NIH Public Access
Author Manuscript

Distance majorization and its applications

Eric C. Chi,
Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

Hua Zhou, and
Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA
hua_zhou@ncsu.edu

Kenneth Lange
Departments of Biomathematics, Human Genetics, and Statistics, University of California, Los Angeles, CA 90095, USA klange@ucla.edu

Abstract

The problem of minimizing a continuously differentiable convex function over an intersection of closed convex sets is ubiquitous in applied mathematics. It is particularly interesting when it is easy to project onto each separate set, but nontrivial to project onto their intersection. Algorithms based on Newton's method such as the interior point method are viable for small to medium-scale problems. However, modern applications in statistics, engineering, and machine learning are posing problems with potentially tens of thousands of parameters or more. We revisit this convex programming problem and propose an algorithm that scales well with dimensionality. Our proposal is an instance of a sequential unconstrained minimization technique and revolves around three ideas: the majorization-minimization principle, the classical penalty method for constrained optimization, and quasi-Newton acceleration of fixed-point algorithms. The performance of our distance majorization algorithms is illustrated in several applications.

Keywords

Constrained optimization; Majorization-minimization (MM); Sequential unconstrained minimization; Projection

1 Introduction

A wide spectrum of problems in applied mathematics and statistics can be formulated as an instance of the convex programming problem

$$\min_{x \in \cap_i C_i} \ell(\boldsymbol{x}), \quad (1)$$

where $\ell(\boldsymbol{x})$ is continuously differentiable and convex and the C_i are closed convex sets in $\mathbb{R}^{\rho}$. At one extreme, problem (1) includes classical least squares. At the other, it includes

finding a feasible point in the intersection of several closed convex sets. In between, the formulation covers a variety of shape restricted regression problems such as fitting a support vector machine (SVM) and projecting an exterior point onto a complicated convex set. Great progress has been made in attacking specific incarnations of problem (1). The projected gradient algorithm and its Newton and quasi-Newton extensions have been very successful when constraints are simple, for example box constraints, and admit a correspondingly simple projection operator [5,28,42,44]. However, there is still room for improvement. In the current paper we present a unified approach to solving a smoothed relaxation of problem (1) via the majorization-minimization (MM) principle [29]. This approach is especially attractive when it is easy to project onto each separate set C_i but nontrivial to project onto their intersection.

Problem (1) can be written as the unconstrained optimization problem

$$\min_{\boldsymbol{x}} \ell(\boldsymbol{x}) + \sum_i \delta_{C_i}(\boldsymbol{x}), \quad (2)$$

where the indicator function $\delta_C(\boldsymbol{x})$ equals 0 if $\boldsymbol{x} \in C$ and ∞ if not. Although problem (2) is now unconstrained, the indicator functions introduce two challenges. The new objective function takes on infinite values and is non-differentiable. This prompts us to replace $\delta_C(\boldsymbol{x})$ by a finite valued smooth approximation $dist(\boldsymbol{x}, C)^2 = \inf_{y \in C} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, where $\|\cdot\|_2$ denotes the standard Euclidean norm. Further progress can be made by solving the related problem

$$\min_{\boldsymbol{x}} \ell(\boldsymbol{x}) + \frac{\mu}{2} \sum_{i=1}^{m} dist(\boldsymbol{x}, C_i)^2, \quad (3)$$

where μ is a positive parameter that penalizes deviation from the original feasible region. The smooth approximation introduced in formulation (3) is an example of the quadratic penalty method [4,38,43]. Problem (3) has many appealing features. The problem is unconstrained with an objective function that is convex and differentiable when $\ell(\boldsymbol{x})$ is convex and differentiable. Consequently optimality conditions can be readily identified. The distance function is closely tied to the projection $P_C(\boldsymbol{x})$ of $\boldsymbol{x}$ onto C; specifically $dist(\boldsymbol{x}, C) = \|\boldsymbol{x} - P_C(\boldsymbol{x})\|_2$, and $\triangledown dist(\boldsymbol{x}, C)^2 = 2[\boldsymbol{x} - P_C(\boldsymbol{x})]$. Thus, a point $\boldsymbol{x}$ solves problem (3) if and only if it satisfies the stationarity condition

$$0 = \nabla \ell(\boldsymbol{x}) + \mu \sum_i \left[\boldsymbol{x} - P_{C_i}(\boldsymbol{x}) \right].$$

Of course finding such an $\boldsymbol{x}$ is often analytically intractable due to the projection term. To solve (3) iteratively, we resort to the MM principle. Because we rely on majorizing $dist(\boldsymbol{x}, C)$, we call our approach distance majorization. A key step in solving the subproblems will be calculating projection operators. Fortunately, many useful projection operators are easy to compute. The best known examples include projection onto: (a) a closed Euclidean ball, (b) a closed rectangle, (c) a hyperplane, (d) a closed halfspace, (e) a vector subspace, (f) the set of positive semidefinite matrices, (g) the unit simplex, (h) a closed ℓ_1 ball, and (i) an isotone

convex cone. While there are no analytic solutions for the last three projections, there are efficient algorithms for computing them [1,20,33,40,46].

