This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning

Ying Sun, Prabhu Babu, and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*—This paper gives an overview of the majorization-minimization (MM) algorithm framework, which can provide guidance in deriving problem-driven algorithms with low computational cost. A general introduction of MM is presented, including a description of the basic principle and its convergence results. The extensions, acceleration schemes, and connection to other algorithm frameworks are also covered. To bridge the gap between theory and practice, upperbounds for a large number of basic functions, derived based on the Taylor expansion, convexity, and special inequalities, are provided as ingredients for constructing surrogate functions. With the pre-requisites established, the way of applying MM to solving specific problems is elaborated by a wide range of applications in signal processing, communications, and machine learning.**

***Index Terms*—Majorization-minimization, upperbounds, surrogate function, non-convex optimization.**

I. Introduction

In the era of big data, we are witnessing a fast development in data acquisition techniques and a growth of computing power. From an optimization perspective, these can result in large-scale problems due to the tremendous amount of data and variables, which cause challenges to traditional algorithms [1]. For example, apart from trivially parallelizable or convex problems where decomposition techniques can be employed, solving a general problem with no structure to exploit calls for a large amount of computational resources (time and storage). Difficulties also arise when data is stored on different computers or is acquired in real-time. In these cases, it can be inefficient or even impossible to first collect the complete data set and then perform centralized optimization. Besides the aforementioned issues caused by the scale, a problem with a complicated form may lead to numerical problems as well. For instance, the second order derivatives, which are required by Newton-type nonlinear programming algorithms, can be costly to compute under this scenario. Facing these obstacles, devising problem-driven algorithms that can take advantage of the problem structure may be a better option than employing a general-purpose solver. This is where MM comes into play.

The MM procedure consists of two steps. In the first majorization step, we find a surrogate function that locally approximates the objective function with their difference minimized at the current point. In other words, the surrogate upperbounds the objective function up to a constant. Then in the minimization step, we minimize the surrogate function. The procedure is shown pictorially in Figure 1. A parallel argument can be made for maximization problems by replacing the upperbound minimization step by a lowerbound maximization step, and is referred to as minorization-maximization.

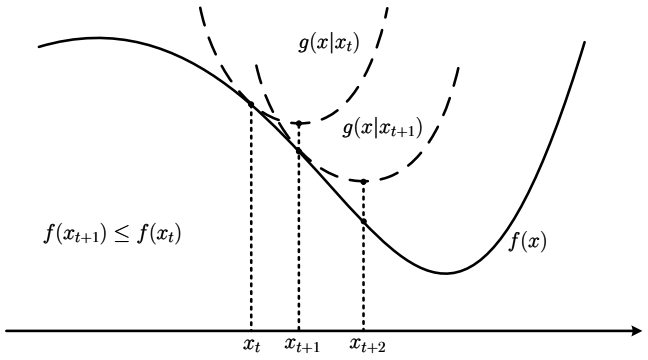

Figure 1. The MM procedure.

MM has a long history that dates back to the 1970s [2], and is closely related to the famous EM algorithm [3] intensively used in computational statistics. As a special case of MM, EM is applied mainly in maximum likelihood (ML) estimation problems with incomplete data, which was systematically introduced in the seminal paper [4] by Dempster, Laird, and Rudin in 1977. MM generalizes EM by replacing the E-step, which calculates the conditional expectation of the log-likelihood of the complete data set, by a minorization step that finds a surrogate function. The surrogate function keeps the key property of the E-step by being a lower bound of the objective function. As a consequence, MM shares most of the convergence results of EM. Compared to EM, which relies on a missing data interpretation of the problem, MM is easier to understand and has a wider scope of applications.

The idea of MM appears in statistics and image processing in early works including [5]–[9], and started taking shape as a general algorithm framework in [10]–[12]. It has been applied to a large number of problems since then [13], including sparse regression with non-convex or discontinuous objective functions [14]–[18], sparse principal component analysis (PCA) with cardinality constraint [19], canonical component analysis (CCA) [20], [21], covariance estimation [22]–[25], and matrix factorization [26], [27] with non-convex objective functions and constraints. It has also been applied to higher level applications such as image processing [28], [29], phase retrieval [30], and design [31], [32], just to name a few.

The key to the success of MM lies in constructing a surrogate function. Generally speaking, surrogate functions with the following features are desired [13]:

- Separability in variables (parallel computing);
- Convexity and smoothness;
- The existence of a closed-form minimizer.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

Consequently, minimizing the surrogate function is efficient and scalable, yielding a neat algorithm that is easy to implement.

Nevertheless, finding an appropriate surrogate function that yields an algorithm with low computational complexity is not an easy task. On one hand, to achieve a fast convergence rate, a surrogate function that tries to follow the shape of the objective function is preferable. On the other hand, it should be simple to minimize to so that the computational cost per iteration is low. Finding the right trade-off between these two opposite goals requires skills in applying inequalities to specific problems. As the main purpose of this paper, we are devoted to presenting surrogate function construction techniques, elaborated by examples in Section III and applications listed in Table I[1].

This paper is organized as follows. Section II serves as an introduction to MM, including a description of the framework, its convergence results, extensions, as well as accelerators. Section III presents the techniques and examples of constructing surrogate functions. Section IV connects MM with some other algorithm frameworks. Section V demonstrates the way of applying the inequalities in Section III to devise MM algorithms for real-world applications. Section VI concludes the overview.

Notation

Italic letters denote scalars, lower case boldface letters denote vectors, and upper case boldface letters denote matrices.

The sequence of nonnegative integers is denoted $\mathbb{N} := \{0, 1, \ldots\}$. Real numbers are denoted $\mathbb{R}$, and complex numbers are denoted $\mathbb{C}$. The Euclidean space of dimension n is denoted $\mathbb{R}^n$. The nonnegative (positive) orthant is denoted $\mathbb{R}^n_+$ ($\mathbb{R}^n_{++}$). The set of symmetric matrices of size $n \times n$ is denoted $\mathbb{S}^n$, and the positive semidefinite (definite) cone is denoted $\mathbb{S}^n_+$ ($\mathbb{S}^n_{++}$).

The elements of vectors and matrices are denoted as follows: scalar x_i stands for the i-th element of vector $\mathbf{x}$, vector $\mathbf{x}_{\mathcal{I}}$ stands for a vector constructed by eliminating all the elements of $\mathbf{x}$ but the x_i's with $i \in \mathcal{I}$, vector $\mathbf{X}_{:,i}$ stands for the i-th column of matrix $\mathbf{X}$, vector $\mathbf{X}_{i,:}$ stands for the i-th row of $\mathbf{X}$, and scalar X_{ij} stands for the ij-th entry of $\mathbf{X}$.

Superscripts $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$, and $(\cdot)^\dagger$ denote the complex conjugate, transpose, conjugate transpose, inverse, and Moore-Penrose pseudoinverse, respectively. The trace and determinant of a matrix $\mathbf{X}$ are denoted $\mathrm{Tr}(\mathbf{X})$ and $\det(\mathbf{X})$, respectively. Vector $\mathrm{vec}(\mathbf{X})$ is constructed by stacking the columns of $\mathbf{X}$. The diagonal matrix $\mathrm{diag}(\mathbf{x})$ is constructed by setting its i-th diagonal element to be x_i. Notation $\mathbf{A} \succeq (\succ) \mathbf{B}$ stands for matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite (definite). The Hadamard product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted $\mathbf{x} \odot \mathbf{y}$. Whenever arithmetic operators such as $\sqrt{}$, $/$, and $^{-1}$ are applied to vectors we mean an element-wise operation.

The magnitude of a scalar x is denoted $|x|$. The ℓ_p-norm of a vector $\mathbf{x}$ is denoted $\|\mathbf{x}\|_p$. The nuclear norm and Frobenius norm for a matrix $\mathbf{X}$ are denoted $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_F$, respectively.

[1]The convergence column indicates the type of convergence discussed in each problem. It should *not* be interpreted as the whole sequence generated by the algorithm converges to the corresponding point, which is a strong conclusion.

Operator $[\cdot]_+ : \mathbb{R}^n \to \mathbb{R}^n_+$ denotes the Euclidean projection of a vector in $\mathbb{R}^n$ to $\mathbb{R}^n_+$. The gradient of a function f is denoted ∇f. The composition of functions f and g is denoted $f \circ g$. The sign function is denoted sgn. The expected value of a random vector $\mathbf{x}$ is denoted $\mathbb{E}(\mathbf{x})$, and its covariance matrix is denoted $\mathrm{Cov}(\mathbf{x})$. Unless otherwise specified, subscript $(\cdot)_t$ in $\mathbf{x}_t$ is reserved for the algorithm iteration that stands for the value of $\mathbf{x}$ at the t-th iteration, and x_i^t stands for the value of the i-th element of $\mathbf{x}_t$, i.e., $(x_i)_t$, for notation simplicity (the same convention applies to vector $\mathbf{x}_i$ and matrix $\mathbf{X}_i$).

II. Algorithm Framework

A. The MM Algorithm

Consider the following optimization problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X}, \end{array} \tag{1}$$

where $\mathcal{X}$ is a nonempty closed set in $\mathbb{R}^n$ and $f : \mathcal{X} \to \mathbb{R}$ is a continuous function. We assume that $f(\mathbf{x})$ goes to infinity when $\mathbf{x} \in \mathcal{X}$ and $\|\mathbf{x}\| \to +\infty$.

Initialized as $\mathbf{x}_0 \in \mathcal{X}$, MM generates a sequence of feasible points $(\mathbf{x}_t)_{t\in\mathbb{N}}$ by the following induction. At point $\mathbf{x}_t$, in the majorization step we construct a continuous surrogate function $g(\cdot|\mathbf{x}_t) : \mathcal{X} \to \mathbb{R}$ satisfying the upperbound property that

$$g(\mathbf{x}|\mathbf{x}_t) \geq f(\mathbf{x}) + c_t, \, \forall \mathbf{x} \in \mathcal{X}, \tag{2}$$

where $c_t = g(\mathbf{x}_t|\mathbf{x}_t) - f(\mathbf{x}_t)$. That is, the difference of $g(\cdot|\mathbf{x}_t)$ and f is minimized at $\mathbf{x}_t$.

Then in the minimization step, we update $\mathbf{x}$ as

$$\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}|\mathbf{x}_t). \tag{3}$$

The sequence $(f(\mathbf{x}_t))_{t\in\mathbb{N}}$ is non-increasing since

$$f(\mathbf{x}_{t+1}) \leq g(\mathbf{x}_{t+1}|\mathbf{x}_t) - c_t \leq g(\mathbf{x}_t|\mathbf{x}_t) - c_t = f(\mathbf{x}_t), \tag{4}$$

where the first inequality follows from (2), and the second inequality follows from (3). We denote the algorithm mapping defined by steps (2) and (3) that sends $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ by $M : \mathbb{R}^n \to \mathbb{R}^n$ in the rest of the paper.

B. Extensions

The MM principle can be combined with other algorithm frameworks, leading to the following extensions.

Instead of computing a minimizer of $g(\cdot|\mathbf{x}_t)$, we can find a point $\mathbf{x}_{t+1}$ that satisfies $g(\mathbf{x}_{t+1}|\mathbf{x}_t) \leq g(\mathbf{x}_t|\mathbf{x}_t)$ (i.e., just making an improvement). This leads to the generalized EM (GEM) algorithm [4]. Point $\mathbf{x}_{t+1}$ can be found by taking a gradient, Newton, or quasi-Newton step. GEM is also closely related to MM acceleration schemes [64]–[66].

Combining with the block coordinate descent algorithm, we can partition the variables into blocks and apply MM to one block while keeping the value of the other blocks fixed. As a benefit, it provides more flexibility in designing surrogate functions. Moreover, in some cases the surrogate function can approximate f better than using a single block, leading to a faster convergence rate [67]. A simple update rule is sweeping the blocks cyclically. It can be generalized to the "essential

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

Construction techniques	Applications	Type of Convergence
First order Taylor expansion	reweighted ℓ_1-norm minimization [15]	stationary point
	robust covariance estimation [22]–[24], [33]	global minimum
	variance component model [34], [35]	objective value
	optimization with projection forms [36]	stationary point
	maximization of a convex objective over a compact set [37]	first order optimal
	sparse eigenvector problem with ℓ_0-norm constraint [19]/penalty [21], [37]	objective value
	edge-preserving regularization in image processing [7], [8], [38]	stationary point (nonconvex objective) global minimum (convex objective)
	ℓ_p-norm minimization [21], [39]–[43]	same as above
Second order Taylor expansion	(sparse) logistic regression [44], [45]	global minimum
	rank constrained matrix quadratic form minimization [46]	objective value
	quartic form minimization [32], [47]	objective value
	sparse linear regression [14], [17], [18], [29], [48], [49]	ℓ_1-norm: global minimum
		concave regularization: stationary point
		ℓ_0-norm: local minimum
	nonnegative least squares [50]–[52]	global minimum
Convexity inequality	robust covariance estimation [24]	global minimum
Special inequalities	signomial proramming, complementary GP [53], [54] (arithmetic-geometric mean inequality)	stationary point
	nonnegative least squares [55] (arithmetic-geometric mean inequality plus first order Taylor expansion)	global minimum
	phase retrieval [47], [56], [57] (Cauchy-Schwartz inequality)	stationary point
	sensor network localization [58]–[61] (Cauchy-Schwartz inequality)	stationary point
	variance component model [25], [62], [63] (Schur complement/convexity inequality)	stationary point

Table I
SUMMARY OF APPLICATIONS IN SECTION V AND THEIR CORRESPONDING SURROGATE FUNCTION CONSTRUCTION TECHNIQUES.

cyclic rule" [68], where each block is updated at least once within a finite number of iterations [57], [69], [70]. Other sweeping schemes include the Gauss-Southwell update rule, maximum improvement update rule, as well as the randomized update rule [70].

An incremental MM was proposed in [71] for minimizing an objective function of the form $f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x})$, which is related to stochastic optimization with f being the empirical average. The algorithm assumes only one of the f_i's is observed at each iteration, and the surrogate function is updated based on the current f_i and the algorithm history recursively.

In [69], the global upperbound requirement of the surrogate function has been relaxed to just being a local upperbound.

In this paper, we restrict our scope to the standard MM with a single block of variables[2]. For a comprehensive analysis of the above-described extensions, we refer the reader to [69], [70], [72], and [73].

C. Convergence of MM

We assume in preliminary that the MM conditions (2) and (3) hold, and $\mathcal{X}$ is convex throughout this subsection. The convexity of $\mathcal{X}$ and continuity of f are minimum assumptions for a unified study of algorithm convergence. In some applications, MM is derived for a problem with a discontinuous objective function or a non-convex constraint set, see [17], [19] for examples. The convergence of these algorithms deserves a case by case study.

In Eq. (4), we have shown that the objective value is non-increasing and converges to a limit $f^\star$ by the assumption that f is bounded below. The next step is to establish the conditions that guarantee $f^\star$ being a stationary value and also the convergence of the sequence $(\mathbf{x}_t)_{t\in\mathbb{N}}$.

1) Unconstrained Optimization: We make the following assumptions on f and g:

(A1) The sublevel set $\mathrm{lev}_{\leq f(\mathbf{x}_0)} f := \{\mathbf{x}\in\mathcal{X} | f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ is compact given that $f(\mathbf{x}_0) < +\infty$;

(A2.1) $f(\mathbf{x})$ and $g(\mathbf{x}|\mathbf{x}_t)$ are continuously differentiable with respect to $\mathbf{x}$;

(A3.1) $g(\mathbf{x}|\mathbf{x}_t)$ is continuous in $\mathbf{x}$ and $\mathbf{x}_t$.

For unconstrained problem (1), the set of stationary points of f is defined as

$$\mathcal{X}^\star = \{\mathbf{x}|\nabla f(\mathbf{x}) = \mathbf{0}\}. \tag{5}$$

Under Assumptions **(A1)**, **(A2.1)**, **(A3.1)**, the following statements hold [74], [75]:

(C1) Any limit point $\mathbf{x}_\infty$ of $(\mathbf{x}_t)_{t\in\mathbb{N}}$ is a stationary point of f;

(C2) $f(\mathbf{x}_t) \downarrow f^\star$ monotonically and $f^\star = f(\mathbf{x}^\star)$ with $\mathbf{x}^\star \in \mathcal{X}^\star$;

(C3) If $f(M(\mathbf{x})) = f(\mathbf{x})$, then $\mathbf{x} \in \mathcal{X}^\star$ and $\mathbf{x} \in \arg\min g(\cdot|\mathbf{x})$;

(C4) If $\mathbf{x}$ is a fixed point of M, then $\mathbf{x}$ is a convergent point of MM and belongs to $\mathcal{X}^\star$.

To establish to convergence of sequence $(\mathbf{x}_t)_{t\in\mathbb{N}}$ to a stationary point, we further require one of the following assumptions:

(A4.1) Set $\mathcal{X}^\star$ is a singleton;

(A4.2) Set $\mathcal{X}^\star$ is discrete and $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| \to 0$;

(A4.3) Set $\mathcal{X}^\star$ is discrete, and $g(\cdot|\mathbf{x})$ has a unique global minimum for all $\mathbf{x} \in \mathcal{X}^\star$.

2) Constrained Optimization with Smooth Objective Function: With $\mathcal{X}$ convex and f continuously differentiable, the set of stationary points is defined as

$$\mathcal{X}^\star = \left\{\mathbf{x}|\nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \geq 0,\ \forall \mathbf{y}\in\mathcal{X}\right\}. \tag{6}$$

[2]There are a few applications in this paper where MM is applied with block alternation. For presentation clarity we only describe the update of one block while treating the other blocks of variables as fixed parameters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

Conclusions **(C1)**-**(C4)** still hold under Assumptions **(A1)**, **(A2.1)** and **(A3.1)** [69]. Moreover, Assumption **(A3.1)** can be replaced by **(A3.2)** stated next.

(A3.2) For all $\mathbf{x}_t$ generated by the algorithm, there exists $\gamma \geq 0$ such that $\forall \mathbf{x} \in \mathcal{X}$, we have

$$\left(\nabla g\left(\mathbf{x}|\mathbf{x}_t\right) - \nabla g\left(\mathbf{x}_t|\mathbf{x}_t\right)\right)^T \left(\mathbf{x} - \mathbf{x}_t\right) \leq \gamma \left\|\mathbf{x} - \mathbf{x}_t\right\|^2.$$

Assumption **(A3.2)** is equivalent to stating that $g\left(\mathbf{x}|\mathbf{x}_t\right)$ can be uniformly upperbounded by a quadratic function with the Hessian matrix being $\gamma\mathbf{I}$, which is easier to verify than **(A3.1)** when $g\left(\cdot|\mathbf{x}_t\right)$ has a complicated form[3].

Convergence of sequence $\left(\mathbf{x}_t\right)_{t\in\mathbb{N}}$ to a stationary point can be proved by further requiring **(A4.1)** or **(A4.2)**.

3) Constrained Optimization with Non-Smooth Objective Function: Finally, we address the case that f and $g\left(\cdot|\mathbf{x}\right)$ are nonsmooth, but their directional derivatives exist for all feasible directions [70]. The set of stationary points is defined as

$$\mathcal{X}^\star = \left\{\mathbf{x}|f'\left(\mathbf{x};\mathbf{d}\right) \geq 0, \forall \mathbf{x}+\mathbf{d} \in \mathcal{X}\right\}, \tag{7}$$

where

$$f'\left(\mathbf{x}_t;\mathbf{d}\right) \coloneqq \liminf_{\lambda\downarrow 0} \frac{f\left(\mathbf{x}_t+\lambda\mathbf{d}\right) - f\left(\mathbf{x}_t\right)}{\lambda} \tag{8}$$

is the directional derivative of f at $\mathbf{x}_t$ in direction $\mathbf{d}$. Accordingly, the gradient consistency assumption **(A2.1)** is modified as follows:

(A2.2) $f'\left(\mathbf{x}_t;\mathbf{d}\right) = g'\left(\mathbf{x}_t;\mathbf{d}|\mathbf{x}_t\right),\ \forall \mathbf{x}_t+\mathbf{d} \in \mathcal{X}$.

