CONVERGENCE AND ERROR ANALYSIS OF DIFFRACTION FIELD ITERATIVE NON-UNIFORM SAMPLING SCHEMES

Vladislav Uzunov, Atanas Gotchev, Karen Egiazarian

Department of Signal Processing, Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, FINLAND

ABSTRACT

Reconstruction of a light field from a set of non-uniformly distributed samples is important for holographic type of 3D scene display. The non-uniform sampling problem has been already addressed and rapidly convergent iterative schemes developed, mainly based on the method of Projection Onto Convex Sets (POCS) and matrix inversion by the method of Conjugate Gradients (CG). This paper presents a derivation of the convergence rates for both of the schemes and analyzes the factors which influence the error decay. The presence of clusters and gaps in the distribution of the given samples turns out to have the most significant influence on the convergence. Based on the analysis, a regularization method is derived for the CG-based reconstruction scheme which proves to be powerful even when the local sample density estimate is rough. The POCS-based scheme is generalized for the case of off-grid known samples. Its error analysis leads to similar conclusions for the convergence impact factors as in the CG-based scheme. In addition, the analysis suggests scenarios where the POCS excels, and others where the CG is much better.

1. INTRODUCTION

The computation of the light field over the entire three-dimensional (3D) space from an abstract 3D scene representation is known as the forward problem in holography. Holographic 3D display systems use the light field information to display a scene, and thus benefit from efficient techniques which solve the forward problem. The development of such techniques is a challenging task, since 3D scenes consist of various shapes and shades. In a general setting, the 3D scene information can be assumed to be available as non-uniformly distributed in space light field points. Thus the forward problem can be formulated as to reconstruct a light field from irregularly distributed samples.

A real-time application needs to reconstruct the field for large data sample sets in fast and accurate manner. Previously, we have developed iterative reconstruction schemes, based mainly on the method of Projection Onto Convex Sets (POCS) [1], and matrix inversion through the Conjugate Gradients (CG) method [2]. The POCS-based scheme assumes that the known field samples belong to a pre-defined regular grid, and utilizes an $O(N \log\ N)$-complex propagation algorithm to iterate between successive lines which define the convex sets. The CG-based scheme builds a discrete model of the light field which represents the reconstruction task as matrix inversion problem, and solves it iteratively through CG. However, for certain sample distributions the involved matrix has high condition number and inversion with CG converges slowly.

The main goal of this paper is to perform a theoretical analysis of the convergence rate of the POCS- and CG-based reconstruction schemes. We aim at identifying and analyzing the main factors which influence the convergence and the computational efficiency, and to illustrate their impact. The results can be used to achieve faster convergence and/or better computational efficiency.

2. PROBLEM FORMULATION

In this section, we formulate the problem of reconstruction of a diffraction field from its sampled version, where the sampling coordinates are given on a non-uniform grid. We start with a brief review of diffraction theory equations we need, then we specify a finite dimensional model of diffraction to be used in the reconstruction.

2.1. A finite-dimensional model of diffraction

Diffraction theory studies phenomena occurring during propagation and reflection of light within, and at the borders of different physical media. In general, these phenomena are described through complex-valued functions representing light (optical) wave fields and their distribution across space. In the general case, these are vector functions confined to the Maxwell equations [3]. When the medium is free space and the light source is monochromatic with wavelength λ, the corresponding light field is accurately described by a scalar function. Under the conditions of linearity, isotropy and homogeneity of the free space, the light field at any spatial point can be related to that on a plane through the Rayleigh-Sommerfeld diffraction integral [4]. The same integral can be used to compute the optical field on a plane given the field on another parallel plane, having linear and shift-invariant relationship between the two functions [4, 5]. Under the assumption that the distance r between the parallel planes satisfies $r >> \lambda$, the Rayleigh-Sommerfeld diffraction inte-

This work was supported by the Academy of Finland, project No. 213462 (Finnish Centre of Excellence program (2006 - 2011).

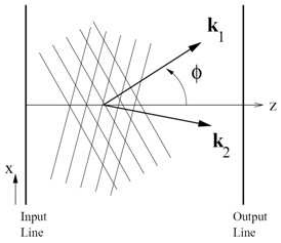

Figure 1. The vectors $\mathbf{k}_1 = [k_{1x}, k_{1z}]$ and $\mathbf{k}_2 = [k_{2x}, k_{2z}]$ are the wave vectors of the plane waves. The field at the output line is equal to the superposition of the plane waves illuminated on the input line according to Eq. 1 [1].

gral can be regarded as an equivalent to the so-called *plane wave decomposition* approach [4, 6, 7].

In the rest of the paper, the plane wave decomposition approach is used. For the sake of simplicity, the discussion is restricted to one transverse dimension only. The diffraction field at any point in the 2D free space is computed as:

$$u(x,z) = \int_{-\frac{2\pi}{\lambda}}^{\frac{2\pi}{\lambda}} a(k_x) e^{j(k_x x + k_z z)} dk_x, \qquad (1)$$

where $u(x,z)$ is the field over 2D space, the x axis is the transverse axis and the z axis is the optical axis along which the field propagates. This integral represents the light field $u(x,z)$ as a superposition of plane waves $e^{j(k_x x + k_z z)}$, where each plane wave is oriented at an angle determined by the wave vector $\mathbf{k} = [k_x, k_z]$, as illustrated in Figure 1. The variable k_z is related to k_x by $k^2 = k_z^2 + k_x^2$ where $k = \frac{2\pi}{\lambda}$ is the length of each wave vector. In the superposition integral each plane wave contributes with complex amplitude $a(k_x)$. This superposition of plane waves bears another interpretation, which is more convenient for our purpose. The variables k_x and k_z are the spatial frequencies of the propagating plane waves along the x and z axes respectively. The angles of the plane waves determine different spatial frequencies k_x along the transversal dimension x. Therefore, different plane waves determine harmonics with different frequencies k_x in the transverse plane. Consequently, the complex amplitudes $a(k_x)$ can be considered as the Fourier spectrum of the of the field $u(x,0)$ on a reference line with $z = 0$. A propagating monochromatic wave with wavelength λ cannot have a harmonic component in the transverse plane with higher frequency than $\frac{2\pi}{\lambda}$ since $k_x = k\cos\phi$ (cf. Figure 1). Therefore, in the case of free space propagation of monochromatic light, the function on the reference line $u(x,0)$ is assumed band-limited. This also determines the limits of the integral in Eq. 1.

