous success of deep learning (DL), DL-based algorithms were proposed for compressed sensing. In particular, there is a great amount of interest in a class of deep unfolding-based algorithms that unfolds an iterative algorithm as a succession of neural network layers. Specifically, in [6], a deep unfolding-based algorithm, named LISTA, was proposed by unfolding the well-known compressed sensing algorithm ISTA [7]. In addition, some other deep unfolding-based methods such as the learned AMP (LAMP) and LVAMP [8] were developed by respectively unfolding the classical AMP [9] and VAMP [10]. The LAMP implements an Onsager correction in residual estimation. As a result, the true signal can be estimated with a higher accuracy and the LAMP converges faster than LISTA. The LVAMP includes a singular-value decomposition (SVD) operation to circumvent a problem that arises with non-i.i.d.-Gaussian matrices in LAMP. All of the above algorithms, however, have a large number of trainable parameters that require lot of training data.

In this paper, an unrolled deep network is proposed for compressed sensing within a variational Bayesian framework. A widely used compressed sensing technique, sparse Bayesian learning (SBL), was introduced in [4] by using the variational Bayesian inference. Nevertheless, SBL needs to calculate the variational posterior distribution of a hidden variable, which requires a large-size matrix inversion at each iteration. In this case, directly unfolding the iterative process of SBL incurs a high computational complexity and meanwhile faces a problem of choosing appropriate learnable parameters to train the network. To overcome the above difficulties, we propose an inverse-free variational Bayesian-based unrolled deep learning architecture, where a relaxed evidence lower bound [11] is introduced to bypass the matrix inversion. Note that such a framework has been utilized for devising unrolled deep networks for MIMO detection [12]. We will show that the proposed unrolled deep network has only handful of learnable parameters, which enables the proposed algorithm to be efficiently trained with fewer training samples.

2. REVIEW OF INVERSE-FREE SBL

We consider the sparse signal recovery problem of estimating a sparse signal $\boldsymbol{y} \in \mathbb{R}^N$ from the underdetermined system of linear measurements

$$\boldsymbol{z} = \boldsymbol{D}\boldsymbol{y} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ ($M \ll N$) stands for the so-called measurement matrix, and $\boldsymbol{n} \in \mathbb{R}^M$ is the observation noise

This work is supported in part by Sichuan Science and Technology Program under Grant 2023ZYD0146.

following a Gaussian distribution $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \gamma^{-1}\boldsymbol{I})$. Our goal is to develop a deep unfolding-based method to solve the above compressed sensing problem. To this end, we first introduce an inverse-free variational Bayesian framework that is aimable for algorithmic unfolding.

As analyzed in [4], placing a two-layer hierarchical prior distribution on each element of $\boldsymbol{y}$ results in a student-t distribution, which has a sharp peak around zero and thus has the potential to promote a sparse solution. The two-layer hierarchical prior model placed on $\boldsymbol{y}$ can be described as

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) = \prod_{n=1}^{N} p\left(y_n|\alpha_n\right) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|0, \alpha_n^{-1}\right) \tag{2}$$

$$p(\boldsymbol{\alpha}) = \prod_{n=1}^{N} \mathrm{Gamma}\left(\alpha_n|a, b\right) \tag{3}$$

where $\boldsymbol{\alpha} \triangleq [\alpha_1, \ldots, \alpha_N]$ obeys the Gamma distribution, the non-negative hyperparameters $\{\alpha_n\}$ are used to control the sparsity of the associated elements $\{y_n\}$, a and b are constants that are set to small values, e.g., 10^{-10}. The precision parameter γ is assigned a Gamma hyperprior to effectively estimate the noise variance, i.e. $p(\gamma) = \mathrm{Gamma}\left(\gamma|c, d\right)$, in which c and d are set to 10^{-10} as well.

