

Bregman methods in quantitative photoacoustic tomography

Hao Gao¹, Hongkai Zhao² and Stanley Osher¹

¹Department of Mathematics, University of California, Los Angeles, CA 90095, USA

²Department of Mathematics, University of California, Irvine, CA 92697, USA

Abstract. The Bregman method is utilized to quantitatively improve the simultaneous reconstruction of absorption and scattering coefficients for both Jacobian-based and gradient-based methods in quantitative photoacoustic tomography with multiple optical excitations. It is the synergistic combination of Bregman method and the total-variation regularization in achieving the desired quantitative improvement. In particular, the Bregman method is directly applied to the minimization of the nonlinear and nonconvex functional in the gradient-based method, which is efficiently solved via split Bregman method with limited-memory BFGS as inner loops. We show in simulations the feasibility of 3D simultaneous reconstruction.

1. Introduction

Diffuse optical tomography (DOT) (Arridge and Scotland 2009) allows the non-invasive mapping of absorption and scattering coefficients through boundary measures. It has great potential for functional imaging, such as brain and breast imaging. In addition, it can provide accurate *in vivo* optical environment to optical molecular imaging for successful visualization of cellular and molecular activities. However, it remains extremely difficult to improve the DOT performance for more accurate and reliable quantification with higher resolution, due to the diffusive light after numerous scattering events.

1.1. Quantitative photoacoustic tomography

Photoacoustic tomography (PAT) (Wang 2009), a synergistic combination of ultrasound and optical imaging, has emerged recently as a potential multi-modality imaging for accurate quantification and high resolution. Through PAT, photoacoustic pressure can be spatially resolved through inversion of the acoustic wave, which has already found a wide range of *in vivo* applications. In quantitative PAT, the absorption coefficient can be further recovered from photoacoustic pressure, which is proportional to the product of absorption coefficient and optical energy density.

Despite significant advances in PAT, the research for quantitative PAT (Cox *et al* 2005, Ripoll and Ntziachristos 2005, Yin *et al* 2007, Banerjee *et al* 2008, Cox *et al* 2009) is still preliminary, especially for simultaneous reconstruction of absorption and scattering coefficients. The difficulty originates both intrinsically and mathematically.

- Physically, the simultaneous reconstruction of both parameters is non-unique in the traditional setting with single optical excitation. As a result, without precisely knowing the underlying scattering property, neither of absorption and scattering coefficients can be accurately reconstructed. Lately, the reconstruction uniqueness and stability was rigorously studied by Bal and Uhlmann (2010), who establishes the stability result when the object is illuminated with two excitations with additional geometric constraints or with $2d$ excitations without additional constraints.
- Computationally, it is far from trivial to recover high-resolution scattering coefficient quantitatively as well as absorption coefficient. In (Gao and Zhao 2010), Bregman method was utilized to achieve and improve the reconstruction of both parameters in two dimensions

(2D). However, it is challenging to extend the method to three dimensions (3D) since it is based on Jacobians, which impose significant requirement on both memory storage and computer speed. In this study, we shall develop the Bregman method for gradient-based reconstruction and show its computational feasibility in 3D.

Let us start to formulate the quantitative PAT problem with the following notations: μ_a for absorption coefficient, μ_s for (reduced) scattering coefficient, k and N for the index and the number of variable pairs $X := (\mu_a, \mu_s)$ on the discretized mesh, i and N_s for the index and the number of optical excitations, j and N_d for the index and the number of measuring locations, ϕ_i for optical energy density, P_j for measuring functional acting on ϕ_i , $Y := \mu_a \phi$ for the data (recovered through PAT or given as priors) with individual component y_{ij} , and $\vec{F}(X)$ for the composite of forward operator and measuring functional with individual component $P_j^T \phi_i$. Here, we assume the measures are specified for each variable pair, i.e., $N_d = N$. Then quantitative PAT can be formulated as to minimize the difference between the model outcome and the data, $f(X) := \frac{1}{2} \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} (P_j^T \phi_i(X) - y_{ij})^2$, which is so called data fidelity,

$$X = \arg \min_X f(X). \quad (1.1)$$

Due to the illposed nature of the inverse problem (1.1) or the noise in the data, the solution needs to be regularized and the data fidelity should be combined with appropriate regularization, which will be introduced later.

1.2. linearized Jacobian-based method

A popular approach for solving (1.1) is to treat it as a nonlinear least-square problem that can be solved with many standard optimization techniques (Nocedal and Wright 1999), among which Levenberg-Marquardt method (LM) is perhaps the most commonly used one, i.e., the Gauss-Newton method with L2 regularization. We shall describe LM briefly as follow.

Given the initial guess X_0 , we linearize the vector-valued function \vec{F} with the Jacobian $J := [\partial_{x_k} \vec{F}_{ij}]$

$$\vec{F}(X) \approx \vec{F}(X_0) + J(X_0)(X - X_0), \quad (1.2)$$

which is indeed a good approximation when X_0 is close to the true value. As a result, (1.1) is transformed to the convex problem

$$\delta X = \arg \min_{\delta X} \| J(X_0) \delta X - (Y - \vec{F}(X_0)) \|_2^2. \quad (1.3)$$

The optimal condition of (1.3) implies that $(J^T J) \delta X = J^T (Y - \vec{F}(X_0))$. However, the solution δX may not be unique due to the ill-conditioning of $J^T J$. Therefore L2 regularization $\lambda_2 \| \delta X \|_2^2$ is added to improve the conditioning, which can also be regarded as to improve the approximation of Hessian. In practice, (1.2) can be solved iteratively for more accurate approximation of (1.1). Hence, the iterative Jacobian-based method via LM is

$$X_{n+1} = X_n + \arg \min_{\delta X} [\| J(X_n) \delta X - (Y - \vec{F}(X_n)) \|_2^2 + \lambda_{2,n} \| \delta X \|_2^2], \quad (1.4)$$

which can be solved through optimal condition $(J^T J + \lambda_{2,n} I) \delta X = J^T (Y - \vec{F}(X_n))$ with adaptive regularization parameter $\lambda_{2,n}$.

The advantages of Jacobian-based method include:

- it is simple to implement and many existing optimization techniques can be utilized to improve the convex problem (1.3) or (1.4), such as Bregman method discussed below;
- it shares the typical quadratic convergence from Newton method when X_n is close to the true value.

The concerns are:

- the computation of Jacobian J can be very expensive, especially in the context of this study as explained in Appendix A;
- the explicit form of J leads to memory issues and the inversion of dense matrix $J^T J$ is expensive, which makes the 3D generalization difficult.

1.3. Nonlinear gradient-based method

Alternatively, (1.1) can also be treated as directly minimizing nonlinear functional f , for example, solved by Quasi-Newton method (Nocedal and Wright 1999), so that the computation of Jacobian J is avoided and only the gradient $\nabla f := [\partial_{x_k} f]$ needs to be computed. Compared with the Jacobian-based method, although the gradient-based method in general has superlinear convergence, it is efficient in computation speed and economic in memory storage, which suits the large-scale 3D problem.

Similarly as in the previous situation, regularization should be incorporated into the algorithm to reduce the ill-conditioning. In the case with L2 regularization, a counterpart of iterative scheme for the gradient-based method as (1.4) for the Jacobian-based method is

$$X_{n+1} = \arg \min_X f(X) + \lambda_{2,n} \|X - X_n\|_2^2, \quad (1.5)$$

where $\lambda_{2,n}$ is decreasing during iterations so that the regularized solution converges to the true solution.

A simple update strategy is $\lambda_{2,n+1} = \lambda_{2,n} / 2$.

A popular Quasi-Newton method for solving (1.5) is limited-memory BFGS method (L-BFGS) with details given in Appendix B.

