An Approximate Message Passing Algorithm for Robust Face Recognition

Guangyu Zhou and Wei Dai
Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom

***Abstract*—This paper focuses on algorithmic approaches to solve the robust face recognition problem where the test face image can be corrupted. The standard approach is to formulate the problem as a sparse recovery problem and solve it using ℓ_1-minimization. As an alternative, the approximate message passing (AMP) algorithm had been tested but resulted in pessimistic results. Our contribution is to successfully solve this problem using the AMP framework. Recently developed adaptive damping technique has been adopted to address the issue that AMP normally only works well with Gaussian matrices. Statistical models are designed to capture the nature of the signal more authentically. Expectation maximization (EM) method has been used to learn the unknown hyper-parameters of the statistical model in an online fashion. Simulations demonstrate that our method achieves better recognition performance than the impressive benchmark ℓ_1-minimization, is robust to the initial values of hyper-parameters, and exhibits low computational cost.**

***Index Terms*—Approximate message passing (AMP); compressed sensing; robust face recognition; sparse signal processing**

I. Introduction

Robust face recognition problem is to recognize a test face image that may be corrupted by arbitrary noise [1], [2]. It has been demonstrated that sparse signal processing can solve this problem with impressive performance. Mathematically, a vector $\boldsymbol{x}$ is sparse if only a small fraction of components in $\boldsymbol{x}$ are significant while the majority of the components are zero or close to zero. Sparse recovery problem is to solve the linear inverse problem

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}, \tag{1}$$

where the observation $\boldsymbol{y} \in \mathbb{R}^m$ and the mixing matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are given, the unknown signal $\boldsymbol{x} \in \mathbb{R}^n$ is assumed to be sparse, and the noise $\boldsymbol{w} \in \mathbb{R}^n$ is often white Gaussian. In the robust face recognition setting [1], [2], the vector $\boldsymbol{y}$ is the test face image, the matrix $\boldsymbol{A}$ is derived from training samples, the sparse vector $\boldsymbol{x}$ contains both feature coefficients and the arbitrary noise (assumed to be relatively sparse).

There are many algorithms to solve the sparse recovery problem. They can be divided into two categories: greedy algorithms and ℓ_1-minimization approaches. Greedy algorithms, such as Orthogonal Matching Pursuit (OMP) [3], Subspace Pursuit (SP) [4], are fast while may not work well in some cases. ℓ_1-minimization is an efficient alternative approach to the sparse recovery problem. It has been a hugely successful approach in the past decade. Despite those existing methods, in this paper, we are particularly interested AMP algorithms [5], [6], [7], which are based on loopy belief propagation. Those alternative algorithms deliver both low computational cost and performance guarantees while $\boldsymbol{A}$ has i.i.d. Gaussian entries of zero mean. Unfortunately, the transform matrix $\boldsymbol{A}$ in the face recognition problem cannot be modeled as a Gaussian matrix. Experiments in [1] gave pessimistic results.

Recently, several variants of Generalized AMP (GAMP) algorithm [7] have been proposed to handle non-Gaussian mixing matrices. The ADMM-GAMP algorithm [8] has provable convergence guarantees with arbitrary measurement matrix. It requires to solve an additional least squares problem in each iteration. That makes the ADMM-GAMP algorithm lacks of computational efficiency. SwAMP [9] offers more robust results as it requires a sequential updating procedure rather in parallel. Also, it is not a fast approach compared with other variant in the literature. Jeremy et. al. [10] proposed an adaptive version of Damped-GAMP [11]. This so called AD-GAMP method adaptively updates the damping coefficient, which is determined by the peak-to-average ratio of the squared singular values in $\boldsymbol{A}$. In AD-GAMP, it partially updates the variables tuned by the damping coefficient. In this paper, we study the AD-GAMP algorithm for the robust face recognition problem.

