Multiplicative Updates for Optimization Problems with Dynamics

Abbas Kazemipour[†⋆], Behtash Babadi[†], Min Wu[†], Kaspar Podgorski[⋆], Shaul Druckmann[⋆]
[†]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
[⋆]Janelia Research Campus, Ashburn, VA, USA
{kaazemi, behtash, minwu}@umd.edu, {podgorskik, druckmanns}@janelia.hhmi.org

***Abstract*—We consider the problem of optimizing general convex objective functions with nonnegativity constraints. Using the Karush-Kuhn-Tucker (KKT) conditions for the nonnegativity constraints we will derive fast multiplicative update rules for several problems of interest in signal processing, including nonnegative deconvolution, point-process smoothing, ML estimation for Poisson Observations, nonnegative least squares and nonnegative matrix factorization (NMF). Our algorithm can also account for temporal and spatial structure and regularization . We will analyze the performance of our algorithm on simultaneously recorded neuronal calcium imaging and electrophysiology data.**

***Index Terms*—NMF, point process smoothing, Poisson image reconstruction, nonnegativity, KKT conditions, latent variables**

I. Introduction

The advent of big data has given rise to new challenges in signal processing. Fast and scalable solvers for solving large optimization problems remains a big challenge of optimization theory. In this paper we consider the problem of solving general optimization problems under nonnegativity constraints. Such optimization problems arise in many applications of interest. Examples include nonnegative matrix factorization for images of objects [1], Poisson image reconstruction [2], point process smoothing for stimulus-response experiments in neurophysiology [3], nonnegative least squares [4] and nonnegative calcium deconvolution [5]. In this paper we will use the KKT conditions [6] to provide a unified framework for solving such optimization problems with nonnegativity constraints. As we will see these conditions naturally lead to multiplicative updates with suitable convergence in many applications.

Multiplicative updates have been used for solving ML and MAP estimation as well as KL-divergence minimization. Many of these algorithms are special cases of the so-called proximal backward-forward scheme [7]. These algorithms try to find fixed points of a set of equations resulting from setting gradients of the objective function to zero. A With the help of parallel computing and graphics processing units (GPUs), these iterative methods can be solved very fast. Therefore, they become increasingly important. An important application of these multiplicative updates is the Richardson-Lucy (RL) algorithm for image deconvolution [8], which is widely used in astronomy and microscopy [9]. The RL algorithm recovers the ML estimate of a sample under Poisson statistics [10].

Multiuplicative updates are commonly contrasted with gradient descent methods. Their update steps do not necessarily follow the direction of the steepest descent. Multiplicative updates are argued to be insensitive to noise and more flexible [11]. Despite fast early convergenece multiplicative updates are claimed to converge slowly in later stages [12]. However, this argument has been refuted for Poisson Image reconstruction [11], the Weiszfeld problem [7] and NMF [13] by showing their equivalence to a Majorization Minimization (MM) algorithm which has linear convergence in iterations [14]. In contrast, both multiplicative updates and gradient descent based algorithms such as the proximal-gradient method have sublinear rate of convergence [7] in general. Moreover, with specific choices of the stepsize, in many cases such as the Weiszfeld problem these algorithms have proven to be equivalent [7] . These findings suggest that slow convergence of multiplicative updates in some cases is due to absence of strong convexity in the objective function.

An advantage of multiplicative updates over gradient descent based algorithms is their flexibility in terms of adapting to the objective functions without the need for calculation dual functions or tuning extra parameters such as the step-size. Despite the recent breakthroughs in choosing these parameters [15], each step in calculation of the step size is usually as costly as an iteration of the algorithm which is not as effective for big data problems. In addition many problems such as image reconstruction and calcium deconvolution [16] are spatially separable and are easily parallelized.

Finally, temporal dynamics and penalization play an important role in signal recovery from noisy data. Examples include state-space estimations, video reconstruction and total variation denoising problems. Apart from special cases, the solutions to these problems are generally batch mode and computationally demanding. In this paper we provide a unified framework for generalizations of multiplicative updates to the problems with nonnegativity constraints and dynamics by adapting the update rules to different forms of penalties. We have empirically found that multiplicative updates show superior convergence properties and speed to gradient descent methods for models that include dynamics and penalization.

