Published in final edited form as:

IEEE Trans Radiat Plasma Med Sci. 2017 September ; 1(5): 426–434. doi:10.1109/TNS.2017.2713521.

Numerical algorithms for scatter-to-attenuation reconstruction in PET: empirical comparison of convergence, acceleration, and the effect of subsets

Yannick Berker [Member, IEEE],

Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104 USA, and the Department of Physics of Molecular Imaging Systems, RWTH Aachen University, 52074 Aachen, Germany

Joel S. Karp [Fellow, IEEE], and

Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104 USA

Volkmar Schulz [Member, IEEE]

Department of Physics of Molecular Imaging Systems, RWTH Aachen University, 52074 Aachen, Germany

Abstract

The use of scattered coincidences for attenuation correction of positron emission tomography (PET) data has recently been proposed. For practical applications, convergence speeds require further improvement, yet there exists a trade-off between convergence speed and the risk of non-convergence. In this respect, a maximum-likelihood gradient-ascent (MLGA) algorithm and a two-branch back-projection (2BP), which was previously proposed, were evaluated.

Methods—MLGA was combined with the Armijo step size rule; and accelerated using conjugate gradients, Nesterov's momentum method, and data subsets of different sizes. In 2BP, we varied the subset size, an important determinant of convergence speed and computational burden. We used three sets of simulation data to evaluate the impact of a spatial scale factor.

Results and discussion—The Armijo step size allowed 10-fold increased step sizes compared to native MLGA. Conjugate gradients and Nesterov momentum lead to slightly faster, yet non-uniform convergence; improvements were mostly confined to later iterations, possibly due to the non-linearity of the problem. MLGA with data subsets achieved faster, uniform, and predictable convergence, with a speed-up factor equivalent to the number of subsets and no increase in computational burden. By contrast, 2BP computational burden increased linearly with the number of subsets due to repeated evaluation of the objective function, and convergence was limited to the case of many (and therefore small) subsets, which resulted in high computational burden.

Conclusion—Possibilities of improving 2BP appear limited. While general-purpose acceleration methods appear insufficient for MLGA, results suggest that data subsets are a promising way of improving MLGA performance.

Index Terms

PET; image reconstruction; algorithms; attenuation; Compton scattering

I. Introduction

In positron emission tomography (PET), reconstruction of the tracer spatial distribution λ is often based on one of two models. In analytic approaches, pre-corrected data is modeled as integrals of λ along straight lines of response, summarized as the Radon transform $\mathcal{R}$; whereas in numerical approaches, the model can describe a more complex, but still linear mapping.

In both models, photon attenuation (pre-)correction (AC) is an essential step of the image reconstruction process, compensating for the loss of detected photon pairs through photoelectric effects and Compton scattering. AC generally requires some information about the spatial distribution of the electron density ρ : most often, this information is provided in the form of linear attenuation coefficients μ . Since AC algorithms are well-established for known μ , the main difficulty lies in determining μ : in fact, this has been extensively studied in all contexts of (multi-modality) PET; in particular, in combination with radionuclide transmission sources, X-ray computed tomography (CT), and magnetic resonance imaging (MRI), as well as using PET emission data [1].

The latter approach, which relies solely on *unscattered* PET emission data, was already studied in the days of standalone PET, for example, in the form of maximum likelihood reconstruction of attenuation and activity (MLAA [2]), and interest was revived in the past five years by improved performance with time-of-flight (TOF) emission data [3]–[7]. These approaches consider a joint (λ, μ) optimization problem, but suffer from two theoretical limitations [4]: first, the PET emission data determines the attenuation sinogram $\mathcal{R}\mu$ only on LORs where $\mathcal{R}\lambda > 0$. Second, the attenuation sinogram is only determined up to an unknown offset, which translates into an unknown scaling factor in the reconstructed λ . Without TOF information, cross-talk between λ and μ is an additional issue. Many combination techniques using PET emission data and other information have been proposed, but are often limited to particular applications in specific multi-modality settings [8].

Some of us have recently proposed a new approach to estimate the electron density ρ from *object-scattered* PET photons having energies below the photo peak [9]. This approach is as universal as other emission-based ones in that it employs only PET detector data and can, in principle, complement any of the aforementioned approaches—in particular, in PET/MRI where transmission information is unavailable. In single photon emission computed tomography (SPECT), other groups have studied a similar question empirically [10] as well as theoretically [11], with the latter paper presenting a local-uniqueness proof of the non-linear inverse problem, assuming small attenuation. What makes this approach challenging, however, both in PET and in SPECT, is that the model of the scatter data is neither linear in

μ nor does it relate to the Radon transform. Approaches which have been pursued so far in PET comprise two likelihood-based methods [12]–[14] and an iterative, two-branch back-projection (2BP) [9]. Based on some of this earlier work, in this study, we focus on evaluating, and improving, the practicality of maximum-likelihood gradient-ascent (MLGA) and 2BP.

A major trade-off concerning the step size in gradient-ascent algorithms is the one between convergence speed and the risk of non-convergence. On the one hand, with larger step sizes, convergence is faster, but the estimates are prone to oscillations and other types of non-convergence. On the other hand, smaller step sizes guarantee convergence, but impact convergence speed. Hence, in this work, we explore options for acceleration of MLGA without increasing the risk of non-convergence. Also, we evaluate the well-described approach of partitioning the data into subsets for both MLGA and 2BP.

After formal introduction of the problem and a description of our data simulation, we will summarize MLGA and 2BP and propose new variants, which we will then evaluate.

