Fast phase retrieval for high dimensions: A block-based approach

Boshra Rajaei[1,5,6], Sylvain Gigan[3,6], Florent Krzakala[2,6], Laurent Daudet[1,4,6]

[1]*Institut Langevin, ESPCI and CNRS UMR 7587, Paris, F-75005, France*
[2]*LPS-ENS, UPMC and CNRS UMR 8550, Paris, F-75005, France.*
[3]*Laboratoire Kastler Brossel, UPMC, ENS, Collège de France, CNRS UMR 8552, Paris, F-75005, France*
[4]*Paris Diderot University, Sorbonne Paris Cité, Paris, F-75013, France*
[5]*Sadjad University of Technology, Mashhad, Iran*
[6] *PSL Research University, F-75005 Paris, France*

This paper addresses fundamental scaling issues that hinder phase retrieval (PR) in high dimensions. We show that, if the measurement matrix can be put into a generalized block-diagonal form, a large PR problem can be solved on separate blocks, at the cost of a few extra global measurements to merge the partial results. We illustrate this principle using two distinct PR methods, and discuss different design trade-offs. Experimental results indicate that this block-based PR framework can reduce computational cost and memory requirements by several orders of magnitude.

INTRODUCTION

Phase retrieval (PR) is the problem of recovering a complex-valued signal $\mathbf{x} \in \mathbb{C}^N$ from the squared magnitude $\mathbf{y} \in \mathbb{R}_+^M$ of its (possibly noisy) projections

$$\mathbf{y} = |\mathbf{H}\mathbf{x}|^2 \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{M \times N}$ is a known matrix called projection (or measurement) matrix. This problem arises in many digital signal processing situations, such as audio source separation, but also in physical sensing / imaging applications where designing an intensity-only detector (such as most optical sensors) is easier, faster and/or cheaper than amplitude-and-phase detectors [1]. Some of these applications include X-ray crystallography [2], X-ray diffraction imaging [3], optical imagers [4, 5] and astronomical imaging [6]. Most PR methods have been designed for Fourier transform or i.i.d. random complex measurement matrices, but in some applications a generic solution to Eqn. (1) without specific restrictions on $\mathbf{H}$ and/or $\mathbf{x}$ may be required. Well-known PR methods include but are not limited to convex relaxation algorithms such as phaseLift [7] and phaseCut [8], iterative non-convex optimization algorithms such as Wirtinger flow (WF) [9] and its truncated version (TWF) [10], iterative projections algorithms such as Gerchberg and Saxton [11], Fienup [12] and variants [13, 14], and spectral recovery method [15].

This study investigates scalability issues for PR algorithms, as a function of the size N of the unknown signal $\mathbf{x}$. To reconstruct the complex signal $\mathbf{x}$ (up to a global phase) using its intensity-only projections, the size M of the measurement vector should be at least $2N$ - it has been established recently that, in a generic case, $M \geq 4N$ measurements are required [16] to recover a unique $\mathbf{x}$. Therefore, the amount of data that a PR algorithm has to handle, for the $\mathbf{H}$ matrix, is at least of the order $O(N^2)$. Besides these memory requirements, the computational complexity of generic PR algorithms scales at least with the same order $O(N^2)$, and possibly worse. This can be a bottleneck for many of the above applications, such as real-time imaging.

There are fundamentally two ways to alleviate these scaling issues : either by making a sparsity assumption on the unknown vector $\mathbf{x}$, or by imposing some extra constraints on the measurement matrix $\mathbf{H}$. In the first case, a sparsity assumption on $\mathbf{x}$ allows a reconstruction with less than $4N$ measurements, and consequently can speed-up the reconstruction. This class of algorithms are mainly referred to as compressive (or compressed) phase retrieval methods in the literature, and are mostly based on a Bayesian framework. Examples of algorithms in this category include Moravec et al. l_1-norm algorithm [17], Mukherjee and Seelamantula PR method [18], GESPAR algorithm by Shechtman et al. [19] and Schniter and Rangan prGAMP algorithm [20]. In the second case, some specific classes of measurements matrices allow PR with reduced complexity, for instance Iwen et al. method based on local correlation measurements [21] and Zhang and Kner phase retrieval using special binary structured matrices [22]. However, in most of the physical scenarios presented above, the entries of the measurement matrix cannot be designed at will, as they correspond to the physical sensing process.