1.4. Regularizations

The appropriate choice of regularizations depends on *a priori* knowledge of X . Although L2 regularization is a natural choice for its simplicity, it is by no means the optimal strategy. For example, when X is sparse, it is well known that L1 regularization is more efficient than L2 regularization in reconstructing the sparse solution (Gao and Zhao 2010a, b).

In this study, piecewise constants are used to approximate optical parameters, which is genuinely appropriate in most situations. For piecewise constant approximation, it is well known that the regularization through bounded total variation (TV) overperforms L2 regularization. TV (Rudin *et al* 1992) has been studied over decades in image processing and has evolved into a fruitful field with many effective and efficient algorithms. With TV, the edges or the shape are well preserved while the artifacts are significantly penalized. Moreover, the power of TV, in terms of improving both accuracy and efficiency, can be significantly enhanced when combined with Bregman methods as introduced later.

Our problem is unique in the sense that TV needs to be defined on unstructured grids. In (Gao and Zhao 2010b), one innovation is to use the coarea formula of TV to acquire the following simple and exact TV on 2D triangular grids or 3D tetrahedral grids

$$\|X\|_{TV} := \sum_{k=1}^{N_e} l_k |x_{k,l} - x_{k,r}|, \quad (1.6)$$

in which the summation is with respect to uniquely-ordered internal edges of elements with the total N_e , l_k is the length or the area of the k th edge, $x_{k,l}$ and $x_{k,r}$ are two adjacent variables sharing

the k th edge. Notice that we assume the grid is conforming, i.e., each internal edge is shared by two elements. In matrix form on the discretized grid, TV is simply

$$\|X\|_{TV} = |MX|, \quad (1.7)$$

where M is a sparse $N_e \times N$ matrix with l_k 's as entries and $|\cdot|$ is the conventional L1 norm.

Similarly, to account for mesh non-uniformity, we use the following weighted L2 norm

$$\|X\|_2^2 := \sum_{k=1}^N w_k x_k^2, \quad (1.8)$$

in which the summation is with respect to elements with the total N , w_k is the area or the volume of the k th element, and x_k is the variable on the k th element. In another form,

$$\|X\|_2^2 = X^T W X, \quad (1.9)$$

where W is a diagonal matrix with w_k 's as entries.

Notice that, since there are two parameters to be recovered in quantitative PAT, each parameter of μ_a and μ_s has a counterpart of (1.7) or (1.9) in the minimization problem we are considering.

1.5. Bregman method

To be consistent with conventional notations, now we consider the minimization problem with the data fidelity $H(X)$ and the regularization $R(X)$, i.e.,

$$X = \arg \min_X H(X) + R(X). \quad (1.10)$$

The motivation for the Bregman method comes from a long-standing denoising problem via Rudin-Osher-Fatemi (ROF) model (Rudin *et al* 1992), which is one of the most successful and popular models, yet can be further improved in terms of both accuracy and computational efficiency. ROF is defined as follow

$$X = \arg \min_X \|X - Y\|_2^2 + \lambda \|X\|_{TV}, \quad (1.11)$$

in which the first term is the data fidelity with the noisy data Y and the second term is TV regularization with parameter λ . Here the use of TV is essential for the recovery of images with edges, which would be impossible for the regularization $\int |\nabla X|^p$ with any $p > 1$.

However, the major weakness of ROF (1.11) is that the contrast can not be fully recovered. One interesting illustrative example given by Meyer (2001) is as follow. Let $Y(x, y) = \alpha \chi_R(x, y)$, where $\chi_R(x, y)$ is a typical support function in 2D such that $\chi_R(x, y) = 1$ if $(x^2 + y^2)^{1/2} \leq R$, $\chi_R(x, y) = 0$ otherwise. It was shown that through ROF (1.11), when $\alpha R \geq \lambda$, $X = (\alpha - \lambda/R) \chi_R$; otherwise, $X = 0$. Notice that the loss of contrast λ/R is independent of the true value α .

Inspired by the improvement of geometric processing via normal maps, the following two-stage iterative TV regularization procedure was proposed by Osher *et al* (2005) with $v_0 = 0$

$$\begin{aligned} X_{n+1} &= \arg \min_X \|X - (Y + v_n)\|_2^2 + \lambda \|X\|_{TV} \\ v_{n+1} &= v_n + Y - X_{n+1} \end{aligned} \quad (1.12)$$

That is, the residual v_n computed through ROF should be added back to the data Y to be corrected in the next ROF. Surprisingly, for the above simple example, when $\alpha R \geq \lambda$, the loss of contrast can be exactly recovered through (1.12) in two iterations! When $\alpha R < \lambda$, there exists the smallest n with $n\alpha R \geq \lambda$, such that $X_k = 0, k < n$, $X_n = (\alpha - \lambda/R) \chi_R$ and $X_k = \alpha \chi_R, k > n$. From this, one can tell that (1.12) has a strongly nonlinear feature in the sense that a sequence of restored images, all of which are totally black, can go from “nothing” to the true result in two steps! In general cases, this add-residual-back algorithm (1.12) can recover the lost contrast from (1.11) to a large degree.

One notable difference in (1.12) is that λ is fixed, while it may be adaptively adjusted in LM (1.4). However, if $\|X\|_{TV}$ were replaced by $\|X\|_2^2$, the algorithm (1.12) coincides with LM (with the fixed regularization parameter) as explained in Appendix C.

In (Osher *et al* 2005), it was also shown that (1.12) is a special case of the following general iterative regularization procedure, which will be observed shortly

$$X_{n+1} = \arg \min_X H(X) + D(X, X_n), \quad (1.13)$$

where D is the Bregman distance defined by $D(X, X_n) := R(X) - R(X_n) - \langle P_n, X - X_n \rangle$ with a subgradient P_n . This is why the algorithm is named after Bregman. In a simpler form, Bregman iterations is

$$\begin{aligned} X_{n+1} &= \arg \min_X H(X) + R(X) - \langle P_n, X \rangle \\ P_{n+1} &= P_n - \partial_X H(X_{n+1}) \end{aligned} \quad (1.14)$$

The well-definedness and the convergence of (1.14) were proved when $R(X)$ is convex and $H(X) = \|AX - Y\|_2^2$ with A as a bounded linear operator whose kernel does not include the space of continuous functions.

With this specific least-square form of H , $\partial_X H = A^T(AX - Y)$. Thus, $P_n = -\sum_{k=1}^n A^T(AX_k - Y)$ and $-\langle P_n, X \rangle = -\sum_{k=1}^n \langle Y - AX_k, AX \rangle$. Hence, (1.14) is transformed to the following add-residual-back form

$$\begin{aligned} X_{n+1} &= \arg \min_X \|AX - (Y + v_n)\|_2^2 + R(X) \\ v_{n+1} &= v_n + Y - AX_{n+1} \end{aligned} \quad (1.15)$$

Notice that, when $R(X) = \lambda \|X\|_{TV}$ and $A = I$, (1.15) is exactly (1.12).

In (Bachmayr and Burger 2009), the Bregman method was analyzed for the nonconvex problem with nonlinear dependence, such as $\vec{F}(X)$ in (1.1). However, it is not clear how the required nonlinearity condition can be satisfied in quantitative PAT.

In the following, we will mainly show the superiority of iterative TV regularization procedure through Bregman method, i.e., (1.14) or (1.15), from numerical and practical aspects. The theoretical justifications will be studied in the future.

Last, it is worthy of mentioning that the similar Bregman ideas can be used in designing a computationally efficient solver of TV-regularized minimization problem, such as the one appearing in (1.15). This is what so called split Bregman method (Goldstein and Osher 2009), which is implemented in this study with necessary modifications. The details of the method are given in Appendix E.

2. Algorithms

2.1. Bregman method for Jacobian-based method

2.1.1. L2 regularization. As discussed in section 1.2, a popular solution strategy for quantitative PAT is via iterative linearizations, in each step of which we first compute the Jacobian J and then minimize a convex regularized least-square problem. The computation of J can be found in Appendix A. The minimization with L2 regularization can be solved similarly as (1.4). However, the algorithm we use is slightly different from (1.4) considering both efficiency and robustness.