Those GAMP based methods assume known prior information about the signal. For example, the sparse signal is Bernoulli-Gaussian (BG), the measurement noise is additive Gaussian, etc. However, the hyper-parameters in those probability distributions are often not known a prior in practice. Gaussian Mixture-GAMP (GM-GAMP) [12] use expectation maximization (EM) method to estimate the hyper-parameters. It assumes the sparse signal is Gaussian mixture distributed and the noise is AWGN. BG-GAMP [13] assumes a BG distributed sparse signal, which is an special case of GM in GM-GAMP [12]. With EM learning, it has been shown that the GAMP based algorithm normally has a better performance.

The main contribution of this paper is to successfully solve the robust face recognition problem using the AMP framework. AD-GAMP is adapted to address the issue that the mixing matrix $\boldsymbol{A}$ in face recognition is far from the standard Gaussian random matrix. Motivated by the nature of Wright et al.'s framework [2], we model the unknown signal $\boldsymbol{x}$ using a statistical model involving Bernoulli-Gaussian priors. The major difference between our model and the benchmark [2] is that in this work the sparse signal is divided into two segments — one corresponds to the feature coefficients and the other is linked to anomalies to achieve robustness — and different segments have different hyper-parameters. Then the EM method is employed to estimate the unknown hyper-parameters associated with the two segments. With the EM

This work was supported in part by Defence Science and Technology Laboratory (Dstl) under Grant No: DSTLX-1000081291.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

and AD-GAMP coupled together, our method achieves better recognition performance than the ℓ_1-minimization benchmarks in the review paper [1], much better than the pessimistic results of the original AMP [1]. Simulation results also demonstrate that the algorithm is quite robust to the initial values of hyper-parameters, and exhibits low computational cost thanks to the efficiency of the AMP framework.

II. PRELIMINARY RESEARCH

A. Robust Face Recognition

Unlike traditional dictionary learning approaches, the authors of [2] let the training samples be the dictionary in the *sparse representation based classification* (SRC) framework. Each testing image is assumed to be a sparse linear combination of the training set. The mathematical model is as follows, $\boldsymbol{y}_0 = [\boldsymbol{A}_1, \boldsymbol{A}_2, ..., \boldsymbol{A}_s] \cdot [\boldsymbol{x}_{1,0}^T, \boldsymbol{x}_{2,0}^T, ..., \boldsymbol{x}_{s,0}^T]^T = \boldsymbol{A}\boldsymbol{x}_0$, where $\boldsymbol{y} \in \mathbb{R}^m$ is the vectorized test image, the sub-matrix $\boldsymbol{A}_i \in \mathbb{R}^{m \times l}$ and each block $\boldsymbol{x}_{i,0} \in \mathbb{R}^l$ for $i \in [1, ..., s]$. Each column of $\boldsymbol{A}$ is an vectorized training image. Here, the $\boldsymbol{A}_i$ contains l different images all for the ith identity. For simplicity, let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ here. A overview of this framework is shown in Fig. 1. In this case, the columns of transform matrix $\boldsymbol{A}$ are correlated, hence the AMP algorithms do not have convergence guarantees.

In [2], [14], the authors consider two fundamental issues in face recognition problem. Firstly, the role of feature extraction. In other words, one is aiming to project high dimensional testing data into low dimensional feature spaces, which is still informative for sparse representation. Secondly, the obstacle of the occlusion. In practice, a fraction of test images are often corrupted. In [2], [14], the robust SRC model is,

$$\boldsymbol{y} = \boldsymbol{y}_0 + \boldsymbol{e}_0 = [\boldsymbol{A},\ \boldsymbol{I}][\ \begin{matrix}\boldsymbol{x}_0 \\ \boldsymbol{e}_0\end{matrix}\ ], \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ is a down sampled vectorized image with sparse occlusion $\boldsymbol{e}_0$, and $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ is an identity matrix. Eq. (2) can then be simplified as,

$$\boldsymbol{y} = \boldsymbol{\Phi}[\ \begin{matrix}\boldsymbol{x}_0 \\ \boldsymbol{x}_1\end{matrix}\ ] = \boldsymbol{\Phi}\boldsymbol{x}, \tag{3}$$

where $\boldsymbol{\Phi} = [\boldsymbol{A},\ \boldsymbol{I}] \in \mathbb{R}^{m \times (n+m)}$ and the lower part of the sparse coefficient $\boldsymbol{x}_1 = \boldsymbol{e}_0 \in \mathbb{R}^m$. Then, it considers the following ℓ_1 problem,

$$\hat{\boldsymbol{x}} = \arg\min \|\boldsymbol{x}\|_1 \ s.t.\ \boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}.$$