In this paper, we propose a conceptually simple but remarkably effective block-based PR framework, that can be used in combination with any PR algorithm, and that not only scales up easily to high dimensions but does not impose any predefined constraint, such as sparsity, on the input signal. The constraint on the measurement matrix is that it can be put in a generalized block diagonal form, but each block may have arbitrary entries. This block-based phase retrieval method starts by splitting the $M \times N$ input problem into K, $m_i \times n_i$ sub-problems, where $\sum_{i=0}^{K-1} n_i = N$, $m_i = \lceil \alpha n_i \rceil$ and $\alpha = M/N$. The K sub-problems are then solved in parallel using any PR method. Finally, all the partial results are merged with a few extra global measurements, by applying a low-dimension global phase tuning step.

Since, as discussed above, the memory requirements and computational complexity of PR algorithms scale at least as $O(N^2)$, breaking down the PR problem into K sub-problems of size N/K results in a reduction of memory/complexity requirements by at least a factor K (neglecting here the cost of the low-dimensional final phase tuning step), even in single-thread mode. Furthermore, the K sub-problems can here be solved in an "embarrassingly parallel" way, opening the way for further gains if multiple computing threads are available. It has to be emphasized that the requirement for the measurement matrix to be put into a block diagonal form is often a mild constraint - much milder than, for instance, imposing constraints on the non-zero entries of the measurement matrix -. In fact, this approach is compatible with many of the physicals systems above ; designing the measurement matrix as block diagonal can be interpreted as the ability to probe a large object by parts, which can be controlled by the source of illumination. Our original motivation for this study arises from such a physical imaging system [23]. More generally, there are a number of optical imaging setups that may benefit from this approach, such as the LED array microscope [24], multiple coherent diffractive imagers (CDI) [25], and single-shot phase imaging with randomized light (SPIRaL) [26].

In summary, the main contributions of this paper are :

- the presentation of a new framework for block-based PR, working with any PR algorithm, and making no assumption on the input signal.

- experimental results with two distinct PR algorithms, showing the computational gains for different signal sizes and choice of parameters.

BLOCK-BASED PR ALGORITHM

For ease of notation, let us consider a noiseless square root version of (1) as

$$\mathbf{y} = |\mathbf{Hx}| \qquad (2)$$

All the equations are convertible to the general case in a straightforward way. Assume $\mathbf{H}$ follows a block structure according to the following definition which is simply an extension of block diagonal matrices to rectangular blocks.

Definition 1. *A $M \times N$ block matrix is called K-rectangular block diagonal (K-RBD) matrix iff it can be partitioned into non-overlapping $m \times n$ blocks with non-zero entries only in blocks containing $(im, in); i = 0, 1, ..., K-1$ entries.*

Here, for the sake of simplicity, we assume equal-length blocks with $m = \frac{M}{K}$ and $n = \frac{N}{K}$ as positive integers.

Therefore, a K-RBD matrix $\mathbf{H}$ has a structure of the form

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_0 & 0 & ... & 0 \\ 0 & \mathbf{H}_1 & ... & 0 \\ & & \ddots & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & ... & \mathbf{H}_{K-1} \end{bmatrix}_{M \times N}. \qquad (3)$$

Note that there is no restriction on the inner structure of the $\mathbf{H}_i$ submatrices. Correspondingly, we split the input vector $\mathbf{x}$ into K equal subvectors of length n

$$\mathbf{x} = [\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x}_{K-1}]^t \qquad (4)$$