II. Data modeling

The aim of this work is attenuation-image reconstruction from object-scattered PET coincidences. Hence, the data model is, physically and mathematically, different from what is known in *activity PET*—reconstruction of the tracer distribution from true coincidences. The two main differences are a curved geometry and non-linearity of the measurement equation.

a) Geometry

We assume that each detected Gamma photon can be classified as scattered or unscattered based on its energy, and restrict our attention to coincidences with exactly one single-scattered photon, the other one being unscattered. In activity PET, each coincidence corresponds to an emission event and thus to an element of the tracer distribution; here, each coincidence also corresponds to a scattering event and thus to an element of the electron-density distribution. Consequently, the notion of a line of response (LOR), which usually represents possible emission locations, is replaced by that of a surface of response (SOR), which represents possible scattering locations. Based on Compton scattering geometry, this set of possible scattering locations can be shown to form an American-football shaped surface [9]. While an LOR can be parameterized by two detector elements, an SOR is described by two detectors elements and the energy of the single-scattered photon. The objective of image reconstruction is to reconstruct the attenuation map, represented by the electron density distribution, from object-scattered coincidences defined on these SORs. Despite these differences, SORs can be represented in a system matrix A just like LORs, which has dimensions of number of SORs (n_j) times number of voxels (n_i).

b) Measurement equation

In Appendix A, we sketch a derivation of a discrete version of the measurement equation from [9], which in turn was derived in detail in [15]. The main complication, compared to activity PET, is the fact that the electron density does not only have a positive impact on the

number of scattered coincidences (by generating scatter), but also a negative one (by removing scatter through attenuation). Hence, the measurement equation is non-linear in the electron density ρ . In addition, it depends on the tracer distribution λ ; here, we assume that λ is known and encode this quantity as weighting factors in the system matrix, consequently called A_λ . Another quantity, an attenuation tensor $\underline{K}$, encodes the attenuation length of each voxel along the path of the scattered radiation; as this depends not only on the SOR, but also on the scattering location (because radiation scattered in different locations propagates along different paths, even if detected on the same SOR), $\underline{K}$ has dimensions of $n_i \times n_j \times n_j$.¹

In summary, we model SORs as described above, and the expected scatter data $\vec{y}$ on these SORs using a nonlinear measurement equation, with a scatter system matrix A_λ (having λ as a parameter) and a constant attenuation tensor $\underline{K}$:

$$\vec{y} = \left(A_\lambda \odot \exp(-\underline{K} \vec{\rho}) \right) \vec{\rho} =: \tilde{A}_{\lambda, \rho} \vec{\rho} \quad (1)$$

Here, $\odot$ and $\exp$ denote element-wise multiplication and exponentiation, respectively. The 2-D matrix A_λ describes the sensitivity of the PET camera on SOR i for radiation scattered in a voxel s (in absence of attenuation); the 3-D tensor $\underline{K}$ is related to the attenuation length of a voxel t for that radiation. The two roles of a voxel (attenuating and scattering) are reflected in the double appearance of $\vec{\rho}$ in the middle expression of (1). Through the final expression of (1), we define the *compound* system matrix $\tilde{A}_{\lambda, \rho}$ for brevity of later expressions.

III. Data simulation

Simple, low-resolution 2-D scatter data was simulated similarly as described in [9]: this involved an object consisting of $n_j = 9 \times 9 = 81$ pixels of $5 \times 5 \text{ cm}^2$ size, and a 40 cm-radius detector equipped with 32 detectors (compare Fig. 1) having 4 energy bins (0 to 127 keV, 128 to 255 keV, 256 to 383 keV, 384 to 511 keV), which resulted in $n_i = 32 \times 32 \times 4 = 4096$ single-scatter SORs across all energy bins. (Note that these energy bins are unrealistic for an implementation in practice as the lower energy windows will have large amounts of multiple scatter.) Thus, in line with the above, each SOR is represented by two detector indices and the energy of the scattered photon; the measured data $\vec{y}$ (comprising n_i measurements) can be understood as a concatenation of 4 sinograms (one for each of the 4 energy bins), allowing for a single update using all of the measured data.

After establishing A (of size $n_i \times n_j$) and $\underline{K}$ ($n_i \times n_j \times n_j$) in memory, data simulation consisted of evaluating (1). Thus, all simulations considered single scattering as well as photon attenuation, without multiple scattering or explicit energy uncertainty (other than the wide energy bins). Considering these physical effects would significantly complicate the geometry and thus the computation of A and $\underline{K}$, yet without changing the shape of the

¹Under the assumption that in each energy bin, the attenuation coefficient μ is proportional to the electron density ρ , the elements of $\underline{K}$, $k_{s,t}^i$, can be defined to consider the dependency of attenuation on energy, up to the effects of energy measurement uncertainty.

measurement equation (1). No noise was added to the simulated data in this study. In order to study the same problem at different scales, which is known to be an important determinant for photon scattering [16], phantom simulations and reconstructions were repeated using smaller FOVs and systems (rabbit-sized: scale factor $u = 0.35$ for a $1.75 \times 1.75 \text{ cm}^2$ pixel size and 14 cm detector radius; and rat-sized: $u = 0.2$, $1 \times 1 \text{ cm}^2$ pixel size, 8 cm detector radius).

IV. Objective functions

We used different objective functions for the two algorithms: the log-likelihood $\mathcal{L}$ of the modeled data, given the measured data, and the sum of the squared differences $\mathcal{S}$ between the modeled and the measured data. When shown, these quantities are plotted against the iteration number to study the monotonicity of the objective function.

