Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

MAX-PLANCK-GESELLSCHAFT

Technical Report No. 170

Non-monotonic Poisson Likelihood Maximization

Suvrit Sra[1], Dongmin Kim[2], Bernhard Schölkopf[1]

2nd June 2008
(Revised: August 2008)

[1] MPI für biologische Kybernetik, AGBS; [2] University of Texas at Austin, USA.

Non-monotonic Poisson Likelihood Maximization

Suvrit Sra, Dongmin Kim, and Bernhard Schölkopf

Abstract. This report summarizes the theory and some main applications of a new non-monotonic algorithm for maximizing a Poisson Likelihood, which for Positron Emission Tomography (PET) is equivalent to minimizing the associated Kullback-Leibler Divergence, and for Transmission Tomography is similar to maximizing the dual of a maximum entropy problem. We call our method non-monotonic maximum likelihood (NMML) and show its application to different problems such as tomography and image restoration. We discuss some theoretical properties such as convergence for our algorithm. Our experimental results indicate that speedups obtained via our non-monotonic methods are substantial.

Contents

1 Introduction **2**
1.1 Background . . . 2
1.1.1 Poisson Maximum Likelihood . . . 2
1.2 Related Work . . . 3

2 Theory and Algorithm **3**
2.1 Algorithm . . . 3
2.2 Theoretical Analysis . . . 4

3 Applications **6**
3.1 Positron Emission Tomography . . . 7
3.1.1 Related Work . . . 7
3.1.2 Theoretical Concerns and Algorithmic Details . . . 8
3.2 Transmission Tomography . . . 8
3.2.1 Maximum Likelihood . . . 9
3.2.2 Related Work . . . 9
3.2.3 Theoretical Concerns and Algorithmic Details . . . 10
3.3 Image Restoration . . . 10
3.3.1 Maximum Entropy Image restoration . . . 10
3.4 Miscellaneous Applications . . . 11
3.4.1 Entropy Maximization . . . 11
3.4.2 KL-Divergence NMA . . . 11
3.4.3 Other applications . . . 11

4 Experimental Results **11**
4.1 PET Experiments . . . 12
4.1.1 Experiments with sparse random matrices . . . 12
4.1.2 Experiments on PET Phantoms . . . 12
4.1.3 Regularized reconstruction . . . 13
4.2 Transmission Tomography . . . 13

5 Conclusions and Future work **15**

Bibliography **16**

1 Introduction

Several physical phenomena involve count-processes that are often modeled using Poisson distributions. Typical examples include physical processes where elementary particles such as electrons or photons are emitted and subsequently measured by detectors after they have travelled some distance. Numerous other common situations can also benefit from modeling data using Poisson distributions. For example, modeling web-server access statistics, tomography, image deconvolution and restoration, distribution of visual receptors in the retina, modeling number of mutations in DNA, and statistical inference, to name a few—the Wikipedia article [41] lists several more interesting scenarios.

1.1 Background

At an abstract level, consider that we have made non-negative measurements $y_1, \dots, y_n$ that may denote frequencies, projection counts, image intensities or other such quantities. Now assume that each of measurement y_i is generated via a Poisson process with mean parameter $[\boldsymbol{A}\boldsymbol{x}_i]$, where the matrix (or operator) $\boldsymbol{A}$ describes how the "true" underlying parameters $\boldsymbol{x}$ are related to each other ($\boldsymbol{A}$ may be viewed as an operator that convolves $\boldsymbol{x}$ to yield the mean parameters). For example, the matrix $\boldsymbol{A}$ may model probabilities of a certain region in space emitting a particle that gets detected by a particular detector—thereby, determining how the count measurements are actually generated. Then, we may write this relationship as

$$y_i \sim \text{Poisson}([\boldsymbol{A}\boldsymbol{x}]_i).$$

This model is naturally highly simplified, and depending on the application characteristics, richer or more diverse models may be considered. Later on in this report we will describe similar models with varying details depending on the problem.

1.1.1 Poisson Maximum Likelihood

Assuming the measurements $\boldsymbol{y}$ to be i.i.d., the Poisson likelihood of observing $\boldsymbol{y}$ given underlying parameters $\boldsymbol{x}$ (that are themselves non-negative for all our applications of interest), is given by

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} e^{-[\boldsymbol{A}\boldsymbol{x}]_i} \frac{[\boldsymbol{A}\boldsymbol{x}]_i^{y_i}}{y_i!}. \tag{1.1}$$

Maximizing the likelihood (1.1) w.r.t. to $\boldsymbol{x} \geq 0$ is equivalent to solving

$$\max_{\boldsymbol{x} \geq 0} \sum_{i=1}^{n} y_i \log[\boldsymbol{A}\boldsymbol{x}]_i - [\boldsymbol{A}\boldsymbol{x}]_i, \tag{1.2}$$

dropping constants for brevity. For the well-known problem of image reconstruction in Positron Emission Tomography (PET) this formulation was proposed and solved by an EM-based algorithm in [38].

We can rewrite (1.2) as the equivalent minimization problem

$$\min_{\boldsymbol{x} \geq 0} \quad f(\boldsymbol{x}) = \text{KL}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x}) = \sum_i y_i \log \frac{y_i}{[\boldsymbol{A}\boldsymbol{x}]_i} - y_i + [\boldsymbol{A}\boldsymbol{x}]_i, \tag{1.3}$$

where KL denotes the unnormalized Kullback-Leibler divergence, and $\boldsymbol{y} \in \mathbb{R}^n_+$ and $\boldsymbol{A} \in \mathbb{R}^{n \times p}_+$ are inputs. Problem (1.3) is a convex optimization problem and in this report we derive a new method for solving it. We also show several important applications and extensions of our method beyond just PET.

Regularized ML. In practice, one does not solve (1.3) directly but rather incorporates a penalty term that enforces smoothness or serves to incorporate prior knowledge. The resulting optimization problem is

$$\min_{\boldsymbol{x} \geq 0} \quad \text{KL}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x}) + \beta R(\boldsymbol{x}), \tag{1.4}$$

where $\beta > 0$ is a penalty parameter and $R(\boldsymbol{x})$ is a regularizing function. Several choices of $R(\boldsymbol{x})$ have been studied in the literature, for e.g., $R(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ is the traditional energy penalty, while $R(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{C}\boldsymbol{x}\|^2$ where $\boldsymbol{C}$ is a finite-differencing matrix provides a first-order roughness penalty. For our algorithm, we only make the requirement that $R(\boldsymbol{x})$ be a differentiable convex function of $\boldsymbol{x}$. To avoid clutter, we omit $R(\boldsymbol{x})$ from the discussion, noting that it can be easily included as long as it satisfies the abovementioned properties.

1.2 Related Work

Problem (1.3) is a simple convex optimization problem, which is especially appealing because of the simple non-negativity constraints. All the applications that we will consider in this report will be essentially of this type—having a convex objective function subject to non-negative constraints. Owing to this simplicity, all of these problems may be in principle solved by modern convex optimization methods. However, generic off-the-shelf software often fails to take advantage of the problem structure or may simply not scale to the problem sizes that we will study in this report. We mention below some typical approaches for solving (1.3) and related problems, noting that application specific related work will be described when we discuss the application.

One of the simplest approaches to solving (1.3) and other similar non-negatively constrained problems would be the projected gradient method of [36], which is a simple and scalable method. However, it is also known that the projected gradient method inherits deficiencies such as slow-convergence from the steepest-descent method. Significant effort has been invested into methods that promise to overcome such deficiencies and quasi-Newton methods are amongst the most successful. The LBFGS-B method of [7], which is an extension of L-BFGS [29] to box-constrained problems is an example of such a procedure. Traditional quasi-Newton methods based on BFGS or Newton methods that require the full Hessian matrix do not scale to the problem sizes considered in this report, and are hence excluded from discussion.

It is not surprising that for all the applications that we discuss in this paper, the associated research communities have developed their own set of methods, each tuned to the specific optimization problem at hand. What is more surprising is that methods grounded in modern optimization theory, such as LBFGS-B or the methods that we derive in this report, handily outperform most of the specific methods still in use in communities alluded to above.

In theoretical and practical spirit, the work closest to this report is our recent report [22], where we discuss a similar non-monotonic method applied to the non-negative least squares problem. In this report, we deal with the Kullback-Leibler divergence and some other related objective functions. These more sophisticated applications are naturally accompanied by associated new theory and experiments.