Using the above definitions, our block-based PR starts by solving K sub-problems of

$$\mathbf{y}_i = |\mathbf{H}_i \mathbf{x}_i|; i = 0 \ldots K-1, \qquad (5)$$

independently. This can be done by any generic PR method. Let us call the first step as the *blocking step* and the resulting estimations as $\hat{\mathbf{x}}_i$. PR methods can only recover the $\mathbf{x}_i$ variables up to a global phase, which means that even under a perfect recovery assumption we have

$$\hat{\mathbf{x}}_i = \mathbf{x}_i e^{j\phi_i}; i = 0 \ldots K-1 \qquad (6)$$

where ϕ_i phase shifts are not necessarily identical. Therefore, to preserve a unique global phase all over the input signal space, the block-based PR goes through a *phase tuning step*. In phase tuning, we employ an extra set of $L = \beta K$ measurements, $\tilde{\mathbf{y}} = |\mathbf{Ax}|$, where $\mathbf{A}$ is a $L \times N$ projection matrix, and β is the measurement oversampling factor, typically larger or equal to 4. Note that, as opposed to the first stage, $\mathbf{A}$ now has to get as many non-zero entries as possible, providing global information on the signal $\mathbf{x}$. By substituting $\mathbf{x}$ from (6) and splitting measurement matrix column-wise into K equal $L \times n$ submatrices, $\mathbf{A} = [\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_{K-1}]$, we have

$$\tilde{\mathbf{y}} = |[\mathbf{A}_0 \hat{\mathbf{x}}_0, \mathbf{A}_1 \hat{\mathbf{x}}_1, \ldots, \mathbf{A}_{K-1} \hat{\mathbf{x}}_{K-1}] \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{K-1} \end{bmatrix}| \qquad (7)$$

where $d_i = e^{-j\phi_i}$. This is a quick low dimension PR problem which can be solved by the same algorithm as first step, or a different one, possibly tking into account the fact that the unknown entries d_i are of modulus one. Eventually, from the phase tuning output $\hat{d} = (\hat{d}_0, \hat{d}_1, \ldots, \hat{d}_{K-1})$, the final estimate of $\mathbf{x}$ is

$$\hat{\mathbf{x}} = [d_0 \hat{\mathbf{x}}_0, d_1 \hat{\mathbf{x}}_1, \ldots, d_{K-1} \hat{\mathbf{x}}_{K-1}]^t \qquad (8)$$

The proposed block-based PR algorithm is summarized in Algorithm 1.

Algorithm 1: Generic block-based PR algorithm

Input : $\mathbf{y}$, $\mathbf{H}$, $\tilde{\mathbf{y}}$, $\mathbf{A}$
Output: $\hat{\mathbf{x}}$
Split $\mathbf{H}$ into $\mathbf{H}_0, \mathbf{H}_1, \ldots, \mathbf{H}_{K-1}$ using Definition 1.
for $i = 0$ **to** $K - 1$ **do**
    Solve PR problem $\mathbf{y}_i = |\mathbf{H}_i \mathbf{x}_i|$
Collect $\hat{\mathbf{x}}_i$ estimations as $\tilde{\mathbf{x}}$.
Solve PR problem $\tilde{\mathbf{y}} = |\mathbf{A}\tilde{\mathbf{x}}\mathbf{d}|$.
$\hat{\mathbf{x}} = [d_0\hat{\mathbf{x}}_0, d_1\hat{\mathbf{x}}_1, \ldots, d_{K-1}\hat{\mathbf{x}}_{K-1}]^t$