In the elaborated version of (1.4), we first consider only the linearization part of the original problem (1.1) via (1.2) to arrive at

$$\begin{aligned} \delta X_{n+1} &= \arg \min_{\delta X} \| J(X_n) \delta X - (Y - \bar{F}(X_n)) \|_2^2 \\ X_{n+1} &= X_n + \delta X_{n+1} \end{aligned} \quad (2.1)$$

Then L2 regularization is applied to the minimization in (2.1) to deal with its illposed nature. To simplify the discussion, let $A = J(X_n)$, $u = \delta X$ and $B = Y - \bar{F}(X_n)$, that is we consider the problem

$$u = \arg \min_u \| Au - B \|_2^2. \quad (2.2)$$

In the case of L2 regularization, (2.2) can be efficiently solved through

$$\begin{aligned} u_{n+1} &= \arg \min_u \| Au - (B + v_n) \|_2^2 + \lambda \| u \|_2^2 \\ v_{n+1} &= v_n + B - Au_{n+1} \end{aligned} \quad (2.3)$$

Notice that (2.3) is a Bregman formulation, however it is equivalent to LM in this case of L2 regularization for reasons explained in Appendix C. The main difference is that λ is fixed in (2.3).

Comparing the new version (2.3) with the traditional version (1.4), the linearization and the regularization are enforced at the same time in (1.4); additionally, the regularization parameter λ is decreased in order for the regularized solution to converge to the true solution. However, (1.4) has two drawbacks.

- First, in early iterations, the large value of λ may prevent the computed solution X_{n+1} as a close approximation of the model, i.e., the data fidelity. Therefore the convergence of algorithm is slow. As a result the computation burden is significantly increased since not only more minimizations need to be solved, but also more Jacobian J needs to be computed, which is very expensive in quantitative PAT, as explained in Appendix A.
- Second, as λ is decreasing, it may happen in latter iterations that the regularization is too weak to be able to regularize the solution, which causes the stability issue.

In contrast, the linearization and the regularization are performed independently through (2.2) and (2.3) in the improved version.

- First, each linearization (2.2) is solved to the extreme so that X_{n+1} is indeed a good approximation of the model. Notice that the minimization in (2.3) is much cheaper than computation of J .
- Second, λ is fixed in (2.3) so that the algorithm is robust.

2.1.2. TV regularization. Despite its popularity, L2 regularization is by no means the optimal one overall. As an alternative, TV regularization (Rudin *et al* 1992) has been studied over decades and its superiority over L2 regularization in recovering images with sharp edges and less artifacts has been well recognized. Despite its existing success and popularity, TV is defect in the sense that the contrast cannot be fully recovered as analyzed in detail (Meyer 2010). To solve this loss-of-contrast issue, Bregman method with TV regularization was proposed (Osher *et al* 2005) and it has been proved to be capable of improving the restored contrast quantitatively. The novel idea is to add the residual, i.e., the error made at the current iteration step, back to the data for the next iteration to be corrected. In the context of quantitative PAT, in contrast of (2.3) for L2-regularized problem, the Bregman method with TV regularization is as follow with $v_0 = 0$

$$\begin{aligned} u_{n+1} &= \arg \min_u \| Au - (B + v_n) \|_2^2 + \lambda \| u \|_{TV} \\ v_{n+1} &= v_n + B - Au_{n+1} \end{aligned} \quad (2.4)$$

That is, to solve the original problem (1.1), we first linearize the problem by (2.1) and solve the linearized subproblem by Bregman method with TV via (2.4) instead of L2 norm via (2.3).

Notice that no similar equivalence between Bregman and LM exists for TV regularization. It is the synergistic combination of TV and Bregman method through (2.4) to achieve the desired quantitative

improvement, which will be shown through simulations later. Actually, when TV were combined with LM, i.e., L2 norm were replaced by TV in (1.4), the common loss-of-contrast pitfall would occur (Gao and Zhao 2010) although the performance would be better than with L2 norm.

On the other hand, TV problem is one of the most difficult optimizations to be solved computationally, which sometimes limits the interest of using TV in practice. Fortunately, besides its capability of improving quantitative accuracy with TV, Bregman method can also be used to design efficient algorithm to solve L1-type problems, including TV, what is so called split Bregman method (Goldstein and Osher 2009). Since the method is similar to what we have discussed, the technical details are given in Appendix E.

2.1.3. Sensitivity scaling. Another computational challenge, which is quite different from abovementioned algorithms, is due to the significant difference between sensitivity of data to the absorption coefficient μ_a and that to the scattering coefficient μ_s . In the Jacobian-based method, this sensitivity difference is reflected through Jacobian J , which is explained in Appendix A.

In order to balance the sensitivity, appropriate scaling needs to be introduced so that the simultaneous reconstruction of μ_a and μ_s is feasible. This can be done through the following

$$J_{ij,k}^{new} = J_{ij,k} / s_k, s_k = (\sum_{i,j} J_{ij,k}^2)^{1/2}, k \leq N. \quad (2.5)$$

Here we use the same notations introduced in section 1.1. That is we scale J for each variable (column) by dividing the L2 norm of that variable (column) to get new Jacobian J^{new} , which has uniform magnitude across all variables. After the scaling of the Jacobian, all the previous algorithms still apply to the following instead of (2.1)

$$\begin{aligned} \delta X_{n+1} &= \arg \min_{\delta X} \| J^{new}(X_n) \delta X - (Y - \bar{F}(X_n)) \|_2^2, \\ X_{n+1} &= X_n + \delta X_{n+1} / S \end{aligned} \quad (2.6)$$

where δX_{n+1} is rescaled accordingly by the scaling vector $S = [s_k]$.

2.1.4. Summary. With multiple optical excitations, the simultaneous reconstruction of absorption and scattering coefficients can be achieved through the aforementioned algorithms with proper sensitivity scaling, such as (2.5). The feasibility was first shown numerically in (Gao and Zhao 2010). The benefit of the Jacobian-based method is that the minimization problem (2.2) after the linearization commonly appears and many existing algorithms, such as TV and Bregman method, can be easily modified to potentially improve the performance. However, there are two major concerns for the Jacobian-based method (2.1), which make its practical use a challenge task from computational perspectives, especially in 3D quantitative PAT.

- First, the computation of J requires solving forward model at least $N_s + N_d$ times. In quantitative PAT, N_d is roughly on the same order as N , the number of variables to be recovered. Therefore, the computation of J is very expensive, even in the case with diffusion approximation as forward model. See Appendix A for detailed explanations.
- Second, J is dense and overdetermined, which imposes significant challenge for memory storage. Besides, the computation involving J is time-consuming.

2.2. Bregman method for gradient-based method

The aforementioned limitation of the Jacobian-based method motivates the study of the gradient-based method. Here the computation of J is not necessary, hence the storage is not an issue any more and the generalization to 3D is possible. Rather we compute the gradient instead of Jacobian and each one

requires the computation of forward solver only $2N_s$ times, which saves a lot on computational time since N_s has single digit or at most the order of tens.

The main purpose of this paper is to study the gradient-based method with TV and Bregman method to reach the similar quantitative improvement as Jacobian-based method. In the following, we will develop an effective and efficient Bregman algorithm for directly solving nonlinear problem (1.1), and the improvement is demonstrated through numerical tests in section 3.