To elucidate the role of the function $a(k_x)$, let us consider the plane waves in Eq. 1 split as $e^{j(k_x x + k_z z)} = e^{j\sqrt{k^2 - k_x^2} z} e^{j k_x x}$. The plane wave decomposition integral takes the form:

$$u(x,z) = \int_{-\frac{2\pi}{\lambda}}^{\frac{2\pi}{\lambda}} a(k_x) e^{j\sqrt{k^2 - k_x^2} z} e^{j k_x x} dk_x. \qquad (2)$$

Now interpret Eq. 2 as inverse Fourier integral where the function $a(k_x)_{z_i} = a(k_x) e^{j\sqrt{k^2 - k_x^2} z_i}$ is the Fourier transform of the field $u(x, z_i)$ at a line slice $z = z_i$. In other words, the field $u(x,z)$ at any particular line $z = z_i$ can be computed from the field on the initial line $u(x,0)$ by three consecutive steps: (1) Fourier transform to find $a(k_x)$; (2) Fourier-domain multiplication by the transfer function of the free space $e^{j\sqrt{k^2 - k_x^2} z}$; and (3) inverse Fourier transform to get back to spatial domain. This interpretation has both theoretical and numerical importance. Theoretically, it clearly shows the dependence between input, system and output. Numerically, it suggests the utilization of an FFT algorithm, under proper sampling [1].

According to Eq. 1, every point of the field $u(x,z)$ can be expressed in terms of the spectrum $a(k_x)$ of the field $u(x,0)$ on the reference line. Consider an *essentially* space-limited function $u(x,0)$ or, more precisely, a function whose space - bandwidth product is finite, and denote the spatial extent of interest by T. Such a function can be periodized, which is equivalent to discretization of $a(k_x)$. A periodic and bandlimited function is equivalently represented by a finite number (M) of discrete spectral components. That is, the discrete values of the frequency k_x are :

$$k_x = \frac{2\pi m}{T}, \qquad (3)$$

where $m = -\lfloor \frac{M}{2} \rfloor, ..., \lfloor \frac{M-1}{2} \rfloor$. Substituting k_x as given by Eq. 3 in Eq. 2 leads to a *finite-dimensional* model of diffraction:

$$u(x,z) = \sum_{m=-\lfloor \frac{M}{2} \rfloor}^{\lfloor \frac{M-1}{2} \rfloor} a_m e^{j\frac{2\pi}{T}\sqrt{\frac{T^2}{\lambda^2} - m^2} z} e^{j\frac{2\pi}{T} m x}, \qquad (4)$$

where $a_m = a(\frac{2\pi}{T} m)$ are the coefficients of the Fourier series expansion of $u(x,0)$. This model assumes the field at the reference line $z = 0$ is a trigonometric polynomial of order M, which is easily transferable to any other z coordinate thought the transfer function $e^{j\sqrt{k^2 - k_x^2} z}$.

2.2. Non-uniformly sampled diffraction fields

The finite-dimensional model in Eq. 4 can be regarded as a function expansion with respect to M kernels formed as product of exponents along x axis and chirps along z axis, weighted by the coefficients a_m. These M generators determine any particular monochromatic field with wavelength λ and spatial extend of the reference line T. In the general case of irregular sampling, the diffraction field $u(x,z)$ is given at a finite set of s sampling points $\{(x_i, z_i)\}_{i=1}^{s}$. The irregular sampling and reconstruction problem can be stated as to find the unknown field generating coefficients a_m, given the samples $\{u(x_i, z_i)\}_{i=1}^{s}$.

Eq. 4 can be written for each point in the irregular sampling set $\{(x_i, z_i)\}_{i=1}^{s}$:

$$u(x_i, z_i) = \sum_{m=-\lfloor \frac{M}{2} \rfloor}^{\lfloor \frac{M-1}{2} \rfloor} a_m e^{j\frac{2\pi}{T}\sqrt{\beta^2 - m^2}\, z_i} e^{j\frac{2\pi}{T} m x_i}, \quad (5)$$

for $i = 1, ..., s$.

These s equations form a linear system for the M unknowns a_m. Speaking linear algebra, this system will have a solution if it is of full rank. To characterize the stability of the sampling set with respect to the field generators, we are interested in having a sampling set bounded by the energy of the field at the first line, as

$$A \|a\|^2 \leq \sum_{i=1}^{s} |u(x_i, z_i)|^2 \leq B \|a\|^2 . \quad (6)$$

Such set of sampling points $\{(x_i, z_i)\}_{i=1}^{s}$ which has bounded energy for any choice of generating coefficients a_m with strictly positive bounds A and B is referred to as *stable sampling set* [8].

In the following we describe and analyze two different approaches to find the coefficients a_m. The first uses iterative algorithm to solve the system of Eq. 5 [2]. The second algorithm iterates from given point to given point and falls in the framework of the method of POCS [9, 10].

3. RECONSTRUCTION BY ITERATIVE MATRIX INVERSION

3.1. Method description

The system in Eq. 5 is linear and can be expressed in a matrix form:

$$\mathbf{u} = \mathbf{R}\mathbf{a}, \quad (7)$$

where $\mathbf{a} = [a_{-\lfloor \frac{M}{2} \rfloor}, a_{-\lfloor \frac{M}{2} \rfloor + 1}, ..., a_{\lfloor \frac{M-1}{2} \rfloor}]^T$ is the unknown vector of the field generating coefficients and $\mathbf{u} = [u(x_1, z_1), u(x_2, z_2), ..., u(x_s, z_s)]^T$ is the vector of given samples. $\mathbf{R}$ is the reconstruction matrix

$$\begin{aligned} \mathbf{R} &= \{r_{p,q}\} \\ &= \{e^{j\frac{2\pi}{T}\sqrt{\beta^2 - (q - \lfloor \frac{M}{2} \rfloor + 1)^2}\, z_p} e^{j\frac{2\pi}{T}(q - \lfloor \frac{M}{2} \rfloor + 1) x_p}\}, \\ &\quad p = 1, ..., s, q = 1, ..., M, \end{aligned} \quad (8)$$

The diffraction field at the point (x_i, z_i) equals the inner product of the i-th row of $\mathbf{R}$ with $\mathbf{a}$.