After one get the estimated sparse coefficient $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_0^T, \hat{\boldsymbol{x}}_1^T]^T$, it is easy to assign the tested object to i^* by applying the Sparsity Concentration Index (SCI) in [2],

$$i^* = \arg\max_i = \frac{s \cdot \|\delta_i(\hat{\boldsymbol{x}}_0)\|_1 \,/\, \|\hat{\boldsymbol{x}}_0\|_1 - 1}{s - 1},$$

where $\delta_i(\cdot)$ is a operator that keeps the i-th block of $(\cdot)$.

B. AMP/GAMP for SRC

The AMP is a powerful tool to solve the ℓ_1 problem since it exhibits both low reconstruction error and low computational complexity compared with benchmarks. However, this mechanism only achieves the desired asymptotical optimal performance when the linear transform is standard Gaussian. The GAMP accommodates more general signal models. Here, we consider GAMP for simplicity. GAMP is also flexible to couple with the EM approach to learn the unknown hyper-parameters such as sparsity values, which is more applicable in real time applications.

In robust face recognition problem, the measurement matrix $\boldsymbol{\Phi}$ violates two assumptions which are critical for AMP approaches. The first one is non-zero mean assumption of the measurement matrix in practice. It has been shown in [15] that even with a small positive mean of the i.i.d measurement matrix, the algorithm may diverge. There are three ways of solving this problem. First, remove the mean of the matrix in pre-processing, which is common in image processing fields. Second, modify the update procedure from parallel to sequential [9][15], since the parallel update is more problematic. Third, modify the mathematical model/measurement matrix to remove the mean, as in [10]. In our case, we remove all the means of the training and testing images. The non-zero mean value of the measurement matrix $\boldsymbol{\Phi}$ is dominated by the identity matrix. In this case, the mean value is roughly $\frac{m}{m(n+m)} = \frac{1}{n+m}$. If the number of training images is fixed to n, a larger sampling size m leads smaller mean value. If n or m is large enough, the non-zero mean issue will not affect the convergence of the algorithm. The second assumption of AMP is that the matrix $\boldsymbol{\Phi}$ is i.i.d, which is impractical in this case, e.g., the columns are correlated. A review of fast ℓ_1-minimization algorithms has been studied in [1]. The authors of [1] also added AMP in comparison. In their i.i.d Gaussian experiments, AMP is shown to be the fastest algorithm with near-machine precision. Not surprisingly, AMP fails as it is not capable of handling the general measurement matrix $\boldsymbol{\Phi}$. We shall address the correlation issue of the measurement matrix using the damping approach AD-GAMP [10] and learn the unknown hyper-parameters using the EM embedded BG-GAMP algorithm [13]. However, simply combined algorithm can not achieve better recognition rate than benchmark algorithms. Adapting to the structure of sparse signal in SRC framework, we designed a new dual updating approach based on the combination of those two algorithms. More details of our method are presented in next Section.

III. AN AMP BASED METHOD FOR ROBUST FACE RECOGNITION

A. Dual BG-GAMP

In this paper, we consider the two segments of the sparse signal in robust SRC model (3) have different hyper-parameters. Furthermore, we assume the two segments sparse coefficients $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are both Bernoulli-Gaussian distributed. Hence, terms dual BG-GAMP here. One can then apply the EM embedded BG-GAMP algorithm [13] to learn the unknown hyper-parameters that associated with the sparse signal, e.g., the sparsities, mean values and the variances.