2.2.1. Formulation. As discussed previously, the minimization (1.1) needs to be regularized due to its ill-posed nature or the noise in the data, which can be realized through L2 regularization (1.5). For the similar reasons explained in section 2.1.1., we actually solve the following Bregman version of (1.5)

$$\begin{aligned} X_{n+1} &= \arg \min_X f(X) + \lambda \|X\|_2^2 - \langle P_n, X \rangle, \\ P_{n+1} &= P_n - \partial_X f(X_{n+1}) \end{aligned} \quad (2.7)$$

which is adapted from the original form of Bregman method (1.14) with f defined in (1.1). Notice that the regularization parameter λ is fixed in (2.7) in contrast of the decreasing λ in (1.5), which is a unique feature of Bregman method. The convergence and stability has been studied (Bachmayr and Burger 2009) that, with certain nonlinear condition on \vec{F} , the regularized solution of (2.7) with a fixed λ converges. Although it is not clear how the condition can be satisfied in our situation, the improvement can be observed in simulations as in section 3. In practice, large λ is preferred for both the stability and the performance of the algorithm.

Similarly, with TV regularization, the Bregman method is

$$\begin{aligned} X_{n+1} &= \arg \min_X f(X) + \lambda \|X\|_{TV} - \langle P_n, X \rangle, \\ P_{n+1} &= P_n - \partial_X f(X_{n+1}) \end{aligned} \quad (2.8)$$

We shall compare (2.7) and (2.8) through simulations in section 3, which will show once again that it is the synergetic combination of TV and Bregman to achieve the desirable quantitative improvement in the gradient-based method.

2.2.2. Algorithms. The solution to the following nonlinear minimization problem with differentiable function $h(X)$ serves as a building block to either of (2.7) and (2.8)

$$X = \arg \min_X h(X). \quad (2.9)$$

Here we choose the limited-memory BFGS (L-BFGS) for solving (2-9), which is suitable for large-scale optimization problem. For the flow of the presentation, the technical details are given in Appendix B. Notice that the proper scaling is again introduced in Appendix B in order for the simultaneous reconstruction of absorption and scattering coefficients similar as in section 2.1.3.

For (2.7) with L2 regularization, $h(X) := f(X) + \lambda \|X\|_2^2 - \langle P_n, X \rangle$ is differentiable and the algorithm for (2.9) can be directly applied to subproblems of (2.7). For (2.8) with TV regularization, the L-BFGS solution for (2.9) can be incorporated in the split Bregman method as a solver of the subproblems appearing in (E.7).

Regarding the implementation difference, the Bregman method is used to solve the linearized subproblems (2.1) in the Jacobian-based method, while it is utilized directly for solving the original nonlinear problem (1.1) in the gradient-based method.

3. Results

The comparison of the Bregman method with LM have been done previously (Gao and Zhao 2010) for the Jacobian-based method with the conclusion that Bregman method gives better quantitative results

than LM with less artifacts in the case of TV regularization. The similar conclusion exists for the gradient-based method, however will not be presented here.

In the following, we will mainly compare “L2”, the Bregman method with L2 regularization, and, “TV”, the Bregman method with TV regularization, for both Jacobian-based and gradient-based methods. Meanwhile, we will show the feasibility of 3D reconstruction using gradient-based method.

Diffusion approximation is used as forward model for simplicity with numerical solution by finite element method (Appendix A). Light density ϕ_i is approximated piecewise-linearly and the variable $X = (\mu_a, \mu_s)$ is approximated piecewise-constantly. The mesh for data generation is different from the one for reconstruction to avoid inverse crime. 5% Gaussian noise is added to the synthetic data. The initial guess X^0 for reconstructions is the same as the true background. Notice that the reconstructions with wrong background value as initial guesses have been reported previously (Gao and Zhao 2010), which shows that the Bregman method is not highly sensitive to the choice of initial guesses. In all figures and tables, the recovered value is the contrast with respect to the true background.

3.1. 2D results

The 2D simulations are performed on a circular phantom (table 1). Four optical excitations are simulated, i.e., the homogenous light sources are assigned at the boundary of each quadrant successively.

First, we compare “L2” (2.3) and “TV” (2.4) for the Jacobian-based method. Notice that “L2” is equivalent to LM with L2 regularization as explained in Appendix C. The reconstruction results are plotted in figure 1 and the recovered mean contrast of inclusions with respect to the true background is summarized in table 1. From the results, “TV” gives better quantitative results than “L2” (or LM) with less artifacts. Therefore, it is the synergistic combination of TV and Bregman method through (2.4) to achieve the desired quantitative improvement.

Next, we compare “L2” (2.7) and “TV” (2.8) for the gradient-based method. The reconstruction results are plotted in figure 2 and the recovered mean contrast of inclusions with respect to the true background is summarized in table 1. Once again, the simulations confirm the quantitative improvement through the synergistic combination of TV and Bregman method via (2.8). Notably, with “TV”, the inclusions are visually separable as in figure 2 in contrast of smoothed inclusions with “L2”.

For the completeness, we will describe the stopping criterions briefly in the following.

In the Jacobian-base method, the stopping criterion for the outer loop, i.e., the linearization through (2.1), is by the mean of ratio difference between successive iterations, i.e., $\|(X^{n+1} - X^n) / X^n\|_2 < 0.01$, and it takes around ten iterations for the presented tests. For inner loop (2.3) or (2.4) for solving each step in (2.1), the number of iterations is fixed empirically for the reasons explained in (Osher *et al* 2005). That is without interference the computed solution would eventually converge to “the true solution” defined by the data with noise. Since no priors are assumed on the noise level, we use large regularization parameter with a fixed number of iterations. In this study, 3 iterations are used with $\lambda = 1$. Besides, in “TV”, there is another loop for solving TV-regularized problems in (2.4) via split Bregman method (Appendix E), for which the algorithm takes around 10 iterations with the similar abovementioned incremental change as the stop criterion.

Next we comment on the gradient-base method. For “L2”, the outer loop, i.e., the Bregman iteration (2.7), is with fixed number of iterations starting with a large regularization parameter for the similar reason explained above. In this study, 6 iterations are used with $\lambda = 0.01$. In the inner loop for solving the nonlinear optimization problem in (2.7) via L-BFGS (Appendix B), $\|\nabla h\|_2 < 0.01$ is used as the stopping criterion and it takes around 30 iterations. For “TV”, there are three major loops instead of two: it is similar for the outer loop (2.8), i.e., 6 iterations with $\lambda = 1$; in the intermediate loop for solving TV-regularized problems in (2.8) via split Bregman method (Appendix E), the algorithm takes around 10 iterations with the same aforementioned relative change as the stop criterion; in the

inner loop for the nonlinear minimization subproblems in Split Bregman method, L-BFGS is used with the same stopping criterion as in “L2” and it takes around 30 iterations.

Regarding the computation efficiency in this study, it is clear that gradient-based method saves a lot of memory due to the circumvention of Jacobian. On the other hand, gradient-based method is faster than Jacobian-based method, which can be justified as follow. Take L2-regularized algorithms in 2D for example. In our simulations, each computation of Jacobian takes thousands of number of computing forward solver while each computation of gradient takes eight computations of forward solver. Although the total number of computing the gradient is about ten times of that of computing the Jacobian, the gradient-based method is still be able to save the computation time by at least 10 times. This gain due to the computation of the gradient instead of the Jacobian would be further announced in 3D. Besides, the minimization (L-BFGS) in the gradient-based method only involves sums or multiplications of vectors while it is expensive to perform operations of the Jacobian in the Jacobian-based method, which is dense and overdetermined. As a sacrifice, it seems that the gradient-based method does not give as accurate results as the Jacobian-based method both visually and quantitatively.

Table 1. 2D reconstruction results. The recorded value (μ_a, μ_s) are the contrast with respect to the true background value (0.01, 1). The circular phantom of diameter 40 is centered at (0, 0) with nine 3-diameter circular inclusions centered at (x, y). The length unit is in millimeter.