We need a fast and numerically stable algorithm to find (approximate) solution $\hat{\mathbf{a}}$ for the unknown vector $\mathbf{a}$ and we limit the consideration to the over-determined case $s \geq M$.

Expressing the residual between the true vector and its approximation

$$\mathbf{g} = \mathbf{u} - \mathbf{R}\hat{\mathbf{a}}, \quad (9)$$

and minimizing the L_2 norm of this residual $\|\mathbf{u} - \mathbf{R}\hat{\mathbf{a}}\|_2$, one ends with solving a *Least Squares (LS)* problem. Reaching an LS solution goes through finding the (pseudo-)inverse of the matrix $\mathbf{R}$, which operation is of cubic complexity in the general case [11]. We opt for the conjugate gradient method (CG), known as one of the most rapidly convergent and numerically stable algorithms for solving LS problems iteratively [12]. However, it requires a Hermitian and positive definite matrix. When the matrix is rectangular, as in our case, we can consider the following equivalent matrix equation that will produce the same solution as Eq. 7:

$$\mathbf{R}^H \mathbf{R}\mathbf{a} = \mathbf{R}^H \mathbf{u}. \quad (10)$$

Now the matrix to be solved by CG is the Hermitian matrix $\mathbf{R}^H\mathbf{R}$. The residual whose norm is minimized by iterating CG in this case is measured as:

$$\mathbf{g} = \mathbf{R}^H \mathbf{u} - \mathbf{R}^H \mathbf{R}\hat{\mathbf{a}}. \quad (11)$$

CG applied on the matrix $\mathbf{R}^H\mathbf{R}$ has a form, where the matrix $\mathbf{R}^H\mathbf{R}$ is never explicitly computed. This form is called CG method on normal equations (CGN) and is outlined below [12]:

1. initialize $\mathbf{b} = \mathbf{R}^H \mathbf{u}\, \hat{\mathbf{a}}_0$ arbitrary, $\mathbf{g}_0 = \mathbf{b} - \mathbf{R}^H\mathbf{R}\hat{\mathbf{a}}_0$ and $\mathbf{d}_0 = \mathbf{g}_0$
2. for $n = 1$ to n_{it}
   (a) $\alpha = \frac{\mathbf{g}_n^H \mathbf{g}_n}{\mathbf{d}_n^H \mathbf{R}^H \mathbf{R} \mathbf{d}_n}$
   (b) $\hat{\mathbf{a}}_{n+1} = \hat{\mathbf{a}}_n + \alpha \mathbf{d}_n$
   (c) $\mathbf{g}_{n+1} = \mathbf{b} - \mathbf{R}^H\mathbf{R}\hat{\mathbf{a}}_{n+1}$
   (d) $\gamma = \frac{\mathbf{g}_{n+1}^H \mathbf{g}_{n+1}}{\mathbf{g}_n^H \mathbf{g}_n}$
   (e) $\mathbf{d}_{n+1} = \mathbf{g}_{n+1} + \gamma \mathbf{d}_n$
   end
3. reconstruct the diffraction field $u(x, z)$ from the estimated field generating coefficient vector $\hat{\mathbf{a}}$ with Eq. 4.

Any iterative algorithm for solving LS problems builds the solution step by step, updating the solution vector each time with a small portion (α) along some search direction $\mathbf{d}_n$ (step 2b). The basic idea of CG is to build the search directions $\mathbf{d}_n$ conjugate to each other, so that after at most M steps the solution will be found. By conjugate is meant that the directions are orthogonal to each other, where the orthogonality is measured with respect to the $\mathbf{R}^H\mathbf{R}$ matrix of the LS problem - $\mathbf{d}_n^H \mathbf{R}^H \mathbf{R} \mathbf{d}_k = 0$. The value of α is chosen in such a manner that the current error $\mathbf{e}_{n+1} = \hat{\mathbf{a}}_{n+1} - \mathbf{a}$ is conjugate to the previous direction $\mathbf{d}_n$ (step 2a). This makes the residual $\mathbf{g}_{n+1}$ orthogonal to all previous search directions. The new direction $\mathbf{d}_{n+1}$ is build from this residual $\mathbf{g}_{n+1}$ as to be conjugate to all previous directions (step 2d-e).

3.2. Convergence analysis

This subsection presents the derivation of the approximation error at the n-th CG iteration in terms of the initial error. Then, factors influencing the final expression can be analyzed with respect to different sampling distributions. The analysis can be used as a base to decrease the impact of the distributions on the convergence rate.

Write the error vector $\mathbf{e}_{n+1}$ in the form:

$$\begin{aligned}\mathbf{e}_{n+1} &= \mathbf{a}_{n+1} - \mathbf{a} = \mathbf{a}_n + \alpha_n \mathbf{d}_n - \mathbf{a} = \mathbf{e}_n + \alpha_n \mathbf{d}_n \\ &= \mathbf{e}_{n-1} + \alpha_{n-1}\mathbf{d}_{n-1} + \alpha_n \mathbf{d}_n = ... \\ &= \mathbf{e}_0 + \sum_{k=0}^{n} \alpha_k \mathbf{d}_k. \end{aligned} \quad (12)$$

This equation shows that $\mathbf{e}_{n+1}$belongs to a space $\mathbf{e}_0 + D_{n+1}$, where $D_{n+1} = span\{\mathbf{d}_0, \mathbf{d}_1, ..., \mathbf{d}_n\}$. The search directions are built from the residuals (step 2e) and therefore $D_{n+1} = span\{\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{g}_n\}$. Now denote $\mathbf{Q} = \mathbf{R}^H\mathbf{R}$ and write each residual $\mathbf{g}_n$ in the form:

$$\begin{aligned}\mathbf{g}_n &= -\mathbf{Q}\mathbf{e}_n = -\mathbf{Q}(\mathbf{e}_{n-1} + \alpha_{n-1}\mathbf{d}_{n-1}) \\ &= \mathbf{g}_{n-1} - \alpha_{n-1}\mathbf{Q}\mathbf{d}_{n-1}. \end{aligned} \quad (13)$$