Under Assumptions **(A1)**, **(A2.2)**, **(A3.1)**, the sequence $\left(\mathbf{x}_t\right)_{t\in\mathbb{N}}$ converges to $\mathcal{X}^\star$, i.e.,

$$\lim_{t\to+\infty} \inf_{\mathbf{x}\in\mathcal{X}^\star} \left\|\mathbf{x}_t - \mathbf{x}\right\|_2 = 0.$$

D. Acceleration Schemes

A drawback of MM is that it can suffer from a slow convergence speed [3], [13], mainly because of the restrictive upperbound condition. To alleviate this shortcoming, MM accelerators are often employed. Various types of accelerators have been proposed in the literature, including those derived based on the multivariate Aitken's method [76], conjugate gradient acceleration [64], Newton and quasi-Newton type acceleration [66], [77], [78], and over-relaxation [79]–[82], see [83, Chap. 4] for an overview in the context of EM.

We begin with the idea of line search type algorithms. To minimize a function f, at the current point $\mathbf{x}_t$ one first determines a descent direction $\mathbf{d}_t$, then a step-size α_t that decreases the objective function. MM can be interpreted in this way by identifying $\mathbf{d}_t \coloneqq M\left(\mathbf{x}_t\right) - \mathbf{x}_t$ and $\alpha_t \coloneqq 1$.

The line search type accelerators modify the value of α_t to achieve a larger decrement of the objective value. For instance, in [84] α_t was determined by the two previous steps based on Aitken's method. This method may, however, destroy the monotonicity of the algorithm. A constant step-size $\alpha_t \equiv \alpha$ was adopted in over-relaxation methods [79]–[82], and the optimal α was provided in [82]. Nevertheless, it is also pointed out that computing the optimal α is generally a difficult problem. To address these issues, α_t was suggested to be computed using line search so that $f\left(\mathbf{x}_{t+1}\right) \leq f\left(\mathbf{x}_t\right)$ is guaranteed [38], [82], [85].

Another class of accelerators also modifies the descent direction $\mathbf{d}_t$. To ensure the objective value is nonincreasing, $\mathbf{x}_{t+1}$ needs not be a global minimizer of $g\left(\cdot|\mathbf{x}_t\right)$. Instead, one can solve (3) inexactly by taking a Newton step. This leads to the EM gradient algorithm [65]. A quasi-Newton accelerator proposed in [66] improves it by adding an approximate of the Hessian of $H\left(\mathbf{x}|\mathbf{x}_t\right) \triangleq f\left(\mathbf{x}\right) - g\left(\mathbf{x}|\mathbf{x}_t\right)$ to $\nabla g^2\left(\mathbf{x}|\mathbf{x}_t\right)$ in the Newton step (assuming both $\nabla^2 H\left(\mathbf{x}|\mathbf{x}_t\right)$ and $\nabla g^2\left(\mathbf{x}|\mathbf{x}_t\right)$ exist). In [64], the generalized gradient algorithm was applied to minimize f by treating $M\left(\mathbf{x}_t\right) - \mathbf{x}_t$ as the generalized gradient. See [85] for an overview and comparison of the above-mentioned accelerators.

Finally, we introduce a class of accelerators based on the idea of finding a fixed point of M, which is a stationary point of f if Assumptions **(A1)**, **(A2)**, and **(A3.1)** hold. Assuming that M is continuously differentiable, it is known that Newton's method enjoys a quadratic convergence rate in the vicinity of a fixed point. Define $F\left(\mathbf{x}\right) = M\left(\mathbf{x}\right) - \mathbf{x}$, a Newton step update of finding a zero of F is given by[4]

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_t - \left(\nabla F\left(\tilde{\mathbf{x}}_t\right)\right)^{-1} F\left(\tilde{\mathbf{x}}_t\right),$$

where ∇F is the Jacobian of F. While $F\left(\tilde{\mathbf{x}}_t\right)$ can be evaluated by the MM step, the Jacobian $\nabla F\left(\tilde{\mathbf{x}}_t\right)$ is hard to obtain in general (unless $M\left(\mathbf{x}\right)$ has an explicit form) and is often approximated based on the previous iterates $\left(\tilde{\mathbf{x}}_{t'}\right)_{0\leq t'\leq t}$. The STEM accelerator proposed in [86] approximates $\nabla F\left(\tilde{\mathbf{x}}_t\right)$ by a scaled identity matrix. The Aitken [76] and SQUAREM accelerators [86] approximate $\nabla F\left(\tilde{\mathbf{x}}_t\right)$ using the secant method. More recently, an accelerator was proposed in [87] that approximates $\nabla F\left(\tilde{\mathbf{x}}_t\right)$ based on the quasi-Newton method.

We point out that Newton type algorithms converge only in the vicinity of a stationary point, therefore accelerators based on Newton's iteration are often executed after a few MM steps so that $\mathbf{x}_t$ falls into the convergence region. It is also worth mentioning that the MM acceleration schemes are developed for unconstrained optimization problems (except the cases where the constraint can be eliminated by reparameterization). For a constrained optimization problem, it is generally not true that the point returned by accelerators will be feasible. In this case, heuristic manipulations such as projection to the feasible set can be employed.

III. Surrogate Function Construction

The key step of applying MM is constructing a surrogate function. While there is no concrete step to follow, some commonly adopted rules that can provide guidance exist. In this section, techniques to find surrogate functions, along with a number of illustrating examples, will be presented. The inequalities provided here will serve as building blocks in finding surrogate functions for more sophisticated objective functions in Section V.

[3]Since the continuously differentiability of f and $g\left(\cdot|\mathbf{x}_t\right)$ and the upperbound condition of MM (Eq. (2)) implies the directional derivative of f and $g\left(\cdot|\mathbf{x}_t\right)$ are equal along all feasible directions (Proposition 1, [70]), the first order consistency condition (R2) in [69] holds automatically.

[4]The sequence $\left(\tilde{\mathbf{x}}_t\right)_{t\in\mathbb{N}}$ should be distinguished from the MM sequence $\left(\mathbf{x}_t\right)_{t\in\mathbb{N}}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

A. First Order Taylor Expansion

Suppose f can be decomposed as

$$f(\mathbf{x}) = f_0(\mathbf{x}) + f_{\text{ccv}}(\mathbf{x}), \tag{9}$$

where f_{ccv} is a differentiable concave function.

Linearizing f_{ccv} at $\mathbf{x} = \mathbf{x}_t$ yields the following inequality:

$$f_{\text{ccv}}(\mathbf{x}) \leq f_{\text{ccv}}(\mathbf{x}_t) + \nabla f_{\text{ccv}}(\mathbf{x}_t)^T (\mathbf{x} - \mathbf{x}_t), \tag{10}$$

thus f can be upperbounded as

$$f(\mathbf{x}) \leq f_0(\mathbf{x}) + \nabla f_{\text{ccv}}(\mathbf{x}_t)^T \mathbf{x} + \text{const.}$$

Example 1. Function $\log(x)$ can be upperbounded as

$$\log(x) \leq \log(x_t) + \frac{1}{x_t}(x - x_t) \tag{11}$$

with equality achieved at $x = x_t$.

Example 2. Function $\log\det(\mathbf{\Sigma})$ can be upperbounded as

$$\log\det(\mathbf{\Sigma}) \leq \log\det(\mathbf{\Sigma}_t) + \text{Tr}\left(\mathbf{\Sigma}_t^{-1}(\mathbf{\Sigma} - \mathbf{\Sigma}_t)\right) \tag{12}$$

with equality achieved at $\mathbf{\Sigma} = \mathbf{\Sigma}_t$.

Example 3. Function $\text{Tr}\left(\mathbf{S}\mathbf{X}^{-1}\right)$ with both $\mathbf{S}$ and $\mathbf{X}$ in $\mathbb{S}_{++}$ can be lowerbounded as

$$\text{Tr}\left(\mathbf{S}\mathbf{X}^{-1}\right) \geq \text{Tr}\left(\mathbf{S}\mathbf{X}_t^{-1}\right) - \text{Tr}\left(\mathbf{X}_t^{-1}\mathbf{S}\mathbf{X}_t^{-1}(\mathbf{X} - \mathbf{X}_t)\right) \tag{13}$$

with equality achieved at $\mathbf{X} = \mathbf{X}_t$.

Example 4 [88]**.** Function $\text{Tr}\left(\mathbf{X}^T\mathbf{Y}^{-1}\mathbf{X}\right)$ with $\mathbf{Y} \in \mathbb{S}_{++}$ can be lowerbounded as

$$\begin{aligned}&\text{Tr}\left(\mathbf{X}^T\mathbf{Y}^{-1}\mathbf{X}\right)\\ \geq& 2\text{Tr}\left(\mathbf{X}_t^T\mathbf{Y}_t^{-1}\mathbf{X}\right) - \text{Tr}\left(\mathbf{Y}_t^{-1}\mathbf{X}_t\mathbf{X}_t^T\mathbf{Y}_t^{-1}\mathbf{Y}\right) + \text{const.}\end{aligned} \tag{14}$$

with equality achieved at $(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_t, \mathbf{Y}_t)$.

Proof: Function $\text{Tr}\left(\mathbf{X}^T\mathbf{Y}^{-1}\mathbf{X}\right)$ is jointly convex in $\mathbf{X}$ and $\mathbf{Y}$, therefore lowerbounded by its linear expansion around $(\mathbf{X}_t, \mathbf{Y}_t)$, which implies (14). ■

Remark 5. We emphasize that the upperbounds derived based on linearizing a concave function are not necessarily linear in the variables, see Eq. (51) for example.

More generally, given a convex, a linear, and a concave function, f_{cvx}, f_{lin}, and f_{ccv}, respectively, if their values and gradients are equal at some $\mathbf{x}_t$, then, for any $\mathbf{x}$,

$$f_{\text{ccv}}(\mathbf{x}) \leq f_{\text{lin}}(\mathbf{x}) \leq f_{\text{cvx}}(\mathbf{x}), \tag{15}$$

as illustrated in Figure 2.

Example 6. Function $|x|^p, 0 < p \leq 1$, which is concave on $(-\infty, 0]$ and $[0, +\infty)$, can be upperbounded as[5]

$$|x|^p \leq \frac{p}{2}|x_t|^{p-2}x^2 + \text{const.}, \tag{16}$$

providing that $x_t \neq 0$.

Inequality (16) plays an important role in iteratively reweighted least squares (IRLS) algorithms, where a quadratic upperbound is preferred to a tighter linear one in the majorization step, with the benefit that the minimization step admits a solution that is easy to compute.

[5]The result also holds for $1 < p \leq 2$ although $|x|^p$ is convex.

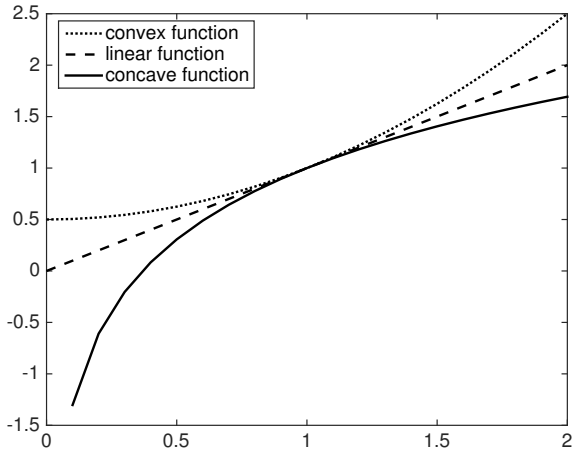

Figure 2. Surrogate function construction technique by first order Taylor expansion: a concave function can upperbound a linear function, which can be upperbounded by a convex function.

In the last example, we show that inequality (15) can be used to construct lowerbounds for maximization problems.

Example 7 [54]**.** A monomial $\prod_{i=1}^n x_i^{\alpha_i}$, where $x_i \geq 0$, $\forall i$, can be lowerbounded as

$$\prod_{i=1}^n x_i^{\alpha_i} \geq \prod_{i=1}^n \left(x_i^t\right)^{\alpha_i} \left(1 + \sum_{i=1}^n \alpha_i \log x_i - \sum_{i=1}^n \alpha_i \log x_i^t\right) \tag{17}$$

with equality achieved at $x_i = x_i^t$.

Proof: Inequality (11) implies that

$$\begin{aligned}\log\left(\prod_{i=1}^n x_i^{\alpha_i}\right) \leq& \log\left(\prod_{i=1}^n \left(x_i^t\right)^{\alpha_i}\right)\\ &+ \left(\prod_{i=1}^n \left(x_i^t\right)^{\alpha_i}\right)^{-1}\left(\prod_{i=1}^n x_i^{\alpha_i} - \prod_{i=1}^n \left(x_i^t\right)^{\alpha_i}\right).\end{aligned}$$

Rearranging the terms we have (17). ■

The surrogate function is separable in the variables, which can be optimized in parallel if the constraints are also separable.

B. Convexity Inequality

For a convex function f_{cvx}, we have the following inequality:

$$f_{\text{cvx}}\left(\sum_{i=1}^n w_i \mathbf{x}_i\right) \leq \sum_{i=1}^n w_i f_{\text{cvx}}(\mathbf{x}_i), \tag{18}$$

where $\sum_{i=1}^n w_i = 1$, $w_i \geq 0. \forall i = 1, \ldots, n$. Equality can achieved if the $\mathbf{x}_i$'s are equal, or for different $\mathbf{x}_i$'s if f_{cvx} is not strictly convex.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

Example 8 (Jensen's Inequality)**.** Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function and $\mathbf{x}$ be a random variable that take values in $\mathcal{X}$. Assuming that $\mathbb{E}(\mathbf{x})$ and $\mathbb{E}(f(\mathbf{x}))$ are finite, then

$$\mathbb{E}(f(\mathbf{x})) \geq f(\mathbb{E}(\mathbf{x})).$$

With Jensen's inequality we can show that EM is a special case of MM (cf. Section IV-A).

Particularizing (18) for the concave function $\log$, we have the following inequality.

Example 9. Function $\sum_{i=1}^{n} \alpha_i \log f_i(x)$ with $\alpha_i > 0$ can be upperbounded as

$$\sum_{i=1}^{n} \alpha_i \log f_i(x) \leq \sum_{i=1}^{n} \alpha_i \log f_i(x_t) + \left(\sum_{i=1}^{n} \alpha_i\right) \log\left(\frac{\sum_{i=1}^{n} \alpha_i \frac{f_i(x)}{f_i(x_t)}}{\sum_{i=1}^{n} \alpha_i}\right), \tag{19}$$

where $f_i(x) > 0$, $\forall i$. Equality is achieved at $x = x_t$.

Inequality (19) creates a concave upperbound for $\sum_{i=1}^{n} \alpha_i \log f_i(x)$ by merging the summation inside the $\log$ function. Recall that by applying inequality (11) we can obtain an alternative upperbound that is linear in the $f_i(x)$'s as

$$\sum_{i=1}^{n} \alpha_i \log f_i(x) \leq \sum_{i=1}^{n} \alpha_i \left(\log f_i(x_t) + \frac{1}{f_i(x_t)}\left(f_i(x) - f_i(x_t)\right)\right). \tag{20}$$

However, the concave upperbound (19) is tighter, thus is preferred to (20) for a faster convergence rate, see Figure 3 as an illustration.

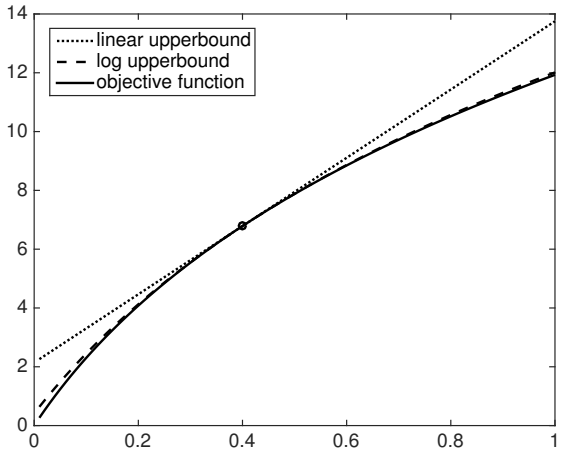

Figure 3. Objective function: $f(x) = 3\log(1+x) + 5\log(1+3x) + 1.5\log(1+6x)$; log upperbound: upperbound given by (19); linear upperbound: upperbound given by (20).

Particularizing inequality (18) for $1/x$ we have the following bound.

Example 10. The function $\frac{1}{\sum_{i=1}^{n} a_i x_i}$ with $a_i > 0$ and $x_i > 0$ can be upperbounded as

$$\frac{1}{\sum_{i=1}^{n} a_i x_i} \leq \frac{\sum_{i=1}^{n} a_i \left(x_i^t\right)^2 x_i^{-1}}{\left(\sum_{i=1}^{n} a_i x_i^t\right)^2} \tag{21}$$

with equality achieved at $x_i = x_i^t$, $\forall i = 1, \ldots, n$.

Generalizing (21) to a convex function f yields the following inequality.

Example 11 [13]**.** The convex function $f\left(\mathbf{a}^T\mathbf{x}\right)$ can be upperbounded as

$$f\left(\mathbf{a}^T\mathbf{x}\right) \leq \sum_{i=1}^{n} \alpha_i f\left(\frac{a_i}{\alpha_i}\left(x_i - x_i^t\right) + \mathbf{a}^T\mathbf{x}_t\right), \tag{22}$$

where $\alpha_i > 0$, $\sum_{i=1}^{n} \alpha_i = 1$. Moreover, if the elements of $\mathbf{a}$ and $\mathbf{x}_t$ are positive, letting $\alpha_i = \frac{a_i x_i^t}{\mathbf{a}^T\mathbf{x}_t}$ yields a different upperbound as

$$f\left(\mathbf{a}^T\mathbf{x}\right) \leq \sum_{i=1}^{n} \frac{a_i x_i^t}{\mathbf{a}^T\mathbf{x}_t} f\left(\frac{\mathbf{a}^T\mathbf{x}_t}{x_i^t} x_i\right). \tag{23}$$

Inequalities (22) and (23) were proposed and applied in medical imaging in [6], [9].

C. Construction by Second Order Taylor Expansion

Lemma 12 (Descent Lemma [89])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with a Lipschitz continuous gradient and Lipschitz constant L (we say that ∇f is L-Lipschitz henceforth). Then, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{24}$$

More generally, if f has bounded curvature, i.e., there exists a matrix $\mathbf{M}$ such that $\mathbf{M} \succeq \nabla^2 f(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{X}$, then the following inequality implied by Taylor's theorem [88] holds:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^T \mathbf{M}(\mathbf{x} - \mathbf{y}). \tag{25}$$

Particularizing (25) for $f(\mathbf{x}) = \mathbf{x}^H\mathbf{L}\mathbf{x}$ gives the following inequality[6].