Inclusion	(x, y)	True Value	Jacobian-based Method		Gradient-based Method	
			L2	TV	L2	TV
1	(-16, 0)	(2, 1)	(1.9, 1.1)	(2.0, 1.0)	(2.0, 1.2)	(2.0, 0.9)
2	(-12, 0)	(2, 1)	(1.9, 1.1)	(2.0, 1.0)	(2.0, 1.1)	(2.0, 0.9)
3	(-8, 0)	(2, 1)	(1.9, 1.0)	(2.0, 1.0)	(2.0, 1.0)	(2.0, 1.0)
4	(-4, 0)	(2, 1)	(1.9, 0.9)	(2.0, 1.0)	(2.0, 1.0)	(2.0, 1.0)
5	(0, 0)	(2, 2)	(1.9, 1.8)	(2.0, 2.3)	(2.1, 1.7)	(2.0, 2.0)
6	(4, 0)	(1, 2)	(1.0, 2.0)	(1.0, 2.5)	(1.0, 1.8)	(1.0, 2.3)
7	(8, 0)	(1, 2)	(1.0, 2.1)	(1.0, 2.1)	(1.0, 2.1)	(1.0, 2.4)
8	(12, 0)	(1, 2)	(1.0, 1.7)	(1.0, 2.0)	(1.0, 1.8)	(1.0, 2.0)
9	(16, 0)	(1, 2)	(1.0, 2.2)	(1.0, 2.1)	(1.1, 1.3)	(1.1, 1.5)

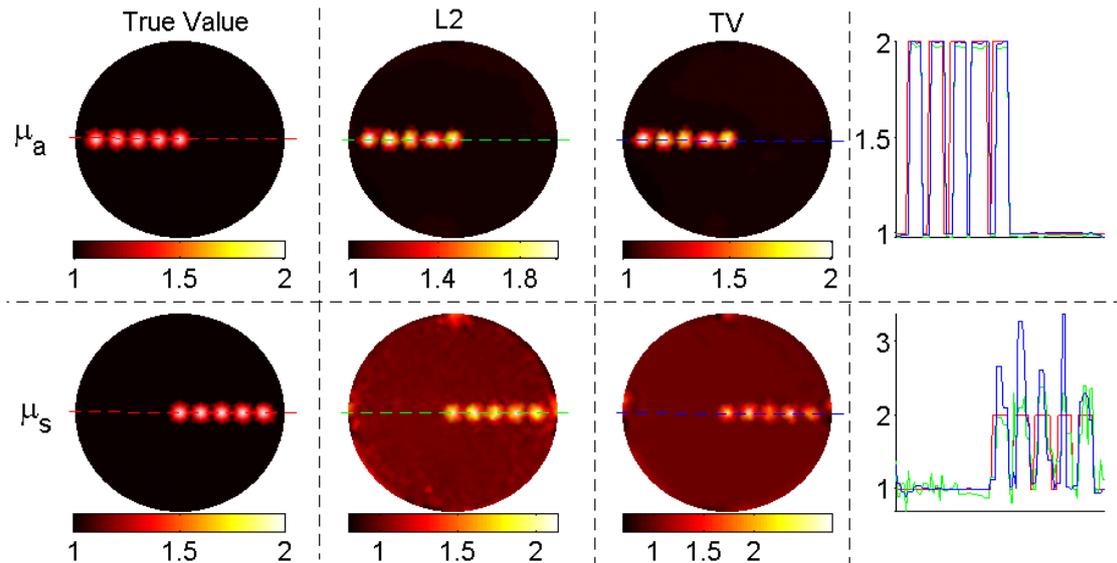


Figure 1. 2D reconstruction results from Bregman method for Jacobian-based method.

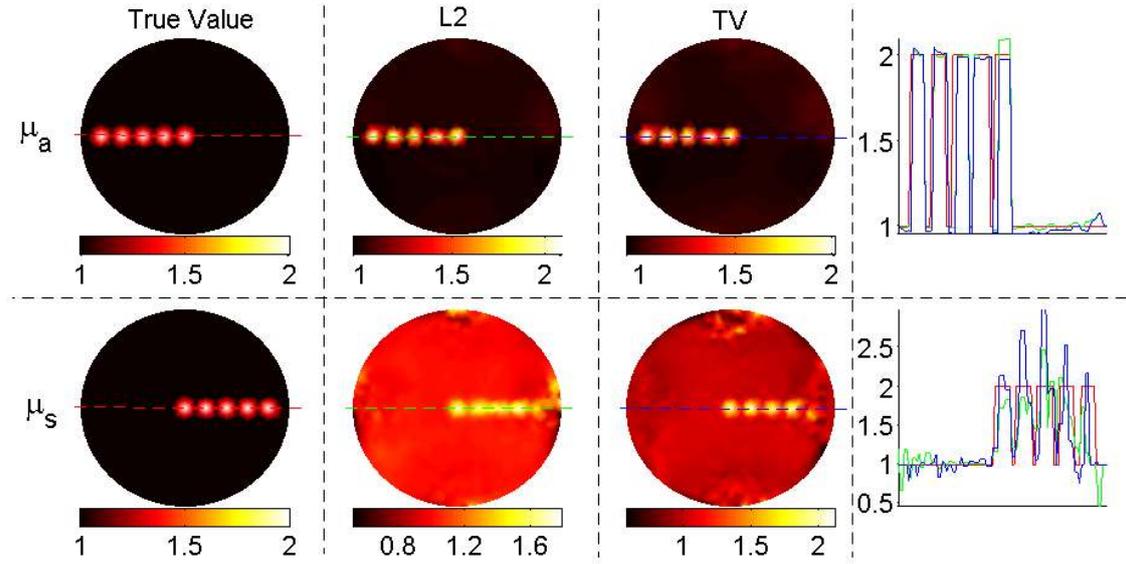


Figure 2. 2D reconstruction results from Bregman method for gradient-based method.

3.2. 3D results

The main purpose here is to show the feasibility of 3D quantitative PAT. In this setting, since it is beyond our ability to run Jacobian-based simulations on a desktop computer due to its memory requirement, we will only show the results with gradient-based method.

The 3D simulations are performed on a cylindrical phantom (table 2). Six optical excitations are simulated, i.e., two homogenous excitations are simulated from the top and the bottom, and four homogenous excitations, similar as in 2D, surround the cylinder with each one covering each quadrant with respect to transverse plane.

Here we compare “L2” (2.7) and “TV” (2.8) for the gradient-based method. The reconstruction results are plotted in figure 3 and the recovered mean contrast of inclusions with respect to the true background is summarized in table 2. The conclusion is consistent with both theory and previous simulations. That is “TV”, Bregman method with TV regularization, is quantitatively better than “L2”.

The 3D reconstruction mesh we use contains approximately 25000 elements and 5000 nodes, and the variables we need to recover are roughly 50000. On a Desktop computer with Intel Core 2 Duo E6850 3.0GHz CPU, with “TV” for the gradient-based method, the computation takes one to two hours in 3D in contrast of minutes in 2D on MATLAB as the simulation platform. On the other hand, the computational time for “L2” is around one tenth of that of “TV”, since solving “TV” brings extra cost through split Bregman method, each iteration of which is equivalent of solving a “L2” problem.

Table 2. 3D reconstruction results. The recorded value (μ_a, μ_s) are the contrast with respect to the true background value (0.01, 1). The cylindrical phantom with diameter 40 and height 40 is centered at (0, 0, 20) with ten 3-diameter spherical inclusions centered at (x, y, z). The length unit is in millimeter.