In our approach, we consider the upper part $\boldsymbol{x}_0$ and the lower part $\boldsymbol{x}_1$ of the sparse coefficient $\boldsymbol{x}$ that are ideally not identically distributed. In other words, we consider each of

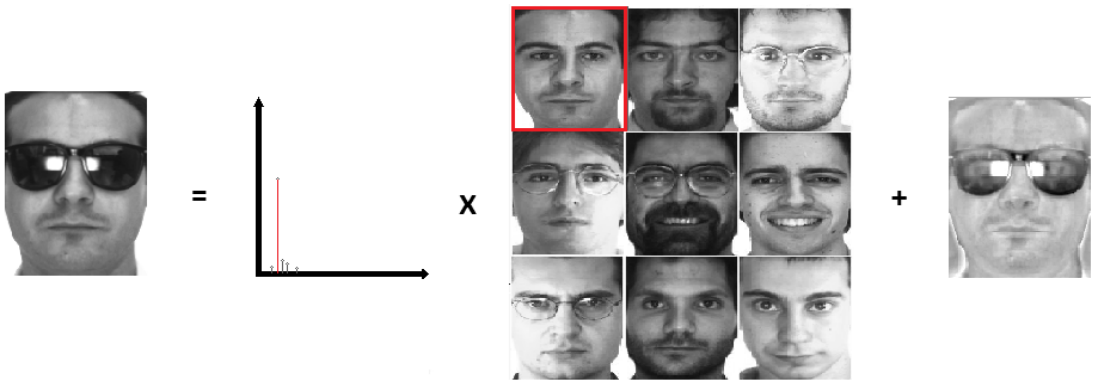

Figure 1. Overview of the SRC framework. The test image (left), which is occluded by an sunglasses. It is equal to the sparse linear combinations of the training images (middle) plus error image (right). The sparse coefficient (red) indicate the corresponding true identity, which is bounded in a red box in the training images (middle). This graph is only for demonstration. There are hundreds or even thousands of training images in the test.

them has different hyper-parameters, i.e., sparsity levels ϵs, mean values θs and variances ϕs. Then, for the signal $\boldsymbol{x} = [\boldsymbol{x}_0^T,\ \boldsymbol{x}_1^T]^T \in \mathbb{R}^{(n+m)}$, which is assumed to be drawn i.i.d from the pdf

$$P_X(x_{jk};\epsilon_k,\theta_k,\phi_k)=(1-\epsilon_k)\delta(x_{jk})+\epsilon_k\mathcal{N}(x_{jk};\theta_k,\phi_k), \quad (4)$$

where $\delta(\cdot)$ denotes the Dirac function, $k = \{0, 1\}$, $j = 1, 2, ...,$ (n+m), and $\mathcal{N}(\cdot;\theta,\phi)$ is the Gaussian pdf. In this paper, we introduce parameter k that it indicates which the element x_{jk} belongs to, either $\boldsymbol{x}_0$ or $\boldsymbol{x}_1$. In particular, if $k = 0$, then $j = 1, ..., n$, otherwise $j = (n + 1), ..., (n + m)$. For the AWGN noise $\boldsymbol{w}$ is assumed to be independent of $\boldsymbol{x}$ with variance ψ, $P_W(w;\psi) = \mathcal{N}(w; 0, \psi)$. In this case, we define the seven unknown hyper-parameters of the prior distribution as $\boldsymbol{q}_k \triangleq [\epsilon_k, \theta_k, \phi_k, \psi]$. It is noteworthy to mention if we drop the subscription k in Eq. (4), it becomes the standard BG in [13].

B. Adaptive Damping

AMP/GAMP approach does not work well while the matrix $\boldsymbol{A}$ is general, e.g., column correlated in robust SRC model. Among the various ways of addressing this issue, we are interested in the AD-GAMP [10] approach. In Damped-GAMP, Rangoon et al. [11] introduced a damping parameter β to adjust the updates of adjacent iterations so that the converges under general transform. It is shown that the damping parameter is proportional to the peak-to-average ratio of the squared singular value of the transform matrix. If this ratio is sufficiently small, the GAMP converges. This scenario explains why AMP performs well when the transaction is large i.i.d Gaussian. The damping approach guarantees the convergence when the transaction is general, while it slows down the progress of convergence. In [10], the authors proposed an adaptive damping GAMP scheme to find a good damping parameter to prevent slowing down the updates procedure too much.