2 Theory and Algorithm

We now describe our solution to (1.3), which we will also extend naturally to other related objective functions. At a high-level, one could view our approach as a specially modified projected gradient method, which overcomes some typical drawbacks of projected-gradient by exploiting the non-monotonic descent procedure of Barzilai and Borwein (BB) [2]. Our approach is based on a similar extension of the BB idea to the non-negative least-squares problem [22].

More specifically, projected-gradient has three main ingredients. First, is the gradient of the objective function, second is a line-search to determine the step-size to use for descent, and the third is the projection step to enforce the constraints. The line-search step can often add substantial running time costs to the method, but in general it seems to be indispensable if one wants to prove convergence guarantees for the resulting algorithm. The main benefit of the BB approach is a closed-form solution to the step-size, thereby circumventing the potentially expensive line-search step. However, our problem (1.3) is *constrained* as opposed to the *unconstrained* scenario considered by [2], a naïve combination of the BB step-size with the projected-gradient approach will not yield a convergent procedure [22]. To develop a convergent algorithm without sacrificing efficiency, some crucial modifications are necessary, and these are described below.

2.1 Algorithm

Forming the Lagrangian of (1.3) and differentiating, we obtain the KKT optimality conditions

$$\nabla f(\boldsymbol{x}) - \boldsymbol{\lambda} = 0, \quad \lambda_i x_i = 0 \;\; \forall i, \quad \boldsymbol{\lambda}, \boldsymbol{x} \geq 0. \tag{2.1}$$

Now, consider the *active* variables, i.e., the variables that are zero at the optimum ($x_i^* = 0$). At any given iteration we can keep track of the currently active variables as an approximation to the final set of active variables, while minimizing the objective by turning an active variable into an inactive one or vice versa. Indeed, this scheme lies at the heart of *active set* methods [3]. However, such schemes can be slow and may under utilize the information available. We refine the active set and replace it by a *fixed set* of variables, which exploits the complementary slackness condition $\lambda_i x_i = 0$ in (2.1). If a given $x_i = 0$, then its corresponding $\lambda_i \geq 0$, and in fact if $\lambda_i > 0$ then $[\nabla f(\boldsymbol{x})]_i > 0$ must hold. Hence, we define:

Definition 1 (Fixed-set). *The Fixed-Set at iteration k is defined to be a subset of the (indices of) active-variables that have a positive gradient. That is,*

$$I_+^k = \{i | x_i^k = 0, [\nabla f(x^k)]_i > 0\}. \tag{2.2}$$

Given that we have a candidate fixed set of variables, we optimize over the *free variables*. At this point, any generic unconstrained minimization method with *line-search* could be applied. However, as previously suggested, the line-search step often turns out to be a bottleneck. In addition, if the generic method of choice were a second order method, then the memory requirements rapidly become prohibitive. Thus, we depart from the generic constrained minimization setup and replace the step-size computation by *closed-form* computations. This circumvention of the potentially expensive line-search step leads to a considerable simplification to the algorithm and associated gains in computational efficiency. It is important to note that the exact form of the step-size is determined by performing descent over only the *free variables*.

Formally, let $\boldsymbol{s}_k = \boldsymbol{x}^k - \boldsymbol{x}^{k-1}$, and $\boldsymbol{y}_k = \boldsymbol{g}^k - \boldsymbol{g}^{k-1}$. In contrast to what an out-of-the-box invocation of the BB method would do, we compute step-sizes using only the "free" parts of $\boldsymbol{s}_k$ and $\boldsymbol{y}_k$. Additionally, we incorporate a positive sequence of parameters β_k such that $\lim_k \beta_k = 0$ and $\lim_k \sum_k \beta_k = \infty$, that rescale the step-sizes α_k. Finally, a user-defined upper bound on the step-sizes is included to ensure that the step-sizes remain bounded and to prevent pathological behavior. In practice, however, we have not found it necessary to include either β_k or τ, thereby considerably alleviating the parameter selection burden.

Let $Z_+(\boldsymbol{x})$ denote the vector $[Z_+(x_i)]$, where $Z_+(x_i) = 0$ for $i \in I_+$, and x_i otherwise. With the introduction of this *zero-out* operator, we can now display the step-size computations:

$$\begin{gathered} \boldsymbol{s}_k \leftarrow Z_+(\boldsymbol{s}_k), \quad \boldsymbol{y}_k \leftarrow Z_+(\boldsymbol{y}_k) \\ \alpha_k = \beta_k \min\left\{ \frac{\langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle}{\langle \boldsymbol{s}_k, \boldsymbol{y}_k \rangle}, \tau \right\}, \quad \text{or} \quad \alpha_k = \beta_k \min\left\{ \frac{\langle \boldsymbol{s}_k, \boldsymbol{y}_k \rangle}{\langle \boldsymbol{y}_k, \boldsymbol{y}_k \rangle}, \tau \right\}. \end{gathered} \tag{2.3}$$

The updates (2.3) are our adapted version of BB-type updates and it is important to note that they operate on only the free-variables. These updates may be viewed as scalar solutions to the quasi-Newton secant equation that arises, or in other words as Rayleigh quotients corresponding to an appropriate interpolated Hessian matrix [22].

Given the step-sizes (2.3) and the current estimate of the solution $\boldsymbol{x}^k$, we have the update

$$\boldsymbol{x}^{k+1} = P_+(\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k), \tag{2.4}$$

where $P_+(x) = \max(0, x)$ projects onto the non-negative orthant to ensure constraints satisfaction.

All these details mentioned above are incorporated into Algorithm 1.

We now proceed onto analyzing important theoretical properties of Algorithm 1 in the next section.

2.2 Theoretical Analysis

Now we analyze some theoretical properties of the algorithm including convergence. In general, for a differentiable function f, the only subgradient of f at $\boldsymbol{x}^k$ is the gradient itself. However, we show that Algorithm 1 with the special constraint $\boldsymbol{x} \geq 0$, generates a subgradient around $\boldsymbol{x}^k$ that differs from the gradient.

Lemma 2.1 (Subgradient Property). *A chopped gradient*

$$\boldsymbol{g}^k = Z_+\left(\nabla f(\boldsymbol{x}^k)\right)$$

is a subgradient of f around $\boldsymbol{x}^k$.

Proof. Since f is convex and differentiable, its Taylor expansion at $\boldsymbol{x}^k$ yields

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}^k) + \left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{z} - \boldsymbol{x}^k \right\rangle.$$

Hence,

$$f(\boldsymbol{z}) - f(\boldsymbol{x}^k) - \left\langle \boldsymbol{g}^k, \boldsymbol{z} - \boldsymbol{x}^k \right\rangle \geq \left\langle \nabla f(\boldsymbol{x}^k) - \boldsymbol{g}^k, \boldsymbol{z} - \boldsymbol{x}^k \right\rangle.$$

Since $[\nabla f(\boldsymbol{x}^k)]_i > 0$ for $i \in I_+$ and $\boldsymbol{z} \geq 0$,

$$\left\langle \nabla f(\boldsymbol{x}^k) - \boldsymbol{g}^k, \boldsymbol{z} - \boldsymbol{x}^k \right\rangle = \sum_{i \in I_+} [\nabla f(\boldsymbol{x}^k)]_i \times z_i \geq 0.$$

Therefore, for all $\boldsymbol{z} \geq 0$, we obtain $f(\boldsymbol{z}) \geq f(\boldsymbol{x}^k) + \left\langle \boldsymbol{g}^k, \boldsymbol{z} - \boldsymbol{x}^k \right\rangle$. □

```
Input: $\boldsymbol{y}$: count data, $\boldsymbol{A}$: system matrix, $\tau$: control param
Output: $x = \operatorname{argmin}_{x \geq 0} f(\boldsymbol{x}) = \mathrm{KL}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x})$
Initialize $k \leftarrow 1$, $\boldsymbol{x}^0 \in \mathbb{R}^n_{++}$,  $\boldsymbol{g}^0 = \nabla f(\boldsymbol{x}^0)$;
$\boldsymbol{x}^1 \leftarrow P_+(\boldsymbol{x}^0 - \alpha_0 \boldsymbol{g}^0)$ for some $\alpha_0 > 0$;
(choose $\alpha_0$ s.t. $f(\boldsymbol{x}^1) < f(\boldsymbol{x}^0)$);
repeat
    $\boldsymbol{g}^k = Z_+\big(\nabla f(\boldsymbol{x}^k)\big)$;
    Compute fixed set $I_+ = \big\{i \big| x_i^k = 0,\ [\boldsymbol{g}^k]_i > 0\big\}$;
    $\boldsymbol{y}_k \leftarrow Z_+(\boldsymbol{g}^k - \boldsymbol{g}^{k-1})$   {Zero out gradient};
    $\boldsymbol{s}_k \leftarrow Z_+(\boldsymbol{x}^k - \boldsymbol{x}^{k-1})$   {Zero out variables};
    if $k \mod 2 = 0$ then
        $\alpha_k \leftarrow \beta_k \min\{\langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle / \langle \boldsymbol{s}_k, \boldsymbol{y}_k \rangle, \tau\}$;
    end
    else
        $\alpha_k \leftarrow \beta_k \min\{\langle \boldsymbol{s}_k, \boldsymbol{y}_k \rangle / \langle \boldsymbol{y}_k, \boldsymbol{y}_k \rangle, \tau\}$;
    end
    $\boldsymbol{x}^{k+1} \leftarrow P_+(\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k)$;
    $k \leftarrow k + 1$;
until Stopping Criteria are met ;
```