Inclusion	(x, y, z)	True Value	Gradient-based Method	
			L2	TV
1	(-16, 0, 28)	(2, 1)	(1.9, 1.2)	(1.9, 1.1)
2	(-12, 0, 28)	(2, 1)	(1.9, 1.1)	(2.0, 1.0)
3	(-8, 0, 28)	(2, 1)	(2.3, 1.0)	(2.0, 1.0)
4	(-4, 0, 28)	(2, 1)	(1.7, 1.0)	(2.0, 1.0)
5	(0, 0, 28)	(2, 1)	(1.8, 1.0)	(2.0, 1.0)
6	(0, 0, 12)	(1, 2)	(1.0, 1.5)	(1.0, 1.7)

7	(4, 0, 12)	(1, 2)	(1.1, 1.9)	(1.0, 2.0)
8	(8, 0, 12)	(1, 2)	(1.0, 2.2)	(1.0, 2.1)
9	(12, 0, 12)	(1, 2)	(1.1, 1.8)	(1.0, 2.0)
10	(16, 0, 12)	(1, 2)	(1.0, 1.7)	(1.0, 1.7)

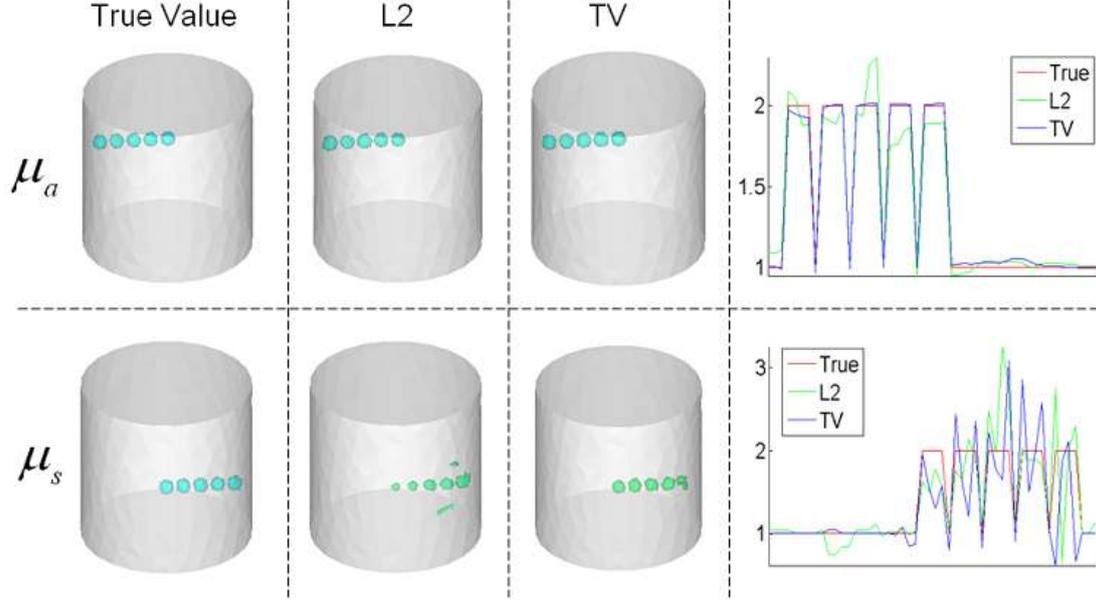


Figure 3. 3D reconstruction results from Bregman method for gradient-based method.

Acknowledgments

This work is partially supported by NSF grant DMS0811254.

Appendix A. Computation of Jacobian and gradient

Let us first formulate the finite element method for diffusion approximation (DA) (Arridge 1999) before computing Jacobian and gradient. Here we consider the following DA as forward model with Robin boundary condition for each optical excitation $q_i, i \leq N_s$,

$$\begin{aligned} -\nabla(D \cdot \nabla)\phi + \mu_a \phi &= q \\ \phi + 2AD\partial\phi/\partial n &= 0 \end{aligned} \quad (\text{A.1})$$

where D is the diffusion coefficient defined by $D := 1/(3\mu_s)$ and A is boundary constant specified by the medium.

Let $\{\varphi_j, j \leq N_p\}$ be the piecewise-linear bases for ϕ and $\{\chi_k, k \leq N\}$ be the piecewise-constant bases for $X = [x_k := (\mu_{a,k}, \mu_{s,k})]$. Then (A.1) is discretized to the following linear system

$$F\phi = Q, \quad (\text{A.2})$$

with $F_{ij}(\mu_a, \mu_s) = \int_{\Omega} [D(\mu_s)\nabla\varphi_j \cdot \nabla\varphi_i + \mu_a\varphi_j\varphi_i] + \int_{\partial\Omega} \varphi_j\varphi_i/2A$ and $Q_j = \int_{\Omega} q\varphi_j$. Differentiating (A.2) with respect to the variable x_k , we have

$$F \frac{\partial\phi}{\partial x_k} = \left(-\frac{\partial F}{\partial x_k}\right)\phi. \quad (\text{A.3})$$

Now let us compute the Jacobian by noticing that the measuring functional P_j is a linear interpolation δ_{jk} of ϕ multiplied by $\mu_{a,k}$. That is the following via (A.3)

$$J_{ij,k} = \frac{\partial(P_j^T \phi_i)}{\partial x_k} = P_j^T \frac{\partial \phi_i}{\partial x_k} + \frac{\partial P_j^T}{\partial x_k} \phi_i = \varphi_j^T \left(-\frac{\partial F}{\partial x_k}\right) \phi_i + \delta_{jk}^T \phi_i, \quad (\text{A.4})$$

where φ_j is a adjoint solution for the j th detector, i.e., $\varphi_j = (F^T)^{-1} P_j$. Here the second term of (A.4) only exists for $\mu_{a,k}$. Therefore, the data is much more sensitive to μ_a than to μ_s , and the proper scaling needs to be introduced in order for the simultaneous reconstruction of both parameters, such as (2.5).

Next, let us compute $\nabla f := [\partial f / \partial x_k]$.

$$\begin{aligned} \frac{\partial f}{\partial x_k} &= \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} (\bar{F}_{ij}(X) - y_{ij}) \frac{\partial \bar{F}_{ij}}{\partial x_k} \\ &= \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} (\bar{F}_{ij}(X) - y_{ij}) \left[P_j^T \frac{\partial \phi_i}{\partial x_k} + \frac{\partial P_j^T}{\partial x_k} \phi_i \right] \\ &= \sum_{i=1}^{N_s} \left\{ \left[\sum_{j=1}^{N_d} (\bar{F}_{ij}(X) - y_{ij}) P_j^T \right] \frac{\partial \phi_i}{\partial x_k} \right\} + \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} (\bar{F}_{ij}(X) - y_{ij}) \delta_{jk}^T \phi_i \end{aligned} \quad (\text{A.5})$$

Let $Q_i^A = \sum_{j=1}^{N_d} (\bar{F}_{ij}(X) - y_{ij}) P_j$ and $\varphi_i = (F^T)^{-1} Q_i^A$ be the adjoint solution for the i th adjoint source.

Using (A.3) and adjoint solution, (A.5) is simplified to

$$\frac{\partial f}{\partial x_k} = \sum_{i=1}^{N_s} \varphi_i^T \left(-\frac{\partial F}{\partial x_k}\right) \phi_i + \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} (\bar{F}_{ij}(X) - y_{ij}) \delta_{jk}^T \phi_i. \quad (\text{A.6})$$

Similarly, the second term of (A.6) exists only for $\mu_{a,k}$. As a result, the special consideration is required in the algorithm for the simultaneous reconstruction of μ_a and μ_s , such as for L-BFGS in Appendix B.

For comparison, the computation of the gradient via (A.6) requires solving linear system (A.2) $2N_s$ times while the computation of the Jacobian via (A.4) requires $N_s + N_d$ times. In quantitative PAT, since $N_s \ll N_d \approx N$, the gradient is computationally much cheaper than the Jacobian.

Appendix B. BFGS and L-BFGS

For notation convenience, we use f to represent the sum of data fidelity and (differentiable) regularization. Notice that the following method requires the differentiability of f . In the case with TV regularization, although the method does not directly apply, it can be utilized as a solver of subproblems in Split Bregman method (Appendix E). The following description of BFGS is adapted from (Nocedal and Wright 1999) with necessary modifications for our purpose.

