Dictionary learning from phaseless measurements
Andreas Tillmann, Yonina Eldar, Julien Mairal

To cite this version:
Andreas Tillmann, Yonina Eldar, Julien Mairal. Dictionary learning from phaseless measurements. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mar 2016, Shanghai, China. pp.4702-4706, 10.1109/ICASSP.2016.7472569. hal-01387416

HAL Id: hal-01387416
https://inria.hal.science/hal-01387416
Submitted on 25 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
DICTIONARY LEARNING FROM PHASELESS MEASUREMENTS

Andreas M. Tillmann*  Yonina C. Eldar†  Julien Mairal○

* TU Darmstadt, Research Group Optimization, Dolivostr. 15, 64293 Darmstadt, Germany
† Department of Electrical Engineering, Technion, Haifa 32000, Israel
○ Inria, Lear Team, 655 Avenue de l’Europe, 38330 Montbonnot, France

ABSTRACT

We propose a new algorithm to learn a dictionary along with sparse representations from signal measurements without phase. Specifically, we consider the task of reconstructing a two-dimensional image from squared-magnitude measurements of a complex-valued linear transformation of the original image. Several recent phase retrieval algorithms exploit underlying sparsity of the unknown signal in order to improve recovery performance. In this work, we consider sparse phase retrieval when the sparsifying dictionary is not known in advance, and we learn a dictionary such that each patch of the reconstructed image can be sparsely represented. Our numerical experiments demonstrate that our proposed scheme can obtain significantly better reconstructions for noisy phase retrieval problems than methods that cannot exploit such “hidden” sparsity.

Index Terms— Phase Retrieval, Dictionary Learning

1. INTRODUCTION

Phase retrieval has been an active research topic for decades [1, 2]. In mathematical terms, it can be formulated as

$$\begin{align*}
\text{find } x & \in \mathcal{X} \subseteq \mathbb{C}^N \text{ s.t. } |f_i(x)|^2 = y_i \quad \forall i = 1, \ldots, M, \\
\end{align*}$$

where the functions $f_i : \mathbb{C}^N \rightarrow \mathbb{C}$ are linear operators and the scalars $y_i = |f_i(\hat{x})|^2$ are nonlinear measurements of the original signal $\hat{x}$ in $\mathcal{X}$, obtained by removing the phase information. A traditional approach is to consider cases in which the solution of (1) is unique, up to global phase, and to devise algorithms to recover $\hat{x}$. Usually, recovering $\hat{x}$ involves oversampling, i.e., taking $M > N$ measurements. The most popular techniques for phase retrieval are based on alternating projections [3, 4, 5]. These methods generally require precise prior information on the signal (such as knowledge of the support set) and often converge to erroneous results. More recent approaches include semidefinite programming [6, 7, 8, 9, 10] and gradient-based methods such as Wirtinger Flow [11, 12].

In recent years, new techniques for (1) were developed when $\hat{x}$ is sparse, i.e., is composed of only few atoms from a known dictionary [7, 13, 12, 14], leading to algorithms with improved recovery performance. The main idea is akin to compressed sensing, where one works with fewer linear measurements than signal components [15, 16]. An important observation that boosted the applicability of sparse recovery techniques is that many classes of signals admit a sparse approximation in some basis or overcomplete dictionary [17, 18, 19]. While sometimes such dictionaries are known explicitly, better results can often be achieved by adapting the dictionary to the data [18]. Numerous algorithms have been developed for this task (e.g., [20, 21]) when the signal measurements are linear.

In this work, we propose a dictionary learning scheme for simultaneously solving the signal reconstruction and sparse representation problems given nonlinear, phaseless and possibly noisy measurements. We demonstrate its ability to achieve significantly improved reconstructions (especially when the oversampling ratio is low and the noise level high) compared to Wirtinger Flow (WF) [11], which cannot exploit sparsity if the dictionary is unknown. Our algorithm, referred to as phase-retrieval dictionary-learning (PRDL), is based on alternating minimization where we iterate between best fitting the data and sparsely representing the recovered image. PRDL combines projected gradient descent like in Wirtinger Flow to update the image, iterative shrinkage to sparsify the image [22] and block-coordinate descent for the dictionary update [21].