Algorithm 1: Non-monotonic maximum likelihood

Now, for notational convenience, let us define the best objective value so far, i.e.,

$$\bar{f}^k = \min\left\{\bar{f}^{k-1}, f(\boldsymbol{x}^k)\right\}.$$

To prove the main theorem we exploit some ideas from the proof of the subgradient method [39], and using the following nonexpansive property of the projection step we obtain our final proof.

Lemma 2.2 (Nonexpansive Property of Projection)**.** *For all $\boldsymbol{x}, \boldsymbol{z}$,*

$$\|P_+(\boldsymbol{x}) - P_+(\boldsymbol{z})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{z}\|_2.$$

Proof. Refer to Proposition B.11 (C) in [3]. □

Theorem 2.3 (ϵ-optimal Convergence)**.** *If f^* denotes the optimal solution to* (1.3)*, and α_k is bounded above, there exists a constant ϵ s.t.,*

$$\lim_{k \to \infty} \bar{f}^k - f^* < \epsilon.$$

Proof. Using Lemma 2.2 and since $P_+(x^*) = x^*$, we have

$$\begin{aligned}
\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|_2^2 = \quad & \|P_+(\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k) - P_+(\boldsymbol{x}^*)\|_2^2 \quad \leq \quad \|\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k - \boldsymbol{x}^*\|_2^2 \\
= \quad & \|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2 - 2\alpha_k \left\langle \boldsymbol{g}^k, \boldsymbol{x}^k - \boldsymbol{x}^* \right\rangle + \alpha_k^2 \|\boldsymbol{g}^k\|_2^2 \\
\leq \quad & \|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2 - 2\alpha_k \big(f(\boldsymbol{x}^k) - f^*\big) + \alpha_k^2 \|\boldsymbol{g}^k\|_2^2 \\
\leq \quad & \|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 - 2\sum_k \alpha_k \big(f(\boldsymbol{x}^k) - f^*\big) + \sum_k \alpha_k^2 \|\boldsymbol{g}^k\|_2^2.
\end{aligned}$$

Consequently,

$$2\sum_k \alpha_k \left(f(\boldsymbol{x}^k) - f^*\right) \leq \|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + \sum_k \alpha_k^2 \|\boldsymbol{g}^k\|_2^2.$$

Further consider the following,

$$\sum_k \alpha_k \big(f(\boldsymbol{x}^k) - f^*\big) \geq \Big(\sum_k \alpha_k\Big) \min_k \big(f(\boldsymbol{x}^k) - f^*\big) = \Big(\sum_k \alpha_k\Big) \big(\bar{f}^k - f^*\big),$$

then we obtain

$$2\Big(\sum_k \alpha_k\Big)\big(\bar{f}^k - f^*\big) \le \|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + \sum_k \alpha_k^2 \|\boldsymbol{g}^k\|_2^2,$$
$$\bar{f}^k - f^* \le \frac{\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + \sum_k \alpha_k^2 \|\boldsymbol{g}^k\|_2^2}{2\sum_k \alpha_k}. \tag{2.5}$$

Now since f is Lipschitz continuous, there exists some constant L such that

$$\|\boldsymbol{g}^k\|_2 \le \|\nabla f(\boldsymbol{x}^k)\|_2 \le L.$$

Depending upon the choice of diminishing sequence β_k, there exist two possible convergence scenarios. First, suppose $\sum_{k=1}^{\infty} \alpha_k = m$, then we also have $\sum_{k=1}^{\infty} \alpha_k^2 = n$, since $\alpha_k > 0$. Substituting the bound on the gradient and the assumption of α_k into (2.5), and taking limits we obtain

$$\begin{aligned}\lim_{k\to\infty} \bar{f}^k - f^* &\le \lim_{k\to\infty} \frac{\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + \sum_k \alpha_k^2 \|\boldsymbol{g}^k\|_2^2}{2\sum_k \alpha_k} \\ &= \frac{\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + nL^2}{2m} = \epsilon,\end{aligned}$$

thereby proving the main claim of the theorem.

When the α_k sequence is not bounded, then using the *forcing* sequence β_k we can still obtain convergence to the optimal. Thus, supposing that the sum $\sum_{k=1}^{\infty} \alpha_k = \infty$, from Algorithm 1 we have $\alpha_k \le \beta_k \tau$ at each iteration k. Furthermore, since $\lim_{k\to\infty} \beta_k = 0$

$$\lim_{k\to\infty} \alpha_k \le \lim_{k\to\infty} \beta_k \tau = 0.$$

Since $\lim_{k\to\infty} \alpha_k = 0$, given an arbitrary $\epsilon > 0$, there exists an integer N_1 such that

$$\alpha_k \le \frac{\epsilon}{L^2},$$

for all $k > N_1$. Further, since $\sum_{k=1}^{\infty} \alpha_k = \infty$, there exists an integer N_2 such that

$$\sum_{i=1}^{k} \alpha_i \ge \frac{1}{\epsilon}\Big(\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + L^2 \sum_{i=1}^{N_1} \alpha_i^2\Big)$$

for all $k > N_2$. Let $N = \max\{N_1, N_2\}$. Then for all $k > N$,

$$\begin{aligned}\bar{f}^k - f^* &\le \frac{\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + L^2\big(\sum_{i=1}^{N} \alpha_i^2 + \sum_{i=N+1}^{k} \alpha_i^2\big)}{2\big(\sum_{i=1}^{N} \alpha_i + \sum_{i=N+1}^{k} \alpha_i\big)} \\ &= \frac{\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + L^2 \sum_{i=1}^{N} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} + \frac{L^2 \sum_{i=N+1}^{k} \alpha_i^2}{2\sum_{i=1}^{N} \alpha_i + 2\sum_{i=N+1}^{k} \alpha_i} \\ &\le \frac{\|x^1 - x^*\|_2^2 + L^2 \sum_{i=1}^{N} \alpha_i^2}{\frac{2}{\epsilon}\big(\|\boldsymbol{x}^1 - \boldsymbol{x}^*\|_2^2 + L^2 \sum_{i=1}^{N} \alpha_i^2\big)} + \frac{L^2 \sum_{i=N+1}^{k} \frac{\epsilon}{L^2} \alpha_i}{2\sum_{i=N+1}^{k} \alpha_i} \\ &= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.\end{aligned}$$

Since ϵ is arbitrary, we obtain $\lim_{k\to\infty} \bar{f}^k = f^*$, concluding the proof. □

3 Applications

We now describe some basic applications of our NMML algorithm and some of its simple variants. Our first example deals with Positron Emission Tomography (PET) image reconstruction, where a Poisson model is of fundamental importance. Then we discuss an adaption of our method to the problem of reconstruction of attenuation maps in Transmission Tomography—traditionally held to be an algorithmically more complicated problem than PET reconstruction. An appropriate modification of our NMML strategy yields a new algorithm for this problem too. We also briefly mention the problem of image restoration or deconvolution, noting some cases where our NMML method can be effective. We would like to stress that the NMML method is fairly general, and therefore is also applicable to situations other than those discussed below. A short list of other potential applications is included in Section 3.4.

3.1 Positron Emission Tomography

Image reconstruction for Poisson Emission Tomography (PET) presents an interesting and important example of a real-world application where maximizing a Poisson likelihood arises naturally. To appreciate how this model arises, it is useful to know a little about the underlying physical phenomena (see [38] for a nice introduction).