II. Notations and Problem Formulation

Throughout the paper we will use the following notation. We use the convention $[T] = \{1, \cdots, T\}$ and $\mathbf{W}_{[T]} = [\mathbf{w}_1, \cdots, \mathbf{w}_T]$, i.e. $\mathbf{w}_k$ represents the kth column of $\mathbf{W}_{[T]}$. $\odot$

and $\oslash$ denote elementwise multiplication and division respectively. Throughout the paper we will use the terms innovations and spikes interchangeably. Unless otherwise stated, a function acts on a vector elementwise. For a matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m\times n}$ its mixed p,q-norm is denoted by $\|\mathbf{A}\|_{p,q}$, i.e.

$$\|\mathbf{A}\|_{p,q} = \left[\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|a_{ij}|^p\right)^{q/p}\right]^{1/q},$$

and $\|\mathbf{x}\|_{\mathbf{\Sigma}}^2 = \mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}$. Finally, for a summation

$$L = \sum_{i=1}^{n} l_i = L^+ - L^-,$$

where $L^+ = \sum_{i=1}^{n}\max\{l_i, 0\}$, $L^- = -\sum_{i=1}^{n}\max\{-l_i, 0\}$.

We consider a convex optimization problem of the form

$$\underset{\mathbf{X}\succeq 0}{\text{minimize}}\, \mathcal{F}(\mathbf{X}) := \mathcal{L}(\mathbf{X}) + \lambda\mathcal{P}(\mathbf{X}), \tag{1}$$

where $\mathcal{L}(.)$ denotes a convex objective function and $\mathcal{P}(.)$ denotes a suitable penalty function. Typically $\mathcal{L}(.)$ is a negative log-likelihood and $\mathcal{P}(.)$ is a smooth norm. Additionally we make the assumption that both $\mathcal{L}$ and $\mathcal{P}$ are differentiable with respect to $\mathbf{X}$ on the positive orthant,

Among the algorithms used for solving (1) one can name the primal-dual algorithm and proximal gradient method. For specific choices of the penalty functions ℓ_1 and ℓ_2 (Tikhonov) regularization several fast algorithms exist. However these algorithms cannot be easily generalized to arbitrary penalties or temporal dynamics. In some cases such as the gradient based methods they require knowledge of the proximal map or have extra parameters such as the step size to be tuned and chosen. Calculation of the step size is usually as costly as a few iterations of the algorithm and could slow them down. However, our approach to solving (1) does not require tuning of extra parameters and is very simple to implement. We will next discuss our solution.

III. Solution to the Main Optimization Problem

In this section we will introduce our solution to (1) via multiplicative updates. The Lagrangian form of (1) is given by

$$\underset{\mathbf{X},\mathbf{S}\succeq 0}{\text{minimize}}\, \mathcal{F}(\mathbf{X}) + \mathbf{S}\odot\mathbf{X}. \tag{2}$$

Assuming convexity and zero duality gap, the KKT conditions for (2) can be expressed as

$$\mathbf{X}^\star \succeq 0,\quad \mathbf{S}^\star \succeq 0, \tag{3}$$
$$\mathbf{S}^\star \odot \mathbf{X}^\star = \mathbf{0}, \tag{4}$$
$$\nabla_{\mathbf{X}}\mathcal{F}(\mathbf{X}) + \mathbf{S} = \mathbf{0}. \tag{5}$$

In the rest of the paper, we drop the subscripts and arguments whenever they can be understood from the context. Multiplying (5) by $\mathbf{X}$ and using (4) we obtain:

$$\nabla\mathcal{F}(\mathbf{X})\odot\mathbf{X} = \mathbf{0}. \tag{6}$$

Our solution to (1) looks for a positive fixed point of (6). Therefore giving us the multiplicative update rule

$$\mathbf{X}^{(k+1)} \leftarrow \left(\nabla\mathcal{F}(\mathbf{X}^{(k)})\right)^{-} \oslash \left(\nabla\mathcal{F}(\mathbf{X}^{(k)})\right)^{+} \odot \mathbf{X}^{(k)}. \tag{7}$$

In all application introduced in this paper we initialize the algorithm with a positive solution, the choice of which depends on the application. The update rule will then ensure the solution remains positive. In order to provide more insight into our algorithm we will next provide several examples and applications.

In applications of interest in this paper we consider temporal dynamics in $\mathbf{X}$, hence referring to our algorithm by FAst DEconvolution (FADE) algorithm. In the spirit of easing reproducibility, we have made MATLAB implementations of our codes publicly available [17].