2. PHASE-RETRIEVAL DICTIONARY-LEARNING

Our focus is on the application to image reconstruction. Therefore, we work in the 2D setting directly; however, all expressions and algorithms could easily be formulated for 1D signals, as in (1).

We wish to recover an image $\hat{X}$ in $[0, 1]^{N_1 \times N_2}$ from noise-corrupted phaseless nonlinear measurements

$$Y := |F(\hat{X})|^2 + N,$$

where $F : \mathbb{C}^{N_1 \times N_2} \rightarrow \mathbb{R}^{M \times M_2}$ is a linear operator, $N$ denotes noise, and the complex modulus and squares are taken component-wise. Signal sparsity is known to aid in phase retrieval, but a sparsifying transform is not always known a priori. This motivates learning a dictionary $D$ in $\mathbb{R}^{s \times n}$ such that each $s_1 \times s_2$ patch $\hat{x}'$ of $\hat{X}$, represented as a vector of size $s = s_1 s_2$, can be approximated by $\hat{x}' \approx Da'$ with a sparse vector $a'$ in $\mathbb{R}^n$. Here, $n$ is chosen a priori and the number of patches depends on whether the patches are overlapping or not. In general, $D$ is chosen such that $n \geq s$.

Before presenting our approach for tackled (2), we define the following notation. We consider the linear operator $\mathcal{E} : \mathbb{C}^{N_1 \times N_2} \rightarrow \mathbb{C}^{s \times p}$ that extracts the $p$ patches $x'$ (which may overlap or not) from an image $X$ and forms the matrix $\mathcal{E}(X) = (x'^1, \ldots, x'^p)$. Similarly, we define the linear operator $\mathcal{R} : \mathbb{C}^{s \times p} \rightarrow \mathbb{C}^{N_1 \times N_2}$ that reverses this process, i.e., builds an image from a matrix containing vectorized patches as its columns. Thus, in particular, we have $\mathcal{R}(\mathcal{E}(X)) = X$. Further, let $A := (a'^1, \ldots, a'^p)$ in $\mathbb{R}^{s \times p}$ be the matrix containing the patch representation coefficient vectors as columns. Then, our desired sparse-approximation relation $x' \approx Da'$ can be expressed as $\mathcal{E}(X) \approx DA$.

With this notation in hand, we now introduce our method, called phase-retrieval dictionary-learning (PRDL), which aims at solving

$$\begin{align*}
\min_{X, D, A} & \frac{1}{2} \|Y - |F(X)|^2\|^2_F + \frac{\lambda}{2} \|\mathcal{E}(X) - DA\|^2_F + \lambda \sum_{i=1}^p \|a'^i\|_1 \\
\text{s.t. } & X \in [0, 1]^{N_1 \times N_2}, \quad D \in D.
\end{align*}$$

(3)
Algorithm 1 Phase-Retrieval Dictionary-Learning Algorithm (PRDL)