The rest of the paper is organized as follows. After a brief review of the MM principle and its place among related iterative minimization schemes, we illustrate the virtues of distance majorization in five different problem areas: (a) finding a point in the intersection of a finite collection of closed convex sets, (b) projection of a point onto the closest point in the intersection of a finite collection of closed convex sets, (c) convex regression, (d) classification via support vector machines, and (e) the facilities location problem. The literature on some of the examples is enormous, so we apologize in advance for omitting relevant references and slighting the ramifications of the various models. After our tour of examples, we present relevant convergence theory in a general algorithmic framework. Our concluding discussion indicates a few extensions and limitations of distance majorization.

2 The MM principle and distance majorization

Although first articulated by the numerical analysts Ortega and Rheinboldt [39], the MM principle currently enjoys its greatest vogue in computational statistics [3,31]. The basic idea is to convert a hard optimization problem (for example, non-differentiable) into a sequence of simpler ones (for example, smooth). The MM principle requires majorizing the objective function $f(\boldsymbol{y})$ by a surrogate function $g(\boldsymbol{y} \mid \boldsymbol{x})$ anchored at the current point $\boldsymbol{x}$. Majorization is a combination of the tangency condition $g(\boldsymbol{x} \mid \boldsymbol{x}) = f(\boldsymbol{x})$ and the domination condition $g(\boldsymbol{y} \mid \boldsymbol{x})$ $f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathbb{R}^\rho$. The iterates of the associated MM algorithm are defined by

$$\boldsymbol{x}_{n+1} := arg \, min_{y} g\left(y \mid \boldsymbol{x}_n\right). \quad (4)$$

Because

$$f\left(\boldsymbol{x}_{n+1}\right) \leq g\left(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n\right) \leq g\left(\boldsymbol{x}_n \mid \boldsymbol{x}_n\right) = f\left(\boldsymbol{x}_n\right), \quad (5)$$

the MM iterates generate a descent algorithm driving the objective function downhill. Strict inequality usually prevails unless $\boldsymbol{x}_n$ is a stationary point of $f(\boldsymbol{x})$.

The most useful majorization of dist$(\boldsymbol{x}, C)$ follows immediately from the observations

$$dist\left(\boldsymbol{x}, C\right) \leq \left\|\boldsymbol{x} - P_C\left(\boldsymbol{x}_n\right)\right\|_2, \quad and \quad dist\left(\boldsymbol{x}_n, C\right) = \left\|\boldsymbol{x}_n - P_C\left(\boldsymbol{x}_n\right)\right\|_2$$

Algorithm 1

Distance majorization

1: Given $\mu_0 > 0$ and a starting point $\boldsymbol{x}_0$.

2: $k \leftarrow 0$

3: **repeat**

4:     $\boldsymbol{y} \leftarrow \boldsymbol{x}_k$

5:     **repeat**

6:          **for** $i = 1, \ldots, m$ **do**

7:              $p_i \leftarrow P_{C_i}(y)$

8:          **end for**

9:           $y \leftarrow \arg\min_u \ell(u) + \mu_k \sum_{i=1}^{m} ||u - p_i||_2^2$

10:         **until** convergence

11:         Choose new penalty parameter $\mu_{k+1} > \mu_k$

12:          $k \leftarrow k + 1$

13:          $x_k \leftarrow y$

14: **until** convergence

for all pairs x and x_n. In practice, majorizing dist$(x, C)^2$ by $||x - P_C(x_n)||_2^2$ leads to more convenient updates than majorizing dist(x, C) by $||x - P_C(x_n)||_2$. Most of our applications can be phrased as minimizing the criterion

$$g(x \mid x_n) = \ell(x) + \frac{\mu}{2}\sum_{i=1}^{m}\gamma_i||x - P_{C_i}(x_n)||_2^2, \quad (6)$$

for a convex loss $\ell(x)$, a collection $\{C_1, \ldots, C_m\}$ of closed convex sets, a positive penalization parameter μ, and a corresponding set of positive weights $\gamma_1, \ldots, \gamma_m$. Without loss in generality, we can require γ_i to sum to one, since scaling of the weights can be absorbed into the overall penalty parameter μ. Uniform weights equally penalize an iterate's violation of each constraint. Nonuniform weights will penalize constraint violations differently. This can be a useful mechanism if it is more important to satisfy some constraints over others in an application. In this paper we consider examples where constraints are all equally important and consequently employ uniform weights. For notational simplicity, we drop the weights from the remainder of our exposition but note that they can be employed in all the examples we cover. The introduction of weights also leaves the convergence analysis presented later untouched. Algorithm 1 shows the pseudocode for the distance majorization algorithm.

We highlight the fact that the algorithm does not require the projection onto the intersection but rather only the projection onto each of the constituent sets C_i. As we will see in our first example, distance majorization can be considered a generalization of the simultaneous projection algorithm for finding a point in the intersection of a collection of closed convex sets. We note, however, that distance majorization is not unique in this regard. For comparison's sake, we will also present a dual ascent algorithm at the end of the next section that employs projections onto the constituent sets. Although the two methods exhibit comparable empirical performance, the distance majorization algorithm is guaranteed to converge under weaker conditions than the dual ascent algorithm.