BFGS is an iterative scheme to minimize f . At the current step X_n , f is approximated quadratically as

$$h_n(\delta X) = f(X_n) + \nabla f_n^T \delta X + \frac{1}{2} \delta X^T B_n \delta X, \quad (\text{B.1})$$

in which ∇f_n is the gradient at X_n , B_n is an positive symmetric definite matrix to mimic the Hessian that will be updated during iterations. From the optimal condition of (B.1), we have

$$\delta X_n = -B_n^{-1} \nabla f_n. \quad (\text{B.2})$$

Then, we update $X_{n+1} = X_n + \alpha_n \delta X_n$, in which α_n is the step length found through some line search methods. The backtracking line search with the Armijo condition is used in this study.

Although Wolfe conditions may accelerate the algorithm, the extra computation of ∇f would be required.

Next we discuss the update of B_{n+1} from B_n . First, by the requirement that the gradient of h_{n+1} should match the gradient of f at both X_n and X_{n+1} , one can derive the well-known secant equation $B_{n+1}s_n = y_n$, with $s_n := X_{n+1} - X_n$ and $y_n := \nabla f_{n+1} - \nabla f_n$. Then, B_{n+1} can be uniquely determined among all symmetric matrices satisfying secant equation through minimizing the difference between B_{n+1} and B_n , for which the weighted Frobenius norm is used due to its simplicity.

The similar approach to update B_n can be used to update the inverse matrix $H_n =: B_n^{-1}$, which can be multiplied directly in (B.2) to get the search direction instead of through matrix inversion. Then with definitions $V_n := I - \rho_n y_n s_n^T$ and $\rho_n := 1 / y_n^T s_n$, we arrive at the following famous BFGS updating formula of H_{n+1} from H_n

$$H_{n+1} = V_n^T H_n V_n + \rho_n s_n s_n^T. \quad (\text{B.3})$$

Despite its computation advantage, BFGS is still not applicable for large-scale computation due to the explicit form of H_n . In L-BFGS, a modification of BFGS, H_n does not appear explicitly and only a few vectors needs to be stored. Briefly, the computation of $H_n \nabla f_n$ for search direction δX_n through (B.2) and (B.3) can be regarded as an iterative process. In L-BFGS, this iterative procedure is truncated involving only the most recent m pairs of (s, y) instead of all pairs. This approximate computation of $H_n \nabla f_n$ can be elegantly realized through a two-loop recursion algorithm, which involves only the multiplication of vectors. In this study, $m = 5$ is fairly robust.

Last but not the least, due to the sensitivity difference between absorption and scattering coefficient in quantitative PAT, that is $\partial_{\mu_a} f$ is much larger than $\partial_{\mu_s} f$, the scattering coefficient μ_s can not be recovered as well as the absorption coefficient μ_a unless the appropriate scaling is taken into consideration. Here and in the following, f denotes only the data fidelity term.

The idea is to balance $\partial_{\mu_a} f$ and $\partial_{\mu_s} f$ through weighting μ_a and μ_s to achieve simultaneous reconstruction of two parameters. The scaling goes as follow:

- Step 1: Given X_0 , compute $w_{0, \mu_a} = \|\partial_{\mu_a} f(X_0)\|_2$ and $w_{0, \mu_s} = \|\partial_{\mu_s} f(X_0)\|_2$, and then scale regularization parameters by $\lambda_{\mu_a}^w = w_{0, \mu_a} \lambda_{\mu_a}$ and $\lambda_{\mu_s}^w = w_{0, \mu_s} \lambda_{\mu_s}$;
- Step 2: at each current step X_n , first compute the weight $w_n := (w_{n, \mu_a}, w_{n, \mu_s})$ with $w_{n, \mu_a} = \|\partial_{\mu_a} f(X_n)\|_2$ and $w_{n, \mu_s} = \|\partial_{\mu_s} f(X_n)\|_2$, and then consider the weighted nonlinear optimization problem with $X_n^w = w_n X_n$, for which the corresponding modifications for L-BFGS are $s_k^w := w_n s_k$ and $y_k^w := y_k / w_n$, $k = n - m, \dots, n - 1$.

Appendix C. LM as a special case of Bregman method with L2 regularization

Here we consider the following Jacobian-based Bregman method with L2 regularization with $v_0 = 0$

$$\begin{aligned} X_{n+1} &= \arg \min_X \lambda \|AX - (Y + v_n)\|_2^2 + \|X\|_2^2 \\ v_{n+1} &= v_n + Y - AX_{n+1} \end{aligned} \quad (\text{C.1})$$

Claim: with $X_0 = 0$, (C.1) is equivalent to the following LM scheme

$$X_{n+1} = X_n + \arg \min_{\delta X} \lambda \|A(X_n + \delta X) - Y\|_2^2 + \|\delta X\|_2^2. \quad (\text{C.2})$$

Proof: First notice that (C.2) is equivalent to

$$X_{n+1} = \arg \min_X \lambda \|AX - Y\|_2^2 + \|X - X_n\|_2^2. \quad (\text{C.3})$$

We therefore only need to prove the equivalence of (C.1) and (C.3), which is by induction.

When $n = 0$, the equivalence is obvious.

Suppose the statement is true for all $k \leq n$, from the optimal condition of (C.3), we have

$$X_k = X_{k-1} + \lambda A^T (Y - AX_k), \text{ for all } k \leq n. \quad (\text{C.4})$$

Thus,

$$X_n = \lambda A^T \sum_{k=1}^n (Y - AX_k) = \lambda A^T v_n. \quad (\text{C.5})$$

The first equality is from the summation of (C.4) up to n and the second equality is from the summation of the second equality of (C.1) up to n .

Plug (C.5) into (C.3), we have

$$\begin{aligned} X_{n+1} &= \arg \min_X \lambda \|AX - Y\|_2^2 + \|X - \lambda A^T v_n\|_2^2 \\ &= \arg \min_X \lambda \|AX - Y\|_2^2 + \|X\|_2^2 - \lambda \langle AX, v_n \rangle. \\ &= \arg \min_X \lambda \|AX - Y - v_n\|_2^2 + \|X\|_2^2 \end{aligned} \quad (\text{C.6})$$

Therefore, we have shown the equivalence of Bregman method (C.1) and LM (C.2) in the case with L2 regularization.

Appendix D. Solving L2-regularized minimization

For the completeness, we shall describe the method for solving L2-regularized problem, which also serves as a build block for L1-type regularization in Split Bregman method (Appendix E). That is we consider the problem

$$u = \arg \min_u \|Au - B\|_2^2 + \lambda \|Mu\|_2^2, \quad (\text{D.1})$$

where M comes from TV or weighted L2 norm and $\|\cdot\|_2$ is understood as the usual non-weighted L2 norm. From optimal condition, the solution of (D.1) is

$$u = (A^T A + \lambda M^T M)^{-1} (A^T B). \quad (\text{D.2})$$

(D.2) is efficient when (D.1) is overdetermined, which is usually the case in quantitative PAT. However, there are many other situations that the system is underdetermined, for which a more efficient solution than (D.2) can be derived as follow.

First, notice that $M^T M$ is symmetric positive definite so that a Cholesky factorization exists, i.e., $M^T M = L^T L$ for a square matrix L . Next let $x = Lu$ and $A' = AL^{-1}$, then (D.1) is transformed into

$$u = L^{-1} [\arg \min_x \|A'x - B\|_2^2 + \lambda \|x\|_2^2]. \quad (\text{D.3})$$

Therefore,

$$\begin{aligned} u &= L^{-1} [(A'^T A' + \lambda I)^{-1} (A'^T B)] \\ &= L^{-1} A'^T [(A' A'^T + \lambda I)^{-1} B] \end{aligned} \quad (\text{D.4})$$

That is we invert $A' A'^T + \lambda I$ instead of $A^T A + \lambda M^T M$ for each L2 minimization problem with a minor additional cost from Cholesky factorization, which can potentially speed up significantly for the underdetermined systems.