Example 13. The quadratic form $\mathbf{x}^H\mathbf{L}\mathbf{x}$, where $\mathbf{L}$ is a Hermitian matrix, can be upperbounded as

$$\mathbf{x}^H\mathbf{L}\mathbf{x} \leq \mathbf{x}^H\mathbf{M}\mathbf{x} + 2\mathrm{Re}\left(\mathbf{x}^H(\mathbf{L} - \mathbf{M})\mathbf{x}_t\right) + \mathbf{x}_t^H(\mathbf{M} - \mathbf{L})\mathbf{x}_t, \tag{26}$$

where $\mathbf{M} \succeq \mathbf{L}$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$.

Example 13 shows that using (26) we can replace $\mathbf{L}$ by $\mathbf{M}$ with some desired structures, such as being a diagonal matrix, so that the surrogate function is separable.

[6]Wirtinger calculus is applied for complex-valued matrix differentials [90].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

D. Arithmetic-Geometric Mean Inequality

The arithmetic-geometric mean inequality states that [88]

$$\prod_{i=1}^{n} z_i^{\alpha_i} \leq \sum_{i=1}^{n} \frac{\alpha_i}{\|\boldsymbol{\alpha}\|_1} z_i^{\|\boldsymbol{\alpha}\|_1}, \tag{27}$$

where z_i and α_i are nonnegative scalars. Equality is achieved when the z_i's are equal.

Letting $z_i = x_i/x_i^t$ for $\alpha_i > 0$ and $z_i = x_i^t/x_i$ for $\alpha_i < 0$ we have the following inequality.

Example 14 [54]**.** A monomial $\prod_{i=1}^{n} x_i^{\alpha_i}$ can be upperbounded as

$$\prod_{i=1}^{n} x_i^{\alpha_i} \leq \left(\prod_{i=1}^{n} \left(x_i^t\right)^{\alpha_i}\right) \sum_{i=1}^{n} \frac{|\alpha_i|}{\|\boldsymbol{\alpha}\|_1} \left(\frac{x_i}{x_i^t}\right)^{\|\boldsymbol{\alpha}\|_1 \mathrm{sgn}(\alpha_i)}. \tag{28}$$

Equality is achieved at $x_i = x_i^t$, $\forall i = 1, \ldots, n$.

Upperbound (28) and lowerbound (17) serve as the basic ingredients for deriving MM algorithms for signomial programming [54].

Example 15 [53]**.** A posynomial $\sum_{i=1}^{n} u_i(\mathbf{x})$, where $u_i(\mathbf{x})$ is a monomial, can be lower bounded as

$$\sum_{i=1}^{n} u_i(\mathbf{x}) \geq \prod_{i=1}^{n} \left(\frac{u_i(\mathbf{x})}{\alpha_i}\right)^{\alpha_i}, \tag{29}$$

where $\alpha_i = \frac{u_i(\mathbf{x}_t)}{\prod_{i=1}^{n} u_i(\mathbf{x}_t)}$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$.

Inequality (29) can be used in solving complementary geometric programming (GP) with the objective function being the ratio of posynomials.

Example 16. The ℓ_2-norm $\|\mathbf{x}\|_2$ can be upperbounded as

$$\|\mathbf{x}\|_2 \leq \frac{1}{2}\left(\|\mathbf{x}_t\|_2 + \|\mathbf{x}\|_2^2 / \|\mathbf{x}_t\|_2\right), \tag{30}$$

given that $\|\mathbf{x}_t\|_2 \neq 0$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$.

E. Cauchy-Schwartz Inequality

Cauchy-Schwartz inequality states that

$$\mathbf{x}^T\mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Equality is achieved when $\mathbf{x}$ and $\mathbf{y}$ are collinear.

Example 17. Function $\left|\mathbf{a}^H\mathbf{x}\right|$ can be lowerbounded as

$$\left|\mathbf{a}^H\mathbf{x}\right| \geq \mathrm{Re}\left(\mathbf{x}_t^H\mathbf{a}\mathbf{a}^H\mathbf{x}\right) / \left|\mathbf{a}^H\mathbf{x}_t\right|, \tag{31}$$

given that $\left|\mathbf{a}^H\mathbf{x}_t\right| \neq 0$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$.

Proof: For two complex numbers $z_1 = u_1 + iv_1$ and $z_2 = u_2 + iv_2$, we have

$$\begin{aligned}\mathrm{Re}(z_1 z_2^*) &= u_1u_2 + v_1v_2 \\ &\leq \sqrt{u_1^2 + v_1^2} \cdot \sqrt{u_2^2 + v_2^2}\end{aligned}$$

by Cauchy-Schwartz inequality. Letting $z_1 = \mathbf{a}^H\mathbf{x}$ and $z_2 = \mathbf{a}^H\mathbf{x}_t$ yields the desired inequality. ■

Example 18. The ℓ_2-norm $\|\mathbf{x}\|_2$ can be lowerbounded as

$$\|\mathbf{x}\|_2 \geq \mathbf{x}^T\mathbf{x}_t / \|\mathbf{x}_t\|_2, \tag{32}$$

given that $\|\mathbf{x}_t\|_2 \neq 0$. Equality is achieved at $\mathbf{x} = \mathbf{x}_t$.

Together with (32), they provide a quadratic upperbound and a linear lowerbound for the ℓ_2-norm on the whole space except the origin.

F. Schur Complement

The Schur complement condition for $\mathbf{C} \succ \mathbf{0}$ states that

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \succeq \mathbf{0}$$

if and only if the Schur complement of $\mathbf{C}$ is in $\mathbb{S}_+$. That is,

$$\mathbf{S} \coloneqq \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T \succeq \mathbf{0}. \tag{33}$$

Inequality (33) provides a way to upperbound the inverse of a matrix.

Example 19 [25]**.** Assuming $\mathbf{P} \succ \mathbf{0}$, the matrix $\left(\mathbf{A}\mathbf{P}\mathbf{A}^H\right)^{-1}$ can be upperbounded as

$$\mathbf{R}_t^{-1}\mathbf{A}\mathbf{P}_t\mathbf{P}^{-1}\mathbf{P}_t\mathbf{A}^H\mathbf{R}_t^{-1} \succeq \left(\mathbf{A}\mathbf{P}\mathbf{A}^H\right)^{-1}, \tag{34}$$

where $\mathbf{R}_t = \mathbf{A}\mathbf{P}_t\mathbf{A}^H$. Equality is achieved at $\mathbf{P} = \mathbf{P}_t$.

Inequality (34) can also be derived based on convexity [63].

Particularizing (34) for $\mathbf{P} = \mathrm{diag}(p_1, \ldots, p_n)$ and $\mathbf{A} = \left[\sqrt{a_1}, \ldots, \sqrt{a_n}\right]$ gives a different derivation for inequality (21).

G. Generalization

With the inequalities provided above we can construct surrogate functions for more complicated objective functions by majorizing f more than once. Specifically, one can find a sequence of functions $g^{(1)}(\cdot|\mathbf{x}_t), \ldots, g^{(k)}(\cdot|\mathbf{x}_t)$ satisfying

$$\begin{aligned} &g^{(i)}(\mathbf{x}_t|\mathbf{x}_t) = g^{(i+1)}(\mathbf{x}_t|\mathbf{x}_t) \\ &g^{(i)}(\mathbf{x}|\mathbf{x}_t) \leq g^{(i+1)}(\mathbf{x}|\mathbf{x}_t), \ \forall \mathbf{x} \in \mathcal{X}, \forall i = 1, \ldots, k-1. \end{aligned}$$

Function $g^{(i)}(\cdot|\mathbf{x}_t)$ usually gets a simpler structure gradually until its minimizer is easy to compute, as illustrated by the applications in Section V.

IV. Connection to Other Algorithm Frameworks

A. The EM Algorithm

Introduced in [4], EM is often employed to derive an iterative scheme for ML estimation problems with latent variables. To be precise, denote the observed variable by $\mathbf{x}$ and the latent variable by $\mathbf{z}$, the maximum likelihood estimator (MLE) of parameter $\boldsymbol{\theta}$ is defined as the maximizer of the log-likelihood function

$$L(\boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta}) = \log \mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}} p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}).$$

In the E-step of EM, we compute

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t} \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}),$$

where $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}_t)$ is the posterior distribution of $\mathbf{z}$ given the current estimate $\boldsymbol{\theta}_t$, and $g(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ is the expected log-likelihood

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

of the complete data set. Then in the M-step, the new estimate $\boldsymbol{\theta}_{t+1}$ is defined as

$$\boldsymbol{\theta}_{t+1} \in \arg\max_{\boldsymbol{\theta}\in\Theta} g\left(\boldsymbol{\theta}|\boldsymbol{\theta}_t\right).$$

Applying Jensen's inequality, we have

$$\begin{aligned} L\left(\boldsymbol{\theta}\right) &= \log \mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}} p\left(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}\right) \\ &= \log \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t} \frac{p\left(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}\right) p\left(\mathbf{z}|\boldsymbol{\theta}\right)}{p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t\right)} \\ &\geq \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t} \log\left(\frac{p\left(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}\right) p\left(\mathbf{z}|\boldsymbol{\theta}\right)}{p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t\right)}\right) \\ &= g\left(\boldsymbol{\theta}|\boldsymbol{\theta}_t\right) + \text{const.}, \end{aligned}$$

which shows that $g\left(\boldsymbol{\theta}|\boldsymbol{\theta}_t\right)$ is a lower bound of $L\left(\boldsymbol{\theta}\right)$. Therefore, EM is a special case of MM [11]. Moreover, theoretical results of EM such as convergence analysis and acceleration schemes can be adapted to MM [3], [10].

We mention that EM can also be viewed as a proximal minimization algorithm by rewriting $g\left(\boldsymbol{\theta}|\boldsymbol{\theta}_t\right)$ as

$$g\left(\boldsymbol{\theta}|\boldsymbol{\theta}_t\right) = \log p\left(\mathbf{x}|\boldsymbol{\theta}\right) - \beta_t I\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}\right)$$

with the proximal term

$$I\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}\right) = \int \log \frac{p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t\right)}{p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}\right)} p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t\right) \mathrm{d}\mathbf{z}$$

being the KL-divergence between $p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t\right)$ and $p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}\right)$, and $\beta_t = 1$ [91]. Ratio $\frac{p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_t)}{p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})}$ is assumed to exist for all $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_t$. This connection suggests that one could tune the penalty parameter β_t to achieve a faster convergence rate.

In addition, EM belongs to the class of cyclic algorithms as well [92]. This can be shown by defining function

$$F\left(\tilde{p},\boldsymbol{\theta}\right) = \mathbb{E}_{\tilde{p}}\left(\log p\left(\mathbf{x},\mathbf{z}|\boldsymbol{\theta}\right)\right) - \mathbb{E}_{\tilde{p}}\left(\log \tilde{p}\left(\mathbf{z}\right)\right)$$

and noticing that the E-step gives the optimal $\tilde{p}\left(\mathbf{z}\right)$ with $\boldsymbol{\theta}$ fixed as $\boldsymbol{\theta}_t$, and the M-step gives the optimal $\boldsymbol{\theta}_{t+1}$ with $\tilde{p}\left(\mathbf{z}\right)$ fixed as $p\left(\mathbf{z}|\mathbf{x},\boldsymbol{\theta}_{t+1}\right)$. The equivalence of MM and cyclic algorithms with finite dimensional variables will be justified in the following subsection.

B. Cyclic Minimization

If there exists an augmented function $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ satisfying

$$f\left(\mathbf{x}\right) = \min_{\mathbf{y}\in\mathcal{Y}} g\left(\mathbf{x},\mathbf{y}\right),$$

then problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f\left(\mathbf{x}\right) \\ \text{subject to} \quad & \mathbf{x}\in\mathcal{X} \end{aligned} \tag{35}$$

can be equivalently reformulated as

$$\min_{\mathbf{x}\in\mathcal{X}} \min_{\mathbf{y}\in\mathcal{Y}} \quad g\left(\mathbf{x},\mathbf{y}\right). \tag{36}$$

The objective function g can be minimized by alternately minimizing it with respect to $\mathbf{x}$ and $\mathbf{y}$. That is, $(\mathbf{x},\mathbf{y})$ is updated as

$$\begin{aligned} \mathbf{y}_{t+1} &\in \arg\min_{\mathbf{y}\in\mathcal{Y}} g\left(\mathbf{x}_t,\mathbf{y}\right) \\ \mathbf{x}_{t+1} &\in \arg\min_{\mathbf{x}\in\mathcal{X}} g\left(\mathbf{x},\mathbf{y}_{t+1}\right). \end{aligned} \tag{37}$$

This method is referred to as cyclic minimization and appears in applications such as [37], [62], [93]–[96].

Here we prove that cyclic minimization and MM are equivalent. First we show that cyclic minimization belongs to MM.

Define $\mathbf{y}^\star\left(\mathbf{x}\right) \in \arg\min_{\mathbf{y}\in\mathcal{Y}} g\left(\mathbf{x},\mathbf{y}\right)$, then

$$g\left(\mathbf{x},\mathbf{y}^\star\left(\mathbf{x}\right)\right) = f\left(\mathbf{x}\right). \tag{38}$$

For any given feasible $\mathbf{x}_t \in \mathcal{X}$, we have

$$g\left(\mathbf{x},\mathbf{y}^\star\left(\mathbf{x}_t\right)\right) \geq g\left(\mathbf{x},\mathbf{y}^\star\left(\mathbf{x}\right)\right) = f\left(\mathbf{x}\right). \tag{39}$$

Eqs. (38) and (39) imply that $g\left(\mathbf{x},\mathbf{y}^\star\left(\mathbf{x}_t\right)\right)$ is a surrogate function of $f\left(\mathbf{x}\right)$, and (37) is an MM iteration with $\mathbf{x}_{t+1} \in \arg\min g\left(\mathbf{x},\mathbf{y}^\star\left(\mathbf{x}_t\right)\right)$.

Conversely, MM can be regarded as cyclic minimization as follows. The MM conditions

$$\begin{aligned} g\left(\mathbf{x}|\mathbf{x}\right) =& f\left(\mathbf{x}\right) \\ g\left(\mathbf{x}|\mathbf{y}\right) \geq& f\left(\mathbf{x}\right) \end{aligned}$$

$\forall \mathbf{x},\mathbf{y}\in\mathcal{X}$ imply that $\mathbf{x} \in \arg\min_{\mathbf{y}\in\mathcal{X}} g\left(\mathbf{x}|\mathbf{y}\right)$. Therefore the MM iteration can be rewritten as

$$\begin{aligned} \mathbf{y}_{t+1} = \mathbf{x}_t &\in \arg\min_{\mathbf{y}\in\mathcal{Y}} g\left(\mathbf{x}_t|\mathbf{y}\right) \\ \mathbf{x}_{t+1} &\in \arg\min_{\mathbf{x}\in\mathcal{X}} g\left(\mathbf{x}|\mathbf{y}_{t+1}\right), \end{aligned}$$

which can be interpreted as minimizing $g\left(\mathbf{x}|\mathbf{y}\right)$ with respect to $\mathbf{x}$ and $\mathbf{y}$ alternately.

C. DC Programming and Concave-Convex Procedure

DC programming problems take the general form

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f_0\left(\mathbf{x}\right) - h_0\left(\mathbf{x}\right) \\ \text{subject to} \quad & f_i\left(\mathbf{x}\right) - h_i\left(\mathbf{x}\right) \leq 0,\ i = 1,\ldots,m, \end{aligned} \tag{40}$$

where $f_i\left(\cdot\right)$ and $h_i\left(\cdot\right)$ for $i = 0,\ldots,m$ are convex functions [97], [98]. We assume that the f_i's and h_i's are differentiable and, without loss of generality, that they are strongly convex.

The concave-convex procedure (CCCP) [99]–[101] developed to reach a local minimum of (40) states that $\mathbf{x}_t$ can be updated by solving the following convex subproblem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & g_0\left(\mathbf{x}|\mathbf{x}_t\right) \\ \text{subject to} \quad & g_i\left(\mathbf{x}|\mathbf{x}_t\right) \leq 0,\ \forall i = 1,\ldots,m, \end{aligned}$$

where

$$g_i\left(\mathbf{x}|\mathbf{x}_t\right) = f_i\left(\mathbf{x}\right) - \left(h_i\left(\mathbf{x}_t\right) + \nabla h_i\left(\mathbf{x}_t\right)^T\left(\mathbf{x}-\mathbf{x}_t\right)\right), \tag{41}$$

for all $i = 0,\ldots,m$.

Approximation (41) satisfies the MM principle and is a tight upperbound of f_i with equality attained at $\mathbf{x} = \mathbf{x}_t$. As a result, CCCP is a special case of MM if $h_i \equiv 0$, $\forall i = 1,\ldots,m$. When there exists some $h_i \neq 0$, the constraint set $\{\mathbf{x}|g_i\left(\mathbf{x}|\mathbf{x}_t\right) \leq 0,\ \forall i = 1,\ldots,m\}$ approximates the original constraint set from inside and is tangent to it at $\mathbf{x} = \mathbf{x}_t$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

D. Proximal Minimization

The proximal minimization algorithm [102]–[104] has a cyclic minimization interpretation, thus also belongs to MM. Specifically, it minimizes $f : \mathcal{X} \to \mathbb{R}$ by introducing an auxiliary variable $\mathbf{y}$ and solving

$$\underset{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{X}}{\text{minimize}} \quad g(\mathbf{x},\mathbf{y}) = f(\mathbf{x}) + \frac{1}{2c}\|\mathbf{y}-\mathbf{x}\|_2^2 .$$

The objective function $g(\mathbf{x},\mathbf{y})$ is minimized alternately with respect to $\mathbf{x}$ and $\mathbf{y}$, leading to the iteration:

$$\begin{aligned} \mathbf{x}_{t+1} &\in \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \frac{1}{2c}\|\mathbf{x}-\mathbf{y}_t\|_2^2 , \\ \mathbf{y}_{t+1} &= \mathbf{x}_{t+1}. \end{aligned} \tag{42}$$

Algorithm (42) can be generalized as:

$$\begin{aligned} \mathbf{x}_{t+1} &= \operatorname{prox}_{\mathbf{A}(\mathbf{x}_t),f}(\mathbf{x}_t) \\ &\coloneqq \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_{\mathbf{A}(\mathbf{x}_t)}^2 , \end{aligned}$$

where $\mathbf{A}(\mathbf{x}_t) \in \mathbb{S}_{++}^n$ and $\|\mathbf{x}\|_{\mathbf{A}(\mathbf{x}_t)}^2 \coloneqq \mathbf{x}^T\mathbf{A}(\mathbf{x}_t)\mathbf{x}$.

E. Variable Metric Splitting Method for Non-Smooth Optimization

Variable metric forward-backward splitting (VMFB) can be derived based on MM for solving problems of the form

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad f(\mathbf{x}) + h(\mathbf{x}),$$

where f is a differentiable function and h is a convex non-smooth function [105]. For presentation clarity we introduce below its simplest version to illustrate the connection.