Recall that $\mathbf{d}_{n-1} \in D_n$ and therefore each new subspace D_{n+1} is constructed from D_n and the subspace $\mathbf{Q}D_n$. Hence,

$$\begin{aligned}D_n &= span\{\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \mathbf{Q}^2\mathbf{g}_0, ..., \mathbf{Q}^{n-1}\mathbf{g}_0\} \\ &= span\{\mathbf{Q}\mathbf{e}_0, \mathbf{Q}^2\mathbf{e}_0, \mathbf{Q}^3\mathbf{e}_0, ..., \mathbf{Q}^n\mathbf{e}_0\}. \end{aligned} \quad (14)$$

The subspaces D_n are known as Krylov subspaces. Recall that the error $\mathbf{e}_n$ belongs to the space $\mathbf{e}_0 + D_n$. Then it can be expressed as a linear combination of the spanning elements $\mathbf{Q}^i\mathbf{e}_0$ of this subspace:

$$\mathbf{e}_n = \left(I + \sum_{i=1}^{n} \phi_i \mathbf{Q}^i\right)\mathbf{e}_0 = P_n(\mathbf{Q})\mathbf{e}_0. \quad (15)$$

$P_n(\mathbf{Q})$ is the polynomial of order n from the parenthesis of the above expression. If $\mathbf{v}$ is an eigenvector of $\mathbf{Q}$ with respective eigenvalue λ, then $\mathbf{Q}^i\mathbf{v} = \lambda^i\mathbf{v}$ and therefore $P_n(\mathbf{Q})\mathbf{v} = P_n(\lambda)\mathbf{v}$. Since the matrix $\mathbf{Q}$ is Hermitian, the eigenvectors can be chosen to form an orthonormal set and the error $\mathbf{e}_0$ can be expressed as a linear combination of these eigenvectors. Inserting this linear combination in the last form of Eq. 15, $\mathbf{e}_n$ can be expressed as similar linear combination of the eigenvectors, where the coefficients of this combination are multiplied by $P_n(\lambda_i)$. Now the $\mathbf{Q}$-norm of the error can be written as [13]:

$$\|\mathbf{e}_n\|_{\mathbf{Q}} \leq \min_{P_n} \max_{\lambda \in \Lambda(\mathbf{Q})} P_n^2(\lambda) \|\mathbf{e}_0\|_{\mathbf{Q}}. \quad (16)$$

CG determines alpha in such a manner which minimizes $\|\mathbf{e}_n\|_{\mathbf{Q}}$ within the space $D_n + \mathbf{e}_0$ (step 2a), or, in other words CG finds the polynomial $P_n(\lambda)$ which minimizes the expression in Eq. 16. However, the convergence is determined by the eigenvalue which gives maximum value of $P_n(\lambda)$.

Since our goal is to analyze and estimate the error reduction after n iterations of CG, we need to find the polynomial $P_n(\lambda)$ explicitly. This means that the coefficients ϕ_i in Eq. 15 must be known. However, they depend on the coefficients α_i and γ_i which are not available before the algorithm is run. More general approach is to find a polynomial which minimize Eq. 16 over the interval $[\lambda_{min}, \lambda_{max}]$ between the smallest and largest eigenvalue, rather than on particular set of eigenvalues $\Lambda(\mathbf{Q})$. Such a polynomial is known to be a relation of two Chebyshev polynomials T_n of order n [12]:

$$P_n(\lambda) = \frac{T_n\left(\frac{\lambda_{max}+\lambda_{min}-2\lambda}{\lambda_{max}-\lambda_{min}}\right)}{T_n\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)}. \quad (17)$$

The polynomial in the numerator has maximal value of 1 inside the interval $[\lambda_{min}, \lambda_{max}]$ and therefore the error $\mathbf{e}_n$ can be estimated as

$$\begin{aligned}\|\mathbf{e}_n\|_{\mathbf{Q}} &\leq \left[T_n\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)\right]^{-1} \|\mathbf{e}_0\|_{\mathbf{Q}} \\ &= \left[T_n\left(\frac{\kappa+1}{\kappa-1}\right)\right]^{-1} \|\mathbf{e}_0\|_{\mathbf{Q}} \\ &= 2\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n\right]^{-1} \|\mathbf{e}_0\|_{\mathbf{Q}}, \end{aligned} \quad (18)$$

where $\kappa = \lambda_{max}/\lambda_{min}$ is the condition number of the matrix $\mathbf{Q}$. The second addend inside the square brackets converges to zero as n grows, so it is common to estimate the error with the weaker inequality

$$\|\mathbf{e}_n\|_{\mathbf{Q}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \|\mathbf{e}_0\|_{\mathbf{Q}}. \quad (19)$$

This equation suggests that the convergence of the CG algorithm highly depends on the condition number of the matrix $\mathbf{Q}$. To relate κ to the bounds a stable sampling set (cf. Eq. 6), the energy of the known samples is can be expressed as:

$$\|\mathbf{u}\|_2^2 = \mathbf{u}^H\mathbf{u} = \mathbf{a}^H\mathbf{R}^H\mathbf{R}\mathbf{a} = \mathbf{a}^H\mathbf{Q}\mathbf{a}. \quad (20)$$

For a stable sampling set the energy of the samples is strictly positive and thus the matrix $\mathbf{Q}$ is strictly positive definite, i.e. its eigenvalues are always positive. Since the energy of the samples is bounded for all vectors $\mathbf{a}$, consider the case when it is an eigenvector $\mathbf{v}$ of $\mathbf{Q}$:

$$\begin{aligned} A\|\mathbf{v}\|_2^2 \leq \mathbf{v}^H\mathbf{Q}\mathbf{v} \leq B\|\mathbf{v}\|_2^2, or \\ A\|\mathbf{v}\|_2^2 \leq \lambda\mathbf{v}^H\mathbf{v} \leq B\|\mathbf{v}\|_2^2. \end{aligned} \quad (21)$$