For a more practically oriented evaluation of reconstruction results, we calculated normalized mean square errors (NMSEs) of ρ in the image domain, relative to the reference in Fig. 1. This analysis embraces the fact that an increase of the objective function does not necessarily correspond to a better (that is, closer to the reference) image. NMSEs are plotted against computation time.

V. Algorithms

In summary, we applied maximum-likelihood gradient-ascent (MLGA) [13], [14] and a two-branch back-projection (2BP) approach [9] to the simulated data. Both of these algorithms are detailed in the following. We studied the following variants:

- MLGA; MLGA with conjugate gradients (MLCGA), enabled by the Armijo step size rule [17]; MLGA with a pre-conditioner (MLPGA); MLPGA with Nesterov acceleration [18]; and MLPGA using data subsets [19].
- 2BP, including with subset sizes different from the previously proposed value 3.

All algorithms are designed for unconstrained optimization; therefore, negative values of ρ were set to 0 after each iteration or subset. Furthermore, we note that conjugate gradients, as well as the Nesterov acceleration scheme, are designed for linear problems, yet regardless, applied to a non-linear problem here. Computations were carried out in MATLAB (R2016b and R2017a; The MathWorks, Natick, MA), using vectorized code and sparse matrix functions.

A. Maximum-likelihood gradient ascent (MLGA)

An in-depth analysis of MLGA, also comparing this algorithm with another likelihood-based approach [12], is being carried out in a separate work [20]. Here, we sketch the necessary detail to follow the implementation of the algorithm.

The Poisson log-likelihood of $\vec{y}$, up to terms only dependent on the given measurement $\vec{y}$, reads

$$\mathcal{L}(\vec{y}) = \vec{y}^\top \log \vec{y} - \vec{1}_y^\top \vec{y}, \quad (2)$$

where $\log$ denotes an element-wise logarithm and $\vec{1}_y$ is a data vector composed of ones. The gradient of the log-likelihood (2), with respect to ρ , uses the gradient of the model of the data (with respect to ρ) and reads

$$\vec{\nabla}_{\vec{\rho}} \mathcal{L} = \left(\vec{\nabla}_{\vec{\rho}} \otimes \vec{y} \right) \left(\vec{y} \odot \vec{y} - \vec{1}_y \right). \quad (3)$$

Using (1), one then finds the gradient of the model of the data,

$$\vec{\nabla}_{\vec{\rho}} \otimes \vec{y} = \left[\mathbf{A}_\lambda \odot \exp(-\mathbf{K} \vec{\rho}) - \vec{\rho}^\top \left\{ \mathbf{K} \odot \left(\left[\mathbf{A}_\lambda \odot \exp(-\mathbf{K} \vec{\rho}) \right] \otimes \vec{1}_\rho \right) \right\} \right]^\top \quad (4a)$$

$$= \left[\tilde{\mathbf{A}}_{\lambda, \rho} - \vec{\rho}^\top \left\{ \mathbf{K} \odot \left(\tilde{\mathbf{A}}_{\lambda, \rho} \otimes \vec{1}_\rho \right) \right\} \right]^\top. \quad (4b)$$

Hence, we have all ingredients to define the update equation, using a potentially non-uniform step size $\vec{s}$, through

$$\vec{\rho}^{\text{new}} = \vec{\rho} + \vec{s} \odot \vec{\nabla}_{\vec{\rho}} \mathcal{L}. \quad (5)$$

We focus on two step sizes here, the constant step size (MLGA)

$$\vec{s}_1 = \alpha \vec{1}_\rho \quad (6)$$

as well as the preconditioned, MLEM-like step size (MLPGA)

$$\vec{s}_3 = \gamma \vec{\rho} \odot \left(\tilde{\mathbf{A}}_{\lambda, \rho}^\top \vec{1}_y \right), \quad (7)$$

where α and γ are empirically determined step-size constants that led to fastest convergence. To avoid over-optimization, for each algorithm, empirical step-size constants were fitted to a power law $a \times b^u$, where u is the scale factor of the experiment (scanner and phantom). As this is still subjective to some extent, we evaluated the Armijo step size rule [17],

$$\vec{s}_1^* = \alpha^* \eta^m \vec{1}_\rho, \quad (8)$$

with the increased step-size constant $\alpha^* = 10\alpha$ in combination with $\eta = 0.5$. In each iteration, we used the current gradient

$$\vec{g} = \vec{\nabla}_{\vec{\rho}} \mathcal{L} \quad (9)$$

and chose for m the smallest nonnegative integer fulfilling the Armijo-Goldstein condition

$$\mathcal{L}(\vec{\rho} + \alpha^* \eta^m \vec{g}) - \mathcal{L}(\vec{\rho}) \geq \sigma \alpha^* \eta^m \vec{g}^\top \vec{g}. \quad (10)$$

This condition requires that the increase of $\mathcal{L}$ in each step (left-hand side) be at least a σ -fraction of the improvement expected from a linear behavior of $\mathcal{L}$, which can be expressed by the step size and gradient (right-hand side). It allows the use of much larger step sizes without increasing the risk of non-convergence, as too large step sizes will be down-regulated. Equation (10) was evaluated using $\sigma = 0.001$.