In PET a radionucleide that is injected into the patients' bloodstream gradually undergoes decay releasing positrons. These positrons travel a short distance within the body before annihilating with electrons and releasing two oppositely traveling gamma-ray photons, which may be subsequently detected by a pair of detectors placed in a ring around the patient. These gamma-ray photons are then assumed to be generated by a Poisson point process, which is natural for such a radioactive phenomenon. This description is of course simplistic and ignores numerous additional physical processes and distortions that the actual photons may undergo (e.g., attenuation, dispersion, scattering) before they are actually detected. Such an abstract simplification is necessary not to detract from the main theme of this paper; modeling and handling each one of these effects forms a separate research area in PET imaging [31, 34].

Given measurements counting the number of coincidentally detected photons, the aim of image reconstruction in PET is to estimate a function whose value is the expected number of positron emissions at each point in space within the body. To facilitate computation one usually discretizes the problem, wherein the space is divided into p pixels (voxels for 3-D data, or for simplicity we can just work slice by slice), and constant radioactivity within each pixel is assumed [30, 38]. Then, one estimates the image vector $\boldsymbol{x} = [x_1, \ldots, x_p]^T$, where each x_j models the activity in pixel j. The measurement process counts coincidences y_i along n detector lines (recall detectors come in radially placed pairs). Let a_{ij} denote the "probability" that a photon emitted by pixel j is detected by detector pair i. Then, the coincidences y_i are samples from a Poisson distribution with expected value $\sum_{j=1}^{p} a_{ij}x_j = [\boldsymbol{Ax}]_i$. Formally, we denote this as

$$y_i \sim \text{Poisson}([\boldsymbol{Ax}]_i).$$

A model that includes random coincidences and non-uniform calibration (see [14]) is given by

$$y_i \sim \text{Poisson}(c_i[\boldsymbol{Ax}]_i + r_i),$$

where r_i denotes the mean of the accidental coincidences counted by the ith detector, and c_i are certain calibration factors reflecting calibration of the ith detector unit. For simplicity, we develop our methods with $r_i = 0$ and $c_i = 1$, noting that our approach can be easily extended to handle both these cases.

Clearly, the maximum likelihood formulation described in Section 1.1.1 is immediately applicable here. Thus, the corresponding optimization problems are (1.3) and (1.4) (penalized likelihood maximization). In fact, penalized likelihood maximization is even more important in real-world PET image reconstruction because without additional regularization, the resulting reconstructions have been observed to be very noisy (a fact that can be partially attributed to the overly simplified Poisson model, and a crude / incompletely available system matrix). Our methods, of course, handle both the penalized as well as unpenalized situations with equal ease. More specific implementation details are described below in Section 3.1.2.

3.1.1 Related Work

A vast number of iterative approaches have been proposed in the PET literature for solving (1.3) and we relate to some of the most well-known ones. Please see the surveys [31, 34] for surveys image reconstruction procedures for PET.

Iterative methods for solving (1.3) may be roughly put into two groups. The first group consists of methods centered around variations of the EM idea—these methods essentially bound the objective by an auxiliary function that is easier to optimize, and they have been the most popular set of methods in the PET community. Some examples include: (i) the basic EM based method called ML-EM or EMML [38], of which the Richardson-Lucy method [35] is a special case, (ii) an accelerated version called ordered subsets EM (OSEM) that performs updates with only a few rows of $\boldsymbol{A}$ at a time and trades convergence guarantees for speed [20]—the OSEM algorithm is one of the most popular method in PET image reconstruction and is often implemented in actual PET scanners too; (iii) several variants of the OSEM scheme and other generalizations based on EM such as C-OSEM, SAGE, BSREM, etc. (for details and a summary of several other related methods, see [14, 34]) (iv) an extension of EMML to deal with regularization [30]. Our work differs from *all* of these methods as we optimize the objective function (1.3) directly without introducing any auxiliary functions, and it does not suffer from the slow convergence exhibited by EM methods.

The second group of methods includes other approaches from convex optimization. For instance, (i) a row-action method (called RAMLA) [5] that goes through the system matrix $\boldsymbol{A}$ one row at a time, (ii) a conjugate gradient based approach [28], (iii) a limited memory quasi-Newton approach LBFGS-B [8], or (iv) even interior-point approaches [21]. RAMLA proceeds through the matrix $\boldsymbol{A}$ row-by-row, however, like other row-action methods it suffers from slow convergence. Approaches based on conjugate-gradient can become slow or even *ad hoc* in the presence of non-negativity constraints, while interior point methods simply do not scale to the large problem sizes that are common in PET. The closest competitor turns out to be the LBFGS-B procedure of [8], which has seen surprisingly little attention in the medical imaging community. However, even LBFGS-B becomes slow when the scale of the problems increases. We remark that the LBFGS-B procedure also seems to outperform the well-established reconstruction methods based on EM, and perhaps because of its associated implementation overhead it has not found wide acceptance in the medical imaging community. Our method is very simple to implement, and not only outperforms the EM based methods but also the LBFGS-B method—thereby, promising a potential practical adoption by the community.

3.1.2 Theoretical Concerns and Algorithmic Details

The convergence proof of our algorithm required $f(\boldsymbol{x})$ to be Lipschitz continuous. However, KL-Divergence as such might fail to be Lipschitz continuous if we permit arbitrary inputs. Therefore, we assume that the system matrix $\boldsymbol{A}$ is designed so that $[\boldsymbol{A}\boldsymbol{x}]_i > 0$ (such an assumption is common in the PET literature, see for e.g., [38]). Additionally, we may assume for all practical purposes that each projection $[\boldsymbol{A}\boldsymbol{x}]_i$ is lower-bounded by a constant ε—this ensures that the resulting objective function is Lipschitz continuous and we can invoke the convergence theory without additional problems. Even though this assumption is practically motivated, and is satisfied by real problem, getting rid of it remains an open issue in our convergence analysis.

Computation: For $f(\boldsymbol{x}) = \mathrm{KL}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x})$ we can simplify individual steps of the algorithm. For example, the gradient computation

$$[\nabla f(\boldsymbol{x})]_j = \sum_i a_{ij} - \sum_i \frac{y_i a_{ij}}{[\boldsymbol{A}\boldsymbol{x}]_i},$$

can be written as

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}^T(\boldsymbol{1} - \hat{\boldsymbol{y}}), \tag{3.1}$$

where $\hat{\boldsymbol{y}}$ is given by the elementwise division $[y_i/[\boldsymbol{A}\boldsymbol{x}]_i]$. In the language of PET image reconstruction (3.1) requires one *forward projection* and one *back projection*—in numerical linear algebra terminology, it requires just two matrix-vector BLAS2 operations that can be computed very efficiently, or even parallelized easily if needed.

Computational Complexity: Given the computation details above, it is easy to see that the resources required by Algorithm 1 are modest. At each iteration computation of the gradient costs two matrix-vector products and this cost dominates the other computations. Thus, the overall running time is $O(2T \cdot nz)$, i.e., where nz is number of nonzero elements in $\boldsymbol{A}$ and T denotes the number of iterations.

3.2 Transmission Tomography

Complementary to Emission Tomography is the technique of Transmission Tomography. Here, instead of injecting radionucleides that emit gamma rays, high-energy X-rays are streamed through a specific region of the patient's body and then measured by a grid of photon detectors on the other side. Different tissue types in the patient's body attenuate the photons differently, and it is the average attenuation rates of the different parts (voxelized) of the body that are now the unknown quantities to be estimated. The goal is thus to construct an attenuation map for the associated regions of the patient's body, given just the measurements counting the number of photons along detector lines ("line-integrals"). For simplicity of exposition we neglect details like multiple angles at which the photons are beamed, or physical affects like scattering.

As for emission tomography we can assume the space to be divided into pixels (voxels) assuming constant attenuation within a pixel. Let b_i be the average number of photons detected by detector i when the patient is not there in the scanner, also known as the "blank-scan factor". When the patient is present, his body leads to attenuation in the number of photons, so that now the mean number of photons detected at detector i may be assumed to be

$$\mu_i = b_i e^{-\sum_j a_{ij} x_j} = b_i e^{-[\boldsymbol{A}\boldsymbol{x}]_i},$$

where a_{ij} denotes the length of the line of projection between the photon source and the ith detector going through the jth pixel (voxel), and x_j denotes the attenuation coefficient of the jth pixel (probability of photon attenuation per unit length). Given this attenuation, the photons counted by each detector may be now assumed to follow a Poisson point process with mean μ_i, so that we have

$$y_i \sim \text{Poisson}(b_i e^{-[\boldsymbol{Ax}]_i}). \tag{3.2}$$

A model that includes the effects of random coincidences is given by

$$y_i \sim \text{Poisson}(b_i e^{-[\boldsymbol{Ax}]_i} + r_i),$$

where r_i denotes the mean number of background events detected by detector i.