Let $(\mathbf{A}_t)_{t\in\mathbb{N}}$ be a sequence of positive definite matrices satisfying

$$\begin{aligned} g^f(\mathbf{x}|\mathbf{x}_t) &= f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^T(\mathbf{x}-\mathbf{x}_t) + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_{\mathbf{A}_t}^2 \\ &\geq f(\mathbf{x}), \end{aligned} \tag{43}$$

i.e., $g^f(\cdot|\mathbf{x}_t)$ is a quadratic function that majorizes f at $\mathbf{x}=\mathbf{x}_t$. Then we can upperbound $f+h$ by

$$\begin{aligned} &g(\mathbf{x}|\mathbf{x}_t) \\ =&f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^T(\mathbf{x}-\mathbf{x}_t) + \frac{1}{2\gamma_t}\|\mathbf{x}-\mathbf{x}_t\|_{\mathbf{A}_t}^2 + h(\mathbf{x}), \end{aligned} \tag{44}$$

where $\gamma_t \in (0,1)$, $\forall t \in \mathbb{N}$. Omitting the constant terms, the update $\mathbf{x}_{t+1}$ is given by

$$\begin{aligned} \mathbf{x}_{t+1} &= \operatorname{prox}_{\gamma_t^{-1}\mathbf{A}_t,h}\left(\mathbf{x}_t - \gamma_t\mathbf{A}_t^{-1}\nabla f(\mathbf{x}_t)\right) \\ &\coloneqq \arg\min_{\mathbf{x}\in\mathcal{X}} \frac{1}{2\gamma_t}\left\|\mathbf{x}-\left(\mathbf{x}_t-\gamma_t\mathbf{A}_t^{-1}\nabla f(\mathbf{x}_t)\right)\right\|_{\mathbf{A}_t}^2 + h(\mathbf{x}). \end{aligned} \tag{45}$$

Steps (44) and (45) can be interpreted as MM naturally.

If ∇f is L-Lipschitz continuous, then $\mathbf{A}_t$ can be set as $\mathbf{A}_t = L\mathbf{I}$ and descent lemma (24) implies condition (43) holds. In this case, VMBF reduces to the proximal gradient algorithm (see [61] and [104] for examples).

Similar to MM, VMFB can also be generalized to blockwise update [57], [106].

F. Successive Convex Approximation (SCA) Algorithms

1) Approximating the Objective Function: Consider the following problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) + h(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x}\in\mathcal{X}, \end{aligned}$$

where $f : \mathcal{X} \to \mathbb{R}$ is smooth with a Lipschitz continuous gradient, and $h : \mathcal{X} \to \mathbb{R}$ is convex possibly non-differentiable.

To arrive at a stationary point, FLEXA [107]–[109] approximates f by a strongly convex function $g(\cdot|\mathbf{x}_t)$ satisfying the property that $\nabla g(\mathbf{x}_t|\mathbf{x}_t) = \nabla f(\mathbf{x}_t)$. The subproblem to be solved is

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & g(\mathbf{x}|\mathbf{x}_t) + \frac{\tau}{2}(\mathbf{x}-\mathbf{x}_t)^T\mathbf{Q}(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t) + h(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x}\in\mathcal{X}, \end{aligned}$$

where $\mathbf{Q}(\mathbf{x}_t) \in \mathbb{S}_{++}$.

The main differences between FLEXA and MM are summarized as follows:

- Applicable problems: To ensure convergence, both FLEXA and MM require the objective function to be continuous, and the set $\mathcal{X}$ to be convex. MM has been applied to some applications with a discontinuous objective function and non-convex set to devise an algorithm with a convergent objective value. The convergence of the iterates, however, needs to be studied separately.
- Approximating function: MM requires the surrogate function to be a global upperbound, not necessarily convex. On the contrary, FLEXA relaxes the upperbound condition, but requires it to be strongly convex.

For the sake of a clearer comparison, the FLEXA algorithm presented here is a simplified version of that proposed in [107] and [108], where blockwise update and parallel computation are incorporated. Extensions of the algorithm to stochastic optimization can be found in [110].

2) Approximating Both the Objective Function and Constraint Set: Consider problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq 0,\ i=1,\ldots,m. \end{aligned}$$

Apart from f_0, we can also approximate the feasible set $\{\mathbf{x}|f_i(\mathbf{x})\leq 0,\ i=1,\ldots,m\}$ at each iteration. As proposed in the early work [111], assuming that f_i is differentiable, we can solve the following convex subproblem at the t-th iteration:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & g_0(\mathbf{x}|\mathbf{x}_t) \\ \text{subject to} \quad & g_i(\mathbf{x}|\mathbf{x}_t) \leq 0,\ i=1,\ldots,m, \end{aligned}$$

where $g_i(\cdot|\mathbf{x}_t)$, $\forall i = 0,\ldots,m$, is a convex function that satisfies

$$\begin{aligned} g_i(\mathbf{x}_t|\mathbf{x}_t) &= f_i(\mathbf{x}_t) \\ g_i(\mathbf{x}|\mathbf{x}_t) &\geq f_i(\mathbf{x}) \\ \nabla g_i(\mathbf{x}_t|\mathbf{x}_t) &= \nabla f_i(\mathbf{x}_t). \end{aligned} \tag{46}$$

The limit of any convergent sequence of $(\mathbf{x}_t)_{t\in\mathbb{N}}$ is a KKT point. In short, the subproblem is constructed by upperbounding the objective function by a convex surrogate function, and approximating the feasible set from inside by a convex set.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

The condition that $g_i\left(\cdot|\mathbf{x}_t\right)$ is a global upperbound can be relaxed to just being the first order convex approximation. In addition, it can be generalized to blockwise update with the blocks updated either sequentially or in parallel. We refer readers to [101] and [112]–[115] for the details and convergence analysis.

G. Subspace MM Algorithm

The descent nature of MM indicates that it can be employed for step-size selection. Recall that line search type nonlinear optimization algorithms with update $\mathbf{x}_{t+1}=\mathbf{x}_t+\alpha_t\mathbf{d}_t$ ($\mathbf{x}_t\in\mathbb{R}^n$) can be described as first finding a gradient-related descent direction $\mathbf{d}_t$, and then the step-size as

$$\alpha_t=\arg\min_{\alpha\geq 0} f\left(\mathbf{x}_t+\alpha\mathbf{d}_t\right). \tag{47}$$

The exact line search criterion (47) can be relaxed by only requiring that $\alpha_t\mathbf{d}_t$ generates a sufficient decrease of the objective value.

MM subspace optimization generalizes the search space to be the column space of a matrix $\mathbf{D}_t=\left[\mathbf{d}_t^1,\ldots,\mathbf{d}_t^m\right]$ ($\mathbf{D}_t$ is usually constructed by the gradient directions of the previous $\mathbf{x}_t$'s), and the step-size to be $\boldsymbol{\alpha}_t\in\mathbb{R}^m$. Given $\mathbf{D}_t$, $\boldsymbol{\alpha}_t$ is found by MM that decreases the objective value.

In the following we assume ∇f is L-Lipschitz. Define $\tilde{f}_t$ as $\tilde{f}_t\left(\boldsymbol{\alpha}\right)=f\left(\mathbf{x}_t+\mathbf{D}_t\boldsymbol{\alpha}\right)$, then

$$\begin{aligned}&\tilde{f}_t\left(\boldsymbol{\alpha}\right)\\ =&f\left(\mathbf{x}_t+\mathbf{D}_t\boldsymbol{\alpha}\right)\\ \leq&\tilde{f}\left(\boldsymbol{\alpha}_t^k\right)+\nabla\tilde{f}\left(\boldsymbol{\alpha}_t^k\right)^T\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}_t^k\right)+\frac{L}{2}\left\|\mathbf{D}_t\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}_t^k\right)\right\|_2^2\\ :=&g\left(\boldsymbol{\alpha}|\boldsymbol{\alpha}_t\right).\end{aligned}$$

The surrogate function $g\left(\boldsymbol{\alpha}|\boldsymbol{\alpha}_t\right)$ is quadratic in $\boldsymbol{\alpha}$, and has a minimizer given by

$$\boldsymbol{\alpha}_t^{k+1}=\boldsymbol{\alpha}_t^k-\left(L\mathbf{D}_t^T\mathbf{D}_t\right)^\dagger\nabla\tilde{f}\left(\boldsymbol{\alpha}_t^k\right).$$

When $m=1$, the method reduces to MM line search [116], [117], and when $m=n$ it recovers the ordinary MM. Analysis of the algorithm convergence and generalizations can be found in [118]–[122].

V. Applications

In this section, we demonstrate applications of MM categorized according to the techniques in Section III.

A. First Order Taylor Expansion

A large number of MM algorithms are derived based on linearizing the concave components in the objective function, as shown in the following applications.

1) Reweighted ℓ_1-norm Minimization: The problem of finding a sparse solution of an underdetermined equation system $\mathbf{y}=\mathbf{A}\mathbf{x}$ can be formulated as

$$\begin{aligned}\underset{\mathbf{x}}{\text{minimize}}\quad&\sum_{i=1}^n\log\left(\epsilon+|x_i|\right)\\ \text{subject to}\quad&\mathbf{y}=\mathbf{A}\mathbf{x},\end{aligned} \tag{48}$$

where the objective function is an approximation of the ℓ_0-norm with $\epsilon>0$ [15].

The reweighted ℓ_1-norm minimization algorithm solves problem (48) by solving

$$\begin{aligned}\underset{\mathbf{x}}{\text{minimize}}\quad&\sum_{i=1}^n\frac{|x_i|}{\epsilon+|x_i^t|}\\ \text{subject to}\quad&\mathbf{y}=\mathbf{A}\mathbf{x}\end{aligned} \tag{49}$$

at the t-th iteration, which is an MM step by applying inequality (11) to the objective function.

2) Robust Covariance Estimation: A robust estimator of covariance matrix $\mathbf{R}$ with zero-mean observations $\{\mathbf{x}_i\}_{i=1}^N$ is formulated as the minimizer of the following problem [33]:

$$\underset{\mathbf{R}\succeq\mathbf{0}}{\text{minimize}}\quad\log\det\left(\mathbf{R}\right)+\frac{K}{N}\sum_{i=1}^N\log\left(\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i\right). \tag{50}$$

By inequality (11) a surrogate function can be found as

$$g\left(\mathbf{R}|\mathbf{R}_t\right)=\log\det\left(\mathbf{R}\right)+\frac{K}{N}\sum_{i=1}^N\frac{\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i}{\mathbf{x}_i^H\mathbf{R}_t^{-1}\mathbf{x}_i}, \tag{51}$$

which is not convex in $\mathbf{R}$, but has a closed-form minimizer given by

$$\mathbf{R}_{t+1}=\frac{K}{N}\sum_{i=1}^N\frac{\mathbf{x}_i\mathbf{x}_i^H}{\mathbf{x}_i^H\mathbf{R}_t^{-1}\mathbf{x}_i}. \tag{52}$$

Notice that we can replace the $\log$ function in $\log\left(\mathbf{x}_i^H\mathbf{R}^{-1}\mathbf{x}_i\right)$ by a continuously differentiable concave function ρ and the same derivation applies. This idea has also been used in [22]–[24] for regularized covariance estimation problems.

In contrast to problems (48) and (50), where discovering a surrogate function is easy, some applications require one to exploit hidden concavity by manipulating the objective function, as illustrated by the following example.

3) Variance Component Model: Consider the signal model

$$\mathbf{x}_i=\mathbf{A}\mathbf{s}_i+\mathbf{n}_i,$$

where the $\mathbf{s}_i$'s and $\mathbf{n}_i$'s are zero mean *i.i.d.* signal and noise with $\mathrm{Cov}\left(\mathbf{s}_i\right)=\mathrm{diag}\left(p_1,\ldots,p_L\right)$ and $\mathrm{Cov}\left(\mathbf{n}_i\right)=\sigma^2\mathbf{I}$, respectively.

Denote $\mathbf{p}=\left[p_1,\ldots,p_L\right]^T$ and $\mathbf{P}=\mathrm{diag}\left(\mathbf{p}\right)$, the covariance $\mathbf{R}$ of $\mathbf{x}_i$ admits the structure

$$\mathbf{R}\quad=\quad\mathbf{A}\mathbf{P}\mathbf{A}^H+\sigma^2\mathbf{I}.$$

Assuming Gaussianity of the observations $\{\mathbf{x}_i\}_{i=1}^N$, a maximum likelihood type estimator $\hat{\mathbf{R}}$ is defined as the solution of the following problem [34], [35]:

$$\begin{aligned}\underset{\mathbf{R},\mathbf{P}\succeq\mathbf{0},\sigma}{\text{minimize}}\quad&\log\det\left(\mathbf{R}\right)+\mathrm{Tr}\left(\mathbf{S}\mathbf{R}^{-1}\right)\\ \text{subject to}\quad&\mathbf{R}=\mathbf{A}\mathbf{P}\mathbf{A}^H+\sigma^2\mathbf{I},\end{aligned} \tag{53}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

where $\mathbf{S}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^H$.

We describe the SBL algorithm that solves (53) derived based on EM [35] using MM. For simplicity, we assume that σ^2 is given. To find a separable surrogate function, we work with the precision matrix $\mathbf{\Gamma}=\mathbf{P}^{-1}$ and rewrite the objective function as

$$\begin{aligned}&L\left(\mathbf{\Gamma}\right)\\=&\log\det\left(\mathbf{\Sigma}^{-1}\right)-\log\det\left(\mathbf{\Gamma}\right)-\sigma^{-4}\mathrm{Tr}\left(\mathbf{S}\mathbf{A}\mathbf{\Sigma}\mathbf{A}^H\right)+\text{const.},\end{aligned}$$

where $\mathbf{\Sigma}=\left(\mathbf{\Gamma}+\sigma^{-2}\mathbf{A}^H\mathbf{A}\right)^{-1}$.

Since $\log\det$ is concave, and the last term of $L\left(\mathbf{\Gamma}\right)$ is convex in $\mathbf{\Sigma}^{-1}$, we construct surrogate function

$$\begin{aligned}&g\left(\mathbf{\Gamma}|\mathbf{\Gamma}_t\right)\\=&\mathrm{Tr}\left(\mathbf{\Sigma}_t\mathbf{\Gamma}\right)-\log\det\left(\mathbf{\Gamma}\right)+\frac{\sigma^{-4}}{N}\sum_{i=1}^{N}\mathbf{x}_i^H\mathbf{A}\mathbf{\Sigma}_t\mathbf{\Gamma}\mathbf{\Sigma}_t\mathbf{A}^H\mathbf{x}_i\end{aligned}$$

by inequalities (12) and (13).

Define $\boldsymbol{\mu}_i=\sigma^{-2}\mathbf{\Sigma}_t\mathbf{A}^H\mathbf{x}_i$, then

$$\begin{aligned}g\left(\mathbf{\Gamma}|\mathbf{\Gamma}_t\right)&=\sum_{j=1}^{L}\Sigma_{jj}^t\Gamma_j-\sum_{j=1}^{L}\log\Gamma_j+\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu}_i^H\mathbf{\Gamma}\boldsymbol{\mu}_i\\&=\sum_{j=1}^{L}\left(\Sigma_{jj}^t+\sum_{i=1}^{N}\left|\mu_{ij}\right|^2\right)\Gamma_j-\sum_{j=1}^{L}\log\Gamma_j,\end{aligned}\tag{54}$$

where μ_{ij} is the j-th element of $\boldsymbol{\mu}_i$. The update of Γ_j (equivalently p_j) can be computed in parallel as

$$\left(\Gamma_j^{t+1}\right)^{-1}=p_j^{t+1}=\Sigma_{jj}^t+\sum_{i=1}^{N}\left|\mu_{ij}\right|^2.\tag{55}$$

4) Optimization with Projection Forms: Projection matrices appear in optimization problems in structured low-rank approximation [123], minimization of MSE criterion [124], etc.

With a slight abuse of terminology, in this subsection, we refer to matrices parameterized as

$$\mathbf{P}\left(\mathbf{X}\right)=\mathbf{L}\left(\mathbf{X}\right)^T\mathbf{Q}\left(\mathbf{X}\right)^{-1}\mathbf{L}\left(\mathbf{X}\right)\tag{56}$$

as projection forms, where $\mathbf{L}\left(\mathbf{X}\right)$ is linear in $\mathbf{X}$ and $\mathbf{Q}\left(\mathbf{X}\right)$ is quadratic in $\mathbf{X}$. Note that $\mathbf{P}$ is a standard projection matrix if $\mathbf{L}\left(\mathbf{X}\right)=\mathbf{X}$ and $\mathbf{Q}\left(\mathbf{X}\right)=\mathbf{X}\mathbf{X}^T$.

By inequality (14), the trace of $\mathbf{P}\left(\mathbf{X}\right)$ can be lowerbounded as

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{P}\left(\mathbf{X}\right)\right)\geq&2\mathrm{Tr}\left(\mathbf{L}\left(\mathbf{X}_t\right)^T\mathbf{Q}\left(\mathbf{X}_t\right)^{-1}\mathbf{L}\left(\mathbf{X}\right)\right)\\&-\mathrm{Tr}\left(\mathbf{Q}\left(\mathbf{X}_t\right)^{-1}\mathbf{L}\left(\mathbf{X}_t\right)\mathbf{L}\left(\mathbf{X}_t\right)^T\mathbf{Q}\left(\mathbf{X}_t\right)^{-1}\mathbf{Q}\left(\mathbf{X}\right)\right)\end{aligned}\tag{57}$$

with equality achieved at $\mathbf{X}=\mathbf{X}_t$.

Let us consider the covariance matrix estimation problem in [36] with the following objective function:

$$L\left(\mathbf{W}\right)=\log\det\left(\tau\mathbf{W}^H\mathbf{W}+\mathbf{I}\right)+\mathbf{z}^H\left(\tau\mathbf{W}\mathbf{W}^H+\mathbf{I}\right)^{-1}\mathbf{z}.\tag{58}$$

The first term is upperbounded as

$$\begin{aligned}&\log\det\left(\tau\mathbf{W}^H\mathbf{W}+\mathbf{I}\right)\\\leq&\mathrm{Tr}\left(\tau\left(\tau\mathbf{W}_t^H\mathbf{W}_t+\mathbf{I}\right)^{-1}\mathbf{W}^H\mathbf{W}\right)+\text{const.}\end{aligned}\tag{59}$$

by inequality (12). As for the second term, we first create a projection form by the matrix inversion lemma as follows:

$$\begin{aligned}&\mathbf{z}^H\left(\tau\mathbf{W}\mathbf{W}^H+\mathbf{I}\right)^{-1}\mathbf{z}\\=&\mathbf{z}^H\mathbf{z}-\underbrace{\mathbf{z}^H\mathbf{W}\left(\tau^{-1}\mathbf{I}+\mathbf{W}^H\mathbf{W}\right)^{-1}\mathbf{W}^H\mathbf{z}}_{\text{projection form}}.\end{aligned}\tag{60}$$

Letting $\mathbf{L}\left(\mathbf{W}\right)=\mathbf{W}$ and $\mathbf{Q}\left(\mathbf{W}\right)=\tau^{-1}\mathbf{I}+\mathbf{W}^H\mathbf{W}$, inequality (57) implies that (60) can be upperbounded as

$$\mathbf{z}^H\left(\tau\mathbf{W}\mathbf{W}^H+\mathbf{I}\right)^{-1}\mathbf{z}\leq\mathrm{Tr}\left(\mathbf{W}\mathbf{H}_t\mathbf{W}^H\right)-2\mathrm{Re}\left(\mathbf{L}_t\mathbf{W}^H\right),\tag{61}$$

where $\mathbf{H}_t$ and $\mathbf{L}_t$ are coefficients given by

$$\mathbf{H}_t=\left(\tau^{-1}\mathbf{I}+\mathbf{W}_t^H\mathbf{W}_t\right)^{-1}\mathbf{W}_t^H\mathbf{z}\mathbf{z}^H\mathbf{W}_t\left(\tau^{-1}\mathbf{I}+\mathbf{W}_t^H\mathbf{W}_t\right)^{-1}$$

and

$$\mathbf{L}_t=\mathbf{z}\mathbf{z}^H\mathbf{W}_t\left(\tau^{-1}\mathbf{I}+\mathbf{W}_t^H\mathbf{W}_t\right)^{-1}.$$

Combining (59) and (61) we arrive at a surrogate function

$$g\left(\mathbf{W}|\mathbf{W}_t\right)=\mathrm{Tr}\left(\mathbf{W}\mathbf{H}\mathbf{W}^H\right)-2\mathrm{Re}\left(\mathbf{L}_t\mathbf{W}^H\right)$$

with

$$\mathbf{H}=\left(\mathbf{W}_t^H\mathbf{W}_t+\tau^{-1}\mathbf{I}\right)^{-1}+\mathbf{H}_t,$$

which has a closed-form minimizer given by $\mathbf{W}_{t+1}=\mathbf{L}_t\mathbf{H}^{-1}$.