Given the above probabilistic model, the core task is to determine the posterior distribution of the hidden variables, which include $\boldsymbol{\vartheta} \triangleq \{\boldsymbol{y}, \boldsymbol{\alpha}, \gamma\}$. To estimate the posterior distribution, the variational Bayesian inference employs the mean field theory to approximate the posterior as $q\left(\boldsymbol{\vartheta}\right) = q_y\left(\boldsymbol{y}\right) q_\alpha\left(\boldsymbol{\alpha}\right) q_\gamma\left(\gamma\right)$, as well as the following evidence lower bound (ELBO) [13]

$$B(q) = \int q(\boldsymbol{\vartheta}) \ln \frac{p(\boldsymbol{z}, \boldsymbol{\vartheta})}{q(\boldsymbol{\vartheta})} d\boldsymbol{\vartheta} \tag{4}$$

where $p(\boldsymbol{z}, \boldsymbol{\vartheta}) = p(\boldsymbol{z}|\boldsymbol{y}, \gamma)p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\gamma)$. By maximizing the ELBO, the divergence between the variational posterior $q\left(\boldsymbol{\vartheta}\right)$ and the posterior distribution $p(\boldsymbol{\vartheta}|\boldsymbol{z})$ can be minimized, which means that $q\left(\boldsymbol{\vartheta}\right)$ can be used as an estimate of $p(\boldsymbol{\vartheta}|\boldsymbol{z})$. The ELBO is maximized via updating the posterior of each hidden variable in an alternate manner. Nevertheless, directly maximizing the ELBO has to invert an $N \times N$ matrix when updating $q_y(\boldsymbol{y})$, which incurs a high computational complexity.

In [11], a novel method was proposed to avoid matrix inversion. The basic idea is to utilize a lower bound of $p(\boldsymbol{z}|\boldsymbol{y}, \gamma)$. By substituting the lower bound into $B(q)$, we obtain a relaxed ELBO that simplifies the solution. Specifically, the relaxed ELBO is obtained by resorting to the following inequality [14]

$$f(\boldsymbol{m}) \leq f(\boldsymbol{n}) + (\boldsymbol{m} - \boldsymbol{n})^T \nabla f(\boldsymbol{n}) + (\boldsymbol{m} - \boldsymbol{n})^T \boldsymbol{T} (\boldsymbol{m} - \boldsymbol{n}) \tag{5}$$

in which $f : \mathbb{R}^n \to \mathbb{R}$ denotes a continuously differentiable function and $\boldsymbol{T} \succcurlyeq \frac{\nabla^2 f(\boldsymbol{x})}{2}, \forall \boldsymbol{x}$. By invoking (5), $p(\boldsymbol{z}|\boldsymbol{y}, \gamma)$ can be lower bounded as follows

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{y}, \gamma) &= \mathit{\Pi} e^{\left\{-\frac{\gamma}{2}\|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{y}\|_2^2\right\}} \geq \mathit{\Pi} e^{\left\{-\frac{\gamma}{2}\mathit{\Phi}(\boldsymbol{y}, \boldsymbol{v})\right\}} \\ &\triangleq \mathit{\Gamma}(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{v}, \gamma) \end{aligned} \tag{6}$$

where $\mathit{\Pi} = \left(\frac{\gamma^{1/2}}{\sqrt{2\pi}}\right)^N$ and, with the definition of $\psi\left(\boldsymbol{y}\right) = \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{y}\|_2^2$, we have

$$\mathit{\Phi}(\boldsymbol{y}, \boldsymbol{v}) \triangleq \psi\left(\boldsymbol{v}\right) + (\boldsymbol{y} - \boldsymbol{v})^T \nabla\psi\left(\boldsymbol{v}\right) + (\boldsymbol{y} - \boldsymbol{v})^T \boldsymbol{T} \left(\boldsymbol{y} - \boldsymbol{v}\right) \tag{7}$$

$$\nabla\psi\left(\boldsymbol{v}\right) = 2\boldsymbol{D}^T \left(\boldsymbol{D}\boldsymbol{v} - \boldsymbol{z}\right) \tag{8}$$

The validity of (6) requires the condition $\boldsymbol{T} \succcurlyeq \frac{\nabla^2 f(\boldsymbol{x})}{2} = \boldsymbol{D}^T\boldsymbol{D}$. It is obvious that matrix inversion can be circumvented if $\boldsymbol{T}$ is set to be a diagonal matrix as

$$\boldsymbol{T} = (\lambda_{\max}(\boldsymbol{D}^T\boldsymbol{D}) + \epsilon)\boldsymbol{I} \tag{9}$$

where ϵ can be a positive constant set to 10^{-10}.