3.2.1 Maximum Likelihood

Assuming that all the measurements y_i are iid, the log-likelihood for (3.2) (ignoring constant factors) is

$$L(\boldsymbol{y}|\boldsymbol{x}) = -\sum_i y_i[\boldsymbol{Ax}]_i - \sum_i b_i e^{-[\boldsymbol{Ax}]_i}, \tag{3.3}$$

so that maximizing the log-likelihood is equivalent to solving the following minimization problem

$$\min_{\boldsymbol{x}\geq 0} \; f(\boldsymbol{x}) = \sum_i y_i[\boldsymbol{Ax}]_i + \sum_i b_i e^{-[\boldsymbol{Ax}]_i}, \tag{3.4}$$

which is a convex optimization problem that is similar in structure to (1.3).

We remark that in practice, we can easily incorporate the mean number of background events r_i detected by detector unit i, into our algorithm. We omit details for brevity. Also, note that with similar ease we can also incorporate a penalization term in (3.4).

3.2.2 Related Work

Several algorithms have been developed for performing maximum likelihood reconstruction for transmission tomography. Owing to the success of EM approaches for the emission case, EM type methods were developed for transmission too [23]. However, several of the recent algorithms are based on direct optimization of the objective function rather than using an EM approach [17, 27, 37]. Our algorithm also solves (3.4) directly, without resorting to an EM type method.

As previously mentioned, for the emission tomography problem the OSEM method has enjoyed considerable success, especially because it is simple to implement and runs quite efficiently. It is but natural to expect an ordered-subsets variant for the transmission problem and Manglos et al. [26] presented such a method. Erdoğan and Fessler [13] introduce a method called separable paraboloidal surrogates (SPS), which is again based on the idea of *surrogate* or *auxiliary* functions as in EM; they then present an ordered subsets *heuristic* of their method, which is claimed to accelerate the "convergence" (their method does not have any convergence guarantees, and can be even non-monotonic). Ahn et al. [1] provides a convergent ordered subsets algorithms for transmission tomography. Other related methods include the papers [11–13, 24], in addition to the references in [13].

Our method for transmission reconstruction has the same subjective features that made OSEM popular for emission tomography, namely:

1. It is simple to implement—a short MATLAB program suffices for practical purposes, making our method a strong candidate for practical adoption by the community
2. It provides orders of magnitude acceleration over basic EM schemes; in fact it improves upon the ordered subsets schemes too
3. It can easily incorporate any type of system model
4. It can accommodate convex penalties on the likelihood without additional difficulty.

Furthermore, our method comes with somewhat more theoretical guarantees than OSEM.

At this point, we again draw the reader's attention to the fact that modern optimization methods such as LBFGS-B [7] generally outperform the EM based methods that are common in the transmission tomography community. A primary reason for the popularity of the EM based methods is the experts' familiarity with them, and the ease of implementation. Our NMML method is extremely simple to implement (much simpler than an LBFGS-B implementation), and can therefore serve as an easy replacement for EM based methods.

3.2.3 Theoretical Concerns and Algorithmic Details

We just invoke our algorithm with $f(\boldsymbol{x})$ given in (3.4). This function is clearly Lipschitz continuous, and the remaining theoretical considerations carry over directly without additional difficulties. The most important new component is the gradient $f(\boldsymbol{x})$ that is computed as

$$[\nabla f(\boldsymbol{x})]_j = \sum_i a_{ij} y_i - \sum_i a_{ij} b_i e^{-[\boldsymbol{Ax}]_i},$$

which may be rewritten as

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}^T(\boldsymbol{y} - \hat{\boldsymbol{b}}), \tag{3.5}$$

where $\hat{b}_i = b_i e^{-[\boldsymbol{Ax}]_i}$. Once again, we see that the computing the gradient requires just one $\boldsymbol{Ax}$ operation and one $\boldsymbol{A}^T\boldsymbol{y}$ operation, resulting in a very efficient algorithm. We just need to replace (3.1) by (3.5) when invoking Algorithm 1. To further optimize the running time, it might be beneficial to implement the exponentiation operation carefully, because it will need to be performed each time the gradient is computed.

3.3 Image Restoration

Image *reconstruction* problems such as those arising in connection with tomography (§§3.1,3.2) are closely related to *image restoration* problems. The former usually have much more complicated physical models as compared to the latter where the inputs are usually simpler. Though the algorithmic techniques or the models used for solving the problems can differ widely, in some formulations there are enough commonalities to permit us to handle image restoration using our reconstruction algorithms.

Typically in image restoration problems one observes a blurred or noisy image and the aim is to recover a good image from it. Astronomical images are furnish one important domain for the application of image restoration. In simple terms, the observed image is assumed to be a convoluted version of the original with some added noise. In symbols

$$\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the observed image, $\boldsymbol{A}$ is the blurring convolution operator, and $\boldsymbol{\varepsilon}$ is the added noise. The blurring operator is usually not known accurately, whereby the actual problem of blind-deconvolution arises. However, for algorithmic simplicity we assume it to be known. We remark that in more realistic scenarios the operator $\boldsymbol{A}$ may be available in a factored form approximating a cascade of convolutions—though from an algorithmic point of view this does not make any difference.

Image restoration (also deblurring) is a vast field in itself, with hundreds of publications. We make no attempt to summarize the field here, and refer the reader to [16, 19] and the references therein for more details. An interesting formulation that leads to qualitatively very different results as compared to tomographic image reconstruction is described in the next section.

3.3.1 Maximum Entropy Image restoration

In the image restoration literature, especially for astronomy images, using maximum entropy priors / regularizers for the objective function are quite popular. Here, a typical problem might be to estimate an image $\boldsymbol{x}$ given the observed image $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{n}$, while ensuring that the entropy of the reconstructed image is as high as possible. A simple formulation is the following regularized non-negative least-squares problem

$$\min_{\boldsymbol{x} \geq 0} \quad \tfrac{1}{2}\|\boldsymbol{Ax} - \boldsymbol{y}\|^2 + \beta \sum_j x_j \log x_j. \tag{3.6}$$

This problem is a regularized version of the non-negative least squares problem of [22]. However, the convergence theory of that paper does not extend to the case with regularizers. Since the entropy function is not necessarily Lipschitz continuous, and in an actual reconstruction we can have $x_j = 0$, our convergence guarantees do not apply out of the box either. However, we can still apply the NMML algorithm to certain cases, though a thorough experimental validation, especially in comparison with other available software for this problem, remains a part of our ongoing work.

For $\beta > 0$, the gradient of the objective function in (3.6) is given by

$$\boldsymbol{A}^T(\boldsymbol{Ax} - \boldsymbol{y}) + \beta(\boldsymbol{1} + \log \boldsymbol{x}), \tag{3.7}$$

where $\log \boldsymbol{x}$ denotes the vector $[\log x_j]$. As for PET and transmission reconstruction, the gradient (3.7) can be computed in time $O(2nz)$, where nz is the number of nonzero entries in $\boldsymbol{A}$, yielding a very fast algorithm.

3.4 Miscellaneous Applications

Here we enlist some other potential applications based on simple extensions of our NMML method. We mention these applications only briefly, and do not present experimental results on them. A closer investigation is a subject of future work.

3.4.1 Entropy Maximization

Entropy maximization is a fundamental convex optimization problem arising in various contexts and applications. A particular instance may be written as

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) = \textstyle\sum_i x_i \log x_i \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \qquad \boldsymbol{x}^T\mathbf{1} = 1, \end{aligned} \tag{3.8}$$

where $\mathrm{dom} f = \mathbb{R}^n_{++}$. Letting $\boldsymbol{\lambda}$ and ν be appropriate Lagrange multipliers, the dual to (3.8) is derived to be (see [4, Chapter 5], for example)

$$\min_{\boldsymbol{\lambda} \geq 0, \nu} \quad g(\boldsymbol{\lambda}, \nu) = \sum_i e^{-[\boldsymbol{A}^T\boldsymbol{\lambda}]_i - \nu - 1} + \boldsymbol{b}^T\boldsymbol{\lambda} + \nu. \tag{3.9}$$

It is interesting to note that the primal problem of likelihood maximization in Transmission Tomography (3.4) is almost the same as (3.9) above. Therefore, we can trivially adapt the algorithm derived for Transmission Tomography to solve the constrained entropy maximization problem—with potential benefits for a large number of problems depending on entropy maximization.