5) Maximizing of A Convex Function Over A Compact Set: Consider problem

$$\begin{aligned}\underset{\mathbf{x}}{\text{maximize}}\quad&f\left(\mathbf{x}\right)\\\text{subject to}\quad&\mathbf{x}\in\mathcal{K},\end{aligned}\tag{62}$$

where $\mathcal{K}$ is a compact set and $f:\mathcal{K}\rightarrow\mathbb{R}$ is convex.

A gradient method has been proposed and analyzed in [37] to solve (62), which falls into the category of MM.

Since f is convex, it can be minorized as

$$f\left(\mathbf{x}\right)\geq f\left(\mathbf{x}_t\right)+\nabla f\left(\mathbf{x}_t\right)^T\left(\mathbf{x}-\mathbf{x}_t\right).$$

The maximization step is then given by

$$\mathbf{x}_{t+1}\in\arg\max_{\mathbf{x}\in\mathcal{K}}\nabla f\left(\mathbf{x}_t\right)^T\mathbf{x}.\tag{63}$$

Since $\mathcal{K}$ is compact, $\mathbf{x}_{t+1}$ is well-defined.

For example, if $\mathcal{K}=\left\{\mathbf{x}\in\mathbb{R}^n|\left\|\mathbf{x}\right\|_2=1\right\}$, then $\mathbf{x}_{t+1}=\nabla f\left(\mathbf{x}_t\right)/\left\|\nabla f\left(\mathbf{x}_t\right)\right\|_2$; and if $\mathcal{K}$ is the Stiefel manifold defined as

$$\mathcal{K}=\left\{\mathbf{X}\in\mathbb{R}^{m\times n}|\mathbf{X}^T\mathbf{X}=\mathbf{I}_n\right\},$$

where $n\leq m$, let the polar decomposition of $\nabla f\left(\mathbf{X}_t\right)$ be $\nabla f\left(\mathbf{X}_t\right)=\mathbf{U}\mathbf{P}$, then $\mathbf{X}_{t+1}=\mathbf{U}$.

6) SEVP with ℓ_0-norm Constraint: The sparse eigenvector problem (SEVP) aims at finding a sparse unit length vector $\mathbf{x}$ that maximizes the quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$, where $\mathbf{A}\in\mathbb{S}_+^n$. It attracts a lot of attention in applications such as bioinformatics, big data analysis, and machine learning, where a parsimonious interpretation of the data set is desired, see references [37], [125]–[127] for examples.

To enforce sparsity, we can include a zero norm constraint on $\mathbf{x}$ and formulate the problem as [19]:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\|_2 = 1 \\ & \|\mathbf{x}\|_0 \le k. \end{aligned} \tag{64}$$

Since $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is convex in $\mathbf{x}$, it can be minorized at $\mathbf{x} = \mathbf{x}_t$ by $g(\mathbf{x}|\mathbf{x}_t) = 2\mathbf{x}_t^T\mathbf{A}\mathbf{x}$.

Define $\mathbf{a} = \mathbf{A}^T\mathbf{x}_t$ for notation simplicity. In the maximization step we need to solve problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \mathbf{a}^T\mathbf{x} \\ \text{subject to} \quad & \|\mathbf{x}\|_2 = 1 \\ & \|\mathbf{x}\|_0 \le k. \end{aligned} \tag{65}$$

Define $\mathcal{I}_k = \{i|x_i \neq 0\}$, then

$$\mathbf{a}^T\mathbf{x} = \mathbf{a}_{\mathcal{I}_k}^T\mathbf{x}_{\mathcal{I}_k} \le \|\mathbf{a}_{\mathcal{I}_k}\|_2 \|\mathbf{x}_{\mathcal{I}_k}\|_2 = \|\mathbf{a}_{\mathcal{I}_k}\|_2,$$

where the inequality follows from the Cauchy-Schwarz inequality, and the last equality follows from the constraint $\|\mathbf{x}\|_2 = 1$ and the definition of $\mathcal{I}_k$. Observe that $\|\mathbf{a}_{\mathcal{I}_k}\|_2$ is maximized when $\mathcal{I}_k$ is the set of indices of a_i with the k largest absolute value, and $\mathbf{a}^T\mathbf{x}$ is maximized when $\mathbf{a}$ and $\mathbf{x}$ are collinear. Sort the elements of $|\mathbf{a}| = [|a_1|, \ldots, |a_n|]^T$ in descending order. That is, we find a permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $|a|_{\pi(1)} \ge \cdots \ge |a|_{\pi(n)}$. The solution $\mathbf{x}^\star$ of the problem (65) is given by

$$\begin{aligned} \tilde{x}_i^\star &= \begin{cases} a_i, & a_i \ge |a|_{\pi(k)} \\ 0, & a_i < |a|_{\pi(k)} \end{cases} \\ \mathbf{x}^\star &= \tilde{\mathbf{x}}^\star / \|\tilde{\mathbf{x}}^\star\|_2. \end{aligned} \tag{66}$$

The algorithm is named the truncated power method as computing vector $\mathbf{a}$ is a power iteration step, and in (66) the smallest $n-k$ elements of $|\mathbf{a}|$ are truncated to zero.

7) SEVP with ℓ_0-norm Penalty: SEVP can also be formulated in penalty form as [21]

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} - \rho\|\mathbf{x}\|_0 \\ \text{subject to} \quad & \|\mathbf{x}\|_2 = 1, \end{aligned} \tag{67}$$

where $\rho \ge 0$ is a parameter that controls the sparsity level. Linearizing the quadratic term we have a minorizing function

$$g(\mathbf{x}|\mathbf{x}_t) = 2\mathbf{x}_t^T\mathbf{A}\mathbf{x} - \rho\|\mathbf{x}\|_0. \tag{68}$$

Denote $\mathbf{a} = 2\mathbf{A}^T\mathbf{x}_t$. Suppose $\|\mathbf{x}\|_0 = k \le \|\mathbf{a}\|_0$, the minimizer $\mathbf{x}^\star$ of $g(\mathbf{x}|\mathbf{x}_t)$ is then given by (66) with $g(\mathbf{x}^\star|\mathbf{x}_t)$ being

$$\sqrt{\sum_{i=1}^{k} |a|_{\pi(i)}} - k\rho.$$

Therefore, update $\mathbf{x}_{t+1}$ has cardinality

$$k^\star = \arg\max_k \sqrt{\sum_{i=1}^{k} |a|_{\pi(i)}} - k\rho,$$

and takes the form (66) with $k = k^\star$.

We introduce in the end algorithms derived by iteratively upperbounding f by a quadratic form, which belongs to the *iteratively reweighted least squares* (IRLS) algorithms. Note that the upperbounds for a convex objective function f are not constructed based on first order Taylor expansion. We put them in this subsection for the integrality of the applications.

8) Edge-Preserving Regularization in Image Processing: Many image restoration and reconstruction problems can be formulated as

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) + \Phi(\mathbf{x}),$$

where f is a quadratic function of the form

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} - 2\mathbf{q}^T\mathbf{x},$$

and $\Phi(\mathbf{x}) = \sum_{i=1}^m \phi(\delta_i)$, where $\delta_i = \left(\mathbf{V}^T\mathbf{x} - \mathbf{w}\right)_i$, is a regularization term with parameters $\mathbf{V} \in \mathbb{R}^{n\times m}$ and $\mathbf{w} \in \mathbb{R}^m$ [38].

Suppose that $\phi : \mathbb{R} \to \mathbb{R}$ satisfies regularity conditions: (1) ϕ is even; (2) ϕ is coercive and continuously differentiable; (3) $\phi\left(\sqrt{\cdot}\right)$ is concave on $\mathbb{R}_+$; and (4) $0 < \phi'(t)/t < \infty$ (cf. Figure 4 for examples of ϕ). Then an upperbound of $\phi(\delta)$ can be derived based on the concavity of $\phi\left(\sqrt{\cdot}\right)$ as follows:

$$\phi(\delta) = \phi\left(\sqrt{\delta^2}\right) \le \frac{1}{2}\frac{\phi'(\delta_t)}{\delta_t}\delta^2 + \text{const.}$$

Letting $\delta_i = \left(\mathbf{V}^T\mathbf{x} - \mathbf{w}\right)_i$ yields the following quadratic surrogate function:

$$g(\mathbf{x}|\mathbf{x}_t) = \mathbf{x}^T\left(\mathbf{Q} + \frac{1}{2}\mathbf{V}\mathbf{D}_t\mathbf{V}^T\right)\mathbf{x} - (2\mathbf{q} + \mathbf{V}\mathbf{D}_t\mathbf{w})^T\mathbf{x},$$

where $\mathbf{D}_t$ is a diagonal matrix with the i-th diagonal element being $d_i = \phi'(\delta_i^t)/\delta_i^t$. Assuming that $2\mathbf{Q} + \mathbf{V}\mathbf{D}_t\mathbf{V}^T$ is invertible, $\mathbf{x}$ is then updated as

$$\mathbf{x}_{t+1} = \left(2\mathbf{Q} + \mathbf{V}\mathbf{D}_t\mathbf{V}^T\right)^{-1}(2\mathbf{q} + \mathbf{V}\mathbf{D}_t\mathbf{w}). \tag{69}$$

Suppose alternatively that ϕ satisfies: (1) ϕ is coercive and continuously differentiable; and (2) ϕ' is L-Lipschitz. Then f can be majorized based on descent lemma (24) by surrogate function

$$g(\mathbf{x}|\mathbf{x}_t) = \mathbf{x}^T\left(\mathbf{Q} + \frac{1}{2L}\mathbf{V}\mathbf{V}^T\right)\mathbf{x} + \left(2\mathbf{q} + \frac{\mathbf{V}(\boldsymbol{\ell}_t + \mathbf{w})}{L}\right),$$

where $\ell_t^i = \delta_i^t - L\phi'(\delta_i^t)$. Assuming that $2\mathbf{Q} + \frac{1}{L}\mathbf{V}\mathbf{V}^T$ is invertible, $\mathbf{x}$ is then updated as

$$\mathbf{x}_{t+1} = \left(2\mathbf{Q} + \frac{1}{L}\mathbf{V}\mathbf{V}^T\right)^{-1}\left(2\mathbf{q} + \frac{\mathbf{V}(\boldsymbol{\ell}_t + \mathbf{w})}{L}\right). \tag{70}$$

Iteration (69) and (70) correspond to the half-quadratic minimization algorithms without over-relaxation as proposed in [7] and [8], respectively. A convergence study of the algorithm with over-relaxation can be found in [38].

9) ℓ_p-Norm Minimization: Optimization problems involving ℓ_p-norm arise frequently in robust fitting and sparse representation problems. When $1 \le p < 2$, $|x|^p$ is convex, and when $1 < p < 1$, the tightest convex upperbound of $|x|^p$ is obtained by linearization. IRLS type algorithms, however, use a quadratic upperbound for $|x|^p$. The idea is to majorize $\|\mathbf{x}\|_p^p$ ($\mathbf{x} \in \mathbb{R}^n$) by $\|\mathbf{x}\|^2$ at each iteration and solve a weighted ℓ_2-norm minimization problem instead. More

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

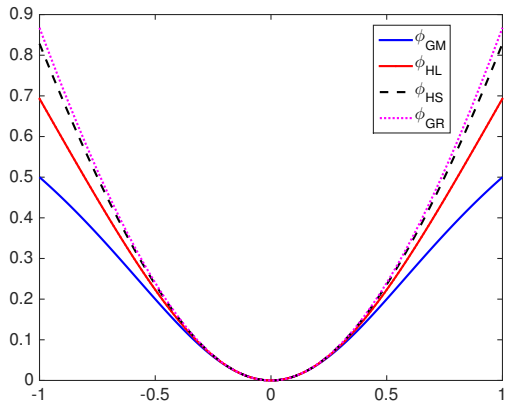

Figure 4. Examples of edge-preserving functions presented in [128] ($\varphi_{GM}(t) = \frac{t^2}{1+t^2}$, $\varphi_{HL} = \log(1+t^2)$, $\varphi_{HS} = 2\sqrt{1+t^2} - 2$, $\varphi_{GR} = 2\log(\cosh(t))$).

precisely, inequality (16) indicates that at $\mathbf{x} = \mathbf{x}_t$, if none of the elements of $\mathbf{x}_t$ are zero, then $\|\mathbf{x}\|_p^p$ can be majorized as

$$\|\mathbf{x}\|_p^p \leq \|\mathbf{x}\|_{\mathbf{W}_t}^2 + \text{const.}, \tag{71}$$

where $\mathbf{W}_t$ is a diagonal matrix with the i-th diagonal element being $w_i^t = \frac{p}{2}|x_i^t|^{p-2}$.

Take the following robust regression problem as an example:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p, \tag{72}$$

where $\mathbf{b} \in \mathbb{R}^m$. With inequality (16) we can construct a quadratic surrogate function:

$$g(\mathbf{x}|\mathbf{x}_t) = \sum_{i=1}^m w_i^t (b_i - \mathbf{A}_{i,:}\mathbf{x})^2,$$

where w_i^t is given by

$$w_i^t = |b_i - \mathbf{A}_{i,:}\mathbf{x}_t|^{p-2}.$$

Function $g(\mathbf{x}|\mathbf{x}_t)$ admits a closed-form minimizer

$$\mathbf{x}^{t+1} = \left(\mathbf{A}^T\mathbf{W}_t\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{W}_t\mathbf{b}.$$

The loss function $|x|^p$ can be generalized to any continuously differentiable concave function for robust fitting [40].

A similar idea has been applied in [41] in solving the sparse representation problem

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1,$$

and in [42] in solving the compressed sensing problem

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned}$$

where $\|\mathbf{x}\|_1$ was upperbounded by a quadratic function using inequality (16) and a least squares problem was solved per iteration.

Note that in [41], convergence was established under the condition that none of the x_i^t's are zero, and in [42] the weight was adaptively modified so that it never goes to infinity. If any x_i^t becomes zero, the algorithm will be ill-posed since weight matrix $\mathbf{W}_t$ will be undefined in the next iteration. The effect of this singularity issue in algorithm convergence has been extensively studied in the literature, see [29], [39], [41] and [42] for examples. A way to circumvent the difficulty is by smoothing the objective function. For example, $\|\mathbf{x}\|_p^p$ was approximated by

$$h^{\epsilon,p}(\mathbf{x}) = \sum_{i=1}^n \left(\epsilon^2 + x_i^2\right)^{p/2}$$

in [43], where ϵ is a positive small number. As a result, the weight at each iteration is always well-defined. The smoothing technique was also adopted in [21] with a different approximation, to solve a sparse generalized eigenvalue problem (GEVP) formulated as

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} - \sum_{i=1}^n |x_i|^p \\ \text{subject to} \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} = 1. \end{aligned} \tag{73}$$

B. Second Order Taylor Expansion (Hessian Bound)

If the Hessian matrix of the objective function f is uniformly bounded, i.e., $\mathbf{M} \succeq \nabla^2 f(\mathbf{x}),\ \forall \mathbf{x} \in \mathcal{X}$, then we can find a quadratic surrogate function using inequality (25). As a benefit, the update usually admits a closed-form solution.

1) Logistic Regression: In a multi-class classification problem, we are given data pairs $(\mathbf{x}_n, \mathbf{t}_n)_{1\leq n\leq N}$, where $\mathbf{x}_n \in \mathbb{R}^m$ is a feature vector and $\mathbf{t}_n$ is a $(K+1)$-dimensional encoding vector with $(\mathbf{t}_n)_i = 1$ if $\mathbf{x}$ belongs to the i-th category and $(\mathbf{t}_n)_i = 0$ otherwise. The task is to train a statistical model that can predict $\mathbf{t}$ based on $\mathbf{x}$ [44], [129]. For notation simplicity we assume there is only one training sample $(\mathbf{x}, \mathbf{t})$.

The problem can be formulated as finding a $\mathbf{w}$, defined as $\mathbf{w} = \left[\mathbf{w}_1^T, \ldots, \mathbf{w}_K^T\right]^T$, that minimizes the negative log-likelihood function:

$$L(\mathbf{w}) = \sum_{j=1}^K -t_j\mathbf{w}_j^T\mathbf{x} + \log\left(1 + \sum_{j=1}^K \exp\left(\mathbf{w}_j^T\mathbf{x}\right)\right). \tag{74}$$

It can be proved that the Hessian of $L(\mathbf{w})$ is uniformly upperbounded by matrix

$$\mathbf{M} = \frac{1}{2}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{K+1}\right) \otimes \left(\mathbf{x}\mathbf{x}^T\right).$$

Therefore, inequality (25) implies that L can be upperbounded by

$$\begin{aligned} g(\mathbf{w}|\mathbf{w}^t) = & \left(\left(\tilde{\mathbf{t}} - \mathbf{p}\left(\mathbf{w}^t\right)\right) \otimes \mathbf{x}\right)^T\left(\mathbf{w} - \mathbf{w}^t\right) \\ & + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^t\right)^T\mathbf{M}\left(\mathbf{w} - \mathbf{w}^t\right), \end{aligned} \tag{75}$$

where $\tilde{\mathbf{t}} \coloneqq [t_1; \ldots; t_K]$ and $\mathbf{p}(\mathbf{w}) \coloneqq [p_1(\mathbf{w}); \ldots; p_K(\mathbf{w})]$ with

$$p_j(\mathbf{w}) = \frac{\exp\left(\mathbf{w}_j^T\mathbf{x}\right)}{1 + \sum_{j=1}^K \exp\left(\mathbf{w}_j^T\mathbf{x}\right)}.$$

The update of $\mathbf{w}$ is then given as

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \mathbf{M}^{-1}\left(\left(\tilde{\mathbf{t}} - \mathbf{p}\left(\mathbf{w}^t\right)\right) \otimes \mathbf{x}\right).$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

Compared to the Newton method that requires computing $\nabla^2 L(\mathbf{w})$ at each iteration, minimizing $g(\mathbf{w}|\mathbf{w}^t)$ only requires pre-computing $\mathbf{M}$ once since it is independent of $\mathbf{w}^t$.

Combining with the technique described in Section V-A9, MM can be derived for sparse logistic regression with the ℓ_1-norm penalty formulated as

$$L(\mathbf{w}) = -\sum_{j=1}^{K} t_j \mathbf{w}_j^T \mathbf{x} + \log\left(1 + \sum_{j=1}^{K} \exp\left(\mathbf{w}_j^T \mathbf{x}\right)\right) + \lambda \|\mathbf{w}\|_1, \tag{76}$$

where $\lambda \geq 0$ is a regularization parameter [45]. At each iteration, a quadratic upperbound for the ℓ_1-norm term was merged with (75), thus still leading to a quadratic surrogate function that has a closed-form minimizer.

Even if the function to be majorized is already quadratic, inequality (25) is still applied in some applications with an $\mathbf{M}$ that is easier to deal with (usually being a diagonal or a scaled identity matrix), as can be seen in Sections V-B2, V-B3, V-B4, and V-B5.