According to (6), $p\left(\boldsymbol{z}, \boldsymbol{\vartheta}\right)$'s lower bound $Q(\boldsymbol{z}, \boldsymbol{\vartheta}, \boldsymbol{v})$ can be expressed as $Q(\boldsymbol{z}, \boldsymbol{\vartheta}, \boldsymbol{v}) = \mathit{\Gamma}(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{v}, \gamma)p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\gamma)$. Consequently, a relaxed ELBO can be obtained as

$$\hat{B}(q, \boldsymbol{v}) = \int q(\boldsymbol{\vartheta}) \ln \frac{Q(\boldsymbol{z}, \boldsymbol{\vartheta}, \boldsymbol{v})}{q(\boldsymbol{\vartheta})} d\boldsymbol{\vartheta} \leqslant B(q) \tag{10}$$

A variational expectation-maximization (EM) can be employed to optimize the relaxed ELBO with respect to the posterior $q(\boldsymbol{\vartheta})$ and the parameter $\boldsymbol{v}$. First, each hidden variable's posterior approximation is calculated by performing expectation in the E-step. Second, the M-step finds the value of $\boldsymbol{v}$ which maximizes $\hat{B}(q, \boldsymbol{v})$ given $q(\boldsymbol{\vartheta})$. Here we omit the derivation and only give the update expressions of each hidden variable and the parameter.

The variational posterior $q_y\left(\boldsymbol{y}\right)$ is a Gaussian distribution with the following covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$

$$\boldsymbol{\Sigma} = \left(\langle\gamma\rangle\boldsymbol{T} + \boldsymbol{\Xi}\right)^{-1} \tag{11}$$

$$\begin{aligned} \boldsymbol{\mu} &= \langle\gamma\rangle \left(\langle\gamma\rangle\boldsymbol{T} + \boldsymbol{\Xi}\right)^{-1} \left(\boldsymbol{D}^T \left(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{v}\right) + \boldsymbol{T}\boldsymbol{v}\right) \\ &= \langle\gamma\rangle\boldsymbol{\Sigma} \left(\boldsymbol{D}^T \left(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{v}\right) + \boldsymbol{T}\boldsymbol{v}\right) \end{aligned} \tag{12}$$

in which $\boldsymbol{\Xi} \triangleq \mathrm{diag}\left(\langle\alpha_1\rangle, ..., \langle\alpha_N\rangle\right)$, $\langle\alpha_n\rangle$ represents the expectation of α_n with respect to $q_\alpha\left(\boldsymbol{\alpha}\right)$.

The variational posterior of the hidden variable γ and each element of $\boldsymbol{\alpha}$ obey Gamma distributions

$$q_\gamma(\gamma) = \mathrm{Gamma}(\gamma; \hat{c}, \hat{d}) \tag{13}$$

$$q_\alpha(\alpha_n) = \mathrm{Gamma}(\alpha_n; \hat{a}, \hat{b}_n) \tag{14}$$

The posterior distribution of $\boldsymbol{\alpha}$ can be expressed as $q_\alpha(\boldsymbol{\alpha}) = \prod_{n=1}^{N} q_\alpha(\alpha_n)$. In addition, the parameters $\{\hat{a}, \hat{b}_n, \hat{c}, \hat{d}\}$ in the Gamma distributions are respectively updated as

$$\begin{cases} \hat{a} = a + \frac{1}{2} \quad \hat{b}_n = b + \frac{1}{2}\ \langle y_n^2\rangle \\ \hat{c} = c + \frac{M}{2} \quad \hat{d} = d + \frac{1}{2}\langle\mathit{\Phi}(\boldsymbol{y}, \boldsymbol{v})\rangle \end{cases} \tag{15}$$