3.4.2 KL-Divergence NMA

The KL-Divergence Non-negative Matrix Approximation (NMA) problem [10, 25] attempts to solve the following non-convex optimization problem

$$\begin{aligned} \min \quad & \mathrm{KL}(\boldsymbol{A}; \boldsymbol{B}\boldsymbol{C}) = \sum_{ij} a_{ij} \log \frac{a_{ij}}{[\boldsymbol{B}\boldsymbol{C}]_{ij}} - a_{ij} + [\boldsymbol{B}\boldsymbol{C}]_{ij} \\ \text{s.t.} \quad & \boldsymbol{B}, \boldsymbol{C} \geq 0. \end{aligned} \tag{3.10}$$

Problem (3.10) is a difficult non-convex problem and a typical approach is to develop an alternating minimization or descent procedure that fixes $\boldsymbol{B}$ while minimizing (or descending) over $\boldsymbol{C}$ and vice-versa. The resulting subproblem is exactly of the form (1.3), whereby our NMML method can be used to solve it, resulting in an alternating KL divergence based NMA algorithm. This particular approach to KL-Divergence NMA is new, and the NMML method makes it practical.

3.4.3 Other applications

Some other applications that could benefit from our methods include:

- Fast variational inference for large-scale internet diagnosis – Platt, NIPS
- Medical imaging, e.g., image registration based on minimizing KL-divergence [18], image-intensity statistics in magnetic resonance imaging [40], or
- Computer Vision, e.g., real-time tracking [9],

4 Experimental Results

We now provide some experimental results to demonstrate the performance of our algorithms. We begin with some basic results for KL-Divergence minimization (1.3) on large sparse matrices (§4.1.1). Then, we show results on simulated and real-world PET data (§4.1), followed by results on simulated data for transmission tomography (§4.2).

4.1 PET Experiments

To demonstrate the effectiveness of our algorithm, we compare it to the well-established methods for likelihood maximization in PET. The algorithms that we compare are,

1. EMML – the baseline method,
2. OSEM – the accelerated version of EM [20],
3. RAMLA – a row-action algorithm [5], and
4. NMML – our non-monotonic maximum likelihood algorithm.

We highlight the fact that in PET literature, OSEM (or its derivatives) is the method of choice, and it is also the method that is implemented in many real-world PET scanners [14]. Thus, our main competitor is OSEM.

To ensure fairness, all the algorithms were carefully implemented in MATLAB while ensuring that each one of them exploits sparsity in the input. This point becomes even more important for algorithms like OSEM and RAMLA because sparse matrices in MATLAB are stored in column oriented format and both OSEM and RAMLA need to access the matrix $\boldsymbol{A}$ in a row oriented manner. Unfortunately, working with $\boldsymbol{A}^T$ alone is not sufficient for the OSEM algorithm because MATLAB does not handle the associated subset level matrix-vector operations efficiently. Many such implementation issues are addressed in the IRT toolkit of [15]; we extracted their OSEM implementation (and simplified it for speed) for our experiments. We also mention in passing that this toolkit trades-off storage for speed, gaining performance at the cost of doubling the storage requirements by essentially storing both $\boldsymbol{A}^T$ and $\boldsymbol{A}$.

4.1.1 Experiments with sparse random matrices

In our first set of experiments we show results with large sparse random matrices to get a feeling for the behavior of method. Dimensions of the matrices used are summarized in Table 1.

Size	# nonzeros ($\times 10^6$)	density
12288×4096	9.12	.1812
17664×8464	14.23	.0952
24576×16384	23.45	.0582
30720×25600	30.84	.0392
49152×65536	28.86	.009
98304×131072	89.88	.007

Table 1: Matrices used for first set of experiments

The aim of this experiment is to demonstrate the rapid convergence of our method compared to the other algorithms. Another aspect is the scalability. Under both these yardsticks, our method excels on this random data.

Figure 1 shows objective function values against the running time (in seconds). We selected problems (without loss of generality) where the true objective function was zero to permit a clearer illustration of the differences between the methods. All algorithms were given a stopping criterion of $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|/\|\boldsymbol{x}^k\| < 10^{-5}$, i.e., when the relative change from one iteration to the next became small. Our NMML algorithm *vastly* outperforms the other standard algorithms. For example, in the first row of plots, the NMML method is seen to converge to several digits of accuracy more than the other methods, that too in time almost negligible in comparison.

4.1.2 Experiments on PET Phantoms

We ran some experiments using both phantom images taken from the PET-Sorteo database [32, 33]. We used two different simulated system matrices $\boldsymbol{A}$, of dimensions 49152×65536 and a density approximately 0.0069. These sizes corresponded to 256×256 images that were projected into 256 radial and 192 angular bins. We simulated one simple system matrix $\boldsymbol{A}$ using the ASPIRE toolkit [15], while the other was generated randomly. We remark that in PET imaging, obtaining an accurate system matrix $\boldsymbol{A}$ is an entire research area [6] in itself, and is thus not treated in this paper.

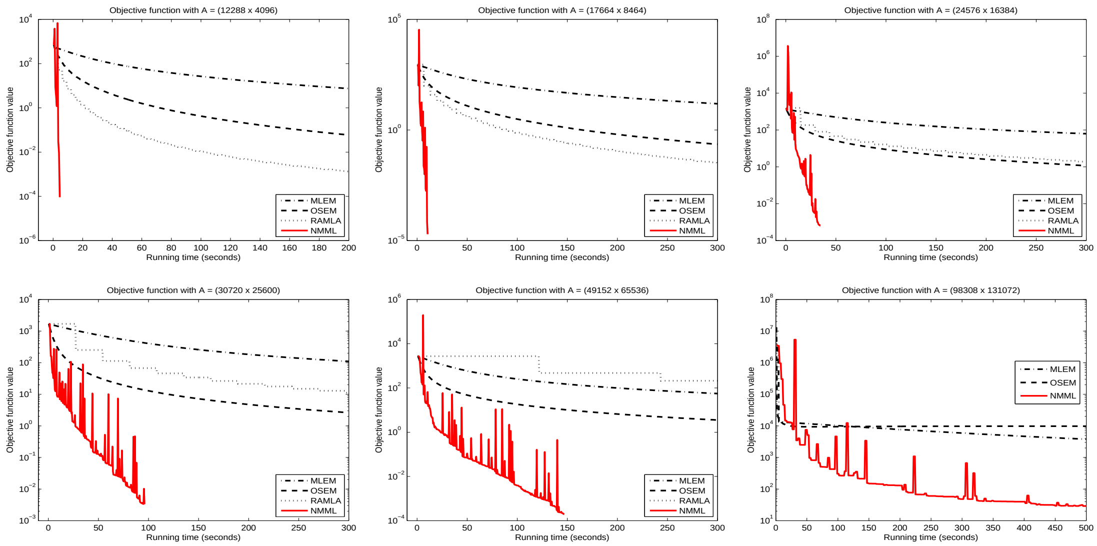

Figure 1: Running time and convergence comparison for the various algorithms (note y-axis is logarithmic). We had to run our own implementation of OSEM for the last experiments as the one derived from IRT [15] ran out of memory (as did RAMLA). To aid visual presentation the plots show at most the first 500 seconds of the individual runs. MLEM and OSEM usually did not converge to accuracies competitive with our method even after 1000 seconds.

4.1.3 Regularized reconstruction

Figure 2 shows reconstruction with the random $\boldsymbol{A}$, and one can see that NMML achieves better results than OSEM in the same time. Figure 3 covers a more realistic scenario and shows results of a regularized reconstruction using $\boldsymbol{A}$ generated with ASPIRE. Again we see that the results from NMML not only look better, but also have smaller objective function value. Both NMML and OSEM were run for the same amount of time.

4.2 Transmission Tomography

For transmission tomography, the EM approach leads to a cumbersome algorithm [24]. We compare the following methods:

1. TRCVX – an approximate iterative algorithm of [24] that ignores the positivity constraints and computes updates using an approximate Newton step to solve the underlying non-linear parameter update equations
2. NMML – our non-monotonic algorithm applied to the transmission problem.

We highlight the fact that for Transmission Tomography the algorithms available in the literature are more complicated than the corresponding methods for Emission Tomography. However, our NMML method retains its simplicity. We implemented the TRCVX method based on the iteration given in the paper [24], though we had to add an additional projection step $\boldsymbol{x} \leftarrow \max\{0, \boldsymbol{x}\}$, without which their proposed iteration was diverging for our datasets.