Input: initial image estimate $X(0) \in [0, 1]^{N_1 \times N_2}$, initial dictionary $D(0) \in D \subset \mathbb{R}^{n \times n}$, parameters $\mu, \lambda > 0$, iteration limits $K_1, K_2$;
Output: Learned dictionary $D = D(K)$, patch representations $A = A(K)$, image reconstruction $X = X(K)$;
1: for $\ell = 0, 1, 2, \ldots, K_1 + K_2$ do
2: choose step size $\gamma^A_\ell$ as explained in Section 3.1 and update $A(\ell+1) \leftarrow S_{\lambda \gamma^A_\ell / \mu} \left( A(\ell) - \gamma^A_\ell D(\ell) A(\ell) - \mathcal{E}(X(\ell)) \right)$;
3: choose step size $\gamma^X_\ell$ and update $X(\ell+1) \leftarrow P_X \left( X(\ell) - \gamma^X_\ell \left( \| F^* (F(X) \odot (|F(X)|^2 - Y)) + \mu R \odot \mathcal{R}(E(X) - DA) \right) \right)$;
4: if $\ell < K_1$ then
5: keep $D(\ell+1) \leftarrow D(\ell)$;
6: else
7: set $B \leftarrow \mathcal{E}(X(\ell)) A(\ell) + C$ and $C \leftarrow A(\ell) A(\ell)^T$;
8: for $j = 1, \ldots, n$ do
9: if $c_{jj} > 0$ then
10: update $j$-th column: $(D(\ell+1))_j \leftarrow \frac{1}{\gamma_j} (B_j - D(\ell) C_j) + (D(\ell))_j$;
11: else
12: reset $j$-th column: e.g., $(D(\ell+1))_j \leftarrow \text{random } \mathcal{N}(0, 1) \text{ vector (in } \mathbb{R}^p)$;
13: normalize $(D(\ell+1))_j \leftarrow \frac{1}{\| D(\ell+1) \|_2} (D(\ell+1))_j$

Here, $\| X \|_F$ denotes the Frobenius matrix-norm, which generalizes the Euclidean norm to matrices. The parameters $\mu, \lambda > 0$ in the objective (3) provide a way to control the trade-off between the data fidelity term from the phase retrieval problem and the approximation of the sparsity of the image patches. To that effect, we use the $\ell_1$-norm, which is well-known to have a sparsity-inducing effect [23]. In order to avoid scaling ambiguities, we also restrict $D$ to be in the subset $D$ of matrices with unit-$\ell_2$-norm columns, and assume $n < p$ (otherwise, each patch is trivially representable by a 1-sparse vector $a^i$ by including $x^i / \| x^i \|_2$ as a column of $D$).

3. ALGORITHMIC FRAMEWORK

Problem (3) is non-convex and difficult to solve, similarly to classical dictionary learning [24, 17, 20, 19]. Therefore, we adopt an algorithm that provides monotonic decrease of the objective. (In fact, convergence to a stationary point of (3) can be obtained under mild assumptions, using results from [25]; we cannot give details here due to space restrictions.)

The algorithmic framework we employ is that of alternating minimization: For each variable $D, A$ and $X$ in turn, we take one step towards solving (3) w.r.t. this variable alone, keeping the other two fixed. We summarize our method in Algorithm 1, where the superscript $*$ denotes the adjoint operator (for a matrix $Z$, $Z^*$ is thus the conjugate transpose), $\mathcal{R}(\cdot)$ extracts the real part of a complex-valued argument, and $\odot$ denotes the Hadamard (element-wise) product of two matrices. The algorithm also involves the classical soft-thresholding operator $S_{\tau}(Z) := \max\{0, |Z| - \tau\} \odot \text{sign}(Z)$ and the Euclidean projection $P_X(Z) := \max\{0, \min\{1, |Z|\}\}$ onto $X := [0, 1]^{N_1 \times N_2}$; here, all operations are meant component-wise.

To avoid training the dictionary on potentially useless early estimates, the algorithm performs two phases—while the iteration counter $\ell$ is smaller than $K_1$, the dictionary is not updated. We next explain the algorithmic steps in more detail.

3.1. Updating the Patch Representation Vectors

Updating $A$ (i.e., (3) with $D$ and $X$ fixed at their current values) calls for decreasing the objective
\[
\sum_{i=1}^p \left( \frac{1}{2} \| D(\ell) a^i - x^i(\ell) \|_2^2 + \frac{1}{\mu} \| a^i \|_1 \right).
\]
(4)