In the next section, we test this framework using two different PR algorithms : a non-convex minimization approach suited for gaussian entries in the measurement matrix, and a Bayesian PR algorithm designed for binary sensing matrices. However, in general the blocking step may be accomplished by any PR method. Since the K sub-problems in this stage are inherently independent (also called "embarrassingly parallel"), in a fully parallel computing configuration, the block-based PR theoretically yields at least a K^2 factor in computational complexity, and in single-thread sequential computing this factor is equal to K. One should notice that this speedup comes at the cost of a phase adjustment step. In a nutshell, assuming the computational complexity of the base PR algorithm as $O(f(N))$, the parallel block-based PR converts this order into $O(\frac{f(N)}{K^2} + f(K))$. This means that, for state of the art PR algorithms with $f(N) \sim N^2$, and by assuming significantly large N, $O(\frac{f(N)}{K^2} + f(K)) < f(N) : \forall N > K$. In addition to the computational complexity of the underlying PR algorithm, the optimal value of K also depends on the number of available processing units in the blocking step. We discuss this more in the next section.

EXPERIMENTAL RESULTS

Block-based PR with truncated Wirtinger flow

To investigate the performance of the proposed block-based phase retrieval approach, we first employ a recent algorithm based on truncated Wirtinger flow (TWF) [10] to solve the PR sub-problems in the blocking step. The TWF method, currently considered amongst the state-of-the-art for generic PR, has been reported to follow $O(MN)$ computational complexity. Since the number of required measurements, M, grows linearly with the number of input samples, N, this order actually resembles $O(N^2)$. Figure 1 represents the effect of employing the block-based approach to improved TWF algorithm using only $K = 4$ blocks. Here, the input signal, $\mathbf{x}$, and the partitions of K-RBD projection matrix, $\mathbf{H_0}, \ldots, \mathbf{H_3}$, have i.i.d. zero-mean complex random Gaussian entries.
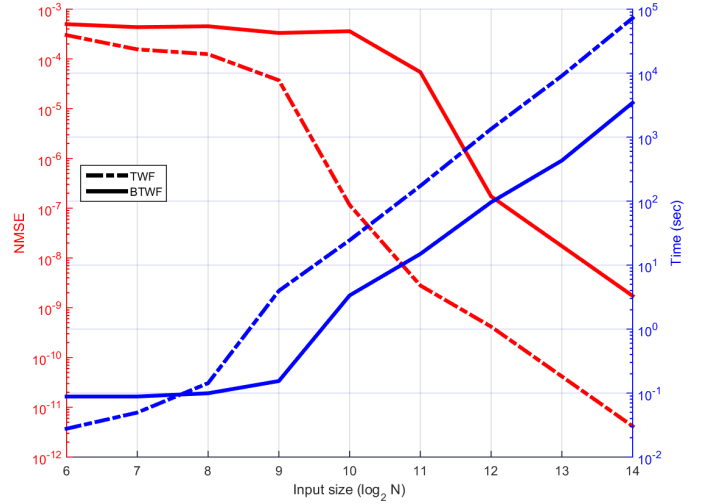

FIG. 1. Comparison of estimation error (in red, left scale) and execution time (in blue, right scale) between TWF (plain lines) and Block-based TWF with $K = 4$ blocks (BTWF - dashed lines), as a function of the input size N.

Gaussian i.i.d. noise is added to the squared magnitude measurements, with SNR=30 dB. In this experiment, $\alpha = 6$ and $\beta = 20$. The blocking step is executed in parallel and a simple PR with alternating projections [27] is employed in the final $K = 4$ dimensional phase tuning step. Each value is the average result over 100 random test inputs, using a 4 cores 3.2 GHz processor with 32 GB of RAM. The performance is measured using the normalized mean square error (NMSE) between original and estimated signals after compensating the global phase shift. The results show up to 20 times speedup with the block-based approach. It has to be noted that this comes at the price of a small loss in precision: local measurements carry information on a smaller number of input coefficients, and are therefore more sensitive to numerical / experimental noise.

As mentioned in the previous section, the optimum number of blocks, K, depends on the computational complexity of the employed PR algorithms in the blocking and phase tuning steps, in addition to the achievable degree of parallelism. Clearly, by increasing K and hence decreasing the block size, we have a faster algorithm in the first step - at the cost of a more complex phase tuning step with K variables.