This last equation implies that $A \leq \lambda \leq B$, for all eigenvalues of $\mathbf{Q}$. Therefore $A \leq \lambda_{min}$ and $\lambda_{max} \leq B$ and consequently the condition number of $\mathbf{Q}$ can be estimated as $\kappa(\mathbf{Q}) \leq B/A$. The constants A and B are determined from the reconstruction function and the sampling set. Suppose that the coefficients $\mathbf{a}$ determine a light field which is a beam centered along the z-axis, and has essentially finite extent along x. Suppose further that the

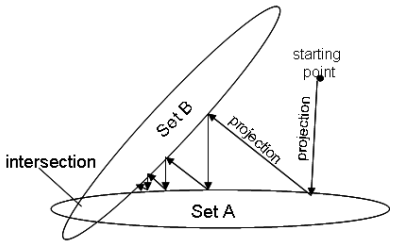

Figure 2. Projection Onto Convex Sets

samples do not cover the extent of the beam, leaving a large gap. The energy of such samples will be very small in comparison with the energy of $\mathbf{a}$, no matter how dense they are outside the gap. This pushes the lower bound A from Eq. 6 down and consequently increases $\kappa(\mathbf{Q})$. Now consider the case when we have a dense cluster of samples lying on the beam. The same local information will be added several times in Eq. 6, which will increase B and $\kappa(\mathbf{Q})$. Thus, the presence of clusters and gaps in the density of the sampling set has significant influence on the condition number of $\mathbf{Q}$.

An add-hoc approach to compensate for the clustering is to weight different samples according to their density. Samples from areas with higher density are assigned with small weights and samples from less populated areas are emphasized as more important with higher weights. This weighting can be represented in a matrix form as multiplying the known samples vector $\mathbf{u}$ with a diagonal weighting matrix $\mathbf{W}$. Then, Eq. 7 is modified as follows:

$$\mathbf{Wu} = \mathbf{WRa}. \tag{22}$$

CG can be used to solve this LS problem by further multiplying both sides by $\mathbf{R}^H$ to obtain a symmetric matrix $\mathbf{R}^H\mathbf{WRa}$ and target vector $\mathbf{R}^H\mathbf{Wu}$.

4. RECONSTRUCTION BY POCS

As a second approach we consider an iterative technique, developed based on the POCS method. POCS is a computational approach for finding an element of a feasible region defined by the intersection of a number of convex constraints, starting with an arbitrary infeasible point [14, 15]. Figure 2 illustrates how convergence to the intersection is achieved by iterative projections onto the individual convex sets. In our earlier work [1], POCS was used to reconstruct a diffraction field from a set of irregularly distributed samples, which belong to a predefined uniform grid. Here, the same algorithm is reformulated to serve the general case when the given field samples have random positions, unrelated to any particular uniform grid.

4.1. Method description

The POCS-based reconstruction method from [1] considers the known data samples as the constraints that determine the convex sets. Naturally, a convex set is defined as all possible diffraction fields that have the given data points on a certain line. This is beneficial only in the case these points belong to a uniform grid with sampling interval T/M, and aligned with the starting sampling grid on the reference line. In such a case, the algorithm makes use of an FFT-based computation of diffraction field from line to line. This was the key-point of the algorithm developed in [1]. There, the algorithm was based on iterating from line to line efficiently by the use of FFT, and re-substituting the values of the known data points at each line, as they were always on positions of an M-point regular grid. However, in the case of no structured samples such an iteration is not possible. Consider, e.g. the case of single point per line, or the case of dense points not aligned with respect to the starting grid.

Here, we suggest a modification to overcome these difficulties. We still want to benefit from using FFT. However, instead of propagating from line to line, we propagate the *new information* from each particular convex set $\mathbf{C}_l, l = 1...L$ to the unknown coefficients a_m. In order to use an M-point IDFT for this propagation,we define the set $\mathbf{C}_l$ as all known samples $u(x_i, z_i)$ with $z_i = z_l$, which can be hosted by an M-point uniform grid $\mathbf{x}_l$. In the general case there might be no more than one sample per line to be considered as belonging to the set $\mathbf{C}_l$. In such a case, the number of convex sets L will coincide with the number of sampling points s and each projection will result to an update caused by a single point. However, if two or more points per line happen to belong to uniform grids, we should benefit from this. Again, in the general case each such grid $\mathbf{x}_l$ will be not aligned to the initial grid and can be assumed centered at some point $\chi_l \neq 0$ and encompassing a spatial interval of length T along x:

$$\mathbf{x}_l = \{(\mathbf{x}_l)_k\}_{k=0}^{M-1} = \chi_l + \{-T/2 + kT/M\}_{k=0}^{M-1} \tag{23}$$

Now the formal definition of a set $\mathbf{C}_l, l = 1...L$ can be stated as:

$$\begin{aligned} \mathbf{C}_l &= \{\forall f(x,z) \in \mathbf{DF}_M : f(\mathbf{p}_l) = u(\mathbf{p}_l)\}, \\ \mathbf{p}_l &= \{(x_i, z_i), i = 1...s : z_i = z_l, x_i \in \mathbf{x}_l\}, \end{aligned} \tag{24}$$

where $\mathbf{p}_l$ is the set of points which are hosted by the grid $\mathbf{x}_l$ and $\mathbf{DF}_M$ denotes the space of diffraction fields, generated by M nonzero coefficients through Eq. 4.

To complete the POCS method, the projections onto the sets have to be determined. Consider a field $f(x, y)$, generated through Eq. 4 by coefficients $a_m^{(f)}$. Its projection P_l onto $\mathbf{C}_l$ can be defined as substituting its samples $f(\mathbf{p}_l)$ by the known samples $u(\mathbf{p}_l)$ which define the set, and reflecting this substitution onto the coefficients $a_m^{(f)}$. The substitution can be represented as just addition of the difference $u - f$ at the positions $\mathbf{p}_l$ of the samples which define the set:

$$P_l f(x,z) = f(x,z) + \sum_{(x_i,z_i)\in\mathbf{p}_l} (u-f)(x_i,z_l)\delta(x-x_i, z-z_l), \tag{25}$$

where $\delta(x, z)$ is a Kronecker delta. The substitution acts only on line $z = z_l$ and on points $\mathbf{p}_l$ which belong to