Since the objective here is separable, we can update all vectors $a^i$ in parallel independently. To do so, we choose to perform one step of the algorithm ISTA (see, e.g., [22]), which has a monotonic decrease property. ISTA is a gradient-based method which performs the following update for each $i = 1, \ldots, p$:
\[
a^i_{\ell+1} = S_{\lambda \gamma^A_\ell / \mu} \left( a^i_{\ell} - \gamma^A_\ell D(\ell) A(\ell) a^i_{\ell} - x^i_{\ell} \right).
\]

This update involves a gradient descent step followed by soft-thresholding. The step size parameter $\gamma^A_\ell$ can be chosen as $1/L_A$, where $L_A$ is an upper bound on the Lipschitz constant of the gradient; here, $L_A = \| D(\ell) \|_2^2$ would do, but a better strategy is to use a backtracking scheme to automatically update $L_A$ [22].

Constructing $A(\ell+1)$ from the $a^i_{\ell+1}$ as specified above is equivalent to Step 2 of Algorithm 1.

3.2. Updating the Image Estimate

With $D = D(\ell)$ and $A = A(\ell+1)$ fixed, updating $X$ consists of decreasing the objective
\[
\frac{1}{2} \| Y - |F(X)|^2 \|_F^2 + \frac{\mu}{2} \| E(X) - DA \|_F^2
\]
(5)

with $X \in X := [0, 1]^{N_1 \times N_2}$. This problem can be seen as a regularized version of the phase retrieval problem (with regularization parameter $\mu$) that encourages the patches of $X$ to be close to the sparse approximation $DA$ obtained during the previous iterate.

Our approach to decrease the value of the objective (5) is by a simple projected gradient descent step. In fact, for $\mu = 0$, this reduces to the so-called Wirtinger Flow method [11], but with necessary modifications to take into account the constraints on $X$ (real-valuedness and variable bounds $[0, 1]$).

The gradient (matrix) of $\varphi(X)$ is $\frac{1}{2} \| Y - |F(X)|^2 \|_F^2$ with respect to $X$ can be computed as
\[
\nabla \varphi(X) = \mathcal{R} \left( F^* (F(X) \odot (|F(X)|^2 - Y)) \right),
\]
by using the chain rule (we omit a more detailed derivation due to space limitations). The gradient of $\psi(X) := \frac{\mu}{2} \| E(X) - DA \|_F^2$ is given by
\[
\nabla \psi(X) = \mu E^* (E(X) - DA) = \mu R \odot \mathcal{R}(E(X) - DA),
\]
where \( R \in \mathcal{X} \) has entries \( r_{ij} \) equal to the number of patches the respective pixel \( x_{ij} \) is contained in. Note that if the whole image is divided into a complete set of non-overlapping patches, \( R \) will just be the all-ones matrix; otherwise, the element-wise multiplication with \( R \) undoes the averaging of pixel values performed by \( \mathcal{R} \) when assembling an image from overlapping patches.

Finally, the gradient w.r.t. \( X \) of the objective in (5) is \( \nabla \varphi(X) + \nabla \psi(X) \in \mathbb{R}^{N_1 \times N_2} \), and the update in Step 3 of Algorithm 1 is indeed shown to be a projected gradient descent step. Typically, a backtracking strategy may be used for choosing the step size \( \gamma^\ell \), as detailed in the section devoted to numerical experiments.

3.3. Updating the Dictionary
To update the dictionary, i.e., to approximately solve (3) w.r.t. \( D \) alone, keeping \( X \) and \( A \) fixed at their current values, we employ one pass of a block-coordinate descent (BCD) algorithm on the columns of the dictionary [21]. The objective to decrease may be written as

\[
\min_D \sum_{i=1}^p \| Da^i_{(\ell+1)} - x^i_{(\ell+1)} \|_2^2 \quad \text{s.t.} \quad D \in \mathcal{D},
\]

and the update rule given by Steps 4–13 corresponds exactly to one iteration of [19, Algorithm 11] applied to (6), which ensures the monotonic decrease of the objective function.

4. NUMERICAL EXPERIMENTS
We first describe some aspects of the concrete models and our implementation and then proceed to discuss our experiments.