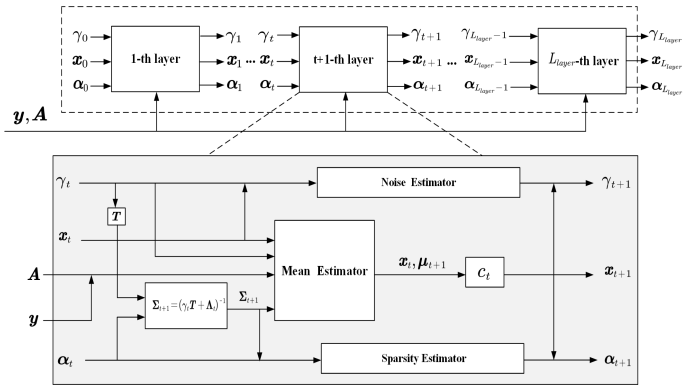

Fig. 1. The neural network constructed by unfolding the iterative IF-SBL.

After executing the M-step, the parameter $\boldsymbol{z}$ is updated as

$$\boldsymbol{v} = \boldsymbol{\mu} \tag{16}$$

Here we summarize the above update formulas into an iterative algorithm called inverse-free SBL (IF-SBL) with two steps at each iteration, namely, a posterior covariance estimation (CE) step and a posterior mean estimation (ME) step:

$$\text{CE} \quad : \boldsymbol{\Sigma}_t = \left(\langle\gamma_{t-1}\rangle \boldsymbol{T} + \boldsymbol{\Xi}_{t-1}\right)^{-1} \tag{17}$$

$$\text{ME} \quad : \boldsymbol{\mu}_t = \langle\gamma_{t-1}\rangle \boldsymbol{\Sigma}_t(-\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\mu}_{t-1} + \boldsymbol{D}^T\boldsymbol{z} + \boldsymbol{T}\boldsymbol{\mu}_{t-1}) \tag{18}$$

where γ_{t-1} and $\boldsymbol{\Xi}_{t-1}$ denote the previous estimate of γ and $\boldsymbol{\Xi}$, respectively, and the current estimate γ_t and $\boldsymbol{\Xi}_t = \text{diag}(\langle\alpha_1^t\rangle, \ldots, \langle\alpha_N^t\rangle)$ can be calculated as

$$\langle\alpha_n^t\rangle = \frac{a + \frac{1}{2}}{b + \frac{1}{2}\left[\left(\mu_n^t\right)^2 + \Sigma_{n,n}^t\right]} \tag{19}$$

$$\langle\gamma_t\rangle = \frac{c + \frac{M}{2}}{d + \frac{1}{2}\Phi\left(\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t-1}\right)} \tag{20}$$

where μ_n^t stands for the nth entry of the $\boldsymbol{\mu}_t$, $\Sigma_{n,n}^t$ represents the nth diagonal entry of $\boldsymbol{\Sigma}_t$, and

$$\begin{aligned}&\Phi\left(\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t-1}\right)\\ &= \left\|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}_{t-1}\right\|_2^2 + 2\left(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\right)^T\boldsymbol{D}^T\left(\boldsymbol{D}\boldsymbol{\mu}_{t-1} - \boldsymbol{z}\right)\\ &+ \text{Tr}\left(\boldsymbol{T}\left(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\right)\left(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\right)^T + \boldsymbol{T}\boldsymbol{\Sigma}_t\right)\end{aligned} \tag{21}$$