This has enabled acceleration by extending the MLGA method towards a conjugate-gradient method (MLCGA): we have chosen MLGA with the Armijo step size rule (8) as a basis, and complemented it with conjugate gradients in accordance with the following update:

$$p_n = \frac{\vec{g}_n^\top (\vec{g}_n - \vec{g}_{n-1})}{\vec{g}_{n-1}^\top \vec{g}_{n-1}} \quad (p_0 = p_{5n} = 0), \quad (11)$$

$$\vec{c}_n = \vec{g}_n + p_n \vec{c}_{n-1} \quad (\vec{c}_0 = \vec{0}), \quad (12)$$

$$\vec{\rho}_{n+1} = \vec{\rho}_n + \vec{s}_1^* \odot \vec{c}_n. \quad (13)$$

$\vec{g}_n$ and $\vec{c}_n$ are the gradients and conjugate gradients in iteration n , respectively; (11) is the Polak–Ribière formula [21]; $\vec{s}_1^*$ is chosen as in (8); and we restarted every 5 iterations by setting $p_{5n} = 0$ for $n \in \mathbb{N}$.

Furthermore, as it was applied to other gradient methods in image reconstruction [18], [22], [23], we have studied acceleration by Nesterov's momentum method. For this part of the study, we have chosen MLPGA (7) as a basis. Since we observed that the value of the step

size constant γ that we used in MLPGA lead to non-convergence, we used half the step size $\vec{s}_3^* - \vec{s}_3/2$:

$$\vec{r}_{n+1} = \vec{\rho}_n + \vec{s}_3^* \odot \vec{g}_n \quad (\vec{r}_0 = \vec{\rho}_0), \quad (14)$$

$$a_{n+1} = \frac{1}{2} \left(1 + \sqrt{1 + 4a_n^2} \right) \quad (a_0 = 1), \quad (15)$$

$$\vec{\rho}_{n+1} = \vec{r}_{n+1} + \frac{a_n - 1}{a_{n+1}} (\vec{r}_{n+1} - \vec{r}_n). \quad (16)$$

Here, $\vec{r}_{n+1}$, representing the updated estimate in unmodified MLPGA [compare (14) with (5)], is used as an intermediate quantity for the accelerated update in (16).

Finally, inspired by Ordered Subset Expectation Maximization (OSEM [19]), we evaluated the impact of subsetization of the measured data, again based on MLPGA (7)—a choice that will be discussed in Section VII-A. I denotes a subset such that $\vec{y}_I$ is a vector with all components other than those in I set to 0. Thus, mathematically, subset operations can be represented by replacing $\vec{y}$ by $\vec{y}_I$ (and similarly for $\vec{1}_y$) throughout equations (1), (3), (4) and (7), emulating zero measurements outside of the current subset. In implementation, this was done more efficiently by replacing all vectors, matrices and tensors by the respective subset quantities of smaller size. We used 2, 4, 8, 16 and 32 subsets of equal size in these experiments, and SORs were randomly partitioned (using each SOR exactly once) into these subsets. We were able to verify that subset sensitivity images, computed by back projecting subset data all set to 1, were reasonably homogeneous across voxels and subsets.

In a first subset experiment, step size constants were not changed; in another one, step sizes were compensated so as to maintain the product of step size constant γ and number of subsets (using half the step size for twice the number of subsets, respectively).

B. Two-branch back-projection (2BP)

In [9], some of us proposed an iterative scatter-to-attenuation approach with additive updates, by back projecting the difference between measured and estimated data into image space. Both a positive and a negative back-projection branch were found necessary to account for attenuating and scattering contributions of $\vec{\rho}$ in (1): in other words, to account for a deficit of estimated coincidences on an SOR, the algorithm either increases the electron density on that SOR (to encourage generation of additional scatter), or decreases the electron density in the volume between the SOR (to decrease attenuation of scatter). Hence, each update step involves computation of two possible updates, and two-fold evaluation of the objective function $\mathcal{S}$ to determine the better update.

Updates are applied per subset of SORs; these subsets are defined, in each iteration, by random partition of all SORs. Using the above notation, we write the 2BP updates as

$$\vec{\rho}^{\text{new}} = \begin{cases} \vec{\rho}^+ := \vec{\rho} + \varepsilon^+ \mathbf{A}_\lambda^\top (\vec{y} - \vec{\bar{y}})_I & \text{if } \mathcal{J}^+ < \mathcal{J}^-, \\ \vec{\rho}^- := \vec{\rho} - \varepsilon^- \left(\vec{1}_\rho^\top \mathbf{K} \right)^\top (\vec{y} - \vec{\bar{y}})_I & \text{otherwise,} \end{cases} \quad (17)$$

where $\vec{1}_\rho$ is an image vector composed of ones, $\varepsilon^\pm$ are step size constants, and $\mathcal{J}^\pm$ is short for $\mathcal{J}(\vec{y}(\vec{\rho}^\pm))$. Note that every evaluation of $\mathcal{J}$ involves the full data; we have experimented with evaluating the objective function only for the current subset I , which often leads to non-convergence. In this work, we varied the subset size from the previous value of 3 (from [9]) by considering the alternatives 1, 2, 5, 7, 10 and 20.

VI. Results

A. Armijo step size and conjugate gradients

As seen in Fig. 2, the Armijo step size increases convergence rates compared to the constant step size. Convergence is still monotonous, yet less uniform from one iteration to the next. The degree of improvement differs across spatial scales.

Conjugate gradients (MLCGA, Fig. 2) contribute an additional improvement in convergence rates. Improvements appear to be confined to later iterations at all spatial scales.