Finally, we note that MM algorithms are often plagued by a slow rate of convergence in a neighborhood of the minimum point. To remedy this situation, we employ quasi-Newton acceleration. MM algorithms can be accelerated via Newton's method just as the classic gradient descent algorithm. Adjusting the direction of steepest descent to account for the

curvature in the objective yields more efficient step directions, and the number of iterations to a minimum can be drastically reduced. Newton's method, however, requires computing and storing a full Hessian matrix, a demanding task in high-dimensional problems. To ease the computational burden, quasi-Newton methods obtain curvature information by approximating the Hessian with secants or differences between successive iterates. Using more secants leads to better approximations of the Hessian initially, but using too many secants can actually lead to a poorer approximation as a smaller collection of secants can adapt more dynamically to changes in the curvature as the iterations proceed. Moreover, using more secants entails additional storage and computation. In the following examples, we use either two or five secants. Using a handful of secants is a modest additional burden in computation and storage but leads to noticeable acceleration in our MM algorithm. For details on the scheme we employed as well as comparisons with alternative acceleration schemes, we direct readers to our earlier paper [49].

2.1 Sequential unconstrained minimization

During the review of this paper, a referee brought to our attention that the MM algorithm is an instance of a broad class of methods termed sequential unconstrained minimization [22]. Consider minimizing $f(x)$: $\mathbb{R}^\rho \to \mathbb{R}$ over a closed, non-empty set $C \subset \mathbb{R}^\rho$. In sequential unconstrained minimization, we generate a sequence of iterates that minimize an unconstrained surrogate

$$x_n = arg \min_{x} G_n(x) := f(x) + h_n(x),$$

where the auxiliary functions $h_n(x)$ encode information about the constraint set C.

When $h_n(x)$ is chosen so that $h_n(x) \quad 0$ for all x and $h_n(x_{n-1}) = 0$, then

$$f(x_n) \le f(x_n) + h_n(x_n) = G_n(x_n) \le G_n(x_{n-1}) = f(x_{n-1}).$$

This is a restatement of the descent property of an MM algorithm. In fact, we can identify $G_n(x) = g(x \mid x_{n-1})$ and $h_n(x) = g(x \mid x_{n-1}) - f(x)$. The tangency and domination conditions of the MM principle can be expressed alternatively as

$$G_n(x) = g(x \mid x_{n-1}) = f(x) + [g(x \mid x_{n-1}) - f(x)] = f(x) + h_n(x),$$

where $h_n(x) \quad 0$ and $h_n(x_{n-1}) = 0$.

Byrne [13] introduced an important subset of sequential unconstrained minimization methods in which the auxiliary functions $h_n(x)$ satisfy

$$G_n(x) - G_n(x_n) \ge h_{n+1}(x) \ge 0. \quad (7)$$

Methods satisfying (7) are examples of sequential unconstrained minimization algorithms (SUMMA) and generate iterates for which $f(x_n)$ converges to $\inf_{x \in C} f(x)$. The SUMMA class includes a wide range of general iterative methods including barrier and penalty function

methods, forward-backward splitting methods, and instances of the expectation maximization (EM) algorithm to name a few. Readers can consult the references [10,13,14] to learn more about the breadth and applicability of the SUMMA class.

Given that examples of the EM algorithms have been shown to belong to the SUMMA class [14] and that EM algorithms are a special case of the MM algorithm [48], it is natural to wonder if MM algorithms, which have now been shown to be SUMMA, belong to the SUMMA class. The answer to this question is not immediately obvious. It is possible to concoct majorizations that fail to meet the SUMMA condition. Rewriting the SUMMA condition (7) in terms of majorizations yields

$$g\left(\boldsymbol{x}\mid\boldsymbol{x}_{n-1}\right)-g\left(\boldsymbol{x}_n\mid\boldsymbol{x}_{n-1}\right)\geq g\left(\boldsymbol{x}\mid\boldsymbol{x}_n\right)-f\left(\boldsymbol{x}\right)\geq 0, \quad (8)$$

for all $\boldsymbol{x}$. Roughly speaking, (8) says that a sequence of majorizations should be hugging $f(\boldsymbol{x})$ uniformly more closely as the iterations proceed. While this is intuitively desirable, it is not necessary to ensure convergence of an MM algorithm.

Nonetheless, it can be non-trivial to declare an iterative algorithm to be outside the SUMMA class, since we must prove that the resulting iterative algorithm could not be derived from some sequence of auxiliary functions that do obey (7). Although majorizations chosen may violate the SUMMA condition, the resulting iterative algorithm may ultimately belong to the SUMMA class. In the "Appendix" we give an example of a convergent MM algorithm with a surrogate function that fails condition (8) globally. Locally the algorithm does belong to the SUMMA class. The ambiguity about the proper domain of an algorithm spills over into selection of starting points and highlights the practical benefits of the MM principle, which leaves the door ajar to less restrictive auxiliary functions. Fortunately, the qualitative features of convergence carry over to this broader set of auxiliary functions.