3. PROPOSED DEEP UNFOLDING-BASED METHOD

In this section, we introduce some trainable parameters to construct a deep neural network which is illustrated in Fig. 1. The deep network is developed by unfolding the iterative IF-SBL into multiple consecutive network layers. We use $\boldsymbol{\Omega} \triangleq \left\{\boldsymbol{T}, \{l_t\}_{t=1}^{L_{\text{layer}}}\right\}$ to denote the trainable parameters in the network and the meaning of the trainable parameters will be discussed later. With these learnable parameters, the $(t+1)$th layer of our proposed network performs the following updates given the input $\{\boldsymbol{z}, \boldsymbol{D}, \boldsymbol{y}_t, \gamma_t, \{\alpha_n^t\}\}$:

$$\text{CE} : \boldsymbol{\Sigma}_{t+1} = \left(\gamma_t\boldsymbol{T} + \boldsymbol{\Xi}_t\right)^{-1} \tag{22}$$

$$\text{ME} : \boldsymbol{\mu}_{t+1} = \gamma_t\boldsymbol{\Sigma}_{t+1}(-\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{y}_t + \boldsymbol{D}^T\boldsymbol{z} + \boldsymbol{T}\boldsymbol{y}_t) \tag{23}$$

$$\text{Update of } \boldsymbol{y}_{t+1} : \boldsymbol{y}_{t+1} = l_t\boldsymbol{\mu}_{t+1} + (1 - l_t)\,\boldsymbol{y}_t \tag{24}$$

where $\boldsymbol{\Xi}_t \triangleq \text{diag}\left(\alpha_t^1, \ldots, \alpha_t^N\right)$. Here with a slight abuse of notations, we use $\boldsymbol{y}_t$ to denote the current estimate of the sparse signal $\boldsymbol{y}$, γ_t is used to represent $\langle\gamma_t\rangle$, and α_n^t is used to represent $\langle\alpha_n^t\rangle$. According to (19) and (20), α_n^{t+1} and γ_{t+1} can be updated as

$$\alpha_n^{t+1} = \frac{a + \frac{1}{2}}{b + \frac{1}{2}\left[\left(y_n^{t+1}\right)^2 + \Sigma_{n,n}^{t+1}\right]} \tag{25}$$

$$\gamma_{t+1} = \frac{c + \frac{M}{2}}{d + \frac{1}{2}\Phi\left(\boldsymbol{y}_{t+1}, \boldsymbol{y}_t\right)} \tag{26}$$

in which

$$\begin{aligned}\Phi\left(\boldsymbol{y}_{t+1}, \boldsymbol{y}_t\right) &= \Phi\left(\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t-1}\right)|_{\boldsymbol{\mu}_{t-1}=\boldsymbol{y}_t}^{\boldsymbol{\mu}_t=\boldsymbol{y}_{t+1}}\\ &= \left\|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{y}_t\right\|_2^2 + 2\left(\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\right)^T\boldsymbol{D}^T\left(\boldsymbol{D}\boldsymbol{y}_t - \boldsymbol{z}\right)\\ &+ \text{Tr}\left(\boldsymbol{T}\left(\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\right)\left(\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\right)^T + \boldsymbol{T}\boldsymbol{\Sigma}_{t+1}\right)\end{aligned} \tag{27}$$

We observe that each layer's update formulas are similar to the iterative formulas of IF-SBL. Nevertheless, there exist two subtle differences which help improve the recovery performance of the proposed deep-unfolding-based network. First, the $\boldsymbol{T}$ is a key factor that affects the performance of the algorithm. The choice of $\boldsymbol{T}$ specified in (9) may not be an optimal choice. Therefore it is desirable to make $\boldsymbol{T}$ a trainable diagonal matrix so that a more suitable choice of $\boldsymbol{T}$ can be found. Second, instead of directly taking $\boldsymbol{\mu}$ as the current estimate of $\boldsymbol{y}$, we use a learnable parameter l_t to obtain the update of $\boldsymbol{y}$ based on the adaptive damping strategy. It has been verified that such a damping scheme helps improve the stability of the developed neural network [15].