In this paper, we consider the combination of the Dual BG-GAMP and the AD-GAMP approaches, which is shown in Algorithm 1. Here we assume the dual hyper-parameters satisfies the Eq. (4), where it is different from the original BG in [13]. In the GAMP, one is aiming to estimate the input $\boldsymbol{x}$ and the noiseless output $\boldsymbol{z} = \boldsymbol{\Phi x}$ of the transform. The probabilistic relationships in the input and output models are defined in [13]. The standard BG input scalar estimation function g_{in} and the AWGN output scalar estimation function g_{out} are already given in [13] and [7], respectively. As one can find in Eq. (6-7), (9-11), $\beta(t)$ is the adaptive damping parameter. At the very end of this algorithm, the damping parameter is tuned according to current estimation $\hat{\boldsymbol{x}}(t+1)$ and the MMSE cost $J(t + 1)$, adaptively. Here, $J(t + 1) = J_{Bethe}(t + 1)$, which is the Bethe Free Energy function. Due to the space limit, we refer [8][10] for more details about AD-GAMP. It is straightforward to obtain BG-GAMP algorithm by letting $\beta(t) = 1$ and ignore the adapting step in Algorithm 1.

C. Dual Expectation Maximization

We use EM algorithm to estimate the hyper-parameters in Algorithm (1). The EM [16][17] is a well-established method for maximum likelihood estimation with hidden variables. An explicit EM algorithm has been given in [13] for BG-GAMP. It updates hyper-parameters sequentially where updating one parameter by fixing all the other parameters simultaneously. The designed algorithm calls the algorithm 1 after each Dual EM update step. In other words, we upgrade the parameters in the outer algorithm (Dual EM) and perform the Dual BG-AD-GMAP using the new parameters in the inner algorithm. In our case, the EM update is,

$$\forall k : \boldsymbol{q}_k^{h+1} = \arg\max_{\boldsymbol{q}} E\{\ln P(\boldsymbol{x}_k, \boldsymbol{w}; \boldsymbol{q}_k) \mid \boldsymbol{y}; \boldsymbol{q}_k^h\},$$

where h denotes the iteration index. It it worth to note that one has to calculate the corresponding hyper-parameters of each $\boldsymbol{x}_k$, separately. Following [13], it is easy to obtain the updates for each hyper-parameters, which are shown in Algorithm 2. Due to the page limit, we therefore only show the differences.

In the Dual EM step, one has to consider the values of the hyper-parameters $\boldsymbol{q}_k$ for different k. In this paper, we proposed to update the $\boldsymbol{q}_k$ according to the structure of the sparse signal, as one can see from the Eq. 2. In our approach, $\boldsymbol{q}_1 \triangleq [\epsilon_1, \theta_1, \phi_1]$ is the hyper-parameters that associated with the feature coefficiation $\boldsymbol{x}_0$, which is linear combination

Algorithm 1 Inner algorithm (Dual BG-AD-GAMP) with AWGN output.

Initialization:

$$\begin{aligned}\forall j : \hat{x}_{jk}(1) &= \int_{\boldsymbol{x}_k} x_{jk} P_X(x_{jk}) \\ \forall j : \mu^x_{jk}(1) &= \int_{\boldsymbol{x}_k} |x_{jk} - \hat{x}_{jk}(1)|^2 P_X(x_{jk}) \\ \forall i : \hat{u}_i(0) &= 0\end{aligned}$$

$\beta(1) = 1$, $\beta_{max} \in (0, 1]$, $\beta_{min} \in (0, \beta_{max}]$, $G_{pass} \geq 1$, $G_{fail} < 1$, $\varepsilon > 0$

for t = 1, 2, 3,...

$$\forall i : \hat{z}_i(t) = \sum_{j=1}^{(n+m)} \boldsymbol{\Phi}_{ij} \hat{x}_{jk}(t) \quad (5)$$

$$\forall j : \tilde{\boldsymbol{x}}_{jk}(t) = \beta(t)\hat{x}_{jk}(t) + (1-\beta(t))\tilde{x}_{jk}(t-1) \quad (6)$$

$$\forall i : \mu^z_i(t) = \beta(t) \sum_{j=1}^{(n+m)} |\boldsymbol{\Phi}_{ij}|^2 \mu^x_{jk}(t) + (1-\beta(t))\mu^z_i(t-1) \quad (7)$$