Beside execution time, another important factor is the estimation error. By increasing the number of blocks in the phase tuning step, the estimation error increases. Suppose we tolerate a 10^{-3} error in terms of NMSE for both the original TWF and its block-based variant. Then, Table I shows the optimal K and the best speedup factor one can achieve using the block-based approach for various input size N. Empirically, the optimal K roughly scales as $N^{0.4}$, which is close to the theoretical prediction

TABLE I. Speedup factor in computation time provided by the block-based PR method, for various input variable size, N. For each N, the best number K of blocks is chosen based on computation time, keeping the relative NMSE below 10^{-3}.

N	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}
K	4	4	8	16	32	64	64
Speed-up	1.2	27	103	343	558	3398	9295

$K^* = \text{argmin}_K c_1 \frac{N^2}{K^2} + c_2 K^2 = c_3 N^{\frac{1}{2}}$, with c_1, c_2 and c_3 as constants.

Block-based Bayesian PR

As a second experiment, we examine the proposed block-based approach on a PR algorithm using binary $\{0,1\}$ measurement matrices. In some physical situations, employing binary projections instead of random complex values makes the measurement setup simpler. However, the corresponding ill-conditioned measurement matrices make PR more challenging, as for instance the TWF typically fails. In [23], the authors suggest a Bayesian-based PR algorithm called prSAMP - for phase retrieval swept approximate message passing. The method which originates from SwAMP [28] and prGAMP [20] algorithms, solves $\mathbf{y} = |\mathbf{Hx} + \mathbf{w}|^2$ problem where $\mathbf{H} \in \{0,1\}^{M \times N}$ is the known binary measurement matrix, $\mathbf{x} \in \mathbb{C}^N$ is the unknown complex signal and $\mathbf{w} \in \mathbb{C}^N$ is the (unknown) noise, assumed i.i.d. complex Gaussian. Even though the algorithm performs well for a real optical imager and strong noise conditions [23], its $O(N^3)$ computational complexity makes it impractical at high dimensions.

Figure 2 compares the execution time of the original prSAMP algorithm and its blocked version at different input sizes $N = 64$ to $65536 = 2^{14}$, for a comparable NMSE. The optimal number of blocks is set using the same approach as in the previous section. As expected, the block-based variant brings very significant speedups to this $O(N^3)$ algorithm. In this experiment, for instance, the block-based prSAMP approach is more than 7000 times faster than the original algorithm at $N = 8192$. Beside the computational complexity, the memory requirement grows as $O(N^2)$ to store the measurement matrix, which in practice is also an important bottleneck. For instance, at $N = 2^{14}$ more than 20 GB of RAM is required to store this matrix in double precision. Due to other temporary variables for AMP messages, the original prSAMP stopped executing at $N = 2^{14}$ on a computer with 32 GB RAM. In reverse, the block-based version could still be run at $N = 2^{16}$ and higher.
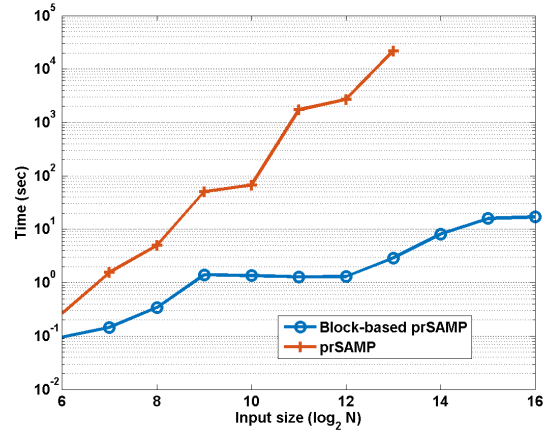

FIG. 2. Computing time (on a log scale) with the prSAMP algorithm, comparing the standard (red) and the block-based approach (blue) using the optimal number of blocks.