the grid $\mathbf{x}_l$. It is sufficient to propagate the information from the grid $\mathbf{x}_l$ back to the coefficients f_m. According to Eq. 4, the Fourier series coefficients of the field $u(x,z)$ at $z = z_l$ are $a_m e^{j\frac{2\pi}{T}\sqrt{\frac{T^2}{\lambda^2}-m^2}\,z_l}$. These coefficients are related through M-point DFT with M regular samples on a grid $\mathbf{x}_0$, centered at the origin. This can be represented in a matrix form as:

$$\mathbf{H}_{z_l}\mathbf{a} = \mathbf{F}u(\mathbf{x}_0, z_l), \tag{26}$$

where $\mathbf{H}_{z_l}^{-1} = diag(e^{j\frac{2\pi}{T}\sqrt{\frac{T^2}{\lambda^2}-m^2}\,z_l})$ and $\mathbf{F}$ is the DFT matrix. Having a sampling grid $\mathbf{x}_l$ for a set $\mathbf{C}_l$ means it is shifted from the origin by χ_l.Alternatively, we can assume shifting the field $u_{\chi_l}(x,z) = u(x+\chi_l, z)$ prior to sampling on $\mathbf{x}_0$. Eq. 4 performs this operation by modulation: $a_m e^{j\frac{2\pi}{T}m\chi_l}$. This changes Eq. 26 to:

$$\mathbf{H}_{z_l}\mathbf{E}_{\chi_l}\mathbf{a} = \mathbf{F}u(\mathbf{x}_l, z_l), \tag{27}$$

where $\mathbf{E}_{\chi_i} = diag(e^{j\frac{2\pi}{T}m\chi_i})$. Eventually,

$$\mathbf{a} = \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}u(\mathbf{x}_l, z_l) \tag{28}$$

is the equation which propagates information from the field line $u(x,z_l)$, sampled with a regular grid $\mathbf{x}_l$, to the coefficients a_m. Denote by $\mathbf{f}$ an s-dimensional vector which contains the samples $f(x_i,z_i), i = 1...s$, ordered in the same way as the known samples $u(x_i,z_i)$ in the vector $\mathbf{u}$ from Eq. 7. The difference $u-f$ sampled at the grid $\mathbf{x}_l$ can be represented with the following equation:

$$(u-f)(\mathbf{x}_l, z_l) = \mathbf{S}_l(\mathbf{u}-\mathbf{f}), \tag{29}$$

where $\mathbf{S}_l$ is a $M \times s$ permutation matrix which takes the samples $(u-f)(x_i,z_i), (x_i,z_i) \in \mathbf{p}_l$ from the difference vector $\mathbf{u}-\mathbf{u}_f$ and positions them properly on the grid $\mathbf{x}_l$. Thus $\mathbf{S}_l$ has value 1 on positions (k,i), where i is such that $(x_i,z_i) \in \mathbf{p}_l$ and k are the positions of these samples on the grid $\mathbf{x}_l$.

Eq. 28 written for the sampled difference $(u-f)(\mathbf{x}_l,z_l)$ from Eq. 29 describes the back-propagation of the difference information to the field generating coefficients $a_m^{(f)}$, ordered in a vector $\mathbf{a}^{(f)}$. Consequently, the projection P_l can be re-defined to act on the field generating coefficient vector $\mathbf{a}^{(f)}$:

$$P_l\mathbf{a}^{(f)} = \mathbf{a}^{(f)} + \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\mathbf{S}_l(\mathbf{u}-\mathbf{f}). \tag{30}$$

The POCS-based reconstruction algorithm proceeds as follows:

1. initialize $\hat{\mathbf{a}}_{0,0}$ arbitrary
2. for $n = 1$ to n_{it}
   for $l = 1$ to L
   (a) predict the field at the points of the given samples $\{(x_i,z_i)\}_{i=1}^{s}$ as $\hat{\mathbf{u}}^{(n,l)} = \mathbf{R}\hat{\mathbf{a}}_{n,l}$.
   (b) calculate $\hat{\mathbf{a}}_{n,l+1} = P_l\hat{\mathbf{a}}_{n,l}$ by projecting the predicted field $\hat{u}^{(n,l)}(x,z)$ through Eq. 30
   end
3. reconstruct the diffraction field $u(x,z)$ from the recovered field generating coefficient vector $\hat{\mathbf{a}}$ with Eq. 4.

4.2. Convergence analysis

We attempt on deriving an expression for the error obtained at the n-th iteration in terms of the error at previous iterations. Such expression can be analyzed to find factors which influence the decay of the error norm through the iterations, and the impact of these factors.

The error vector $\mathbf{e}_{n,l+1}$ which is obtained at the n-th iteration, after projecting on the set $\mathbf{C}_l$ can be derived as follows:

$$\begin{aligned}
\mathbf{e}_{n,l+1} &= \mathbf{a} - \hat{\mathbf{a}}_{n,l+1} = \mathbf{a} - P_l\hat{\mathbf{a}}_{n,l} \\
&= \mathbf{a} - \hat{\mathbf{a}}_{n,l} - \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\mathbf{S}_l(\mathbf{u}-\hat{\mathbf{u}}^{(n,l)}) \\
&= \mathbf{e}_{n,l} - \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\mathbf{S}_l\mathbf{R}(\mathbf{a}-\hat{\mathbf{a}}_{(n,l)}) \\
&= \left(\mathbf{I} - \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\mathbf{S}_l\mathbf{R}\right)\mathbf{e}_{(n,l)}.
\end{aligned} \tag{31}$$

In order to ensure convergence, that is $\|\mathbf{e}_{n,l+1}\| \le \|\mathbf{e}_{n,l}\|$, the iteration matrix $\left(\mathbf{I} - \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\mathbf{S}_l\mathbf{R}\right)$ must be *non-expansive* i.e. it does not increase the the norm of a vector, when applied to it. A *strict* non-expansiveness of the matrix is sufficient condition and a non-strict non- expansiveness is a necessary condition for convergence. The norm of the iteration matrix is a measure for expansiveness:

$$\begin{aligned}
&\left\|\mathbf{I} - \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\mathbf{S}_l\mathbf{R}\right\| = \\
&\left\|\mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\left(\mathbf{F}^{-1}\mathbf{H}_{z_l}\mathbf{E}_{\chi_l} - \mathbf{S}_l\mathbf{R}\right)\right\| \le \\
&\left\|\mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\right\|\left\|\mathbf{F}^{-1}\mathbf{H}_{z_l}\mathbf{E}_{\chi_l} - \mathbf{S}_l\mathbf{R}\right\| = \\
&\left\|\mathbf{F}^{-1}\mathbf{H}_{z_l}\mathbf{E}_{\chi_l} - \mathbf{S}_l\mathbf{R}\right\|.
\end{aligned} \tag{32}$$