B. Nesterov momentum

Fig. 3 shows the improvement in convergence rate by acceleration through Nesterov's momentum method, compared to MLPGA (MLGA with an MLEM-like step size). Nesterov acceleration with a non-compensated step size $\vec{s}_3$ was found not to converge, so Fig. 3 shows a comparison of the slowed-down version with MLPGA at the same step size constant $\vec{s}_3^*$, stressing the gain in convergence rate. MLPGA with Nesterov acceleration slightly outperforms non-compensated MLPGA, although this improvement is rather modest and, as that of conjugate gradients, confined to later iterations.

C. Subset size and number of subsets

Fig. 4 details the behavior of 2BP for subsets of different sizes, with 3 being the subset size as originally published [9]. Convergence (per iteration) is fastest with smallest subsets (of size 1) and slows down, yet irregularly, with increasing subset size. In fact, there is hardly a difference between 1 and 2 SORs per subsets, while the NMSE decreases non-monotonously with as few as 7 SORs per subset, and convergence is not guaranteed with 10 or more SORs per subset. This is similar across all spatial scales (Figs. 4a to 4c).

The effect of the number of subsets on MLGA performance is shown in Fig. 5. Likewise, a higher number of subsets (that is, a smaller subset size) increases convergence rates, which is again true at all spatial scales (Figs. 5a to 5c). In contrast to 2BP and similar to MLEM/ OSEM in emission tomography, MLPGA generally converges for fewer, larger subsets and

becomes unstable only for more, that is, smaller subsets—compare Fig. 5c. In our experiments with only 32 detectors, 32 subsets turned out to converge at human and rabbit scale, while 16 was the maximum number of MLPGA subsets for convergence at the rat scale. However, in this latter case, 16 subsets yielded an NMSE of 0.01 as early as after a single iteration, which more subsets can hardly improve.

In order to quantify the effect of subsets, we decreased the 2BP step size for larger subsets until monotonous convergence was achieved for all subset sizes. As seen in Fig. 6a, the convergence rates of image NMSE (as a function of computation times) are similar for several of these variants, with a subset size of 5 appearing optimal in this example.

In MLPGA, the subset gain is approximately linear: the gain can be quantified indirectly in two ways. First, each doubling of the number of subsets is almost exactly compensated by halving of the number of iterations (Fig. 5); second, halving the step size, as indicated in Fig. 6b, almost exactly compensates doubling the number of subsets. Note that contrary to Fig. 6a, which compares several algorithms at the edge of nonconvergence, Fig. 6b compares algorithms artificially slowed down by varying degrees.

D. Subsets and computational burden

Fig. 7 shows the run times of 2BP and MLGA as a function of subset sizes and numbers of subsets, respectively, both in the phantom geometry (Fig. 1), as well as in a fully populated geometry (all voxels containing non-zero tracer activity and attenuation). All run times are independent of the spatial scale and are higher with the fully populated geometry. Note that according to the results in Section VI-C, 2BP converges only for subsets of size up to 7, while MLGA converges across a wide range of subset numbers.

In 2BP, the computational effort per iteration grows with increasing number of subsets (decreasing subset size), as every subset requires two evaluations of the objective function. In fact, run times increase approximately 20-fold for a 20-fold increase in the number of subsets (Fig. 7a).

By contrast, as known for OSEM, computational efforts for MLGA do not increase with the number of subsets, as all tensor and matrix computations can be efficiently restricted to the relevant SORs in each subset. MLGA run times even decreased by a factor of 0.8 for a 32-fold subset-number increase (Fig. 7b).

VII. Discussion

In this work, we have evaluated acceleration strategies for maximum-likelihood gradient-ascent (MLGA) and a simple back projection-based (2BP) algorithm for scatter-to-attenuation reconstruction.

A. Convergence properties

The main result regarding 2BP is that convergence critically depends on the number of SORs per subset, with the unusual result that a small number of SORs per subset (that is, a high number of subsets) is required for convergence. This is due to the fact that for each

subset, a single update branch [positive or negative; compare (17)] is chosen and applied to all SORs in that subset. However, different SORs in the subset may call for different update directions. According to our results, not only are updates with more SORs per subsets less effective; apparently, the larger the number of SORs per subsets, the higher the chance that *no* effective update exists. This effect can be moderated by reducing step sizes, yet at the price of decreased convergence rates.

A similar trade-off in MLGA is the one between convergence speed and the risk of non-convergence through the choice of step size constants. Introduction of the Armijo step size rule into MLGA was found to allow the use of 10-fold increased (maximum) step sizes without impacting convergence. This suggests that optimization of step size constants is much less important in this case for achieving optimal performance. Choosing even excessively high step sizes will have some impact on computational burden through added evaluations of the Armijo-Goldstein condition (10), but not so much on convergence of the algorithm.

The Armijo step size further enabled the use of conjugate gradients. This, however, lead to more non-uniform convergence, which makes performance less predictable. In addition, during our preliminary optimization we observed that this algorithm was sensitive towards changes of the numerous parameters, such as those of the Armijo-Goldstein condition or the specific update equation of the conjugate gradients. We also found the restart parameter (5 iterations in our case) to be an important factor for good convergence rates.

This was not true for MLPGA when accelerated using Nesterov's momentum method, which has no additional parameters; however, this algorithm required a smaller choice of step sizes for convergence compared to non-accelerated MLPGA. Again, convergence was non-uniform, that is, challenging to predict.