3 Examples of distance majorization

3.1 Finding a feasible point

When the intersection $C=\cap_{j=1}^{m}C_j$ is nonempty, majorization can be employed to locate a point in C. The general idea is to drive the convex combination

$$f\left(\boldsymbol{x}\right)=\sum_{i=1}^{m}dist(\boldsymbol{x},C_i)^2 \quad (9)$$

to 0. Minimization of the surrogate function

$$g\left(\boldsymbol{x}\mid\boldsymbol{x}_n\right)=\sum_{i=1}^{m}||\boldsymbol{x}-P_{C_i}\left(\boldsymbol{x}_n\right)||_2^2$$

leads to the well-known simultaneous projection algorithm

$$x_{n+1} = \sum_{i=1}^{m} P_{C_i}(x_n).$$

The earliest version of this algorithm is attributed to Cimmino [17]. It does not necessarily find the closest point in C to x. The evidence suggests that simultaneous projection converges more slowly than alternating projection [15,24]. However, simultaneous projection enjoys the advantage of being parallelizable. One can invoke the theory of paracontractive operators to prove the convergence of both simultaneous and alternating projections [11].

The alternating projection algorithm can also be derived by distance majorization. The least distance between two closed convex sets C_1 and C_2 can be found by minimimizing dist(x, C_2)^2 over $x \in C_1$. If we majorize dist(x, C_2) by the surrogate function $\|x - y_n\|^2$, where $y_n = P_{C_2}(x_n)$, then the minimum of the surrogate occurs at $P_{C_1}(y_n) = P_{C_1}[P_{C_2}(x_n)]$. When the two sets intersect, the least distance of 0 is achieved at any point in the intersection. Thus, the MM principle provides very simple and direct derivations of the simultaneous and alternating projection algorithms.

Distance majorization can be generalized by replacing Euclidean distances with Bregman divergences. For simplicity we limit our discussion to Bregman divergences generated by strictly convex twice differentiable functions $\varphi(x)$. The Bregman divergence

$$D_\phi(y \mid x) = \phi(y) - \phi(x) - \langle \nabla\phi(x), y - x \rangle.$$

is a convex function of y anchored at x and majorizing 0. For instance, the four convex functions $\varphi_1(y) = \|y\|^2$, $\varphi_2(y) = -\sum_i \log y_i$, $\varphi_3(y) = \sum_i y_i \ln y_i$, and $\varphi_4(y \mid x) = y^t M y$ generate the Bregman divergences

$$\begin{aligned}
D_{\phi_1}(y \mid x) &= \|y - x\|^2 \\
D_{\phi_2}(y \mid x) &= \sum_i \left[\frac{y_i}{x_i} - \log\left(\frac{y_i}{x_i}\right) - 1 \right] \\
D_{\phi_3}(y \mid x) &= \sum_i y_i \ln\left(\frac{y_i}{x_i}\right) - \sum_i (y_i - x_i) \\
D_{\phi_4}(y \mid x) &= (y - x)^t M (y - x).
\end{aligned}$$

The matrix M in the definition of $\varphi_4(y)$ is assumed positive definite.

The Bregman projection $P_C^\phi(x)$ onto a closed convex set C is defined as

$$P_C^\phi(x) = \arg\min_{y \in C} D_\phi(y \mid x).$$

Under suitable additional hypotheses, the Bregman projection exists. It is unique because $\varphi(\boldsymbol{x})$ is strictly convex. Moreover, $P_C^\phi(\boldsymbol{x}) = \boldsymbol{x}$ (equivalently $D_\phi\left[P_C^\phi(\boldsymbol{x}), \boldsymbol{x}\right] = 0$) exactly when $\boldsymbol{x} \in C$. The analogue of the proximity function (9) is the proximity function

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} D_{\phi_i}\left[P_{C_i}^{\phi_i}(\boldsymbol{x}) \mid \boldsymbol{x}\right]. \quad (10)$$

If we abbreviate $\boldsymbol{y}_n^i = P_{C_i}^{\phi_i}(\boldsymbol{x}_n)$, then the function

$$\varphi(\boldsymbol{x} \mid \boldsymbol{x}_n) = \sum_i D_{\phi_i}\left(\boldsymbol{y}_n^i \mid \boldsymbol{x}\right),$$

majorizes $f(\boldsymbol{x})$. The MM principle suggests that we minimize $\varphi(\boldsymbol{x} \mid \boldsymbol{x}_n)$. A brief calculation produces the stationarity condition

$$0 = \nabla\varphi(\boldsymbol{x} \mid \boldsymbol{x}_n) = \sum_i \nabla^2\phi_i(\boldsymbol{x})\left(\boldsymbol{x} - \boldsymbol{y}_n^i\right),$$

where $\nabla^2\varphi_i(\boldsymbol{x})$ denotes the Hessian of $\varphi_i(\boldsymbol{x})$. Readers can consult [12] for a more in depth and thorough treatment of minimizing the proximity function (10).