$$\forall i : \hat{p}_i(t) = \hat{z}_i(t) - \mu^z_i(t)\hat{u}_i(t-1) \quad (8)$$

$$\forall i : \hat{u}_i(t) = \beta(t) g_{out}(y_i, \hat{p}_i(t), \mu^z_i(t)) + (1-\beta(t))\hat{u}_i(t-1) \quad (9)$$

$$\forall i : \mu^u_i(t) = \beta(t)(-g'_{out}(y_i, \hat{p}_i(t), \mu^z_i(t))) + (1-\beta(t))\mu^u_i(t-1) \quad (10)$$

$$\forall j : \mu^r_{jk}(t) = \beta(t)(\sum_{j=1}^{(n+m)} |\boldsymbol{\Phi}_{ij}|^2 \mu^u_i(t))^{-1} + (1-\beta(t))\mu^r_{jk}(t-1) \quad (11)$$

$$\forall j : \hat{r}_{jk}(t) = \tilde{x}_{jk}(t) + \mu^r_{jk}(t) \sum_{i=1}^{m} \boldsymbol{\Phi}^*_{ij} \hat{u}_i(t) \quad (12)$$

$$\forall j : \mu^x_{jk}(t+1) = \mu^r_{jk}(t) g'_{in}(\hat{r}_{jk}(t), \mu^r_{jk}(t)) \quad (13)$$

$$\forall j : \hat{x}_{jk}(t+1) = g_{in}(\hat{r}_{jk}(t), \mu^r_{jk}(t)) \quad (14)$$

$$J(t+1) = J_{Bethe}(t+1) \quad (15)$$

if $J(t+1) \leq \max\{J(\triangle t), ..., J(t)\}$ or $\beta(t) = \beta_{min}$
then if $\|\hat{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t+1)\| / \|\hat{\boldsymbol{x}}(t+1)\| < \varepsilon$
then stop
else $\beta(t+1) = \min\{\beta_{max}, G_{pass}\beta(t)\}$
$t = t + 1$
else $\beta(t) = \min\{\beta_{min}, G_{fail}\beta(t)\}$
end
end

coefficiation of the training images. For $\boldsymbol{q}_2 \triangleq [\epsilon_2, \theta_2, \phi_2]$, it is determined by the down sampling methods and the noise of the test images $\boldsymbol{x}_1$. It is natural to guess that $\boldsymbol{q}_1 \neq \boldsymbol{q}_2$. In order to compare the performances of the BG-AD-GMAP based algorithms, we set the all the initial value of the sparse vectors to be the same. We present the comparison in next section.

IV. EXPERIMENTS

In this section, we present two experiments to compare the performances of the method in this paper with the benchmark

Algorithm 2 The outer (Dual EM) Algorithm.

$$\begin{aligned}\epsilon_1^{h+1} &= \frac{1}{n}\sum_{j=1}^{n} \pi(\hat{r}_{j1}, \mu^r_{j1}; \boldsymbol{q}_1^h) \\ \epsilon_2^{h+1} &= \frac{1}{m}\sum_{j=n+1}^{n+m} \pi(\hat{r}_{j2}, \mu^r_{j2}; \boldsymbol{q}_2^h) \\ \theta_1^{h+1} &= \frac{1}{n\epsilon_1^{h+1}}\sum_{j=1}^{n} g_{in}(\hat{r}_{j1}, \mu^r_{j1}; \boldsymbol{q}_1^h) \\ \theta_2^{h+1} &= \frac{1}{m\epsilon_2^{h+1}}\sum_{j=n+1}^{n+m} g_{in}(\hat{r}_{j2}, \mu^r_{j2}; \boldsymbol{q}_2^h) \\ \phi_1^{h+1} &= \frac{1}{n\epsilon_1^{h+1}}\sum_{j=1}^{n} \pi(\hat{r}_{j1}, \mu^r_{j1}; \boldsymbol{q}_1^h) \\ &\quad \cdot(\left|\theta_1^h - \gamma(\hat{r}_{j1}, \mu^r_{j1}; \boldsymbol{q}_1^h)\right|^2 + \nu(\hat{r}_{j1}, \mu^r_{j1}; \boldsymbol{q}_1^h)) \\ \phi_2^{h+1} &= \frac{1}{m\epsilon_2^{h+1}}\sum_{j=n+1}^{n+m} \pi(\hat{r}_{j2}, \mu^r_{j2}; \boldsymbol{q}_2^h) \\ &\quad \cdot(\left|\theta_2^h - \gamma(\hat{r}_{j2}, \mu^r_{j2}; \boldsymbol{q}_2^h)\right|^2 + \nu(\hat{r}_{j2}, \mu^r_{j2}; \boldsymbol{q}_2^h)) \\ \psi^{h+1} &= \frac{1}{m}\sum_{i=1}^{m}(|y_i - \hat{z}_i|^2 + \mu^z_i)\end{aligned}$$