An observation common to both acceleration techniques (conjugate gradients and Nesterov momentum) is that their benefit is only seen at later iterations. We hypothesize that this is due to the non-linearity of the problem: in particular, nonlinear conjugate gradients methods, an extension of conjugate gradients derived for linear problems, may perform better in close proximity of the optimum than far away from it, where a linear approximation to the objective function fails to properly represent the optimum. This may also explain the need for restarts, as some kind of memory about earlier environments is kept in the conjugate gradients and may require invalidation after a couple of iterations. Applying a restarting scheme to Nesterov's method would be possible; however, considering that Nesterov acceleration manifests in consistently faster convergence rates after 5 iterations, we do not expect a benefit from additional restarts.

Acceleration through subsets is more promising than either of the two previous approaches, due to several advantages:

- The approach has been subject to evaluation in many contexts since first publication [19] and is now well understood. In particular, various subsetization strategies have been proposed based on acquisition time, geometry, or TOF information [24]. In this work, the random subsets partitioning scheme was

chosen since other, geometrically established subsetization rules do not immediately apply to the extended measurement space of SORs, due to the different shapes of different SORs. While the subset sensitivities were reasonably balanced in the cases we verified, optimized ordering schemes might help eliminate the uncertainty associated with our random selection [25].

- There are no additional parameters except the number of subsets. In fact, compared to MLPGA without subsets, not even step size constants were changed. We have mentioned before that mathematically, the subset variant can be formulated by replacing $\vec{y}$ and $\vec{1}_y$ by their subset variants (with zero-valued entries outside of the current subset). This includes $\vec{1}_y$ in the denominator of the step size (7); as a consequence, each element of the step size scales approximately linearly with the number of subsets. This compensates the inverse proportionality between the elements of the log-likelihood gradient (3) and the number of subsets. Hence, as in OSEM, each update of ρ^{new} amounts to the same order of magnitude as in the no-subsets case, with as many updates per iteration as subsets. This is true for $\vec{s}_3$ (7), while not so for $\vec{s}_1$ (6), and explains our choice of MLPGA as a basis for the subset algorithm in Section V-A.
- Convergence is uniform, and convergence rates are quantitatively predictable. In fact, doubling the number of subsets has the same effect as doubling the number of iterations or doubling the step size, yet without the risk of non-convergence (within limits).
- Subsets are easy to implement, with only minor modifications compared to MLGA without subsets.

B. Computational burden

In addition, MLGA with subsets is computationally efficient, with efforts similar to MLGA without subsets, and straight-forward to implement in loop-based languages such as C++. In this study using MATLAB, a matrix-based language, care had to be taken to avoid repeated slicing of the full system matrix to extract partial system matrices for each subset, which results in an undesirable constant computational effort per subset. In this study, this was achieved by pre-arranging the data and the system matrix in memory in slices according to the known number of subsets. As a result, regular MATLAB matrix computations could be used, simplifying the efficient use of sparse matrices.

Interestingly, we have measured a slight decrease in computation times per iteration for more subsets (Fig. 7b). This decrease is supposedly related to more efficient memory operations when operating on smaller tensors, matrices and vectors.

The computational burden of 2BP, by contrast, scales linearly with the number of subsets, which in combination with the upper limit on subset sizes leads to an explosion of the computational burden required for convergence. An explanation is given in the following.

While the individual use of 3 SORs per subset in the 2BP update (17) may appear as an example of subsetization that is neutral in terms of computation time, it is in fact most

inefficient. The reason is that 2BP requires two evaluations of the objective function (that is, two full forward projections) for each subset to determine the correct back-projection branch, which dominate computational burden. Hence, more subsets increase the computational burden.

A similar statement holds for a combination of, for example, the Armijo step size rule with data subsets in MLGA: the computational burden would, again, increase linearly with the number of subsets, due to objective function evaluations in each subset as part of the Armijo-Goldstein condition (10).

In this regard, the previously used number of 3 SORs per subset in 2BP appears close to optimal, at least for problems of this size. Note, however, as 2BP uses a fixed *number* (as opposed to a fixed *fraction*) of the total number of SORs in each subset, increasing the number of SORs further increases the number of subsets and thus the computational burden.

C. Limitations

This study has three main limitations. First, the use of simple physics simulation and the size of the problems, which makes validation in more realistic scenarios (energy uncertainty, 3-D scattering, multiple scattering) necessary; we note, however, that many, more complicated models can still be represented by (1), except for multiple photon scattering.

Second, the inverse crime, as no noise was simulated and reconstruction was performed with the same model as the simulation; again, this demands further validation, but does not invalidate the results obtained in this study, in particular, in terms of computational burden. Also, each instance of non-convergence can be directly attributed to the reconstruction problem and the unfitness of the respective algorithm, e.g., due to non-linearity. No algorithm should perform better on noisy data than on noiseless data, so what we have found to be unsuitable in this work would be with noisy data, too.

Finally, we assumed knowledge of λ to focus on algorithms that reconstruct ρ . This assumption is easily justified in applications such as Compton scatter imaging with external sources [26], where the sources are exactly known. In the PET application at hand, extension towards joint reconstruction of (λ, ρ) may follow the lines of schemes such as MLAA [2]. In particular, this study should be interpreted similar to studies of the maximum likelihood transmission (MLTR) algorithm [27] in the context of MLAA: like MLGA, the MLTR component of MLAA assumes knowledge of the correct activity distribution λ and does not exploit TOF information, yet is a crucial component of the very promising TOF-MLAA joint reconstruction scheme.