IV. Examples and Application to Real Data

In this Section we will provide examples of the multiplicative updates in different applications of interest.

A. Nonnegative Deconvolution

In its simplest form the nonnegative deconvolution problem can be formalized by considering the state-space model given by

$$\mathbf{x}_t = \mathbf{\Theta}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \tag{8}$$

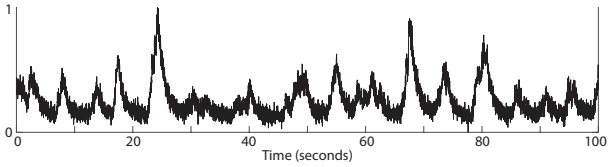
where $\mathbf{w}_t \succeq 0$ models the innovations at time $t \in [T]$. Usually, the observation noise is assumed to be i.i.d normal, i.e. $\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ and the measurement matrices $\mathbf{A}_t$ are assumed to conserve positivity. For this problem we can identify $\mathbf{W} = \mathbf{W}_{[T]}$ and

$$\mathcal{L}(\mathbf{W}) = \sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|_{\mathbf{\Sigma}_t}^2 = \sum_{t=1}^{T}\left\|\mathbf{y}_t - \mathbf{A}_t\sum_{\tau=0}^{t-1}\mathbf{\Theta}^\tau\mathbf{w}_{t-\tau}\right\|_{\mathbf{\Sigma}_t}^2,$$

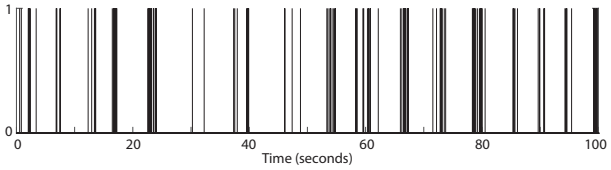
from which we can calculate

$$\left(\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{W})\right)^+ = \sum_{\tau\geq t}\left(\mathbf{\Theta}^{\tau-t}\right)^T\mathbf{A}_\tau^T\mathbf{\Sigma}_\tau^{-1}\mathbf{y}_\tau,$$
$$\left(\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{W})\right)^- = \sum_{\tau\geq t}\left(\mathbf{\Theta}^{\tau-t}\right)^T\mathbf{A}_\tau^T\mathbf{\Sigma}_\tau^{-1}\mathbf{A}_\tau\mathbf{x}_\tau.$$

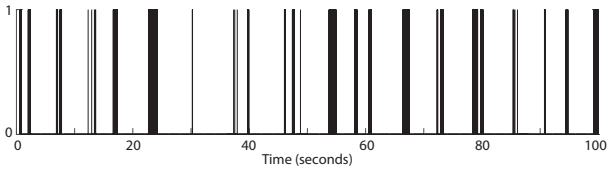
Typically one can use a smooth norm in order to enforce prior assumptions on the spikes, for example one can use a sparsity inducing prior $\mathcal{P} = \|\mathbf{W}\|_{1,1}$, for which $(\nabla\mathcal{P})^+ = \mathbf{1}$ and $(\nabla\mathcal{P})^- = \mathbf{0}$. The choice of the penalty function on the spikes is arbitrary and could differ from application to application. In applications where such information is not readily available, one would like to enforce minimal assumptions on the spikes and hence would want to enforce non-informative priors. The most famous example of such priors is known as Jeffrey's prior [18]. However this problem is an active area of research as there is no unanimously agreed upon choice of non-informative priors.



(a) Normalized calcium traces



(b) Ground-truth spikes



(c) Deconvolved spikes

Fig. 1: Application of the FADE algorithm to calcium deconvolution problem.

B. Application to Calcium Deconvolution

Calcium imaging is used to visualize currents associated with action potentials in living neurons. This is done using fluorescent molecules that change their fluorescence properties upon binding calcium, and using a one- or two-photon fluorescence microscope to record these changes [19], [20]. Inferring action potentials (spikes) from calcium recordings, referred to as calcium deconvolution, is an important problem in neural data analysis. For the special case of calcium imaging we have $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, $\mathbf{A}_t = \mathbf{I}$ and $\mathbf{\Theta} = \theta\mathbf{I}$. Here the baseline is assumed to have been estimated and subtracted separately, but can be estimated similarly. We refer to [5] for details on estimation of the unknown parameters σ^2 and θ and a list of methods used for calcium deconvolution. These approaches require solving convex optimization problems, which do not scale well with the temporal dimension of the data.