3.2 Projection onto the intersection of closed convex sets

We next consider how distance majorization can be used to find the closest point in the intersection C to a point $\boldsymbol{y}$. This involves minimizing the strictly convex function

$$f_\mu(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x} - \boldsymbol{y}||^2 + \frac{\mu}{2}\sum_{i=1}^{m} d_{C_i}(\boldsymbol{x})^2$$

for μ large. The solution $\boldsymbol{x}(\mu)$ tends to the optimal point as μ tends to ∞. The MM update for minimizing the surrogate function

$$g_\mu(\boldsymbol{x} \mid \boldsymbol{x}_n) = \frac{1}{2}||\boldsymbol{x} - \boldsymbol{y}||^2 + \frac{\mu}{2}\sum_{i=1}^{m}||\boldsymbol{x} - P_{C_i}(\boldsymbol{x}_n)||^2$$

is the convex combination

$$\boldsymbol{x}_{n+1} = \frac{1}{1+\mu}\boldsymbol{y} + \frac{\mu}{1+\mu}\sum_{i=1}^{m} P_{C_i}(\boldsymbol{x}_n). \quad (11)$$

The corresponding algorithm map $\psi(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} g_\mu(\boldsymbol{u} \mid \boldsymbol{x})$ is strictly contractive with contraction constant $c = \mu/(1 + \mu)$. According to the contraction mapping theorem, the iterates converge to the unique fixed point at geometric rate c. This fixed point coincides with the minimum point of the function $f_\mu(\boldsymbol{x})$. Indeed, $f_\mu(\boldsymbol{x})$ is differentiable with gradient

$$\nabla f_\mu\left(\boldsymbol{x}\right)=\boldsymbol{x}-\boldsymbol{y}+\mu\sum_{i=1}^{m}\left[\boldsymbol{x}-P_{C_i}\left(\boldsymbol{x}\right)\right].$$

Rearrangement of the stationarity condition $\nabla f_\mu(\boldsymbol{x})=\boldsymbol{0}$ gives the fixed point condition

$$\boldsymbol{x}=\frac{1}{1+\mu}\boldsymbol{y}+\frac{\mu}{1+\mu}\sum_{i=1}^{m}P_{C_i}\left(\boldsymbol{x}\right).$$

One can generalize these results in various ways. For instance, if we replace Euclidean loss by weighted Euclidean loss $\frac{1}{2}\sum_{i=1}^{p}w_i(\boldsymbol{x}_i-\boldsymbol{y}_i)^2$, then the MM update of the penalized loss has components

$$\boldsymbol{x}_{n+1,i}=\frac{w_i}{w_i+\mu}\boldsymbol{y}_i+\frac{\mu}{w_i+\mu}\sum_{k=1}^{m}P_{C_k}\left(\boldsymbol{x}_n\right)_i.$$

The quadratic penalty method suffers from roundoff errors and numerical instability for large μ. These are mitigated in the MM algorithm since its updates (11) rely on stable projections and avoid matrix inversion. The slow rate $\mu/(1+\mu)$ of convergence for large μ is an issue. In practice one can improve the rate of convergence by starting μ small and gradually increasing it to its target value. For a fixed μ one can also accelerate the MM iterates by systematic extrapolation. For instance, our quasi-Newton acceleration [49] often reduces the required number of iterations by one or two orders of magnitude.

3.2.1 Projection as a dual program—For the sake of comparison, we describe a dual algorithm for solving the projection problem. This alternative algorithm can be accelerated by Nesterov's method [2,37]. The unaccelerated dual algorithm is a variation of Dykstra's algorithm [21], which solves the dual problem by block descent.

To derive the dual problem, we first observe that the primal problem consists of minimizing

$$\frac{1}{2}||\boldsymbol{x}-\boldsymbol{y}||_2^2+\sum_{i=1}^{m}\delta_{C_i}\left(\boldsymbol{x}_i\right)$$

subject to $\boldsymbol{x}_1=\boldsymbol{x},\ldots,\boldsymbol{x}_m=\boldsymbol{x}$. The Lagrangian for the primal problem is

$$\begin{aligned}&\mathscr{L}\left(\boldsymbol{x},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_m\right)\\&=\frac{1}{2}||\boldsymbol{x}-\boldsymbol{y}||_2^2-\left(\sum_{i=1}^{m}\boldsymbol{z}_i\right)^t\boldsymbol{x}+\sum_{i=1}^{m}\left[\delta_{C_i}\left(\boldsymbol{x}_i\right)+\boldsymbol{z}_i^t\boldsymbol{x}_i\right].\end{aligned}$$

If $\boldsymbol{z}=(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_m)$ denotes the concatenation of the dual variables $\boldsymbol{z}_i$, then the dual function

$$\mathscr{D}\left(\boldsymbol{z}\right) = \inf_{\boldsymbol{x},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m} \mathscr{L}\left(\boldsymbol{x},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_m\right)$$

reduces to

$$\mathscr{D}\left(z\right) = -\frac{1}{2}||\boldsymbol{s}||_2^2 - \boldsymbol{s}^t\boldsymbol{y} - \sum_{i=1}^{m} \sup_{\boldsymbol{x}_i \in C_i} \left(-z_i^t \boldsymbol{x}_i\right),$$

where $\boldsymbol{s} = \sum_{i=1}^{m} z_i$. The dual function can be maximized by the proximal gradient algorithm

$$\begin{aligned}
\boldsymbol{x}^n &\leftarrow \boldsymbol{y} + \sum_{i=1}^{m} z_i^n \\
z_i^{n+1} &\leftarrow z_i^n + \left[P_{C_i}\left(\boldsymbol{x}^n - z_i^n\right) - \boldsymbol{x}^n \right].
\end{aligned}$$

For the sake of clarity, we have adopted novel notation in this derivation; $\boldsymbol{x}_i^n$ and $\boldsymbol{z}_i^n$ denote the nth iterate of the ith primal and dual variables respectively.