Other methods such as the OS-SPS algorithm of [13], both of which are available in the IRT toolkit [15], remain to be tested. We chose TRCVX because of its simplicity and scalability. For transmission we include only results on simulated data matrices. An expanded experimental treatment, that includes penalized likelihoods will appear in a different publication. We note in passing that our NMML method incorporates convex penalty functions without any particular difficulty, but several of the other scalable methods, including TRCVX and OS-SPS deal with regularization in a more *ad hoc* fashion leading to corresponding numerical difficulties.

We work with the same sized data matrices as mentioned in Table 1. However, the data was generated to correspond to the Poisson model (3.2). We note that we had to normalize the projection counts $\boldsymbol{y}$, as well as the columns of matrix $\boldsymbol{A}$ for numerical stability.

In all our experiments, we initialized the TRCVX method with an appropriately scaled all ones vector, while NMML was initialized using one iteration of TRCVX. We found the latter to be helpful in ensuring competitive convergence of the method.

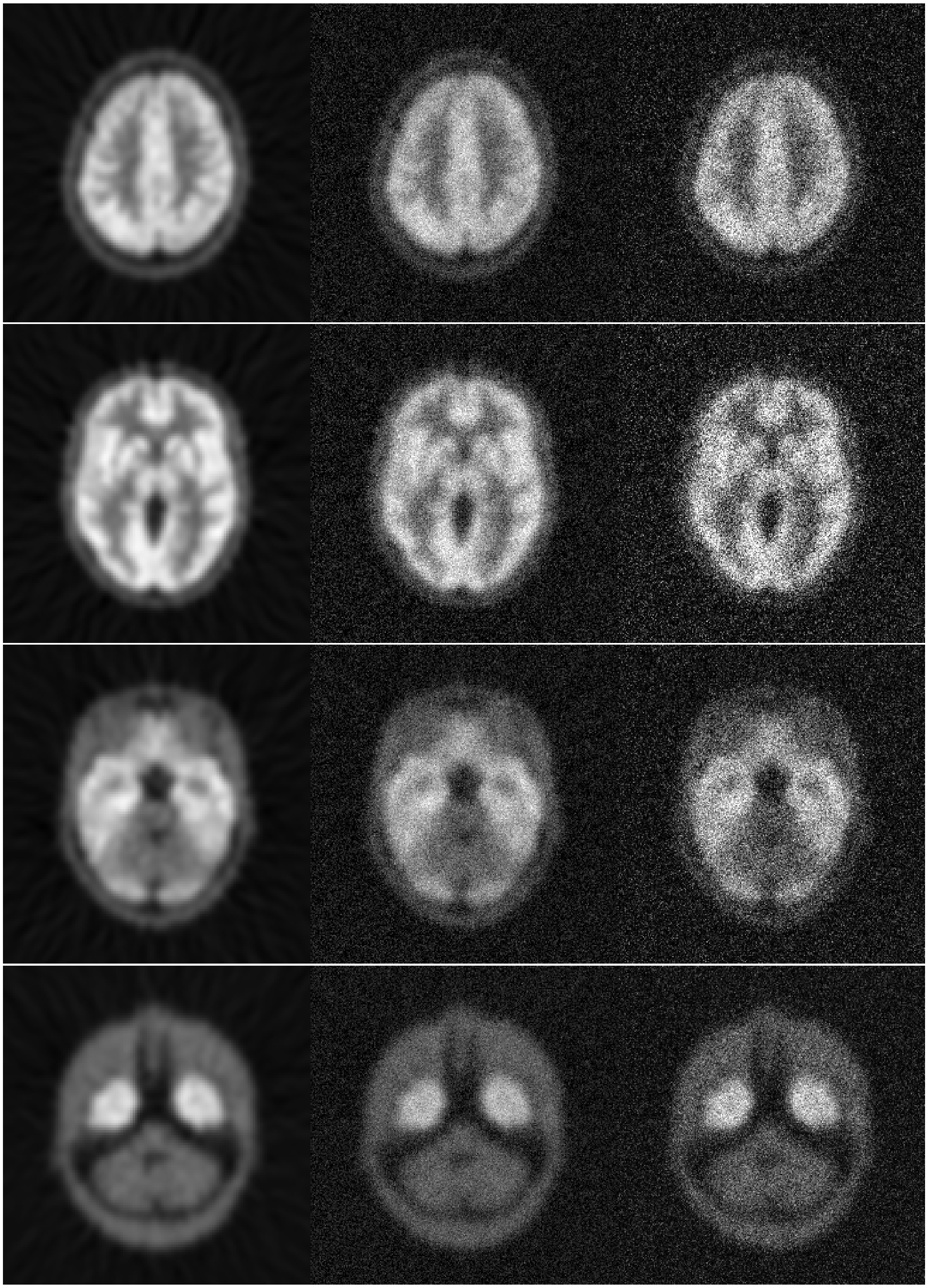

Figure 2: Image reconstruction via NMML and OSEM for brain image phantoms (of size 256×256). Noise was added to the true image and then removed via NMML and OSEM; both algorithms ran for 20 seconds. OSEM results appear noisier than NMML.

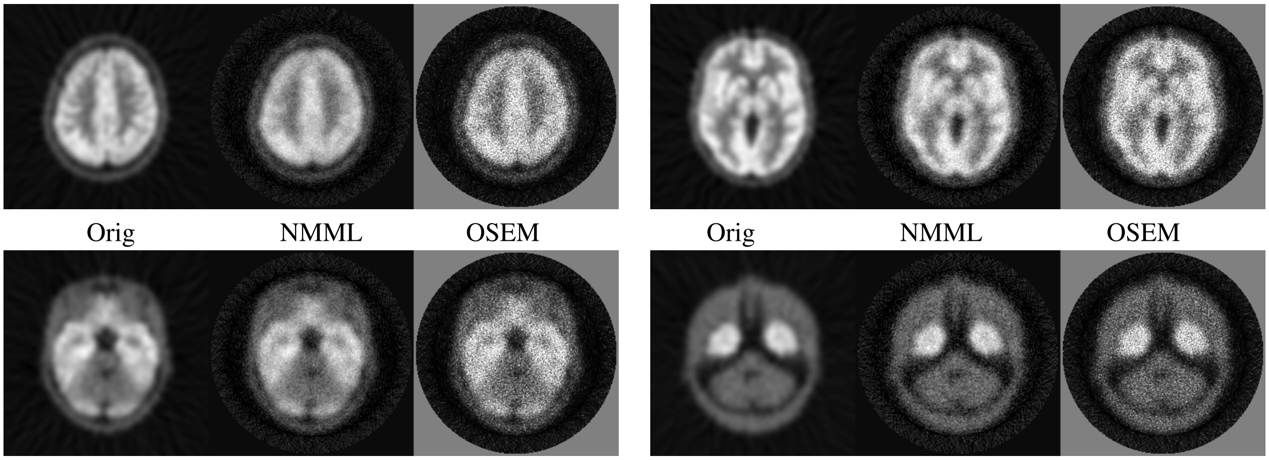

Figure 3: Regularized image reconstruction via NMML and OSEM for brain image phantoms. Regularizer was $\beta\|\boldsymbol{x}\|^2$ with $\beta = 1$. Noise was added to the true image and then removed via NMML and OSEM. For each row the corresponding objective function values are displayed on the right. NMML images are smoother, which is to be expected because of better handling of the regularizer.