2) Matrix Quadratic Form Minimization with Rank Constraint: Consider problem

$$\begin{array}{ll} \underset{\mathbf{X}}{\text{minimize}} & \text{vec}(\mathbf{X})^T \mathbf{Q} \text{vec}(\mathbf{X}) + \text{vec}(\mathbf{L})^T \text{vec}(\mathbf{X}) \\ \text{subject to} & \text{rank}(\mathbf{X}) \leq r, \end{array} \tag{77}$$

where $\mathbf{Q}$ is a symmetric square matrix with its maximum eigenvalue positive and $\mathbf{X}, \mathbf{L} \in \mathbb{R}^{m \times n}$ $(m \geq n)$.

Observe that if $\mathbf{Q}$ is a scaled identity matrix, then problem (77) can be written as

$$\begin{array}{ll} \underset{\mathbf{X}}{\text{minimize}} & \|\mathbf{X} + c\mathbf{L}\|_F^2 \\ \text{subject to} & \text{rank}(\mathbf{X}) \leq r \end{array} \tag{78}$$

with c being some positive constant, which has a closed-form minimizer based on the singular value decomposition (SVD) of $\mathbf{L}$.

For this reason, we construct the following surrogate function by applying inequality (26) to the first term of the objective function:

$$g(\mathbf{X}|\mathbf{X}_t) = \lambda \|\mathbf{X} - \mathbf{Y}\|_F^2 + \text{const.}, \tag{79}$$

where $\lambda = \lambda_{\max}(\mathbf{Q})$ and $\text{vec}(\mathbf{Y}) = -(\mathbf{Q}/\lambda - \mathbf{I})\,\text{vec}(\mathbf{X}_t) - \text{vec}(\mathbf{L})/(2\lambda)$.

Let the thin SVD of $\mathbf{Y}$ be $\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ with $\mathbf{S} = \text{diag}(\sigma_1, \ldots, \sigma_n)$ and $\sigma_1 \geq \cdots \geq \sigma_n$, $\mathbf{X}_{t+1} \in \arg\min_{\text{rank}(\mathbf{X}) \leq r} g(\mathbf{X}|\mathbf{X}_t)$ is then given by

$$\mathbf{X}_{t+1} = \mathbf{U}\mathbf{S}_r\mathbf{V}^T,$$

where $\mathbf{S}_r$ is obtained by thresholding the smallest $(n-r)$ elements of the diagonal of $\mathbf{S}$ to zero. We refer to this procedure as singular value hard thresholding.

Many problems can be cast in the form of (77). Two examples are given as follows.

Problem 20. The weighted low rank approximation problem is formulated as

$$\begin{array}{ll} \underset{\mathbf{X}}{\text{minimize}} & \|\mathbf{X} - \mathbf{R}\|_{\mathbf{Q}}^2 \\ \text{subject to} & \text{rank}(\mathbf{X}) \leq r, \end{array} \tag{80}$$

where $\mathbf{R} \in \mathbb{R}^{n \times m}$ is the matrix to be approximated, and $\|\mathbf{X}\|_{\mathbf{Q}} \coloneqq \text{vec}(\mathbf{X})^T \mathbf{Q} \text{vec}(\mathbf{X})$ with $\mathbf{Q} \in \mathbb{S}_+$ being a weight matrix. Problem (80) is an instance of problem (77), thus can be solved accordingly.

Problem 21. The low rank matrix completion problem is formulated as

$$\begin{array}{ll} \underset{\mathbf{X}}{\text{minimize}} & \|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{R})\|_F^2 \\ \text{subject to} & \text{rank}(\mathbf{X}) \leq r, \end{array} \tag{81}$$

where

$$P_\Omega(\mathbf{R})_{ij} = \begin{cases} R_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise} \end{cases}.$$

By defining $\mathbf{Q} = \text{diag}(\mathbf{q})$ with $q_i = 1$ if $\text{vec}(P_\Omega(\mathbf{R}))_i \neq 0$ and $q_i = 0$ otherwise, problem (81) is a special case of (80).

Moreover, since q_i is either 0 or 1, we have $\lambda_{\max}(\mathbf{Q}) = 1$. Consequently, $\text{vec}(\mathbf{Y}) = \mathbf{Q}(\text{vec}(\mathbf{R}) - \text{vec}(\mathbf{X}_t)) + \text{vec}(\mathbf{X}_t)$ and the expression of $\mathbf{Y}$ can be simplified as

$$Y_{ij} = \begin{cases} R_{ij}, & (i,j) \in \Omega \\ X_{ij}^t, & \text{otherwise} \end{cases}. \tag{82}$$

We randomly generate a matrix $\mathbf{R} \in \mathbb{R}^{500 \times 600}$ of rank $r = 10$, and create $P_\Omega(\mathbf{R})$ by uniformly randomly deleting 70% of the entries of $\mathbf{R}$. Figure 5 plots the objective value evolution curve and the recovery error $\|\mathbf{X}_t - \mathbf{R}\|_F$ versus iterations. It can be seen that in 100 iterations the objective value (approximation error) decreases to a value below 10^{-8} and the recovery error decreases to 10^{-4}.

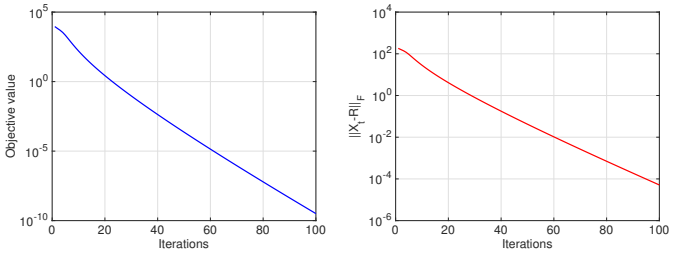

Figure 5. Matrix completion problem. Left: objective value $\|P_\Omega(\mathbf{X}_t) - P_\Omega(\mathbf{R})\|_F^2$ versus the number of iterations; right: recovery error measured by $\|\mathbf{X}_t - \mathbf{R}\|_F$ versus the number of iterations.

Similar to Section V-A7, the penalty form of problem (77), which relaxes the rank constraint to the objective function, can be handled with minor modifications of the above-described derivation.

3) Minimization of Quartic Forms: Minimizing a quartic function is closely related to minimizing matrix quadratic forms as discussed in Section V-B2. The objective function takes the form

$$f(\mathbf{x}) = \sum_{i=1}^{N} \left(\mathbf{x}^H \mathbf{A}_i \mathbf{x} - y_i\right)^2, \tag{83}$$

where $\mathbf{A}_i$ is Hermitian positive definite.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

The idea is to reduce the order of f by a change of variables. To this end, we define the "lifting matrix" $\mathbf{X} = \mathbf{x}\mathbf{x}^H$. Then $f(\mathbf{x})$ can be written as

$$\begin{aligned} f(\mathbf{X}) =& \sum_{i=1}^{N} \left(\mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}\right) - y_i\right)^2 \\ =& \mathrm{vec}\left(\mathbf{X}\right)^H \left(\sum_{i=1}^{N} \mathrm{vec}\left(\mathbf{A}_i\right)\mathrm{vec}\left(\mathbf{A}_i\right)^H\right)\mathrm{vec}\left(\mathbf{X}\right) \\ & - 2\sum_{i=1}^{N} y_i \mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}\right) + \sum_{i=1}^{N} y_i, \end{aligned} \tag{84}$$

which is quadratic in $\mathbf{X}$.

Although the order of the objective function has been reduced from quartic to quadratic, we have introduced the constraint that $\mathbf{X}$ is rank-one, i.e., $\mathbf{X} = \mathbf{x}\mathbf{x}^H$.

Denote $\mathbf{A} = \sum_{i=1}^{N} \mathrm{vec}\left(\mathbf{A}_i\right)\mathrm{vec}\left(\mathbf{A}_i\right)^H$, minimizing f is then equivalent to solving

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \mathrm{vec}\left(\mathbf{X}\right)^H \mathbf{A}\mathrm{vec}\left(\mathbf{X}\right) - 2\sum_{i=1}^{N} y_i \mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}\right) \\ \text{subject to} \quad & \mathrm{rank}\left(\mathbf{X}\right) = 1, \end{aligned}$$

which is a special case of (77) with the identification that $\mathbf{Q} = \mathbf{A}$ and $\mathbf{L} = -2\sum_{i=1}^{N} y_i\mathbf{A}_i$.

Problem 22. The phase retrieval problem aims at recovering signal $\mathbf{x}$ from phaseless measurements $y_i = \left|\mathbf{a}_i^H\mathbf{x}\right|^2$, $i = 1, \ldots, N$. The problem can be formulated as

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{i=1}^{N} \left(y_i - \left|\mathbf{a}_i^H\mathbf{x}\right|^2\right)^2. \tag{85}$$

Defining matrix $\mathbf{A}_i = \mathbf{a}_i\mathbf{a}_i^H$, problem (85) is of the form (83) and MM algorithms can be derived accordingly [47].

Problem 23. The sequence design problem considered in [31] aims at finding a length N complex-valued unimodular sequence $(x_n)_{1\le n\le N}$ with low autocorrelation sidelobes. The associated optimization problem takes the form

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \sum_{p=1}^{2N} \left(\mathbf{x}^H\mathbf{a}_p\mathbf{a}_p^H\mathbf{x}\right)^2 \\ \text{subject to} \quad & |x_i| = 1,\ i = 1, \ldots, N. \end{aligned} \tag{86}$$

The objective function is a special case of that of problem (83) with $\mathbf{A}_p = \mathbf{a}_p\mathbf{a}_p^H$ and $y_p = 0$.

To deal with the unit-modulus constraint $|x_i| = 1$, we observe that $\mathrm{Tr}\left(\mathbf{X}^H\mathbf{X}\right)$ and $\mathbf{x}^H\mathbf{x}$ are constants in the set $\mathcal{X} = \{\mathbf{x} \,|\, |x_i| = 1,\ i = 1, \ldots, N\}$. Therefore, we can apply inequality (25) twice with $\mathbf{M}$ being a scaled identity matrix, yielding a linear surrogate function that has a closed-form minimizer in $\mathcal{X}$ [31].

Remark 24. Upperbounding a quadratic function by a linear function is not possible if the constraint set is the entire Euclidean space. However, restricting it to the set $\mathcal{X}$ makes it possible. Figure 6 visualizes how a linear function upperbounds a convex quadratic one on the unit circle.

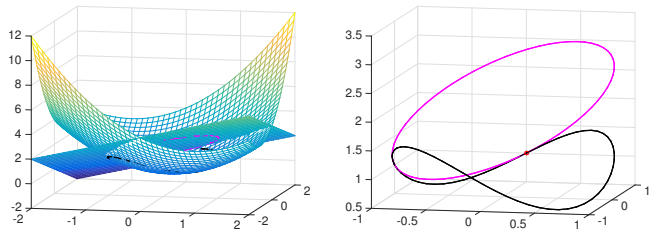

Figure 6. Linear upperbound for a quadratic function on unit circle (black curve: intercept of the quadratic function on the unit circle; magenta curve: intercept of the linear upperbound on the unit circle; red dot: point at which the value of the functions are equal).

We test the performance of MM in designing a sequence of length $N = 1024$. The algorithm is initialized with a Golomb sequence as a reasonably good starting point, and the SQUAREM accelerator is employed to achieve a fast convergence rate. Figure 7 shows the evolution curve of the objective value versus the number of iterations and the correlation level of the resulting sequence at convergence.

Extensions of the problem to the design of a sequence minimizing a weighted integrated sidelobe level criterion and the design of a sequence set using MM can be found in [32], [130].

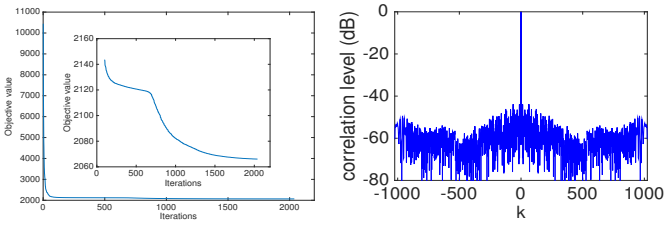

Figure 7. Sequence design problem. Left: objective value versus iterations; right: correlation level of sequence of length $N = 1024$.

4) Sparse Linear Regression: The sparse linear regression problem can be formulated as

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \rho h\left(\mathbf{x}\right), \tag{87}$$

where h is a penalty function that promotes a sparse $\mathbf{x}$, and $\rho \ge 0$ is the regularization parameter. We assume that h is separable and even, i.e., $h\left(\mathbf{x}\right) = \sum_{i=1}^{n} h_i\left(|x_i|\right)$, and h_i is concave and nondecreasing on $\mathbb{R}_+$.

The idea is to decouple the objective function, so that optimizing $\mathbf{x}$ can be done element-wise [18]. To this end, we resort to inequality (26) and upperbound the first term as

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \le \lambda\mathbf{x}^T\mathbf{x} - 2\mathbf{y}_t^T\mathbf{x} + \mathrm{const.},$$

where $\lambda = \lambda_{\max}\left(\mathbf{A}^T\mathbf{A}\right)$, and $\mathbf{y}_t = \mathbf{A}^T\mathbf{b} - \left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right)\mathbf{x}_t$. Then for each x_i, the problem boils down to finding a minimizer of

$$g^{(1)}\left(x_i|\mathbf{x}_t\right) = \lambda x_i^2 - 2y_i x_i + h_i\left(|x_i|\right). \tag{88}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

For example, when $h(\mathbf{x}) = \|\mathbf{x}\|_1$ the update is given by the soft-thresholding operator as [49]:

$$x_i^{t+1} = \mathcal{S}_{\rho/\lambda}\left(\frac{y_i}{\lambda}\right) = \begin{cases} \frac{y_i}{\lambda} - \frac{\rho}{2\lambda}, & \frac{y_i}{\lambda} > \frac{\rho}{2\lambda} \\ \frac{y_i}{\lambda} + \frac{\rho}{2\lambda}, & \frac{y_i}{\lambda} < \frac{\rho}{2\lambda} \\ 0, & \text{otherwise.} \end{cases} \tag{89}$$

When h_i is concave, further applying inequality (10) we can upperbound $h_i(|x_i|)$ as

$$h_i(|x_i|) \leq h_i'\left(|x_i^t|\right)|x_i| + \text{const.}$$

Together with (88) we arrive at the surrogate function

$$g^{(2)}(x_i|\mathbf{x}_t) = \lambda x_i^2 - 2y_i x_i + \rho h_i'\left(|x_i^t|\right)|x_i|$$

with a minimizer given by

$$x_i^{t+1} = \mathcal{S}_{\rho h_i'(|x_i^t|)/\lambda}\left(\frac{y_i}{\lambda}\right).$$

This method has been applied in image restoration in [29] and [48], where the problem is formulated as a high-dimensional penalized least square problem.

In the end, we present a special case that $h(|x_i|) = \|x_i\|_0$, which is discontinuous [16], [17]. The minimizer of $g^{(1)}(x_i|\mathbf{x}_t)$ has a closed-form given by

$$x_i^{t+1} = \begin{cases} y_i/\lambda, & y_i^2/\lambda > \rho \\ 0, & \text{otherwise,} \end{cases} \tag{90}$$

which is an iterative hard thresholding algorithm.

5) Nonnegative Least Squares: The nonnegative least squares (NLS) problem is a least squares fitting problem that requires the regressor to be nonnegative. The problem is stated as

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{subject to} & \mathbf{x} \geq \mathbf{0}. \end{array} \tag{91}$$

To obtain a closed-form update of $\mathbf{x}$ under the constraint $\mathbf{x} \geq \mathbf{0}$, we construct a separable surrogate function. The simplest way is to apply inequality (26) with $\mathbf{M} = \lambda\mathbf{I}$, which gives the following surrogate function:

$$g(\mathbf{x}|\mathbf{x}_t) = \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\left(\mathbf{x}_t - \frac{1}{\lambda}\left(\mathbf{A}^T\mathbf{A}\mathbf{x}_t - \mathbf{A}^T\mathbf{b}\right)\right),$$

where $\lambda \geq \lambda_{\max}\left(\mathbf{A}^T\mathbf{A}\right)$. Consequently, the update of $\mathbf{x}$ is given by

$$\mathbf{x}_{t+1} = \left[\mathbf{x}_t - \frac{1}{\lambda}\left(\mathbf{A}^T\mathbf{A}\mathbf{x}_t - \mathbf{A}^T\mathbf{b}\right)\right]_+, \tag{92}$$

which is a gradient projection algorithm.

If further assuming that $\mathbf{A} \in \mathbb{R}_{++}^{m\times n}$, $\mathbf{b} \in \mathbb{R}_+^m$, and $\mathbf{b} \neq \mathbf{0}$, it has been proven in [51] and [131] that

$$g(\mathbf{x}|\mathbf{x}_t) = \mathbf{x}^T\mathbf{M}_t\mathbf{x} + 2\mathbf{x}^T\left(\left(\mathbf{A}^T\mathbf{A} - \mathbf{M}_t\right)\mathbf{x}_t - \mathbf{A}^T\mathbf{b}\right)$$

with

$$\mathbf{M}_t = \text{diag}\left(\frac{\left(\mathbf{A}^T\mathbf{A}\mathbf{x}_t\right)_1}{x_1^t}, \ldots, \frac{\left(\mathbf{A}^T\mathbf{A}\mathbf{x}_t\right)_n}{x_n^t}\right)$$

is a valid surrogate function since $\mathbf{M}_t \succeq \mathbf{A}^T\mathbf{A}$. The update of $\mathbf{x}$ is then given by

$$\mathbf{x}_{t+1} = \left(\mathbf{A}^T\mathbf{b}/\mathbf{A}^T\mathbf{A}\mathbf{x}_t\right) \odot \mathbf{x}_t. \tag{93}$$

Initialized at $\mathbf{x}_0 > \mathbf{0}$, we can see that $(\mathbf{x}_t)_{t\in\mathbb{N}}$ remains nonnegative if the elements of $\mathbf{A}$ and $\mathbf{b}$ are nonnegative. Although both derived based on separable quadratic surrogate functions, (92) is an additive update while (93) is multiplicative. Iteration (93) was studied in a more general context, namely as an instance of multiplicative iterative algorithms for convex problems, in [50].

We mention that another surrogate function can be constructed based on (17) and (28), whose derivation is postponed to Section V-D3.

C. Convexity Inequality

In this subsection we show the application of inequality (19) to a robust mean-covariance estimation problem formulated in [24] as

$$\begin{aligned} \underset{\boldsymbol{\mu},\mathbf{R}\succeq\mathbf{0}}{\text{minimize}} \quad & \frac{K+1}{N}\sum_{i=1}^N \log\left(1 + (\mathbf{x}_i - \boldsymbol{\mu})^H\mathbf{R}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right) \\ & + \alpha\left(\log\det(\mathbf{R}) + K\log\text{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right) \\ & + \gamma\log\left(1 + (\boldsymbol{\mu} - \mathbf{t})^H\mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{t})\right) + \log\det(\mathbf{R}). \end{aligned} \tag{94}$$

Similar to (50), Problem (94) can be solved by linearizing the log function. Here we provide an alternative solution by exploiting convexity.