ℓ_1 algorithms in the comprehensive review paper [1][1]. The first experiment is designed to compare the recognition rate of all tested algorithms. The second experiment corresponds to comparison of the computational costs in order to achieve the best recognition rates in first experiment.

In the experiments, we explore the designed and benchmark algorithms using the Extended Yale B Face Database [18]. We choose 722 (19 images for each person) normal lighting conditioned images as the training data and another 266 images as the test data, which has more extreme lighting conditions. The images are down-sampled from 192×168 to 32×28. A percentage of randomly chosen pixels from each of the test images are corrupted/replaced with i.i.d uniform distribution (e.g., uniform over [0,255] for the 8-bit images). We vary the percentages of corrupted pixels c from 10 to 90 percent. In this experiment, $\boldsymbol{\Phi} \in \mathbb{R}^{896\times 1618}$, which has a high sampling rate to keep the mean value as small as possible. We test all the benchmark algorithms for all corruption cases within a fixed time limit, which is 80 seconds in our experiment. We use the SCI to calculate the recognition rate.

For the first experiment, we compared our approach with the Dual Augmented Lagrangian Method (DALM) [1], Primal Augmented Lagrangian Method (PALM) [1], Primal-dual Interior-point Algorithm (PDIPA) [19] and Truncated Newton Interior-point Method (TNIPM, known as L1LS) [20] in the review paper [1]. We also compared our Dual BG-AD-GAMP (initialize $\epsilon_1 = \epsilon_2 = 0.08$ with Dual EM update) approach with the BG-AD-GAMP (initialize $\epsilon = 0.08$ with EM update) algorithm. In other words, we update the hyper-parameters of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ separately. In the experiments, we set both algorithms have maximum 25 inner iterations and 200

[1]All the benchmark algorithms are available as Matlab toolbox: http://www.eecs.berkely.edu/~yang/software/l1benchmark/.

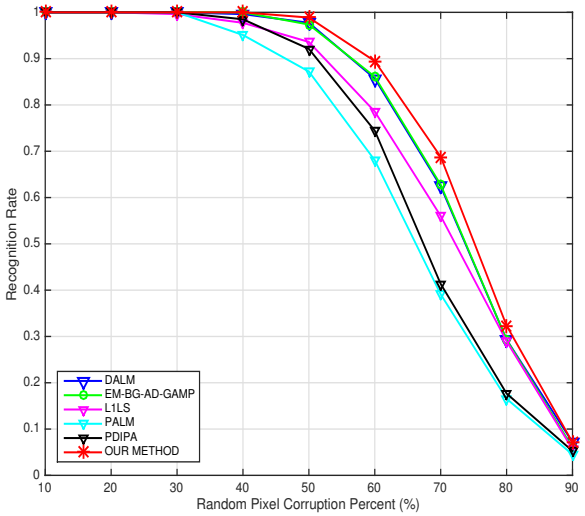

Figure 2. Recognition rates for different algorithms under different fractions of noise corruptions.