VIII. Conclusion

In conclusion, 2BP appears impractical due to the computational burden, and possibilities of improving 2BP appear limited. While general-purpose acceleration methods such as conjugate gradients and Nesterov's momentum method prove insufficient for MLGA, possibly because they are inappropriate for these particular non-linear problems, results suggest that data subsets are a promising way of improving MLGA performance. Future

work may be directed towards studying geometrically backed subsetization strategies, taking into account the special structure of the measurement space (where the shape of each SOR depends on the energy of the scattered photon), as well as towards applying subset-MLGA to realistic data.

Acknowledgments

Y.B. would like to thank Dr. Yusheng Li (University of Pennsylvania) for helpful discussions.

The work leading to this publication was supported by the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF) and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007–2013) under REA grant agreement n° 605728 (P.R.I.M.E. – Postdoctoral Researchers International Mobility Experience) [\[6\]](#), and by the National Cancer Institute of the National Institutes of Health under award number R01CA113941. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Two patent applications relevant to this work have been filed by Koninklijke Philips Electronics N.V. with Y.B. and V.S. as inventors.

Appendix

Physical role of system matrix quantities A_λ and $\underline{K}$

By mapping [9, Eq. (5)] into the discrete domain, we find that the expected number of coincidences on SOR i is

$$\bar{y}^i = \sum_s \rho^s \cdot \left(\sum_e b_{s,e}^i \cdot \lambda^e \right) \cdot \exp \left(- \sum_t k_{s,t}^i \cdot \rho^t \right), \quad (18)$$

where the indices e , s , and t represent emitting, scattering, and transmitting voxels, respectively. We remind that i defines a detector pair and the energy of the scattered photon; then, $b_{s,e}^i$ is related to the probability, per unit electron density in s , that radiation from a unit radioactivity in e is scattered in s and detected on i (that is, the scattered photon is detected, with the given energy, in the first detector, and the unscattered photon is detected in the second detector defined by i). That probability is thus scaled by λ_e and ρ_s and summed over all possible emission and scattering locations. The final term in (18) describes attenuation by the voxels traversed by either of the photons; therefore, $k_{s,t}^i$ is proportional to the radiological path length of voxel t for radiation on SOR i , scattered in the particular location s .

We now incorporate $\vec{\lambda}$ into the scatter matrix by defining

$$a_s^i := \sum_e b_{s,e}^i \cdot \lambda^e \quad (19)$$

to obtain A_λ composed of the a_s^i . Likewise, we represent $k_{s,t}^i$ by the three-dimensional tensor $\underline{K}$ and abbreviate

$$(\vec{K} \vec{\rho})_s^i := \sum_t k_{s,t}^i \cdot \rho^t \quad (20)$$

to find an element-wise expression of (1).

References

1. Keereman V, Mollet P, Berker Y, Schulz V, Vandenberghe S. Challenges and current methods for attenuation correction in PET/MR. *Magn. Reson. Mater. Phys. Biol. Med.* Feb; 2013 26(1):81–98.
2. Nuyts J, Dupont P, Stroobants S, Banninck R, Mortelmans L, Suetens P. Simultaneous maximum a posteriori reconstruction of attenuation and activity distributions from emission sinograms. *IEEE Trans. Med. Imaging.* May; 1999 18(5):393–403. [PubMed: 10416801]
3. Salomon A, Goedicke A, Schweizer B, Aach T, Schulz V. Simultaneous reconstruction of activity and attenuation for PET/MR. *IEEE Trans. Med. Imaging.* Mar; 2011 30(3):804–813. [PubMed: 21118768]
4. Defrise M, Rezaei A, Nuyts J. Time-of-flight PET data determine the attenuation sinogram up to a constant. *Phys. Med. Biol.* Feb; 2012 57(4):885–899. [PubMed: 22290428]
5. Rezaei A, et al. Simultaneous reconstruction of activity and attenuation in time-of-flight PET. *IEEE Trans. Med. Imaging.* Dec; 2012 31(12):2224–2233. [PubMed: 22899574]
6. Defrise M, Rezaei A, Nuyts J. Transmission-less attenuation correction in time-of-flight PET: analysis of a discrete iterative algorithm. *Phys. Med. Biol.* Feb; 2014 59(4):1073–1095. [PubMed: 24504259]
7. Rezaei A, Defrise M, Nuyts J. ML-reconstruction for TOF-PET with simultaneous estimation of the attenuation factors. *IEEE Trans. Med. Imaging.* Jul; 2014 33(7):1563–1572. [PubMed: 24760903]
8. Berker Y, Li Y. Attenuation correction in emission tomography using the emission data: a review. *Med. Phys.* Feb; 2016 43(2):807–32. [PubMed: 26843243]
9. Berker Y, Kiessling F, Schulz V. Scattered PET data for attenuation-map reconstruction in PET/MRI. *Med. Phys.* Oct. 2014 41(10):102502. [PubMed: 25281977]
10. Cade SC, Arridge S, Evans MJ, Hutton BF. Use of measured scatter data for the attenuation correction of single photon emission tomography without transmission scanning. *Med. Phys.* Aug. 2013 40(8):082506. [PubMed: 23927351]
11. Courdurier M, Monard F, Osses A, Romero F. Simultaneous source and attenuation reconstruction in SPECT using ballistic and single scattering data. *Inverse Probl.* Aug. 2015 31(9):095002.
12. Zhang, G., Sun, H., Pistorius, S. Electron density reconstruction from scattered coincidences for attenuation correction in positron emission tomography; *Proc. 13th Intern. Meeting Fully Three-Dimens. Image Reconstr. Radiol. Nucl. Med.*; Jun. 2015 p. 526–529.
13. Berker, Y., Schulz, V., Karp, JS. Comparison of numerical algorithms for scatter-to-attenuation reconstruction in PET [oral presentation]. Paper presented at PSMR 2016, 5th Conference on PET/MR and SPECT/MR; Cologne, Germany. 2016.
14. Berker, Y., Schulz, V., Karp, JS. Discrete iterative algorithms for scatter-to-attenuation reconstruction in PET; 2016 IEEE Nucl. Sci. Symp. Med. Imaging Conf; Oct–Nov. 2016 p. M10A-5.
15. Berker, Y., Schulz, V. Scattered PET data for attenuation-map reconstruction in PET/MRI: Fundamentals; 2014 IEEE Nucl. Sci. Symp. Med. Imaging Conf; Nov. 2014 p. 6
16. Konik A, Madsen MT, Sunderland JJ. GATE simulations of human and small animal PET for determination of scatter fraction as a function of object size. *IEEE Trans. Nucl. Sci.* Oct; 2010 57(5):2558–2563.
17. Armijo L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pac. J. Math.* Nov; 1966 16(1):1–3.