Derivation of this algorithm and its Nesterov acceleration (FISTA) appear in the "Appendix". The dual updates, which are essentially projection steps, can be computed in parallel. Thus, the dual algorithm matches the MM algorithm in this regard.

3.2.2 Projecting onto the set of doubly nonnegative matrices—As a numerical example, consider the problem of projecting a symmetric matrix onto the set of doubly nonnegative matrices, namely the intersection of the set of nonnegative matrices with the set of positive semi-definite matrices. Many covariance matrices – for example, kinship matrices in statistical genetics – have nonnegative entries. Projection onto each of the component sets is relatively easy while projection onto the intersection is not. Projecting onto the set of nonnegative matrices is accomplished by setting all negative entries of a matrix to zero. Projecting onto the set of positive semi-definite matrices is accomplished by truncating the eigenvalue decomposition of the matrix and rejecting all outer products with negative eigenvalues.

As a test case, we generated a 200-by-200 matrix with independent and identically distributed (i.i.d.) entries drawn from a standard normal distribution. After projecting the simulated matrix onto the space of symmetric matrices, we compared the distance majorization algorithm to its quasi-Newton acceleration (2 secants), the dual proximal gradient algorithm, and its FISTA acceleration. We implemented the MM algorithm with the geometrically increasing sequence $\mu_i = 2^i - 1$ of penalty constants μ. The decision to switch to the next larger μ was based on the ratio

$$\frac{||\boldsymbol{x}^{n+1} - \boldsymbol{x}^n||_2}{||\boldsymbol{x}^n||_2 + 1}. \quad (12)$$

Whenever this ratio fell below $\rho = 10^{-4}$, we updated μ. To track the progress of each algorithm, we calculated two measures of constraint violation by the current matrix: (a) the absolute value of the most negative eigenvalue, and (b) the absolute value of the most negative entry. Figure 1 plots the maximum of the two constraint violations on a log scale at each iteration. The abrupt transitions in the MM and quasi-Newton MM paths reflect the switch points for the penalty constant μ. Obviously, the amount of work done in each iterate varies across the methods. For a more direct comparison, Table 1 records several statistics, including run times in seconds. In the table, the distance column conveys the Frobenius norm of the difference between the simulated matrix and the fitted matrix. The two featured algorithms perform about equally well. As expected, their accelerated versions do much better.

3.2.3 Shape-restricted regression—Isotone regression minimizes the least squares criterion $\frac{1}{2}\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \boldsymbol{x}_i)^2$ subject to the isotonic constraint $x_1 \quad \ldots \quad x_n$. This problem is readily amenable to the projection algorithm. Projection onto the isotone convex cone

$$C = \{\boldsymbol{x} : \boldsymbol{x}_1 \leq \boldsymbol{x}_2 \leq \cdots \leq \boldsymbol{x}_n\}$$

is rapidly accomplished by the pool adjacent violators algorithm [1,40,46]. More complicated order restrictions such as $x_i \quad x_j$ for all arcs (i, j) in a directed graph can be handled as well. In this setting all components of a vector $\boldsymbol{x}$ projected on the convex set $C_{ij} = \{\boldsymbol{x} : x_i \quad x_j\}$ are left untouched except components x_i and x_j, These are left untouched when $x_i \quad x_j$. Both x_i and x_j are replaced by their average when $x_i > x_j$.

We considered the problem of fitting a nondecreasing function to the data shown in Fig. 2 (black dots). Each observed pair (x_i, y_i) was generated as follows. The x_i are equally spaced points between 1 and 3, and the y_i satisfy

$$\boldsymbol{y}_i = \boldsymbol{x}_i^2 + \epsilon_i,$$

where the ϵ_i are i.i.d. standard normal deviates. For the MM algorithms we used the geometrically increasing sequence of penalty constants μ_i featured in the previous example and two secant conditions for the quasi-Newton acceleration. We switched to the next value of μ whenever the stopping criterion (12) fell below $\rho\, 10^{-6}$. A looser threshold $\rho = 10^{-4}$ resulted in unacceptably poor fits for these data.

To track the progress of each algorithm, we measured the constraint violation of an iterate as the maximum absolute constraint violation between two successive parameters. Figure 2 shows that all four algorithms return similar solutions under the specified stopping rule. Figure 3 plots the constraint violation for each method on a log scale. Table 2 compares timing results and constraint violations at convergence. In the table the distance column conveys the Euclidean norm of the difference between observed points and fitted points. Compared to the previous example, we see an even greater improvement in the performance in the accelerated versions of the two algorithms. In general, it is safe to conclude that

distance majorization is a viable alternative to its most likely fastest competitor in non-smooth convex optimization.