CONCLUSION

We have introduced a framework for block-based PR, allowing substantial speed and memory savings for large signals. This comes of course at a price : first, it can only work if one is able to design the measurement matrix in a general block-diagonal manner - this is the case in any physical systems where one can probe the whole object by parts. Then, for a given number of measurements the approximation error is slightly increased. Finally, a small number of extra measurements is needed, but this number scales as the number of blocks, K, and does not depend on the signal dimension, N.

Although, depending on the application, these may be seen as strong limitations, one should be reminded that, due to the fundamentally harsh $O(N^2)$ scaling laws of generic PR, using these block-based PR might not just be a matter of mere computing time : in practice, it may be the only way to achieve PR on very large signals.

ACKNOWLEDGMENT

[1] A. Drémeau, A. Liutkus, D. Martina, O. Katz, C. Schülke, F. Krzakala, S. Gigan, and L. Daudet, Opt. Express **23**, 11898 (2015).
[2] R. Harrison, J. Opt. Soc. Amer. A **10**, 1046 (1993).

[3] O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, D. K. Satapathy, and J. F. Veen, Acta Crystallograph. Sect. A: Found. Crystallogr. **63**, 306 (2007).

[4] A. Walther, Opt. Acta. **10**, 41 (1963).

[5] A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron, Sci. Rep. **4** (2014).

[6] C. Fienup and J. Dainty, Image Recovery: Theory and Application , 231 (1987).

[7] E. J. Candes, T. Strohmer, and V. Voroninski, Communications on Pure and Applied Mathematics **66**, 1241 (2013).

[8] I. Waldspurger, A. d'Aspremont, and S. Mallat, Mathematical Programming **149**, 47 (2015).

[9] E. J. Candes, X. Li, and M. Soltanolkotabi, IEEE Trans. on Info. Theory **61**, 1985 (2015).

[10] Y. Chen and E. Candes, in *Advances in Neural Information Processing Systems* (2015) pp. 739–747.

[11] R. W. Gerchberg, Optik **35**, 237 (1972).

[12] J. R. Fienup, Optics letters **3**, 27 (1978).

[13] S. Marchesini, Review of scientific instruments **78**, 011301 (2007).

[14] P. Netrapalli, P. Jain, and S. Sanghavi, IEEE Trans. on Signal Proc. **63**, 4814 (2015).

[15] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, SIAM Journal on Imaging Sciences **7**, 35 (2014).

[16] B. G. Bodmann and N. Hammen, Advances in computational mathematics **41**, 317 (2015).

[17] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, in *Optical Engineering+ Applications* (International Society for Optics and Photonics, 2007) pp. 670120–670120.

[18] S. Mukherjee and C. S. Seelamantula, IEEE Trans. on Signal Proc. **62**, 4659 (2014).

[19] Y. Shechtman, A. Beck, and Y. C. Eldar, IEEE Trans. on Signal Proc. **62**, 928 (2014).

[20] P. Schniter and S. Rangan, IEEE Trans. Signal Proc. **63**, 1043 (2015).

[21] M. Iwen, A. Viswanathan, and Y. Wang, arXiv preprint arXiv:1501.02377 (2015).

[22] X. Zhang and P. Kner, in *SPIE BiOS* (International Society for Optics and Photonics, 2015) pp. 93350U–93350U.

[23] B. Rajaei, E. W. Tramel, S. Gigan, F. Krzakala, and L. Daudet, arXiv preprint arXiv:1510.01098 (2015).

[24] L. Tian and L. Waller, Optica **2**, 104 (2015).

[25] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, Signal Processing Magazine, IEEE **32**, 87 (2015).

[26] R. Horisaki, R. Egami, and J. Tanida, Optics express **24**, 3765 (2016).

[27] H. H. Bauschke, P. L. Combettes, and D. R. Luke, JOSA A **19**, 1334 (2002).

[28] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, arXiv preprint arXiv:1406.4311 (2014).