4.1. Experimental Setup
We consider several linear operators \( F \) corresponding to different types of measurements. We denote by \( F \) the (normalized) 2D-Fourier transform, and introduce two complex Gaussian matrices \( G \in \mathbb{C}^{M_1 \times N_1}, H \in \mathbb{C}^{M_2 \times N_2} \), whose entries are i.i.d. samples from the distribution \( \mathcal{N}(0, I/2) \) or \( i\mathcal{N}(0, I/2) \). Then, we experiment with the operators \( \mathcal{F}(X) = GX, \mathcal{F}(X) = GXG^*, \mathcal{F}(X) = GH \), and the coded diffraction model

\[
\mathcal{F}(X) = \begin{bmatrix} F(M_1 \odot X) \\ \vdots \\ F(M_m \odot X) \end{bmatrix}, \quad \mathcal{F}^* (Z) = \sum_{j=1}^m (M_j \odot F^* (Z)),
\]

where the \( M_j \)'s are admissible coded diffraction patterns (CDPs), see for instance [11, Sect. 4.1]; in our experiments we used ternary CDPs, such that each \( M_j \) is in \( \{0, \pm 1\}^{N_1 \times N_2} \). Note that the operator \( F \) is implemented using fast Fourier transforms.

To reconstruct \( X \), we choose an oversampling setting where \( M_1 = 4N_1, M_2 = 4N_2 \) and/or \( m = 2 \), respectively. Moreover, we corrupt our measurements with additive white Gaussian noise \( N \) such that \( \text{SNR}(Y, |\mathcal{F}(X)|^2 + N) = 10 \text{dB} \) for the Gaussian-type, and \( 20 \text{dB} \) for CDP measurements, respectively. Note that this settings yields a quite heavy noise level for the Gaussian cases, and a relatively low oversampling ratio for the CDPs.

4.2. Implementation Details
We choose to initialize our algorithm with a simple random image \( X_{(0)} \) in \( \mathcal{X} \) to demonstrate the robustness of our approach with respect to its initialization. Nevertheless, other choices are possible. For instance, one may also initialize \( X_{(0)} \) with a power-method scheme similar to that proposed in [11], modified to account for the required real-valuedness and box-constraints. The dictionary is initialized as \( D_{(0)} = (I, F_0) \in \mathbb{R}^{s \times 2s} \), where \( F_0 \) is the 2D discrete cosine transform (see, e.g., [18]).

To update \( A \), we use the ISTA implementation from the SPAMS package [21] with its integrated backtracking line search (for \( L_\lambda \)).

Regarding the step sizes \( \gamma^\ell \) for the update of \( X \) (Step 3 of Alg. 1), we adopt the following simple strategy: Whenever the gradient step leads to a reduction in the objective function value, we accept it; otherwise, we recompute the step with \( \gamma^\ell \) halved until a reduction was achieved (no more than 100 trials). Regardless of whether \( \gamma^\ell \) was reduced or not, we reset its value to \( 1.68\gamma^0 \) for the next round; the initial step size is 1.

Finally, we consider non-overlapping \( 8 \times 8 \) patches (though our code can also handle overlap). We run PRDL (Algorithm 1) with \( K_1 = 25 \) and \( K_2 = 50 \); the regularization/penalty parameter values can be read from Table 1 (here, \( N_{11} \) is the number of elements of \( Y \)). We remark that these parameter values were empirically found to work well for the instances considered here but are not the outcome of rigorous parameter benchmarking, which is a subject of future investigations.

4.3. Computational Experiments
We test our method on a collection of typical (grayscale) test images used in the literature, namely cameraman, house and peppers of size \( 256 \times 256 \), and lena, barbara, boat, fingerprint and mandrill of size \( 512 \times 512 \). All experiments were carried out on a Linux 64-bit quad-core machine (2.8 GHz, 8 GB RAM) running Matlab R2015a.