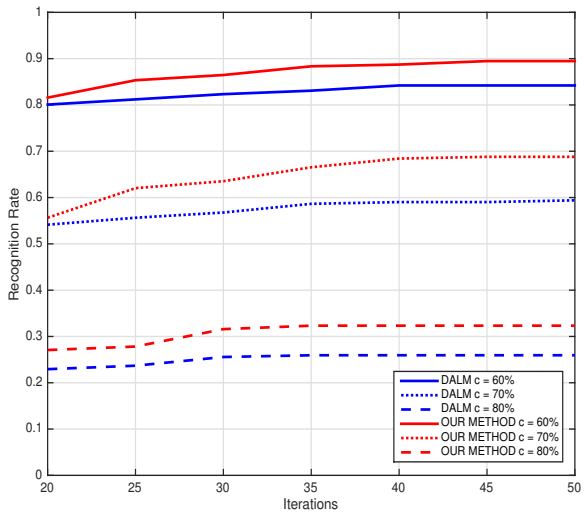

Figure 3. Comparison of our algorithm with DALM under different fractions of corrupted entries c = 60%, 70%, 80%. Our method: red lines. DALM algorithm: blue lines.

outer iterations (EM step). In all the other algorithms, we let number of iterations to 5000. The results of the recognition rates of the benchmark algorithms are shown in fig. 2. As one can see from the figure, our algorithm has the best performance among all benchmarks in terms of recognition rate. For the BG-AD-GAMP, it has similar recognition rate DALM. Interestingly, Compared with our method, BG-AD-GAMP has lower recognition rate since it does not update the hyper-parameters separately.

In the second experiment, we test the computational cost in in terms of recognition rate. For the lake of space, we only show the comparison of our method and DALM (best algorithm in the review paper [1]) under different fractions of corruptions c = 60%, 70%, 80% in Fig. 3. Our method achieves the best recognition rate as shown in Fig. 2 in 50 iterations. However, the DALM algorithm requires more iterations to reach its best.

REFERENCES

[1] A. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, "Fast ℓ_1 -minimization algorithms for robust face recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3234–3246, Aug 2013.

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[3] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, Nov 1993, vol. 1, pp. 40–44.

[4] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[5] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 45, pp. 18914–18919, 2009.

[6] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *IEEE Information Theory Workshop*, 2010.

[7] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proceedings of IEEE International Symposium on Information Theory*, 2011, pp. 2168–2172.

[8] S. Rangan, A. K. Fletcher, P. Schniter, and U. Kamilov, "Inference for Generalized Linear Models via Alternating Directions and Bethe Free Energy Minimization," *arXiv:1501.01797*, pp. 1–20, 2015.

[9] A. Manoel, F. Krzakala, E.W. Tramel, and Z. Lenka, "Sparse Estimation with the Swept Approximated Message-Passing Algorithm," *Arxiv:1406.4311*, pp. 1–11, 2014.

[10] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborova, "Adaptive Damping and Mean Removal for the Generalized Approximate Message Passing Algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2021–2025.

[11] S. Rangan, P. Schniter, and A. K. Fletcher, "On the Convergence of Approximate Message Passing with Arbitrary Matrices," *arXiv:1402.3210*, pp. 1–10, 2014.

[12] J. Vila and P. Schniter, "Expectation-Maximization gaussian-mixture approximate message passing," *IEEE Transactions on Signal Processing*, vol. 61, no. 1–19, pp. 4658–4672, 2013.

[13] J. Vila and P. Schniter, "Expectation-maximization Bernoulli-Gaussian approximate message passing," in *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov. 2011, pp. 799–803.

[14] J. Wright, A. Ganesh, A. Yang, Z. Zhou, and Y. Ma, "Sparsity and robustness in face recognition," *arXiv:1111.1014v1*, pp. 1–12, 2011.

[15] F. Caltagirone, F. Krzakala, and L. Zdeborová, "On Convergence of Approximate Message Passing," *Arxiv:1401.6384*, pp. 1–5, 2014.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of The Royal Statitisal Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[17] R. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. 1998, pp. 355–368, Kluwer Academic Publishers.

[18] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, Jun 2001.

[19] M. Kojima, N. Megiddo, and S. Mizuno, "Theoretical convergence of large-step primal dual interior point algorithms for linear programming," *Mathematical Programing*, vol. 59, no. 1, pp. 1–21, 1993.

[20] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l1-regularized least squares," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, Dec 2007.