Specifically, applying inequality (19) to the sum of the log terms, the objective function can be majorized by

$$\begin{aligned} & g(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t) = \\ & (1+\alpha)\log\det(\mathbf{R}) + (K + 1 + \gamma + \alpha K) \times \\ & \log\left(\sum_{i=1}^N \frac{K+1}{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t)\left(1 + (\mathbf{x}_i - \boldsymbol{\mu})^H\mathbf{R}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)\right. \\ & + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)\left(1 + (\boldsymbol{\mu} - \mathbf{t})^H\mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{t})\right) \\ & \left. + \frac{\alpha K}{\text{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}\text{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right). \end{aligned} \tag{95}$$

Proposition 25. *The surrogate function $g(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t)$ has a closed-form minimizer given by*

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &= \frac{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t, \mathbf{R}_t)\mathbf{x}_i + \gamma N w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)\mathbf{t}}{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t, \mathbf{R}_t) + \gamma N w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)} \\ \mathbf{R}_{t+1} &= \beta\mathbf{S}_t, \end{aligned} \tag{96}$$

where

$$\begin{aligned} \mathbf{S}_t = & \sum_{i=1}^N \frac{K+1}{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}\right)^H \\ & + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)\left(\boldsymbol{\mu}_{t+1} - \mathbf{t}\right)\left(\boldsymbol{\mu}_{t+1} - \mathbf{t}\right)^H + \frac{\alpha K}{\text{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}\mathbf{T}. \end{aligned} \tag{97}$$

and

$$\beta = \frac{1+\gamma}{1+\alpha}\left(\sum_{i=1}^N \frac{K+1}{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t) + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)\right)^{-1}. \tag{98}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

Proof: See Appendix. ∎

Note that the MM update (96) turns out to be the accelerated MM algorithm (without convergence proof) provided in [24]. Figure 4 and Table II in [24] show that the number of iterations required for algorithm (96) to converge is significantly smaller than MM derived based on linearization, which can be explained by the fact that the former algorithm has a tighter surrogate function (see Figure 3 for example).

D. Geometric and Signomial Programming

An unconstrained standard GP takes the form

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) \coloneqq \sum_{j=1}^{J} c_j \prod_{i=1}^{n} x_i^{a_{ij}}, \tag{99}$$

where $\forall i,j,\ a_{ij} \in \mathbb{R}$, $c_j > 0$, and $x_i > 0$. Function $c_j \prod_{i=1}^n x_i^{a_{ij}}$ is a monomial, and the sum of monomials is a posynomial. When some of the c_j's are negative, f is a signomial.

1) Signomial Programming: We apply inequality (28) to the summands of f with positive c_j's and inequality (17) to those with negative c_j's, which leads to the following surrogate function [54]:

$$g(\mathbf{x}|\mathbf{x}_t) = \sum_{i=1}^{n} g_i(x_i|\mathbf{x}_t)$$

$$g_i(x_i|\mathbf{x}_t) = \sum_{j:c_j>0} c_j \left(\prod_{k=1}^{n} \left(x_k^t\right)^{a_{kj}}\right) \frac{|a_{ij}|}{\|\mathbf{a}_j\|_1} \left(\frac{x_i}{x_i^t}\right)^{\|\mathbf{a}_j\|_1 \operatorname{sgn}(a_{ij})} + \sum_{j:c_j<0} c_j \left(\prod_{k=1}^{n} \left(x_k^t\right)^{a_{kj}}\right) a_{ij} \log x_i,$$

where $\mathbf{a}_j = [a_{1j}, a_{2j}, \ldots, a_{nj}]^T$. Surrogate function $g(\mathbf{x}|\mathbf{x}_t)$ is separable, and minimizing $g_i(x_i|\mathbf{x}_t)$ is a uni-variate optimization problem and can be done in parallel.

Having discussed how to minimize a signomial, we move to the problem of minimizing the ratio of two posynomials.

2) Complementary GP: Consider the following minimization problem:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \frac{f(\mathbf{x})}{g(\mathbf{x})}, \tag{100}$$

where f and g are posynomials. The idea is to lowerbound g by a monomial, so that the resulting surrogate function becomes a posynomial.

Write $g(\mathbf{x})$ as $g(\mathbf{x}) = \sum_{j=1}^{J} u_j(\mathbf{x})$, where u_j is a monomial. Invoking inequality (29), the objective function is majorized by

$$g(\mathbf{x}|\mathbf{x}_t) = f(\mathbf{x}) \Big/ \left(\prod_{j=1}^{J} \left(\frac{u_j(\mathbf{x})}{\alpha_j}\right)^{\alpha_j}\right),$$

where $\alpha_j = \frac{u_j(\mathbf{x}_t)}{\prod_{j=1}^{J} u_j(\mathbf{x}_t)}$. As a result, minimizing the surrogate function becomes a standard GP [53].

At this point, we can either solve the GP directly or further upperbound $g(\mathbf{x}|\mathbf{x}_t)$ by a separable surrogate function using the techniques in Section V-D1.

3) Nonnegative Least Squares Revisit: An alternative approach to find a separable surrogate function for NLS problem (91) hinges on inequalities (17) and (28) for monomials.

To lighten the notation, denote $\mathbf{A}^T\mathbf{A}$ by $\mathbf{Q}$ and $-\mathbf{A}^T\mathbf{b}$ by $\mathbf{q}$. Problem (91) can be written equivalently as

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x} \\ \text{subject to} & \mathbf{x} \geq \mathbf{0}. \end{array}$$

To find a separable surrogate function, we need to take care of the cross terms $Q_{ij}x_ix_j$.

Notice that $|Q_{ij}|\, x_ix_j$ is a monomial, and we have separable upper and lower bounds for a monomial given by inequalities (28) and (17), respectively. To be precise, for the terms x_ix_j with $Q_{ij} > 0$, we have

$$x_ix_j \leq \frac{1}{2}\left(\frac{x_j^t}{x_i^t}x_i^2 + \frac{x_i^t}{x_j^t}x_j^2\right), \tag{101}$$

and for the terms x_ix_j with $Q_{ij} < 0$, we have

$$x_ix_j \geq \left(x_i^tx_j^t\right)\left(1 + \log x_i + \log x_j - \log x_i^t - \log x_j^t\right). \tag{102}$$

Define $\mathbf{Q}^+ = \max[\mathbf{Q}, \mathbf{0}]$ and $\mathbf{Q}^- = -\min[\mathbf{Q}, \mathbf{0}]$, where the maximum and minimum are taken element-wise. The bounds (102) and (101) lead to the following surrogate function:

$$\begin{aligned} &g(\mathbf{x}|\mathbf{x}_t) \\ =&\frac{1}{2}\sum_i \frac{\left(\mathbf{Q}^+\mathbf{x}_t\right)_i}{x_i^t}x_i^2 - \sum_i x_i\left(\mathbf{Q}^-\mathbf{x}_t\right)_i \log x_i + \sum_i q_ix_i. \end{aligned}$$

Setting its gradient to zero gives the multiplicative update [55]

$$x_i^{t+1} = x_i^t\left(\frac{-q_i + \sqrt{q_i^2 + 4\left(\mathbf{Q}^+\mathbf{x}_t\right)_i\left(\mathbf{Q}^-\mathbf{x}_t\right)_i}}{2\left(\mathbf{Q}^+\mathbf{x}_t\right)_i}\right). \tag{103}$$

Remark 26. If $\mathbf{A} \in \mathbb{R}_{++}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}_+^m$, $\mathbf{b} \neq \mathbf{0}$, we can see that $\mathbf{Q} \in \mathbb{R}_{++}^{n\times n}$ and $-\mathbf{q} \in \mathbb{R}_{++}^n$. In this case, iteration (103) coincides with (93).

Figure 8 shows the performance of MM iterations (92), (93), and (103). We have also included the accelerated algorithm (92) by modifying the step-size according to the Armijo line search rule. $\mathbf{A}$ and $\mathbf{x}$ are generated randomly so that each of the elements follows a uniform distribution in $[0, 1]$, and the dimensions are set to be $m = 60$ and $n = 100$, and we assume the noiseless case, i.e., $\mathbf{b} = \mathbf{A}\mathbf{x}$.

Remark 27. Combining with the techniques for sparse linear regression in Section V-B4, an alternating MM algorithm can be derived for the matrix factorization problem, possibly with the nonnegativity constraint and sparsity penalty. We refer the readers to [27] and [132]–[134] for the details.

E. Cauchy-Schwartz Inequality

Cauchy-Schwartz Inequality can be used to lower bound the ℓ_2-norm by a linear function, which is applied in the following applications.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

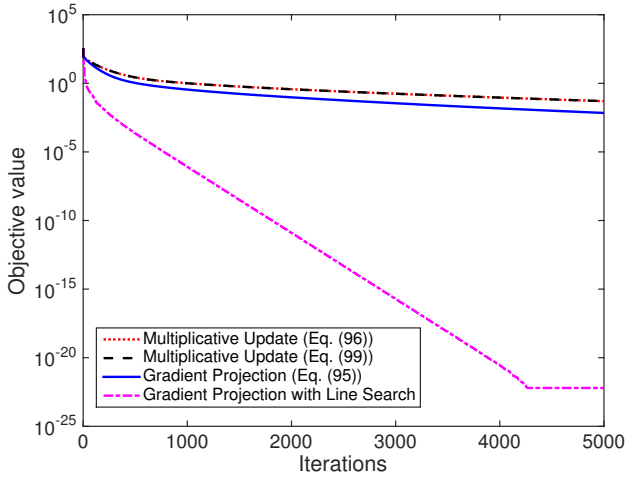

Figure 8. Objective value evolution curve of MM algorithms for the NLS problem. Red: algorithm (93), black: algorithm (103), blue: algorithm (92), magenta: accelerated algorithm (92).

1) Phase Retrieval Revisit: The phase retrieval problem considered in Section V-B3 can be alternatively formulated by magnitude matching as [56], [135]–[137]

$$\underset{\mathbf{x}}{\text{minimize}} \quad \left\| \sqrt{\mathbf{y}} - \left| \mathbf{A}^H \mathbf{x} \right| \right\|_2^2, \tag{104}$$

where $\sqrt{\cdot}$ is applied element-wise.

Expanding the squares and applying inequality (31) to the cross term (assuming $\left|\mathbf{A}^H \mathbf{x}_t\right| \neq 0$) leads to the following surrogate function:

$$g^{(1)}\left(\mathbf{x}|\mathbf{x}_t\right) = \left\| \mathbf{C}_t \sqrt{\mathbf{y}} - \mathbf{A}^H \mathbf{x} \right\|_2^2,$$

where $\mathbf{C}_t = \operatorname{diag}\left(e^{j \arg\left(\mathbf{A}^H \mathbf{x}_t\right)}\right)$, which has a minimizer

$$\mathbf{x}_{t+1} = \left(\mathbf{A}\mathbf{A}^H\right)^{-1} \mathbf{A}\mathbf{C}_t \sqrt{\mathbf{y}}. \tag{105}$$

Algorithm (105) turns out to be the famous Gerchberg-Saxton algorithm [56].

Restricting $\mathbf{x}$ to be real-valued, we can further majorize $g^{(1)}\left(\cdot|\mathbf{x}_t\right)$ using inequality (26) [57].

Let us write $\left\|\mathbf{A}^H \mathbf{x}\right\|^2$ as $\left\|\mathbf{A}^H \mathbf{x}\right\|^2 = \sum_{i=1}^N \left|\mathbf{a}_i^H \mathbf{x}\right|^2 = \sum_{i=1}^N \left|\sum_{j=1}^p a_{ij} x_j\right|^2$, where a_{ij} is the j-th element of $\mathbf{a}_i^H$ and p is the dimension of $\mathbf{x}$. For simplicity we assume that $a_{ij} \neq 0$. Applying Jensen's inequality we have

$$\begin{aligned}
&\left|\sum_{j=1}^p a_{ij} x_j\right|^2 \\
=&\left(\sum_{j=1}^p \operatorname{Re}\left(a_{ij}\right) x_j\right)^2 + \left(\sum_{j=1}^p \operatorname{Im}\left(a_{ij}\right) x_j\right)^2 \\
=&\left(\sum_{j=1}^p V_R^{ij} \frac{\operatorname{Re}\left(a_{ij}\right)}{V_R^{ij}} x_j\right)^2 + \left(\sum_{j=1}^p V_I^{ij} \frac{\operatorname{Im}\left(a_{ij}\right)}{V_I^{ij}} x_j\right)^2 \\
\leq& \sum_{j=1}^p \frac{\operatorname{Re}\left(a_{ij}\right)^2}{V_R^{ij}} x_j^2 + \sum_{j=1}^p \frac{\operatorname{Im}\left(a_{ij}\right)^2}{V_I^{ij}} x_j^2,
\end{aligned} \tag{106}$$

where $V_R^{ij} = \frac{|\operatorname{Re}(a_{ij})|}{\sum_{j'=1}^p |\operatorname{Re}(a_{ij'})|}$ and $V_I^{ij} = \frac{|\operatorname{Im}(a_{ij})|}{\sum_{j'=1}^p |\operatorname{Im}(a_{ij'})|}$.

Summing Eq. (106) over indices i we have

$$\begin{aligned}
&\left\|\mathbf{A}^H \mathbf{x}\right\|^2 \\
\leq& \sum_{i=1}^N \left(\sum_{j=1}^p \frac{\operatorname{Re}\left(a_{ij}\right)^2}{V_R^{ij}} x_j^2 + \sum_{j=1}^p \frac{\operatorname{Im}\left(a_{ij}\right)^2}{V_I^{ij}} x_j^2 \right) \\
=& \sum_{i=1}^N \left(\sum_{j=1}^p \left(\left|\operatorname{Re}\left(a_{ij}\right)\right| \sum_{j'=1}^p \left|\operatorname{Re}\left(a_{ij'}\right)\right| \right) x_j^2 \right. \\
&\left. + \sum_{j=1}^p \left(\left|\operatorname{Im}\left(a_{ij}\right)\right| \sum_{j'=1}^p \left|\operatorname{Im}\left(a_{ij'}\right)\right| \right) x_j^2 \right) \\
:=& \mathbf{x}^H \mathbf{M} \mathbf{x},
\end{aligned}$$

where $\mathbf{M}$ is a diagonal matrix with its j-th diagonal entry being

$$\sum_{i=1}^N \left(\left|\operatorname{Re}\left(a_{ij}\right)\right| \sum_{j'=1}^p \left|\operatorname{Re}\left(a_{ij'}\right)\right| + \left|\operatorname{Im}\left(a_{ij}\right)\right| \sum_{j'=1}^p \left|\operatorname{Im}\left(a_{ij'}\right)\right| \right).$$

2) Sensor Network Localization: We introduce the localization problem described in [138], where a sensor network is modeled by a graph $G\left(V, E \cup \bar{E}\right)$. The nodes V are partitioned into a set of m anchor nodes $V_a = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ with known location, and the rest $V_x = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are n sensors with unknown location. An edge in the set $E = \{(i,j) \,|\, i,j \in V_x\}$ is associated with d_{ij} representing the distance between sensors i and j, and an edge in the set $\bar{E} = \{(k,j) \,|\, k \in V_a, j \in V_x\}$ is associated with $\bar{d}_{kj}$ representing the distance between anchor k and node j.

To estimate the location of all nodes, we formulate the problem as

$$\min_{\{\mathbf{x}_i\}_{i=1}^n} \sum_{(i,j)\in E} \left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2 - d_{ij}\right)^2 + \sum_{(k,j)\in \bar{E}} \left(\left\|\mathbf{x}_j - \mathbf{a}_k\right\|_2 - \bar{d}_{kj}\right)^2. \tag{107}$$

Expanding the squares we can see that the cross terms are concave and thus destroy the convexity of the objective function.

Invoking inequality (32), we can get the following quadratic surrogate function [58]–[61]:

$$\begin{aligned}
&g\left(\{\mathbf{x}_i\}_{i=1}^n \,|\, \{\mathbf{x}_i^t\}_{i=1}^n\right) \\
=& \sum_{(i,j)\in E} \left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 - 2 d_{ij} \frac{\left(\mathbf{x}_i^t - \mathbf{x}_j^t\right)^T \left(\mathbf{x}_i - \mathbf{x}_j\right)}{\left\|\mathbf{x}_i^t - \mathbf{x}_j^t\right\|_2} \right) \\
&+ \sum_{(i,j)\in \bar{E}} \left(\left\|\mathbf{x}_j - \mathbf{a}_k\right\|_2^2 - 2 \bar{d}_{kj} \frac{\left(\mathbf{x}_j^t - \mathbf{a}_k\right)^T \left(\mathbf{x}_j - \mathbf{a}_k\right)}{\left\|\mathbf{x}_j^t - \mathbf{a}_k\right\|_2} \right),
\end{aligned}$$

which has a closed-form minimizer.

Problem (107) has many variants. For example, to achieve robustness, the authors of [59] replaced the squared loss function by the ℓ_1-norm loss function, and the authors of [139] employed the Huber's loss function. In both cases, inequality (16) has been applied together with (32) to arrive at a quadratic surrogate function.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2601299, IEEE Transactions on Signal Processing

F. Schur Complement

We revisit the variance component model problem (53) and show that an alternative MM algorithm can be derived based on inequality (34).

Define $\tilde{\mathbf{A}} = [\mathbf{A}; \mathbf{I}_K]$ and

$$\tilde{\mathbf{P}} = \operatorname{diag}(p_1, \ldots, p_L, \underbrace{\sigma^2, \ldots, \sigma^2}_{K}),$$

then $\mathbf{R} = \tilde{\mathbf{A}}\tilde{\mathbf{P}}\tilde{\mathbf{A}}^H$ and the objective function can be rewritten as

$$L(\mathbf{P}) = \log\det\left(\tilde{\mathbf{A}}\tilde{\mathbf{P}}\tilde{\mathbf{A}}^H\right) + \operatorname{Tr}\left(\mathbf{S}\left(\tilde{\mathbf{A}}\tilde{\mathbf{P}}\tilde{\mathbf{A}}^H\right)^{-1}\right).$$

To find a surrogate function separable in $\mathbf{P}$, we apply inequality (12) to the first term, which leads to the first step majorization with surrogate function

$$\begin{aligned} g^{(1)}(\mathbf{R}|\mathbf{R}_t) &= \operatorname{Tr}\left(\mathbf{R}_t^{-1}\mathbf{R}\right) + \operatorname{Tr}\left(\mathbf{S}\mathbf{R}^{-1}\right) \\ &\triangleq \mathbf{w}_t^H\mathbf{p} + \operatorname{Tr}\left(\mathbf{S}\left(\tilde{\mathbf{A}}\tilde{\mathbf{P}}\tilde{\mathbf{A}}^H\right)^{-1}\right), \end{aligned}$$

where $w_j^t = \mathbf{a}_j^H\mathbf{R}_t^{-1}\mathbf{a}_j$ with $\mathbf{a}_j$ being the j-th column of $\mathbf{A}$.

In the next step, we find a separable upperbound for $\operatorname{Tr}\left(\mathbf{S}\left(\tilde{\mathbf{A}}\tilde{\mathbf{P}}\tilde{\mathbf{A}}^H\right)^{-1}\right)$. By inequality (34) we have

$$\begin{aligned} &g^{(1)}(\mathbf{R}|\mathbf{R}_t) \le g^{(2)}(\mathbf{R}|\mathbf{R}_t) \\ &= \mathbf{w}_t^H\mathbf{p} + \operatorname{Tr}\left(\tilde{\mathbf{P}}_t\mathbf{A}^H\mathbf{R}_t^{-1}\mathbf{M}_t\mathbf{R}_t^{-1}\mathbf{A}\tilde{\mathbf{P}}_t\tilde{\mathbf{P}}^{-1}\right) \\ &= \mathbf{w}_t^H\mathbf{p} + \sum_{j=1}^{L}\left(p_j^t\right)^2\mathbf{a}_j^H\mathbf{R}_t^{-1}\mathbf{M}_t\mathbf{R}_t^{-1}\mathbf{a}_j p_j^{-1} + \text{const.} \end{aligned}$$

The update of p_j can be obtained in closed-form as

$$p_j^{t+1} = \sqrt{\frac{\mathbf{a}_j^H\mathbf{R}_t^{-1}\mathbf{M}_t\mathbf{R}_t^{-1}\mathbf{a}_j}{\mathbf{a}_j^H\mathbf{R}_t^{-1}\mathbf{a}_j}}p_j^{t+1}. \tag{108}$$

Iteration (108) is similar to the LIKES algorithm presented in [93], but executes the outer loop iteration only once. Detailed numerical comparisons between the SBL and LIKES algorithm can be found in [140].