18. Kim D, Ramani S, Fessler JA. Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction. *IEEE Trans. Med. Imaging*. Jan; 2015 34(1):167–178. [PubMed: 25163058]
19. Hudson HM, Larkin RS. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans. Med. Imaging*. Dec; 1994 13(4):601–609. [PubMed: 18218538]
20. Berker Y, Schulz V, Karp JS. Scatter-to-attenuation reconstruction algorithms in PET: theoretical analysis, small-scale comparison, and use in joint reconstruction of activity and attenuation. *Phys. Med. Biol.* revising for resubmission.
21. Polak E, Ribière G. Note sur la convergence de méthodes de directions conjuguées. *Rev. Fr. Inform. Rech. Oper.* 1969; 3(16):35–43.
22. Wang AS, Stayman JW, Otake Y, Vogt S, Kleinszig G, Siewerdsen JH. Accelerated statistical reconstruction for C-arm cone-beam CT using Nesterov's method. *Med. Phys.* May; 2015 42(5): 2699–2708. [PubMed: 25979068]
23. Rezaei A, Michel C, Casey ME, Nuyts J. Simultaneous reconstruction of the activity image and registration of the CT image in TOF-PET. *Phys. Med. Biol.* Feb; 2016 61(4):1852–1874. [PubMed: 26854817]
24. Mehranian A, Kotasidis F, Zaidi H. Accelerated time-of-flight (TOF) PET image reconstruction using TOF bin subsetization and TOF weighting matrix pre-computation. *Phys. Med. Biol.* Jan; 2016 61(3):1309–1331. [PubMed: 26796770]
25. Kim SM, et al. Fully three-dimensional OSEM-based image reconstruction for Compton imaging using optimized ordering schemes. *Phys. Med. Biol.* Sep; 2010 55(17):5007–5027. [PubMed: 20702926]
26. Hussein EMA, Enjilela E. Transmission-like calibration-free tomographic reconstruction with Compton-scattered photons. *IEEE Trans. Nucl. Sci.* Oct; 2016 63(5):2586–2598.
27. Nuyts J, De Man B, Dupont P, Defrise M, Suetens P, Mortelmans L. Iterative reconstruction for helical CT: a simulation study. *Phys. Med. Biol.* Apr; 1998 43(4):729–737. [PubMed: 9572499]

Fig. 1.

Reference attenuation map (linear attenuation coefficients μ in 1/cm) with voxel and detector indices (human scale, $u = 1$).

Fig. 2.

Comparison of algorithms I: (a)–(c) data log-likelihood as a function of iteration number for MLGA with constant step size, MLGA with Armijo step size, and MLCGA with Armijo step size; (d)–(f) for the same algorithms, NMSE of ρ in the image domain as a function of computation time. (a), (d) human-sized FOV; (b), (e) rabbit-sized FOV; (c), (f) rat-sized FOV.

Fig. 3.

Comparison of algorithms II: (a)–(c) data log-likelihood as a function of iteration number for MLPGA, MLPGA with acceleration through Nesterov momentum and half the step size, and native MLPGA with half the step size; (d)–(f) for the same algorithms, NMSE of ρ in the image domain as a function of computation time. (a), (d) human-sized FOV; (b), (e) rabbit-sized FOV; (c), (f) rat-sized FOV.

Fig. 4.

Influence of number of 2BP subsets: sums of squared data differences for 2BP with different subset sizes. (a) human-sized FOV, (b) rabbit-sized FOV, (c) rat-sized FOV.

Fig. 5.

Influence of number of MLGA subsets: data log-likelihood as a function of iteration number for MLPGA with different numbers of subsets. (a) human-sized FOV, (b) rabbit-sized FOV, (c) rat-sized FOV.

Fig. 6.

Subsets vs. step sizes: (a) Ensuring convergence of 2BP with larger subset sizes through smaller step sizes, (b) Compensating subsets in MLPGA through step sizes to maintain constant number-of-subsets—step-size-constant products. Image NMSE as a function of computation time (rabbit-sized FOV).

Fig. 7.

Computational burden with subsets: run times per iteration for (a) 2BP (in s) (b) MLPGA (in ms) as a function of number of subsets. Regression lines represent unweighted linear fits; factors in the legends denote the fold change from lowest to highest number of subsets, according to fits.