We further elaborate the proposed IFSBL-NET's advantage by comparing it with the state-of-the-art deep unfolding-based compressed sensing method called LAMP [8]. The LAMP performs the following updates at each layer:

$$\hat{\boldsymbol{y}}_{t+1} = \beta_t\boldsymbol{\eta}_{\text{st}}\left(\hat{\boldsymbol{y}}_t + \boldsymbol{B}\boldsymbol{r}_t; \frac{\alpha_t}{\sqrt{M}}\left\|\boldsymbol{r}_t\right\|_2\right) \tag{28}$$

$$\boldsymbol{r}_{t+1} = \boldsymbol{z} - \boldsymbol{D}\hat{\boldsymbol{y}}_{t+1} + \frac{\beta_t}{M}\left\|\hat{\boldsymbol{y}}_{t+1}\right\|_0\boldsymbol{r}_t \tag{29}$$

where both the matrix $\boldsymbol{B} \in \mathbb{R}^{N\times M}$ and the parameters $\{\alpha_t, \beta_t\}$ are trainable parameters, and $\boldsymbol{\eta}_{\text{st}}(\cdot;\lambda) : \mathbb{R}^N \to \mathbb{R}^N$

represents the soft thresholding shrinkage function. We see that the LAMP has a total number of $N \times M + 2L_{\text{layer}}$ trainable parameters. For comparison, the number of learnable parameters in IFSBL-NET is $N + L_{\text{layer}}$, which is far fewer than that of the LAMP. This enables the proposed IFSBL-NET to be more efficiently trained when the number of training samples is limited.

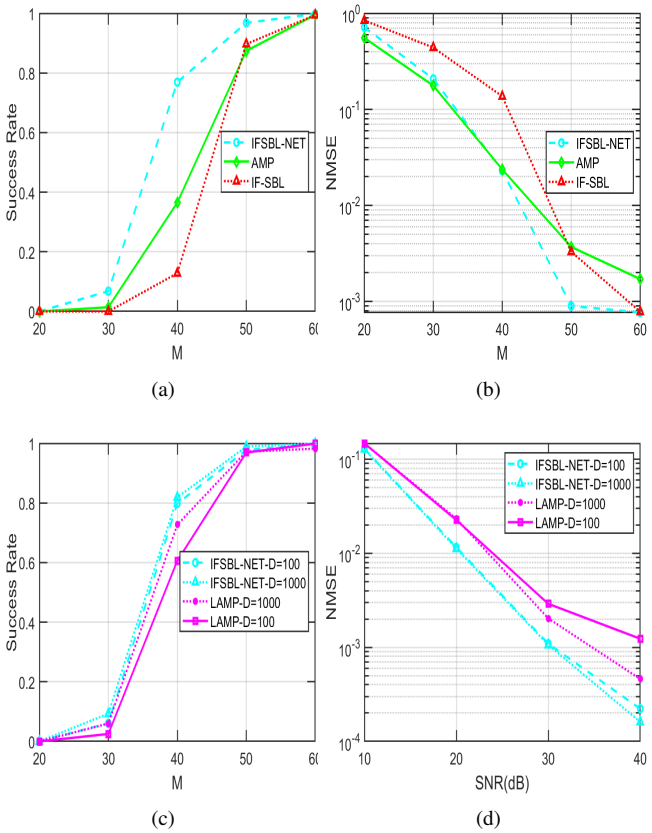

Fig. 2. (a). Success rates versus M in noiseless case; (b). NMSEs versus M with SNR $=$ 30dB; (c). Success rates versus M with $D = 100$ and $D = 1000$; (d). NMSEs versus SNR with $D = 100$ and $D = 1000$.

4. SIMULATION RESULTS

The simulation results are provided in this section to show the sparse signal recovery performance of the proposed unrolled network IFSBL-NET. During training, we use the stochastic gradient descent to optimize the learnable parameters with the help of auto-differentiable machine-learning framework Pytorch, in which the following squared error is used as the loss function

$$f_{\text{loss}} = \frac{1}{L_{\text{layer}}} \sum_{t=1}^{L_{\text{layer}}} \left\| \boldsymbol{y}_t - \boldsymbol{y}^{\text{true}} \right\|_2^2 \qquad (30)$$