VI. Conclusions

In this overview, we have presented the MM principle and its recent developments. From a theoretical perspective, we have introduced the general algorithm framework, its convergence conditions, as well as acceleration schemes. We have also related MM to several algorithm frameworks, namely EM, cyclic minimization algorithms, CCCP, proximal minimization, VMFB, SCA, and subspace MM. More importantly, a large part of the article has been devoted to presenting the techniques of constructing surrogate functions and applying MM to problems in signal processing, communications, and machine learning. A wide range of applications have been covered in this overview such as sparse regression, matrix completion, phase retrieval, sparse PCA, covariance estimation, sequence design, and sensor network localization. In the end, we mention that although MM has been proven to be an effective tool for many applications, practitioners should also be aware of the following issues. One is that MM algorithms can get stuck at stationary points for nonconvex problems, therefore the performance of the convergent point (whether it satisfies application design criterion) should be studied either theoretically or empirically. Another problem is that MM can suffer from a slow convergence rate. In this situation, either the surrogate function should be tightened, or an MM accelerator needs to be employed (at the cost of losing convergence guarantees).

Acknowledgement

The authors would like to thank professor Wing-Kin (Ken) Ma for his comments on the connection between the MM algorithm and the cyclic minimization algorithm.

Appendix A
Proof of Proposition 25

To find a minimizer of the surrogate function (95), we first set the gradient of $g(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t)$ with respect to $\boldsymbol{\mu}$ to zero, which leads to the minimizer

$$\boldsymbol{\mu}_{t+1} = \frac{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t, \mathbf{R}_t)\mathbf{x}_i + \gamma N w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)\mathbf{t}}{(K+1)\sum_{i=1}^N w_i(\boldsymbol{\mu}_t, \mathbf{R}_t) + \gamma N w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)}.$$

Substituting the optimal $\boldsymbol{\mu}$ back into $g(\boldsymbol{\mu}, \mathbf{R}|\boldsymbol{\mu}_t, \mathbf{R}_t)$ and setting the gradient of it with respect to $\mathbf{R}$ to zero leads to the fixed-point equation

$$\mathbf{R} = \frac{(K + (1+\gamma)/(1+\alpha))\mathbf{S}_t}{\left(\sum_{i=1}^N \frac{K+1}{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t) + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)\right) + \operatorname{Tr}(\mathbf{S}_t\mathbf{R}^{-1})}, \tag{109}$$

where $\mathbf{S}_t$ is given by (97). Similar to the proof of Theorem 10 in [23], it can be shown by contradiction that if (109) has a solution, it is unique.

Since equation (109) indicates that $\hat{\mathbf{R}}$ should be proportional to $\mathbf{S}_t$, we let the solution $\hat{\mathbf{R}}$ be $\beta\mathbf{S}_t$. To get the value of β, we substitute $\hat{\mathbf{R}}$ back into (109), which leads to the following equation of β:

$$\begin{aligned} &\frac{K+1+\gamma+\alpha K}{(1+\alpha)\beta} \\ &= \left(\sum_{i=1}^N \frac{K+1}{N} w_i(\boldsymbol{\mu}_t, \mathbf{R}_t) + \gamma w_{\mathbf{t}}(\boldsymbol{\mu}_t, \mathbf{R}_t)\right) + K\beta^{-1}. \end{aligned} \tag{110}$$

The solution of (110) is given by (98).

References

[1] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.

[2] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, 1970, vol. 30.

[3] T. T. Wu and K. Lange, "The MM alternative to EM," *Statistical Science*, vol. 25, no. 4, pp. 492–505, 2010.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[5] J. De Leeuw, "Convergence of the majorization method for multidimensional scaling," *Journal of Classification*, vol. 5, no. 2, pp. 163–180, 1988.

[6] K. Lange and J. A. Fessler, "Globally convergent algorithms for maximum a posteriori transmission tomography," *IEEE Transactions on Image Processing*, vol. 4, no. 10, pp. 1430–1438, 1995.

[7] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 3, pp. 367–383, 1992.

[8] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[9] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography." *IEEE Transactions on Medical Imaging*, vol. 14, no. 1, pp. 132–137, 1994.

[10] M. P. Becker, I. Yang, and K. Lange, "EM algorithms without missing data," *Statistical Methods in Medical Research*, vol. 6, no. 1, pp. 38–54, 1997.

[11] W. J. Heiser, "Convergent computation by iterative majorization: theory and applications in multidimensional data analysis," *Recent Advances in Descriptive Multivariate Analysis*, pp. 157–189, 1995.

[12] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of computational and graphical statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[13] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[14] T. Blumensath, M. Yaghoobi, and M. E. Davies, "Iterative hard thresholding and l_0 regularisation," in *Proceedings of the 2007 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 3. IEEE, 2007, pp. III–877.

[15] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[16] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[17] G. Marjanovic, M. O. Ulfarsson, and A. O. Hero III, "Mist: l_0 sparse linear regression with momentum," *arXiv preprint arXiv:1409.7193*, 2014.

[18] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[19] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 899–925, 2013.

[20] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Machine Learning*, vol. 85, no. 1-2, pp. 3–39, 2011.

[21] J. Song, P. Babu, and D. P. Palomar, "Sparse generalized eigenvalue problem via smooth optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 7, pp. 1627–1642, 2015.

[22] A. Wiesel, "Unified framework to regularized covariance estimation in scaled Gaussian models," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 29–38, 2012.

[23] Y. Sun, P. Babu, and D. P. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5143–5156, 2014.

[24] ——, "Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3096–3109, June 2015.

[25] ——, "Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions," *arXiv preprint arXiv:1506.05215*, 2015.

[26] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Transactions on Signal Processing*, vol. 57, no. 6, pp. 2178–2191, 2009.

[27] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 1980–1983.

[28] J. M. Bioucas-Dias, M. A. Figueiredo, and J. P. Oliveira, "Total variation-based image deconvolution: a majorization-minimization approach," in *Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 2006, pp. II–II.

[29] M. A. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2980–2991, 2007.

[30] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[31] J. Song, P. Babu, and D. P. Palomar, "Optimization methods for designing sequences with low autocorrelation sidelobes," *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 3998–4009, Aug 2015.

[32] ——, "Sequence design to minimize the weighted integrated and peak sidelobe levels," *arXiv preprint arXiv:1506.04234*, 2015.

[33] D. Tyler, "A distribution-free M-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[34] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[35] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[36] Y. Sun, A. Breloy, P. Babu, D. P. Palomar, F. Pascal, and G. Ginolhac, "Low-complexity algorithms for low rank clutter parameters estimation in radar systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1986–1998, April 2016.

[37] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *The Journal of Machine Learning Research*, vol. 11, pp. 517–553, 2010.

[38] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1130–1142, 2006.

[39] H. W. Kuhn, "A note on Fermat's problem," *Mathematical Programming*, vol. 4, no. 1, pp. 98–107, 1973.

[40] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, 1974.

[41] J. J. Fuchs, "Convergence of a sparse representations algorithm applicable to real or complex data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 598–605, 2007.

[42] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[43] 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.

[44] D. Böhning, "Multinomial logistic regression algorithm," *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 197–200, 1992.

[45] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[46] J. T. Chi and E. C. Chi, "Getting to the bottom of matrix completion and nonnegative least squares with the mm algorithm," StatisticsViews.com, March 2014.

[47] T. Qiu, P. Babu, and D. P. Palomar, "PRIME: Phase retrieval via majorization-minimization," *arXiv preprint arXiv:1511.01669*, 2015.

[48] M. A. Figueiredo and R. D. Nowak, "A bound optimization approach to wavelet-based image deconvolution," in *Proceedings of the 2005 IEEE International Conference on Image Processing (ICIP)*, vol. 2. IEEE, 2005, pp. II–782.

[49] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[50] P. Eggermont, "Multiplicative iterative algorithms for convex programming," *Linear Algebra and its Applications*, vol. 130, pp. 25–42, 1990.

[51] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[52] K. Lange, E. C. Chi, and H. Zhou, "A brief survey of modern optimization for statisticians," *International Statistical Review*, vol. 82, no. 1, pp. 46–70, 2014.

[53] M. Chiang, C. W. Tan, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, pp. 2640–2651, July 2007.

[54] K. Lange and H. Zhou, "MM algorithms for geometric and signomial programming," *Mathematical Programming*, vol. 143, no. 1-2, pp. 339–356, 2014.

[55] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," in *Advances in Neural Information Processing Systems*, 2002, pp. 1041–1048.

[56] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, p. 237, 1972.

[57] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "A block coordinate variable metric forward-backward algorithm," *Journal of Global Optimization*, pp. 1–29, 2013.

[58] J. A. Costa, N. Patwari, and A. O. Hero III, "Distributed weighted-multidimensional scaling for node localization in sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 1, pp. 39–64, 2006.

[59] P. Oğuz-Ekim, J. P. Gomes, J. Xavier, and P. Oliveira, "Robust localization of nodes and time-recursive tracking in sensor networks using noisy range measurements," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3930–3942, 2011.

[60] A. Beck, M. Teboulle, and Z. Chikishev, "Iterative minimization schemes for solving the single source localization problem," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1397–1416, 2008.

[61] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery problems," in *Convex Optimization in Signal Processing and Communications*, D. P. Palomar and Y. C. Eldar, Eds. Cambridge, UK: Cambridge University Press, 2010, ch. 2.

[62] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 629–638, 2011.

[63] H. Zhou, L. Hu, J. Zhou, and K. Lange, "MM algorithms for variance components models," *arXiv preprint arXiv:1509.07426*, 2015.

[64] M. Jamshidian and R. I. Jennrich, "Conjugate gradient acceleration of the EM algorithm," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 221–228, 1993.

[65] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 425–437, 1995.

[66] ——, "A quasi-Newton acceleration of the EM algorithm," *Statistica sinica*, vol. 5, no. 1, pp. 1–18, 1995.

[67] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2664–2677, 1994.

[68] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

[69] M. W. Jacobson and J. A. Fessler, "An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2411–2422, 2007.

[70] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[71] J. Mairal, "Incremental majorization-minimization optimization with application to large-scale machine learning," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 829–855, 2015.

[72] ——, "Optimization with first-order surrogate functions," *arXiv preprint arXiv:1305.3120*, 2013.

[73] J. Bolte and E. Pauwels, "Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs," *Mathematics of Operations Research*, 2016.

[74] C. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, pp. 95–103, 1983.

[75] F. Vaida, "Parameter convergence for EM and MM algorithms," *Statistica Sinica*, pp. 831–840, 2005.

[76] T. A. Louis, "Finding the observed information matrix when using the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 226–233, 1982.

[77] I. Meilijson, "A fast improvement to the EM algorithm on its own terms," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 127–138, 1989.

[78] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proceedings of Conference on Information Sciences and Systems*, 1993, pp. 611–616.

[79] R. M. Lewitt and G. Muehllehner, "Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation," *IEEE Transactions on Medical Imaging*, vol. 5, no. 1, pp. 16–22, 1986.

[80] D. P. Helmbold, R. E. Schapire, Y. Singer, and M. K. Warmuth, "A comparison of new and old algorithms for a mixture estimation problem," *Machine Learning*, vol. 27, no. 1, pp. 97–119, 1997.

[81] E. Bauer, D. Koller, and Y. Singer, "Update rules for parameter estimation in bayesian networks," in *Proceedings of the Thirteenth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1997, pp. 3–13.

[82] R. Salakhutdinov and S. Roweis, "Adaptive overrelaxed bound optimization methods," in *Proceedings of the International Conference on Machine Learning*, 2003, pp. 664–671.

[83] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[84] N. Laird, N. Lange, and D. Stram, "Maximum likelihood computations with repeated measures: application of the EM algorithm," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 97–105, 1987.

[85] M. Jamshidian and R. I. Jennrich, "Acceleration of the EM algorithm by using quasi-Newton methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 3, pp. 569–587, 1997.

[86] R. Varadhan and C. Roland, "Simple and globally convergent methods for accelerating the convergence of any EM algorithm," *Scandinavian Journal of Statistics*, vol. 35, no. 2, pp. 335–353, 2008.

[87] H. Zhou, D. Alexander, and K. Lange, "A quasi-Newton acceleration for high-dimensional optimization algorithms," *Statistics and computing*, vol. 21, no. 2, pp. 261–273, 2011.

[88] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 1995.

[89] D. P. Bertsekas, *Nonlinear programming*. Athena scientific, 1999.

[90] A. Hjørungnes, *Complex-valued matrix derivatives: with applications in signal processing and communications*. Cambridge University Press, 2011.

[91] S. Chrétien and A. O. Hero III, "Kullback proximal algorithms for maximum-likelihood estimation," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1800–1810, 2000.

[92] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*. Springer, 1998, pp. 355–368.

[93] P. Stoica and P. Babu, "SPICE and LIKES: Two hyperparameter-free methods for sparse-parameter estimation," *Signal Processing*, vol. 92, no. 7, pp. 1580–1590, 2012.

[94] M. M. Naghsh, M. Soltanalian, and M. Modarres-Hashemi, "Radar code design for detection of moving targets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 4, pp. 2762–2778, 2014.

[95] P. Stoica, H. He, and J. Li, "New algorithms for designing unimodular sequences with good correlation properties," *IEEE Transactions on Signal Processing*, vol. 57, no. 4, pp. 1415–1425, 2009.

[96] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.

[97] R. Horst and N. V. Thoai, "DC programming: overview," *Journal of Optimization Theory and Applications*, vol. 103, no. 1, pp. 1–43, 1999.

[98] P. D. Tao *et al.*, "The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems," *Annals of Operations Research*, vol. 133, no. 1-4, pp. 23–46, 2005.

[99] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," *Advances in Neural Information Processing Systems*, vol. 2, pp. 1033–1040, 2002.

[100] T. Lipp and S. Boyd, "Variations and extension of the convex-concave procedure," *Optimization and Engineering*, pp. 1–25, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11081-015-9294-x

[101] T. D. Quoc and M. Diehl, "Sequential convex programming methods for solving nonlinear optimization problems with DC constraints," *arXiv preprint arXiv:1107.5841*, 2011.

[102] D. P. Bertsekas and P. Tseng, "Partial proximal minimization algorithms for convex pprogramming," *SIAM Journal on Optimization*, vol. 4, no. 3, pp. 551–572, 1994.

[103] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[104] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2011, pp. 185–212.

[105] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function," *Journal of Optimization Theory and Applications*, vol. 162, no. 1, pp. 107–132, 2014.

[106] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet, "Euclid in a taxicab: Sparse blind deconvolution with smoothed regularization," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 539–543, 2015.

[107] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 641–656, 2014.

[108] F. Facchinei, S. Sagratella, and G. Scutari, "Flexible parallel algorithms for big data optimization," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7208–7212.

[109] F. Facchinei, G. Scutari, and S. Sagratella, "Parallel selective algorithms for nonconvex big data optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 7, pp. 1874–1889, 2015.

[110] Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento, "A parallel decomposition method for nonconvex stochastic multi-agent optimization problems," *IEEE Transactions on Signal Processing*, vol. 64, no. 11, pp. 2949–2964, June 2016.

[111] B. R. Marks and G. P. Wright, "Technical note-A general inner approximation algorithm for nonconvex mathematical programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, 1978.

[112] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Distributed methods for constrained nonconvex multi-agent optimization-Part I: Theory," *arXiv preprint arXiv:1410.4754*, 2014.

[113] Q. T. Dinh and M. Diehl, "Local convergence of sequential convex programming for nonconvex optimization," in *Recent Advances in Optimization and its Applications in Engineering*. Springer, 2010, pp. 93–102.

[114] F. Palacios-Gomez, L. Lasdon, and M. Engquist, "Nonlinear optimization by successive linear programming," *Management Science*, vol. 28, no. 10, pp. 1106–1120, 1982.

[115] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[116] C. Labat and J. Idier, "Convergence of conjugate gradient methods with a closed-form stepsize formula," *Journal of Optimization Theory and Applications*, vol. 136, no. 1, pp. 43–60, 2008.

[117] E. Chouzenoux, S. Moussaoui, and J. Idier, "Majorize-minimize linesearch for inversion methods involving barrier function optimization," *Inverse Problems*, vol. 28, no. 6, p. 065011, 2012.

[118] E. Chouzenoux, J. Idier, and S. Moussaoui, "A majorize-minimize strategy for subspace optimization applied to image restoration," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1517–1528, 2011.

[119] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, pp. 563–591, 2013.

[120] A. Florescu, E. Chouzenoux, J.-C. Pesquet, P. Ciuciu, and S. Ciochina, "A majorize-minimize memory gradient method for complex-valued inverse problems," *Signal Processing*, vol. 103, pp. 285–295, 2014.

[121] E. Chouzenoux and J.-C. Pesquet, "A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation," *arXiv preprint arXiv:1512.08722*, 2015.

[122] ——, "Convergence rate analysis of the majorize-minimize subspace algorithm," *arXiv preprint arXiv:1603.07301*, 2016.

[123] I. Markovsky, "Structured low-rank approximation and its applications," *Automatica*, vol. 44, no. 4, pp. 891–909, 2008.

[124] J.-J. Xiao, S. Cui, Z.-Q. Luo, and A. J. Goldsmith, "Linear coherent decentralized estimation," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 757–770, 2008.

[125] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[126] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007.

[127] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, p. kxp008, 2009.

[128] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 298–311, 1997.

[129] D. Böhning and B. G. Lindsay, "Monotonicity of quadratic-approximation algorithms," *Annals of the Institute of Statistical Mathematics*, vol. 40, no. 4, pp. 641–663, 1988.

[130] J. Song, P. Babu, and D. P. Palomar, "Sequence set design with good correlation properties via majorization-minimization," *arXiv preprint arXiv:1510.01899*, 2015.

[131] A. R. De Pierro, "On the convergence of the iterative image space reconstruction algorithm for volume ect," *IEEE Transactions on Medical Imaging*, vol. 6, no. 2, pp. 174–175, June 1987.

[132] J. D. Lee, B. Recht, N. Srebro, J. Tropp, and R. R. Salakhutdinov, "Practical large-scale optimization for max-norm regularization," in *Advances in Neural Information Processing Systems*, 2010, pp. 1297–1305.

[133] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 366–373.

[134] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[135] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[136] J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," *Optics letters*, vol. 3, no. 1, pp. 27–29, 1978.

[137] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4814–4826, 2015.

[138] D. Shamsi, N. Taheri, Z. Zhu, and Y. Ye, "Conditions for correct sensor network localization using SDP relaxation," in *Discrete Geometry and Optimization*. Springer, 2013, pp. 279–301.

[139] S. Korkmaz and A.-J. Van der Veen, "Robust localization in sensor networkswith iterative majorization techniques," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 2049–2052.

[140] P. Babu and P. Stoica, "Sparse spectral-line estimation for nonuniformly sampled multivariate time series: SPICE, LIKES and MSBL," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. Bucharest: IEEE, Aug. 2012, pp. 445–449.