The matrices $\mathbf{H}_{z_l}$ and $\mathbf{E}_{x_l}$ represent propagation of the coefficients and shift in spatial domain along x by χ_l. According to the definition of the matrix norm, the derivation from Eq. 32 continues as:

$$\begin{aligned}
&\left\|\mathbf{F}^{-1}\mathbf{H}_{z_l}\mathbf{E}_{\chi_l} - \mathbf{S}_l\mathbf{R}\right\| = \\
&\max_{\mathbf{a},\|\mathbf{a}\|=1}\left\|\mathbf{F}^{-1}\mathbf{H}_{z_l}\mathbf{E}_{\chi_l}\mathbf{a} - \mathbf{S}_l\mathbf{R}\mathbf{a}\right\| = \\
&\max_{\mathbf{a},\|\mathbf{a}\|=1}\left\|u(\mathbf{x}_l,z_l) - \mathbf{S}_l\mathbf{u}\right\|.
\end{aligned} \tag{33}$$

The operation $\mathbf{S}_l\mathbf{u}$ selects the samples $u(x_i,z_i), (x_i,z_i) \in \mathbf{p}_l$ which determine the set $\mathbf{C}_l$ from the vector $\mathbf{u}$, and puts them on the respective positions of the grid $\mathbf{x}_l$. Therefore the difference $u(\mathbf{x}_l,z_l) - \mathbf{S}_l\mathbf{u}$ is $u(\mathbf{x}_l,z_l)$ with zeros instead of the set determining samples $u(x_i,z_i), (x_i,z_i) \in \mathbf{p}_l$. For L_2 norm, it can be written as:

$$\|u(\mathbf{x}_l,z_l) - \mathbf{S}_l\mathbf{u}\|_2^2 = \|u(\mathbf{x}_l,z_l)\|_2^2 - \|u(\mathbf{p}_l)\|_2^2. \tag{34}$$

As the norm of the coefficient vector $\mathbf{a}$ is unity, the norm of the M-point uniformly sampled field $u(\mathbf{x}_l,z_l)$ on the line z_l is unity as well. Therefore, the norm of the difference $u(\mathbf{x}_l,z_l) - \mathbf{S}_l\mathbf{u}$ is always less than or equal to unity.

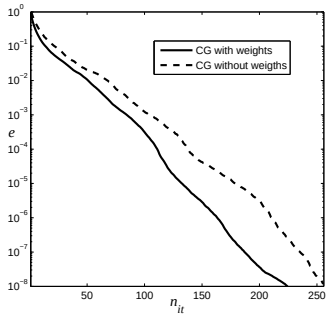

Figure 3. Performance of the CG-based reconstruction on data with (solid line) cluster with and without (dashed line) using weights for regularization.

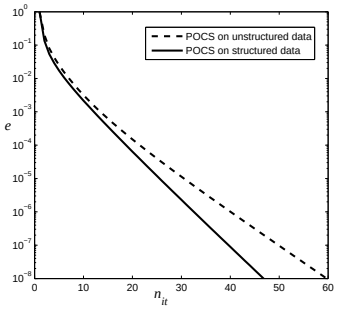

Figure 4. Performance of the POCS-based reconstruction method on data which defines structured (solid line) and unstructured (dashed line) convex sets.

Hence the iteration matrix $\left(\mathbf{I} - \mathbf{E}_{\chi_l}^{-1}\mathbf{H}_{z_l}^{-1}\mathbf{F}\mathbf{S}_l\mathbf{R}\right)$ is non-expansive and the algorithm is convergent.

Eq. 34 suggest that if samples forming the set $\mathbf{C}_l$ contain large portion of the energy of $u(\mathbf{x}_l, z_l)$ then the convergence is faster. Consider the case when $u(x, z)$ is a beam, with finite extent along x at the line $z = z_l$. If this beam falls into a place where the samples $u(\mathbf{p}_l)$ form a gap, then these samples will have small energy. Increasing their density outside the gap will not decrease the norm in Eq. 34 and therefore will not speed up the convergence. Thus, taking the maximal value of this norm in Eq. 33 over all possible field generators $\mathbf{a}$ will produce a value close to 1 in the case when there are clusters or gaps in the samples $u(\mathbf{p}_l)$. On the other hand, if these samples are spread along the whole spatial extent T, then for any $\mathbf{a}$ they will take significant part of the energy of $u(\mathbf{x}_l, z_l)$ resulting in lower norm in Eq. 33. Clusters along z are also undesired. Closely related points define closely spaced convex sets. A projection from a set to set will not change much the projected signal and will bring little new information.

The POCS performance highly depends on the structure of the given samples. Fully arbitraty sample positions determine large number of sets (up to s) while samples on few lines and on regural grids determine low number of sets and more efficient projectctions from set to set.

5. EXPERIMENTS

The theoretical analysis on the convergence of the described approaches is illustrated by three different experiments. The given samples for the experiments were generated by Eq. 4, where $M = 256$ non-zero field generating coefficients a_m were chosen as a Gaussian pulse centered at the origin. Assessment of the results is based on the normalized error between the original $\mathbf{a}$ and reconstructed $\hat{\mathbf{a}}$ coefficient vectors - $e = \|\mathbf{a} - \hat{\mathbf{a}}\|_2 \,/\, \|\mathbf{a}\|$.