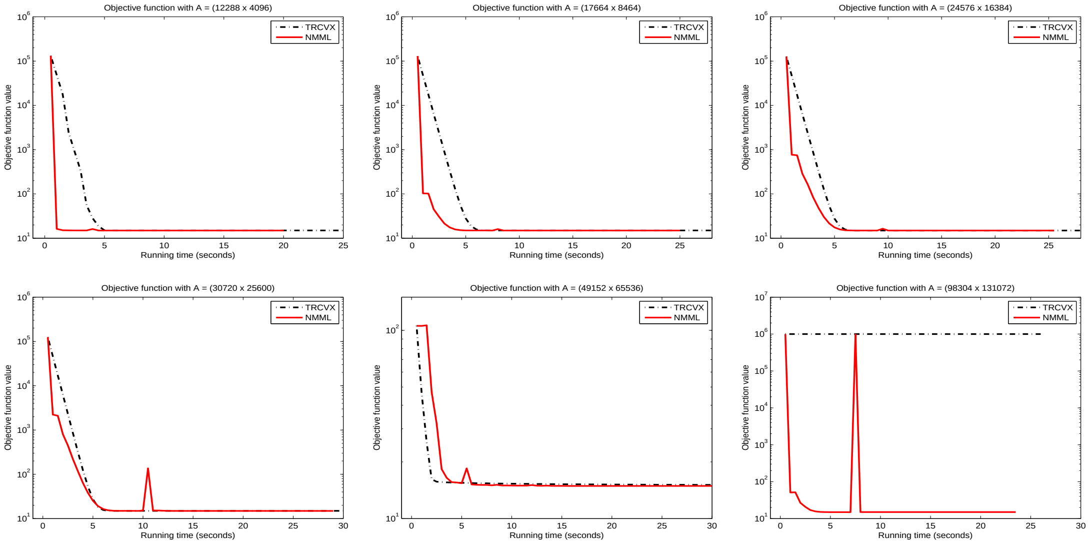

Figure 4: Running time and convergence comparison for TRCVX and NMML (note y-axis is logarithmic). Notice that in the last experiment, the TRCVX method essentially failed to make progress and converged to a much higher objective function value—this behavior is easily explained because TRCVX includes hacks to make it converge.

From Figure 4 one can see that almost always NMML converges faster to the the optimum solution, though the improvements are not as dramatic as in the case of Emission Tomography experiments (Figure 1). It is also interesting to note that in this case, NMML exhibits much lesser non-monotonic behavior—indicating that the amount of non-monotonicity might be related to convergence advantages. This intuition is further strengthened by the last plot in Figure 4, where the TRCVX method *fails* to converge to the correct solution, while NMML exhibits a huge non-monotonic step before converging. A theoretical investigation of the relation between the degree of non-monotonicity and rate of convergence lies outside the scope of the analysis presented in this paper, and remains a piece of our future work.

5 Conclusions and Future work

In this report we extended the non-monotonic optimization algorithm of [22] to the case of maximizing Poisson likelihood and related variants. Our extension included an initial proof of convergence, under fairly simple and easy to satisfy assumptions. We showed application of our algorithm to Positron Emission Tomography (PET), Transmission Tomography, and Image Restoration problems, in addition to enlisting several other important applications that can benefit from our methods.

We showed initial experimental results for our three main applications observing considerable speedups in comparison with standard methods for these problems. However, several important directions of future work and extensions do remain open at this point, and we are continuing to address them as a part of of research. For example,

- a better understanding of the convergence properties of the algorithm combined with a sharper analysis of the rate of convergence
- scaling up the algorithm to even larger problems, e.g., by parallelization
- making the method more robust to initialization
- automatically handling ill-conditioning in the data either via regularization, preconditioning or other approaches.
- Fine-tuning our approach to each particular application mentioned in the paper, along with a more thorough experimental validation are important facets of our ongoing work.

References

[1] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero. Convergent incremental optimization transfer algorithms: application to tomography. *IEEE Tran. Med. Imag.*, 25(3):283–296, March 2006.

[2] J. Barzilai and J. M. Borwein. Two-Point Step Size Gradient Methods. *IMA J. Num. Analy.*, 8(1), 1988.

[3] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, 1999.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787.

[5] J. Browne and A. B. de Pierro. A row-action alternative to the EM algorithm for maximizing likelihood in emission tomography. *IEEE Tran. Med. Imag.*, 15(5):687–699, 1996.

[6] I. Buvat and I. Castiglioni. Monte Carlo Simulations in SPET and PET. *Q. J. Nucl. Med.*, 46:48–61, 2002.

[7] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[8] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[9] H.-Z. Chen and T.-L. Liu. Trust-region methods for real-time tracking. In *ICCV*, pages 717–722, 2001.

[10] I. S. Dhillon and S. Sra. Generalized Nonnegative Matrix Approximations with Bregman Divergences. In *NIPS 18*, Vancouver, Canada, 2006.

[11] H. Erdoğan and J. A. Fessler. Accelerated monotonic algorithms for transmission tomography. In *Prof. IEEE ICIP*, volume 2, pages 680–684, 1998.

[12] H. Erdoğan and J. A. Fessler. Monotonic algorithms for transmission tomography. *IEEE Tran. Med. Imag.*, 18, 1999.

[13] H. Erdoğan and J. A. Fessler. Ordered subsets algorithms for transmission tomography. *Phys. Med. Biol.*, 44 (11), 1999.

[14] J. Fessler. *Image reconstruction: Algorithms and Analysis*. 2008. Book preprint.

[15] J. Fessler. ASPIRE & IRT Software Toolkits. http://www.eecs.umich.edu/ fessler/, 2008.

[16] J. A. Fessler. *Image reconstruction: Algorithms and analysis*. Under Preparation, 2008.

[17] J. A. Fessler, E. P. Ficaro, N. H. Clinthorne, and K. Lange. Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction. *IEEE Tran. Med. Imaging*, 16, 1997.

[18] A. Gholipour, N. Kehtarnavaz, R. W. Briggs, and K. S. Gopinath. Kullback-Leibler distance optimization for non-rigid registration of echo-planar to structural magnetic resonance brain images. In *Proceedings ICIP 2007*, pages 221–224, Oct. 2007.

[19] P. C. Hansen, J. G. Nagy, and D. P. O'Leary. *Deblurring images: Matrices, Spectra and Filtering*. SIAM, 2006.

[20] H. M. Hudson and R. S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Tran. Med. Imag.*, 13(4):601–609, 1994.

[21] C. A. Johnson, J. Seidel, and A. Sofer. Interior-point methodology for 3-D PET reconstruction. *IEEE Tran. Med. Imag.*, 19(4):271–285, 2000.

[22] D. Kim, S. Sra, and I. S. Dhillon. A Non-monotonic Gradient Projection Method for the Non-negative Least Squares Problem. Technical report, Univ. of Texas at Austin, June 2008.

[23] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assit. Tomography*, 8(2):306–316, 1984.

[24] K. Lange and J. A. Fessler. Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Tran. Imag. Proc.*, 4, 1995.

[25] D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In *NIPS*, pages 556–562, 2000.

[26] S. H. Manglos, G. M. Gagne, A. Krol, F. D. Thomas, and R. Narayanaswamy. Transmission maximum-likelihood reconstruction with ordered subsets for cone beam CT. *Phsy. Med. Biol.*, 40, 1995.

[27] E. U. Mumcuoglu, R. Leahy, S. R. Cherry, and Z. Zhou. Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images. *IEEE Tran. Med. Imag.*, 13, 1994.

[28] E. U. Mumcuoglu, R. Leahy, and S. R. Cherry Z. Zhenyu. Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images. *IEEE Tran. Med. Imag.*, 13(4):687–701, 1994.

[29] J. Nocedal. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation*, 95:339–353, 1980.

[30] A. R. De Pierro. A modified Expectation Maximization Algorithm for Penalized Likelihood Estimation in Emission Tomography. *IEEE Tran. Med. Imag.*, 14(1):132–137, 1995.

[31] A. J. Reader and H. Zaidi. Advances in PET Image Reconstruction. *PET Clinics*, 2(2):173–190, 2007.

[32] A. Reilhac. PET-SORTEO. http://sorteo.cermep.fr/home.php.

[33] A. Reilhac, C. Lartizien, N. Costes, S. Sans, C. Comtat, R. N. Gunn, , and A. C. Evans. PET-SORTEO: A Monte Carlo-based simulator with high count rate capabilities. *IEEE Trans. Nucl. Sci.*, 51(1), 2004.

[34] R. M. Rewitt and S. Matej. Overview of methods for image reconstruction from projections in emission computed tomography. *Proc. IEEE*, 91(10):1588–1611, 2003.

[35] W. H. Richardson. Bayesian-based iterative method of image restoration. *J. Opt. Society Amer.*, 62:55–59, 1972.

[36] J.B. Rosen. The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1):181–217, 1960.

[37] K. Sauer and C. Bouman. A local update strategy for iterative reconstruction from projections. *IEEE Tran. Signal Proc.*, 41, 1993.

[38] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Tran. Medical Imaging*, 1:113–122, 1982.

[39] N. Z. Shor. *Minimization Methods for Non-differentiable Functions*. Springer, 1985.

[40] N. I. Weisenfeld and S. K. Warfield. Normalization of joint image-intensity statistics in MRI using the Kullback-Leibler divergence. In *IEEE ISBI*, 2004.

[41] Wikipedia. Poisson Distribution. http://en.wikipedia.org/wiki/Poisson_distribution, 2008.