Figure 1 shows application of the FADE algorithm to simultaneously recorded imaging and electrophysiology data. The algorithm has covnverged (less than 0.5% change in spikes) in 28 iterations. The data is a 100 second interval from the spikefinder challenge [21] (dataset 3, neuron 1). We have used an AR(2) model and an $\ell_{0.5,1}$ penalty on the spikes in order to enforce temporal sparsity. The spikes have been obtained by simply thresholding the deconvolved spikes at 3σ, where σ is the estimated standard deviation of the observation noise. A comparison of the performance of our algorithm with many other methods is provided on the spikefinder challenge website [21].

One can use spatial regularization on elements of $\mathbf{w}_t$ in this setup as well as compressive sensing regimes for when $\mathbf{A}$ satisfies the restricted isometry property RIP [5]. We refer to [5] for a more detailed discussion.

C. Poisson Image Reconstruction and Point Process Smoothing

State-space models with Poisson observations have also been studied in many applications of interest. In neuroscience, temporal dynamics of stimulus-response experiments in neurophysiology have been modeled using a Poisson state-space model. In emission tomography, dynamics of the photons hitting the detectors can be modeled with Poisson noise models. Without loss of generality we consider the state-space model given by

$$\mathbf{x}_t = \mathbf{\Theta}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t \sim \text{Poisson}\left(\phi\left(\mathbf{A}\mathbf{x}_t + \mathbf{b}_t\right)\right), \quad (9)$$

where $\mathbf{w}_t \succeq 0$ and $\mathbf{b}_t \succeq 0$ model the spikes and baseline rates at time $t \in [T]$ respectively and $\phi(.)$ is a bijective convex function. Common examples include $\phi(x) = \exp(x)$, $\phi(x) = \frac{\exp(x)}{1+\exp(x)}$ and $\phi(x) = x$. We assume the latter in our derivations due to space considerations.

Several approaches have been proposed in the literature for finding the MAP solution to (9). We refer to [22] for a detailed list of these methods. In [3] the authors used the maximum a posteriori derivation of the Kalman filter and proposed an approximate expectation maximization (EM) approach to this problem by Gaussian approximations of the posterior likelihood. This EM approach has several shortcomings. First, it requires solving a nonlinear system of equations which could potentially be computationally costly. Second, it only accounts for Gaussian spikes. Third its performance heavily depends on the Poisson rate model, especially when the rates are small, which is the usual case for spiking activities. In these cases usually $\phi(x) = \exp(x)$ is considered for stability of approximations. Moreover due to nonlinear recursive filtering nature of the problem, the performance of the Gaussian approximation quickly degrades as the dimension of the latent space goes beyond 2 or 3. Similarly, in [22] the authors proposed SPIRAL which uses a Gaussian approximation to $\mathcal{L}$ and is a gradient-based solution to (9). Except for the special cases of ℓ_1 and TV penalties, calculation of the Gaussian model is tedious leading to slow convergence. In [23] the authors introduce a variational auto-encoder (gradient descent based) model to retrieve the low-dimensional temporal factors.

In applications such as fluorescence microscopy, it is also common to use to use variance stabilizing transforms [22] such as square root filtering [24] in order to make Gaussian approximations to the Poisson distribution. In the high photon regime such transformations are not necessary as one can use infinite divisibility property of the Poisson distribution for Gaussian approximations. However one would then need to deal with complications arising from equality of the mean and the covariance matrices for such approximations. In contrast, our algorithm gives an exact solution, is fast, can account for any rate model and suitably scales with the problem dimensions.

The Gaussian approximations could then be used as an input to a Kalman smoother if the innovations (spikes) follow a

half-normal or Gaussian distribution. Despite the fact that our solutions are faster, exact and do not involve approximations, for Gaussian state-spaces the Kalman smoother provides a smoothed estimate of the covariances which could be used for building confidence intervals, whereas the covariances are not a direct output of the multiplicative updates.