The goal of the first experiment is to verify the benefit of using adaptive weights to regularize the CG-based method for the case when there are clusters in the known samples. The number of the given samples is chosen to be $s = 1.5M$ such that the system in Eq. 5 is not underdetermined. M samples are chosen to be randomly scattered within a spatial rectangle, centered at the origin and lying in the half-plane $z > 0$, while the rest $0.5M$ samples are chosen to form a cluster. The CG-based reconstruction method is run with and without adaptive weights on the samples for regularization. The error convergence is shown in Figure 3 with solid and dashed lines respectively. The condition numbers of the weighted and unweighted matrices reflects the benefit of using the weights - $\kappa(\mathbf{R}^H\mathbf{W}\mathbf{R}) = 2826, \kappa(\mathbf{R}^H\mathbf{R}) = 8031$. A simple approach is used to measure the clustering of the samples and assign the adaptive weights in $\mathbf{W}$. The spatial rectangle where the known samples are situated is subdivided into rectangular cells with a coarse grid. The number of samples n_c inside each cell c is counted, and each sample which falls inside the cell c is weighted by $1/n_c$. While this method is not very precise, as it does not adapt to the cluster shapes, it is used here only to demonstrate the regularization power of the adaptive weights iterative approach.

The second experiment illustrates the rapid convergence of the POCS-based reconstruction method when structured convex sets can be formed from the known samples. Two sets of $s = 1.5M$ scattered light field samples were generated. The samples for the first set are chosen to be randomly scattered within a spatial rectangle, centered at the origin and lying in the half-plane $z > 0$. The samples for the second set are chosen inside the same spatial rectangle, but they are selected so that they form 4 structured convex sets. Each of these sets contains points which lie on the same distance z, and are irregularly scattered such that an M-point regular grid is able to host them. The error convergence of the POCS-based reconstruction method is shown in Figure 4 with a solid line for the structured data set and dashed line for the completely random data set.

In general, the CG-based reconstruction algorithm converges faster than the POCS-based algorithm. However, when the known samples can be used to form structured convex sets, the POCS-based algorithm uses much less computations. This can be illustrated by choosing irregularly scattered data points which form structured data sets, in the same manner as in the previous experiment. For such data, the error convergence rate in terms of number of

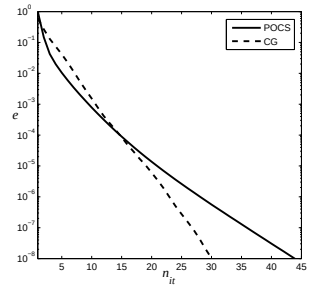

Figure 5. Performance in terms of number of iterations n_{it} for the POCS-based (solid line) and CG-based (dashed line) reconstruction methods on structured data.

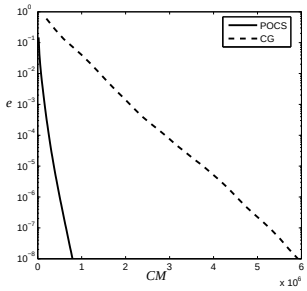

Figure 6. Performance in terms of complex multiplications (CM) for the POCS-based (solid line) and CG-based (dashed line) reconstruction methods on structured data.

iterations is similar for both reconstruction methods (Figure 5). However, one iteration of CG has computational complexity of $O(Ms)$, while the POCS-based algorithm has complexity of $O(M \log M)$ for the projection from one structured set to another, resulting in $O(LM \log M)$ per iteration for L sets. In this experiment the structured sets are $L = 4$, and the benefit of using POCS instead of CG can be seen in Figure 6.

6. CONCLUSION

In this work, we aimed at analyzing the convergence of the two preferred iterative schemes for diffraction field reconstruction from non-uniformly distributed samples. The convergence of the CG-based reconstruction method is highly dependent on the condition number of the non-uniform sample generating matrix. For clustered sample distributions, this matrix has a high condition number which can be reduced by adaptively weighting the samples. This method shows efficiency even when the weight selection method is rather rough. The POCS-based reconstruction method is generalized for the case when the known samples do not belong to a pre-defined regular grid. A detailed theoretical analysis shows that if structured convex sets can be defined from the samples distribution, the reconstruction method is rapidly convergent and efficient implementation exists. Numerical simulations show smaller computational costs than those required by the CG-based approach.

7. REFERENCES

[1] G. B. Esmer, V. Uzunov, L. Onural, H. M. Ozaktas, and A. Gotchev, "Diffraction field computation from arbitrarily distributed data points in space," *Signal Processing: Image Communication*, vol. 22, pp. 178–187, February 2007.

[2] V. Uzunov, A. Gotchev, G. B. Esmer, L. Onural, and H. M. Ozaktas, "Non-uniform sampling and reconstruction of diffraction field," in *Proceedings of The 2006 SMMSP Workshop*, Florence, Italy, 2007, pp. 191–197.

[3] E. Hecht, *Optics*, Addison Wesley, 1998.

[4] J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1996.

[5] B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, John Wiley and Sons, Inc., 1991.

[6] É. Lalor, "Conditions for the validity of the angular spectrum of plane waves," *J. Opt. Soc. Am.*, vol. 58, pp. 1235–1237, 1968.

[7] G. C. Sherman, "Application of the convolution theorem to rayleigh's integral formulas," *J. Opt. Soc. Am.*, vol. 57, pp. 546–547, 1967.

[8] K. Gröchenig and T. Strohmer, "Numerical and theoretical aspects of non-uniform sampling of band-limited images," in *Nonuniform Sampling: Theory and Practice*, F. Marvasti, Ed. Kluwer, 2001.

[9] R. Aharoni and Y. Censor, "Block iterative projection methods for parallel computation of solutions to convex feasibility problems," *Linear Algebra Applications*, vol. 120, pp. 165–175, 1989.

[10] L. G. Gubin, B. T. Polyak, and E. V. Raik, "The method of projections for finding the common point of convex sets," *USSR Comput Math Math Phys*, vol. 7, pp. 1–24, 1967.

[11] Åke Björck, *Numerical Methods for Least Squares Problems*, SIAM, Amsterdam, Holland, 1990.

[12] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, USA, 2003.

[13] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," http://www.cs.cmu.edu/ quake-papers/painless-conjugate-gradient.pdf.

[14] R. Aharoni and Y. Censor, "Block iterative projection methods for parallel computation of solutions to convex feasibility problems," *Linear Algebra Applicat*, vol. 120, pp. 165–175, 1989.

[15] L. G. Gubin, B. T. Polyak, and E. V. Raik, "The method of projections for finding the common point of convex sets," *USSR Comput Math Math Phys*, vol. 7, pp. 1–24, 1967.