Appendix E. Split Bregman method for TV-regularized minimization

The problem we are considering is as follow:

$$X = \arg \min_X H(X) + \lambda \|X\|_{TV}. \quad (\text{E.1})$$

Due to the simplification of TV via coarea formulation (1.7), we are to minimize

$$X = \arg \min_X H(X) + \lambda |MX|. \quad (\text{E.2})$$

The motivation of split Bregman method (Goldstein and Osher 2009) comes from two aspects: first the functional (E.2) is non-differentiable; second the solution of 1D problem $\min_x \frac{1}{2}(x-y)^2 + \lambda |x|$ can be explicitly computed through “the shrinkage”, i.e.,

$$\text{shrink}(y, \lambda) = \frac{y}{|y|} \max(|y| - \lambda, 0). \quad (\text{E.3})$$

A natural way is to “split” the data fidelity term and the non-differentiable TV regularization.

In the first step, we introduce the variable $d := MX$ to modify (E.3) into a constrained optimization problem, and then penalize the constraints through

$$(X, d) = \arg \min_{(X, d)} H(X) + \lambda |d| + \frac{\alpha}{2} \|d - MX\|_2^2, \quad (\text{E.4})$$

where α is the penalization parameter. As $\alpha \rightarrow \infty$, the solution of (E.4) goes to that of the original problem (E.2). Now, instead of increasing α , we equivalently apply the Bregman method with a fixed value for α ($\alpha = 1$ in this study) through a similar add-residual-back iterative two-stage algorithm

$$(X_{n+1}, d_{n+1}) = \arg \min_{(X, d)} \frac{\alpha}{2} \|MX - d - v_n\|_2^2 + H(X) + \lambda |d| \quad (\text{E.5})$$

$$v_{n+1} = v_n + d_{n+1} - MX_{n+1}$$

The minimization subproblems of (E.5) can be iteratively solved by splitting into the minimizations of X and d separately as

$$X_{n+1} = \arg \min_X \frac{\alpha}{2} \|MX - d_n - v_n\|_2^2 + H(X) \quad (\text{E.6})$$

$$d_{n+1} = \arg \min_d \frac{1}{2} \|MX_{n+1} - d - v_n\|_2^2 + \frac{\lambda}{\alpha} |d|$$

Notice that the second minimization of (E.6) can be solved efficiently through the component-wise shrinkage (E.3), i.e., $d_{n+1} = \text{shrink}(MX_{n+1} - v_n, \lambda / \alpha)$.

For Jacobian-based approach, $H(X) = \frac{1}{2} \|JX - Y\|_2^2$ formally. From the optimal condition,

$X_{n+1} = (J^T J + \alpha M^T M)^{-1} [J^T Y + \alpha M^T (d_n + v_n)]$ for the first minimization subproblem of (E.6). However, this may not be optimal when J is underdetermined, and we refer the readers to Appendix D for a discussion of the efficient solution method to L2 problem.

In implementation, it is not necessary to compute the exact (X_{n+1}, d_{n+1}) by solving (E.6) iteratively until the solution converges. The convergence, efficiency and robustness of the algorithm with only one iteration of (E.6) have been rigorously proven (Cai *et al* 2009) and also practically observed (Goldstein and Osher 2009). This suggests the following three-stage iterative split Bregman algorithm combining (E.5) and (E.6) for solving (E.2)

$$X_{n+1} = (J^T J + \alpha M^T M)^{-1} [J^T Y + \alpha M^T (d_n + v_n)]$$

$$d_{n+1} = \text{shrink}(MX_{n+1} - v_n, \lambda / \alpha) \quad (\text{E.7})$$

$$v_{n+1} = v_n + d_{n+1} - MX_{n+1}$$

For gradient-based approach, $H(X) = f(X)$. Similarly to (E.7), we have

$$\begin{aligned}
X_{n+1} &= \arg \min_X \frac{\alpha}{2} \|MX - d_n - v_n\|_2^2 + f(X) \\
d_{n+1} &= \mathit{shrink}(MX_{n+1} - v_n, \lambda / \alpha) \\
v_{n+1} &= v_n + d_{n+1} - MX_{n+1}
\end{aligned} \tag{E.8}$$

The only difference in (E.8) from (E.7) is the first nonlinear minimization problem, which can be solved through the aforementioned L-BFGS method with proper scaling. Algorithm (E.8) with L-BFGS is numerically robust, despite the lack of convergence proof, which will be studied in the future work.

References

- S. R. Arridge, "Optical tomography in medical imaging," *Inverse Problems* **15**, R41-93 (1999).
- S. R. Arridge and J. C. Schotland, "Optical tomography: forward and inverse problems," *Inverse Problems* **25**, 123010-1-59 (2009).
- M. Bachmayr and M. Burger, "Iterative total variation schemes for nonlinear inverse problems," *Inverse Problems* **25**, 105004-1-26 (2009).
- G. Bal and G. Uhlmann, "Inverse diffusion theory of photoacoustics", Preprint.
- B. Banerjee, S. Bagchi, R. M. Vasu, and D. Roy, "Quantitative photoacoustic tomography from boundary pressure measurements: noniterative recovery of optical absorption coefficient from the reconstructed absorbed energy map," *J. Opt. Soc. Am. A* **25**, 2347-2356 (2008).
- J.-F. Cai, S. Osher and Z. Shen, "Split Bregman Methods and Frame Based Image Restoration," *Multiscale Model. Simul.*, **8**, 337-369 (2009).
- B. T. Cox, S. R. Arridge, K. Köstli and P. Beard, "Quantitative photoacoustic imaging: fitting a model of light transport to the initial pressure distribution," *Proc. SPIE* **5697**, 49-55 (2005).
- B. T. Cox, S. R. Arridge and P. C. Beard, "Estimating chromophore distributions from multiwavelength photoacoustic images," *J. Opt. Soc. Am. A* **26**, 443-455 (2009).
- H. Gao and H. Zhao, "Simultaneous reconstruction of absorption and scattering with multiple optical excitations in quantitative photoacoustic tomography via Bregman method," Submitted to *JBO* (2010).
- H. Gao and H. Zhao, "Multilevel bioluminescence tomography based on radiative transfer equation Part 1: l1 regularization," *Optics Express* **18**, 1854-1871 (2010a).
- H. Gao and H. Zhao, "Multilevel bioluminescence tomography based on radiative transfer equation Part 2: total variation and l1 data fidelity," *Optics Express* **18**, 2894-2912 (2010b).
- T. Goldstein and S. Osher, "The split bregman method for l1 regularized problems," *SIAM J. Imaging Sci.* **2**, 323-343 (2009).
- Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations* (AMS, Providence, RI, 2001).
- J. Nocedal and S. J. Wright, *Numerical Optimization* (Springer, 1999).
- S. Osher, M. Burger, D. Goldfarb, J. Xu and W. Yin, "An iterated regularization method for total variation based image restoration," *SIAM Multiscale Model. Simul.* **4**, 460-489 (2005).
- J. Ripoll and V. Ntziachristos, "Quantitative point source photoacoustic inversion formulas for scattering and absorbing media," *Phys. Rev. E* **71**, 031912-1-9 (2005).
- L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *J. Phys. D* **60**, 259-268 (1992).
- L. V. Wang, "Multiscale photoacoustic microscopy and computed tomography," *Nat. Photonics* **3**, 503-509 (2009).
- L. Yin, Q. Wang, Q. Zhang, and H. Jiang, "Tomographic imaging of absolute optical absorption coefficient in turbid media using combined photoacoustic and diffusing light measurements," *Opt. Lett.* **32**, 2556-2558 (2007).