Considering the MAP estimator for $\mathbf{W} = \mathbf{W}_{[T]}$ we can identify

$$\mathcal{L}(\mathbf{W}) = \sum_{t=1}^{T} \mathbf{1}^T \left(\mathbf{A}\mathbf{x}_t + \mathbf{b_t}\right) - \mathbf{y}_t^T \log\left(\mathbf{A}\mathbf{x}_t + \mathbf{b_t}\right), \quad (10)$$

for which we have

$$\left(\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{W})\right)^+ = \sum_{\tau \geq t} \left(\mathbf{\Theta}^{\tau - t}\right)^T \mathbf{A}_\tau^T \mathbf{1},$$

and

$$\left(\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{W})\right)^- = \sum_{\tau \geq t} \left(\mathbf{\Theta}^{\tau - t}\right)^T \mathbf{A}_\tau^T \left(\mathbf{y}_t \oslash \left(\mathbf{A}\mathbf{x}_t + \mathbf{b_t}\right)\right).$$

The penalty function and the corresponding terms can be calculated similar to the nonnegative deconvolution problem. A similar update rule can be derived for the baseline. The special case of $\mathbf{\Theta} = \mathbf{0}$ (no dynamics with the convention $\mathbf{0}^0 = \mathbf{I}$) and $\lambda = 0$ (no penalization) is known as the Richardson-Lucy (RL) iterations. The RL algorithm has also been used with TV seminorm regularization in [25]. Similar to the RL algorithm we can use FADE for blind deconvolution, when the measurement matrix $\mathbf{A}$ is unknown. In this setup one can alternatively update $\mathbf{A}$ and $\mathbf{X}$. We can also used FADE, for estimation of GLM models for self-exciting point process models [26].

D. Combination with Other Constraints

In many applications of interest the optimization problem could also include several inequality constraints. For example in fluorescence microscopy the maximum changes of the fluorescence level with respect to baseline (also referred to as $\frac{\mathbf{\Delta F}}{\mathbf{F}}$) is controlled by the properties of the indicator in use. In these situations we need to satisfy the KKT conditions for the extra constraints. Here we will introduce an adaptive method in order to achieve this goal. Consider the modified problem setup of Section IV-C given by

$$\begin{aligned} \left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t &= \mathbf{\Theta}\left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_{t-1} + \mathbf{w}_t \\ \mathbf{y}_t &\sim \text{Poisson}\Big(\mathbf{A}b_t(1 + \left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t)\Big) \end{aligned}, \quad (11)$$

where $b_t \geq 0$ denotes the known baseline fluorescence at time t, on top of which $\left(\frac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t$ lies. In addition to nonnegativity constraints we need to account for the following constraints

$$\left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t \preceq c_f \quad \text{for all } t \ . \quad (\star)$$

The constant c_f is a characteristic of the indicator used and is assumed to be known. In order to enforce $(\star)$ we proceed as in Algorithm 1.

Algorithm 1 Multiplicative Updates with Adaptive Regularization

procedure MULTIPLICATIVE UPDATES
  Initialize: $\mathcal{P}\left(\frac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t = \|\left(\frac{\mathbf{\Delta F}}{\mathbf{F}}\right)_{[T]}\|_{\infty,\infty}$, $\lambda = 0$, $\lambda_0 = 0.01$, $i = 0$.
  repeat
    if $\max\left(\frac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t \geq c_f$ and $i = 0$ **then**
      $\lambda \leftarrow \lambda_0$, $i \leftarrow 1$
    end if
    if $\lambda > 0$ **then**
      Set $\lambda \leftarrow \lambda \frac{\|\left(\frac{\mathbf{\Delta F}}{\mathbf{F}}\right)_{[T]}\|_{\infty,\infty}}{c_f}$
    end if
    Update $\mathbf{W}$.
  until convergence criteria met
end procedure

The main idea behind Algorithm 1 is that when the constraints are violated the complimentary slackness condition should be met for the optimal dual variable λ in Lagrangian form of the problem, meaning that the optimal solution should satisfy $\|\left(\frac{\mathbf{\Delta F}}{\mathbf{F}}\right)_{[T]}\|_{\infty,\infty} = c_f$, which is equivalent to finding a fixed point of updates for the dual (regularization) variable λ.