3.2.4 Least squares fitting with convex functions—Given responses y_i, predictor vectors $\boldsymbol{x}_i$ in $\mathbb{R}^\rho$, and case weights w_i, convex regression seeks to minimize the sum of squares of residuals

$$\frac{1}{2}\sum_{i=1}^{n}w_i(y_i-\theta_i)^2$$

subject to the constraints $\boldsymbol{\xi}_i^t\,(\boldsymbol{x}_j-\boldsymbol{x}_i)\leq\theta_j-\theta_i$ for every ordered pair (i,j) [9]. In effect, θ_i is viewed as the value of the regression function $\theta(\boldsymbol{x})$ at the point $\boldsymbol{x}_i$. The unknown vector $\boldsymbol{\xi}_i\in\mathbb{R}^p$ serves as a subgradient of $\theta(\boldsymbol{x})$ at $\boldsymbol{x}_i$. Because convexity is preserved by maxima, the formula

$$\theta\left(\boldsymbol{x}\right)=\max_{j}\left[\theta_j+\boldsymbol{\xi}_j^t\left(\boldsymbol{x}-\boldsymbol{x}_j\right)\right]$$

defines a convex function with value θ_i at $\boldsymbol{x}=\boldsymbol{x}_i$. In concave regression the opposite constraint inequalities are imposed. Interpolation of predicted values in this model is accomplished by simply taking minima or maxima. Estimation reduces to a positive semidefinite quadratic program involving $n(p+1)$ variables and $n(n-1)$ inequality constraints. Note that the feasible region is nontrivial because it contains the point $(\theta,\Xi)=(\boldsymbol{0},\boldsymbol{0})$, where $\Xi=[\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n]$.

The penalized objective function is

$$f_\mu\left(\theta,\boldsymbol{\Xi}\right)=\frac{1}{2}\sum_{i=1}^{n}w_i(y_i-\theta_i)^2+\frac{\mu}{2}\sum_{j\neq k}d_{C_{ik}}^2\left(\theta,\boldsymbol{\Xi}\right),$$

where $C_{jk}=\left\{(\boldsymbol{\theta},\boldsymbol{\Xi}):\boldsymbol{\xi}_k^t\left(\boldsymbol{x}_j-\boldsymbol{x}_k\right)\leq\theta_j-\theta_k\right\}$. Let $P_{C_{jk}}(\boldsymbol{\theta},\Xi)_i$ and $P_{C_{jk}}(\boldsymbol{\theta},\Xi)^\ell$ denote the components of $P_{C_{jk}}(\boldsymbol{\theta},\Xi)$ relevant to θ_i and $\boldsymbol{\xi}_l$, respectively. The surrogate function

$$g_\mu\left(\theta,\boldsymbol{\Xi}\mid\theta_{\mathbf{m}},\boldsymbol{\Xi}_{\mathbf{m}}\right)=\frac{1}{2}\sum_{i=1}^{n}w_i(y_i-\theta_i)^2\frac{\mu}{2}\sum_{i=1}^{n}\sum_{j\neq k}||\theta_i-P_{C_{jk}}(\theta_m,\boldsymbol{\Xi}_m)_i||_2^2$$
$$+\frac{\mu}{2}\sum_{l=1}^{n}\sum_{j\neq k}||\boldsymbol{\xi}_l-P_{C_{jk}}(\theta_m,\boldsymbol{\Xi}_m)^l||_2^2$$

admits the minimizer

$$\theta_{m+1,i}=\frac{w_i}{w_i+n(n-1)\mu}y_i+\frac{\mu}{w_i+n(n-1)\mu}\sum_{j\neq k}P_{C_{jk}}(\theta_m,\boldsymbol{\Xi}_m)_i$$
$$\boldsymbol{\xi}_{m+1,l}=[n\left(n-1\right)]^{-1}\sum_{j\neq k}P_{Cjk}(\theta_m,\boldsymbol{\Xi}_m)^l.$$

The projection operator $P_{C_{jk}}$ is easy to compute because C_{jk} is a half-space. Furthermore, if we define the quantities

$$\boldsymbol{r}_{jj}=0 \quad and \quad \boldsymbol{r}_{jk}=\left[\frac{(\boldsymbol{x}_j-\boldsymbol{x}_k)^t\boldsymbol{\xi}_k-\theta_j+\theta_k}{2+||\boldsymbol{x}_j-\boldsymbol{x}_k||_2^2}\right]_+ \quad for \quad j \neq k,$$

then the sums entering the MM updates reduce to

$$\sum_{j\neq k}P_{C_{jk}}(\boldsymbol{\theta},\boldsymbol{\Xi})_i=n(n-1)\theta_i+\sum_{k=1}^n\boldsymbol{r}_{ik}-\sum_{j=1}^n\boldsymbol{r}_{ji}$$
$$\sum_{j\neq k}P_{C_{jk}}(\boldsymbol{\theta},\boldsymbol{\Xi})^l=n(n-1)\xi_l-\sum_{j=1}^n\boldsymbol{r}_{jl}(\boldsymbol{x}_j-\boldsymbol{x}_l)$$

evaluated at $\boldsymbol{\theta}=\boldsymbol{\theta}_m$ and $\boldsymbol{\Xi}=\boldsymbol{\Xi}_m$.