We evaluate our approach with the following question in mind: Can we improve upon the quality of reconstruction compared to standard phase retrieval algorithms? Standard methods cannot exploit sparsity if the underlying basis or dictionary is unknown; as we will see, the introduced (patch)-sparsity indeed allows for better recovery results (at least in the oversampling and noise regimes considered here).

To evaluate the achievable sparsity, we look at the average number of nonzeros in the columns of \( A \) after running our algorithm. Generally, smaller values indicate an improved suitability of the learned dictionary for sparse patch coding (high values often occur if the regularization parameter \( \lambda \) is too small and the dictionary is learning the noise, which is something we would like to avoid).

To assess the quality of the image reconstruction, we consider two standard measures, namely the peak signal-to-noise ratio (PSNR) of the reconstructed \( X \) as well as its structural similarity index (SSIM) [26]. For PSNR, larger values are better, and SSIM-values closer to 1 (always ranging between 0 and 1) indicate better visual quality.

Table 1 displays the CPU times, PSNR- and SSIM-values and mean patch representation vector sparsity levels obtained for the various measurement types, averaged over the instance and the same size. The concrete example in Figure 1 shows the results from PRDL and plain Wirtinger Flow (WF: the real-valued, [0, 1]-box constrained variant, which corresponds to running Algorithm 1 with \( \mu = 0 \) and omitting the updates of \( A \) and \( D \)) for CDP measurements of the mandrill image. (In all tests, we let the Wirtinger Flow method run for the same number of iterations (75) and use the same starting points as for the PRDL runs.)

The PRDL method consistently provides better image reconstructions than WF, which clearly shows that our approach successfully introduces sparsity into the phase retrieval problem and exploits

\[\text{http://spams-devel.gforge.inria.fr/}.\]
it in the solution process. As can be seen from Table 1, the obtained
dictionaries allow for rather sparse representation vectors, with the
effect of making better use of the information provided by the mea-
surements, and also denoising the image along the way. The latter
effect of making better use of the information provided by the mea-
surements can be seen in the example (Fig. 1) and also inferred from the
measurements, and also denoising the image along the way. The latter
result. Some preliminary experiments (not reported here) indicate
that both approaches produce comparable results in the noisefree
setting, while our numerical results demonstrate a denoising feature
of our algorithm that the simple approach would obviously lack.
Another interesting aspect to evaluate is by how much reconstruc-
tion quality and achievable sparsity degrade due to the loss of phase
(or, more generally, measurement nonlinearity).

5. DISCUSSION AND CONCLUSION

We believe that our experiments demonstrate that dictionary learn-
ing for phase retrieval with a patch-based sparsity is a promising
direction, especially for cases where the original Wirtinger Flow ap-
proach fails. For the future, we are planning to conduct larger-scale
experiments to answer some questions that we have left open.

For example, we expect further improvements by rigorously
benchmarking the algorithmic parameters (in particular, to find
good default settings). Several variants of our algorithm may also
be developed—for instance, we successfully used ℓ1-constraints in-
stead of the ℓ1-penalty, by combining Orthogonal Matching Pursuit
(OMP) [27] with our framework. Finally, to evaluate the quality of
the learned dictionary, one might also ask how PRDL compares with
the straightforward approach to first run (standard) phase retrieval
and then learn dictionary and sparse patch representations from the
result. Some preliminary experiments (not reported here) indicate
that both approaches produce comparable results in the noisefree
setting, while our numerical results demonstrate a denoising feature
of our algorithm that the simple approach would obviously lack.
Another interesting aspect to evaluate is how much reconstruc-
tion quality and achievable sparsity degrade due to the loss of phase
(or, more generally, measurement nonlinearity).

Acknowledgements

The work of J. Mairal was funded by the French National Research
Agency [Macaron project, ANR-14-CE23-0003-01]. The work of Y.
Eldar was funded by the European Union’s Horizon 2020 research
and innovation programme under grant agreement ERC-BNYQ, and
by the Israel Science Foundation under Grant no. 335/14.
6. REFERENCES