V. Other Examples

A. Dynamic Nonnegative Least Square (NLS)

The NLS problem can in general be formulated

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{V}, \qquad \mathbf{V} \sim \mathcal{N}(0, \sigma^2\mathbf{I}),$$
$$\mathcal{L}(\mathbf{X}) = \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_2^2, \ \left(\nabla\mathcal{L}\right)^+ = \mathbf{A}^T\mathbf{Y}, \ \nabla\left(\mathcal{L}\right)^- = \mathbf{A}^T\mathbf{A}\mathbf{Y}.$$

The most famous algorithm for solving the NLS problem is the active set method [4] which does not account for temporal dynamics in $\mathbf{x}_t$ or other forms of penalty. In these settings our update rules are very similar to the nonnegative deconvolution problem. A very useful example from the compressed sensing literature is the Multiple Measurement Vector (MMV) problem (without the positivity constraint) [27]. A commonly used penalty in this setup is the $\|\mathbf{X}\|_{2,1}$ which enforces row sparsity.

B. Dynamic Nonnegative Matrix Factorization (NMF)

The NMF problem is very similar to the NLS problem except that the matrix $\mathbf{A}$ is not known. In this case we can alternatively update our estimates of $\mathbf{A}$ and $\mathbf{X}$ [28].

$$\begin{aligned} &\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{V}, \qquad \mathbf{V} \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \\ &\mathcal{L}(\mathbf{X}) = \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_2^2, \\ &\left(\nabla_{\mathbf{X}}\mathcal{L}\right)^+ = \mathbf{A}^T\mathbf{Y}, \quad \left(\nabla_{\mathbf{X}}\mathcal{L}\right)^- = \mathbf{A}^T\mathbf{A}\mathbf{Y} \\ &\left(\nabla_{\mathbf{A}}\mathcal{L}\right)^+ = \mathbf{Y}\mathbf{X}^T, \quad \left(\nabla_{\mathbf{A}}\mathcal{L}\right)^- = \mathbf{Y}\mathbf{X}^T\mathbf{X} \end{aligned}$$

In the absence of penalization or dynamics we recover the multiplicative updates of [1]. Our update rules can also account for the dynamic case where

$$\begin{aligned} \mathbf{X}_t &= \alpha\mathbf{X}_{t-1} + \mathbf{W}_t, \qquad \mathbf{W}_t \succeq \mathbf{0} \\ \mathbf{Y}_t &= \mathbf{A}\mathbf{X}_t + \mathbf{V}_t, \qquad \mathbf{V}_t \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \end{aligned}$$

For example one can account for sparsely changing temporal factors by considering a Laplacian distribution on $\mathbf{W}_t$.

VI. Extensions and Future Work

In this paper we considered convex optimization problems with nonnegativity constraints and provided unified multiplicative updates for them using the KKT conditions. These updates are easy to implement and parallelizable on a CPU. They do not require tuning of extra parameters such as the step size, exhibit fast convergence in practice and can account for temporal dynamics and smooth penalties without slowing down.

Although in the absence of convexity the KKT conditions no longer hold, we have empirically observed that our updates exhibit good performance when the problem has simple non-convexities. As an example one can model calcium saturation in the calcium deconvolution problem by adopting the calcium hill model given by $\mathbf{y}_t = \alpha \frac{\mathbf{x}_t}{\mathbf{x}_t + \mathbf{c}} + \mathbf{v}_t$ [29]. These observations suggest that suitable initializations result in convergence to a suitable local minimum. As another example one can combine the multiplicative updates with the IRLS algorithm [30] for ℓ_q, $q < 1$ minimization problems. The convergence of the IRLS algorithm was shown in the literature by showing an equivalence to a special case of the EM algorithm [31]. We applied this generalization to calcium imaging data using a nonconvex penalty.

Finally, the positivity constraint can easily be relaxed in the general form of the problems in two ways: First, any variable $\mathbf{X}$ can be decomposed into $\mathbf{X} = \mathbf{X}^+ - \mathbf{X}^-$, where both $\mathbf{X}^+$ and $\mathbf{X}^-$ are positive. Second, generalized positivity and negativity could be defined with respect to the boundary of the convex set of feasible solutions, i.e. any point point inside/outside the feasibility set could be considered as positive/negative. Generalized positive and negative terms in the decompositions could be redefined similarly. Therefore by looking for a generalized positive fixed point of the gradient of the log-likelihood, the multiplicative updates can be generalized to a larger class of problems with not necessarily positivity constraints. We leave full details of these extensions and examples and their convergence properties to future work.

References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] R. M. Willett, Z. T. Harmany, and R. F. Marcia, "Poisson image reconstruction with total variation regularization," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 4177–4180.