We compare our proposed IFSBL-NET with the classical compressed sensing methods AMP [9] and IF-SBL [11]. Also, the deep unfolding-based method LAMP is included for comparison. We generate the training and test samples according to (1), where the measurement matrix is an i.i.d. Gaussian matrix and the sparse signal is randomly generated with the locations of the nonzero entries uniformly distributed and values of the nonzero entries drawn from a normal distribution. For IFSBL-NET and LAMP, the number of samples used for network training is set to D. In our experiments, we consider two choices of D, i.e. $D = 100$ and $D = 1000$, in order to show the behavior of deep unfolding-based methods under different numbers of training samples. We define the signal-to-noise ratio (SNR) as $E\left[\left\|\boldsymbol{D}\boldsymbol{y}\right\|_2^2\right] / E\left[\left\|\boldsymbol{n}\right\|_2^2\right]$.

Fig. 2(a) shows recovery success rates of different algorithms as the number of measurements M varies, where we consider a noiseless case and set $N = 100$ and $K = 10$. We observe that our proposed IFSBL-NET exhibits better recovery performance than the iterative methods AMP and IF-SBL. For the noisy scenario, the normalized mean square errors (NMSEs) is used to evaluate the recovery performance. Fig. 2(b) plots NMSEs versus M, in which we set SNR $=$ 30dB while keep other experimental settings unchanged. For a moderate M/N ratio, it can be observed that IFSBL-NET outperforms the other two approaches. These results indicate that the deep unfolding-based method is superior to the traditional iterative methods.

In Fig. 2(c), we plot the success rates of the proposed IFSBL-NET and LAMP versus M, where the number of training samples D is set to $D = 100$ and $D = 1000$, respectively. We see that the LAMP's performance suffers visible degradation as the number of training samples is decreased from 1000 to 100. Meanwhile, the IFSBL-NET achieves similar performance for different choices of D. The reason is that the IFSBL-NET has much fewer learnable parameters and therefore it can be efficiently trained even in case of insufficient training samples. In Fig. 2(d) we let $M = 50$, $N = 100$ and $K = 10$ to show the NMSEs of IFSBL-NET and the LAMP, where the SNR varies from 10dB to 40dB. We observe that the proposed IFSBL-NET shows better performance over the LAMP across different noise levels.

5. CONCLUSIONS

In this paper, we developed an unrolled deep network based on the inverse-free variational Bayesian framework for compressed sensing. The proposed network, known as IFSBL-NET, has a small number of learnable parameters and can therefore be trained effectively with fewer training samples. Simulation results has shown that the proposed network can outperform both traditional sparse signal recovery methods and the state-of-the-art deep learning based networks such as LAMP.

6. REFERENCES

[1] Joel A. Tropp and Anna C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[2] E.J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[3] David Wipf and Srikantan Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[4] Michael E Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. 3, pp. 211–244, 2001.

[5] Shihao Ji, Ya Xue, and Lawrence Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 1 2008.

[6] Karol Gregor and Yann LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th international conference on international conference on machine learning*, 2010, pp. 399–406.

[7] Ingrid Daubechies, Michel Defrise, and Christine De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[8] Mark Borgerding, Philip Schniter, and Sundeep Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4293–4308, 2017.

[9] David L Donoho, Arian Maleki, and Andrea Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*. Cairo, Egypt, 2010, pp. 1–5.

[10] Sundeep Rangan, Philip Schniter, and Alyson K Fletcher, "Vector approximate message passing," *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[11] Huiping Duan, Linxiao Yang, Jun Fang, and Hongbin Li, "Fast inverse-free sparse Bayesian learning via relaxed evidence lower bound maximization," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 774–778, 2017.

[12] Qian Wan, Jun Fang, Yinsen Huang, Huiping Duan, and Hongbin Li, "A variational Bayesian inference-inspired unrolled deep network for MIMO detection," *IEEE Transactions on Signal Processing*, vol. 70, pp. 423–437, 2022.

[13] Dimitris G. Tzikas, Aristidis C. Likas, and Nikolaos P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, pp. 131–146, Nov. 2008.

[14] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[15] Sundeep Rangan, Philip Schniter, Alyson K. Fletcher, and Subrata Sarkar, "On the convergence of approximate message passing with arbitrary matrices," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5339–5351, 2019.