[3] A. Smith and E. N. Brown, "Estimating a state-space model from point process observations," *Neural Comp.*, vol. 15, no. 5, pp. 965–991, 2003.

[4] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1995.

[5] A. Kazemipour, J. Liu, K. Solarana, D. Nagode, P. Kanold, M. Wu, and B. Babadi, "Fast and stable signal deconvolution via compressible state-space models," *IEEE Transactions on Biomedical Engineering*, 2017.

[6] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[7] D. P. Palomar and Y. C. Eldar, *Convex optimization in signal processing and communications*. Cambridge university press, 2010.

[8] L. B. Lucy, "An iterative technique for the rectification of observed distributions," *The astronomical journal*, vol. 79, p. 745, 1974.

[9] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE transactions on medical imaging*, vol. 1, no. 2, pp. 113–122, 1982.

[10] W. H. Richardson, "Bayesian-based iterative method of image restoration," *JOSA*, vol. 62, no. 1, pp. 55–59, 1972.

[11] M. Yan, A. Bui, J. Cong, and L. A. Vese, "General convergent expectation maximization (em)-type algorithms for image reconstruction," *Inverse Problems and Imaging*, vol. 7, no. 3, pp. 1007–1029, 2013.

[12] R. L. White, "Image restoration using the damped richardson-lucy method," in *1994 Symposium on Astronomical Telescopes & Instrumentation for the 21st Century*. International Society for Optics and Photonics, 1994, pp. 1342–1348.

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[14] C. Wu, C. Yang, H. Zhao, and J. Zhu, "On the convergence of the em algorithm: From the statistical perspective," *arXiv preprint arXiv:1611.00519*, 2016.

[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang *et al.*, "Simultaneous denoising, deconvolution, and demixing of calcium imaging data," *Neuron*, vol. 89, no. 2, pp. 285–299, 2016.

[17] *MATLAB implementation of the FCSS algorithm*. Available on GitHub Repository: https://github.com/kaazemi/FADE, 2016.

[18] H. Jeffreys, "An invariant form for the prior probability in estimation problems," in *Proceedings of the Royal Society of London a: mathematical, physical and engineering sciences*, vol. 186, no. 1007. The Royal Society, 1946, pp. 453–461.

[19] D. Smetters, A. Majewska, and R. Yuste, "Detecting action potentials in neuronal populations with calcium imaging," *Methods*, vol. 18, no. 2, pp. 215–221, 1999.

[20] C. Stosiek, O. Garaschuk, K. Holthoff, and A. Konnerth, "In vivo two-photon calcium imaging of neuronal networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 7319–7324, 2003.

[21] *SpikeFinder Challenge*. Available at: http://spikefinder.codeneuro.org/, 2016.

[22] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "Spiral out of convexity: Sparsity-regularized algorithms for photon-limited imaging," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2010, pp. 75 330R–75 330R.

[23] D. Sussillo, R. Jozefowicz, L. Abbott, and C. Pandarinath, "Lfads-latent factor analysis via dynamical systems," *arXiv preprint arXiv:1608.06315*, 2016.

[24] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.

[25] N. Dey, L. Blanc-Feraud, C. Zimmer, P. Roux, Z. Kam, J.-C. Olivo-Marin, and J. Zerubia, "Richardson–lucy algorithm with total variation regularization for 3d confocal microscope deconvolution," *Microscopy research and technique*, vol. 69, no. 4, pp. 260–266, 2006.

[26] A. Kazemipour, M. Wu, and B. Babadi, "Robust estimation of self-exciting generalized linear models with application to neuronal modeling," *IEEE Transactions on Signal Processing*, 2017.

[27] J. Chen and X. Huo, "Sparse representations for multiple measurement vectors (mmv) in an over-complete dictionary," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 4. IEEE, 2005, pp. iv–257.

[28] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[29] R. Yasuda, E. A. Nimchinsky, V. Scheuss, T. A. Pologruto, T. G. Oertner, B. L. Sabatini, and K. Svoboda, "Imaging calcium concentration dynamics in small neuronal compartments," *Sci STKE*, vol. 2004, no. 219, p. pl5, 2004.

[30] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*. IEEE, 2008, pp. 3869–3872.

[31] D. Ba, B. Babadi, P. L. Purdon, and E. N. Brown, "Convergence and stability of iteratively re-weighted least squares algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 183–195, 2014.