Figure 4 displays a randomly generated data set with 51 data points and the corresponding least squares fit with convexity constraints. We employed the same geometrically increasing sequence of μ used earlier, took five secant conditions for the quasi-Newton acceleration, and set the stopping criterion (12) to $\rho=10^{-8}$. The MM algorithm requires 8,940 iterations and 4.12s in total to achieve the objective value of 1.0709 and the maximal constraint violation at order of 7×10^{-9}.

3.2.5 Support vector machine—Given data $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, where $y_i \in \{-1, 1\}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$, the goal of discriminant analysis is to choose classification labels y_i using the p-dimensional predictor $\boldsymbol{x}_i$. The SVM [47] is one of the most popular classifiers and potentially benefits from distance penalization. Here the problem is to minimize the quadratic loss function

$$f(\boldsymbol{\theta}, \mathbf{b}, \epsilon) =\sum_{i=1}^n\epsilon_i+\frac{\lambda}{2}||\boldsymbol{\theta}||^2$$

subject to the inequality constraints

$$1 - y_i\left(b+\boldsymbol{x}_i^t\boldsymbol{\theta}\right) \leq \epsilon_i$$

using slack variables ϵ_i 0. See Example 15.5.2 of the book [30] for further details about problem formulation and passing to the dual. In the following we assume that the first element of $\boldsymbol{x}_i$ is 1, and thus the intercept b is absorbed in the parameter $\boldsymbol{\theta}$. Then the penalized objective function is

$$f_\mu(\epsilon, \boldsymbol{\theta}) =\sum_{i=1}^n\epsilon_i+\frac{\lambda}{2}||\boldsymbol{\theta}||_2^2+\frac{\mu}{2}\sum_{j=1}^nd_{C_j}^2(\epsilon, \boldsymbol{\theta})$$

where $C_j = \left\{ (\epsilon, \boldsymbol{\theta}) : \epsilon_j + y_j \boldsymbol{x}_j^t \boldsymbol{\theta} \geq 1 \right\}$. Minimizing the surrogate function

$$g_\mu \left(\epsilon, \boldsymbol{\theta} | \epsilon_{\mathbf{m}}, \boldsymbol{\theta}_{\mathbf{m}} \right) = \sum_{i=1}^n \epsilon_i + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2 + \frac{\mu}{2} \sum_{j=1}^n ||\epsilon - \mathbf{P}_{\mathbf{C_j}} \left(\epsilon_{\mathbf{m}}, \boldsymbol{\theta}_{\mathbf{m}} \right)_\epsilon ||_2^2$$
$$+ \frac{\mu}{2} \sum_{j=1}^n ||\boldsymbol{\theta} - \mathbf{P}_{\mathbf{C_j}} \left(\epsilon_{\mathbf{m}}, \boldsymbol{\theta}_{\mathbf{m}} \right)_\theta ||_2^2$$

subject to the non-negativity of ϵ_i yields the next iterate

$$\epsilon_{m+1} = \frac{1}{n} \left[\sum_{j=1}^n P_{C_j} \left(\epsilon_m, \boldsymbol{\theta}_m \right)_\epsilon - \mu^{-1} 1_n \right]_+$$
$$\boldsymbol{\theta}_{m+1} = \frac{\mu}{\lambda + n\mu} \sum_{j=1}^n P_{C_j} \left(\epsilon_m, \boldsymbol{\theta}_m \right)_\theta.$$

Because C_j is a half-space,

$$P_{C_i} \left(\epsilon_m, \boldsymbol{\theta}_m \right) = \begin{pmatrix} \epsilon_m \\ \boldsymbol{\theta}_m \end{pmatrix} + \left[\frac{1 - \epsilon_{mj} - y_j \boldsymbol{x}_j^t \boldsymbol{\theta}_m}{1 + y_j^2 ||\boldsymbol{x}_j||_2^2} \right]_+ \begin{pmatrix} \boldsymbol{e}_j \\ y_j \boldsymbol{x}_j \end{pmatrix},$$

where the vector $\boldsymbol{e}_j$ has all entries equal to 0 except for a 1 at entry j.

We report the results on an example SVM problem with a training data set of $n = 1,371$ observations and $p = 7$ features. We employed the same geometrically increasing sequence of μ and the same stopping criterion ρ used in the previous example. At $\lambda = 10$, the MM algorithm takes 14,432 iterations and 2.69s to achieve the objective value 489.0058 and the maximal constraint violation 8.6×10^{-9}.

As a generalization, consider the kernel SVM [45] attractive in handling $p \gg n$ problems. The optimization problem is to minimize

$$\sum_{i=1}^n \epsilon_i + \frac{\lambda}{2} \sum_{i=1}^n \sum_{j=1}^n \theta_i \theta_j K \left(\boldsymbol{x}_i, \boldsymbol{x}_j \right)$$

subject to

$$1 - y_i \left[b + \sum_{j=1}^n \theta_j K \left(\boldsymbol{x}, \boldsymbol{x}_j \right) \right] \leq \epsilon_i \quad and \quad \epsilon_i \geq 0 \quad for\ all\ i.$$

Common choices of kernels include the polynomial kernel

$$K \left(\boldsymbol{x}_i, \boldsymbol{x}_j \right) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^\gamma$$

and the Gaussian kernel

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = exp\left\{-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\sigma^2}\right\}.$$

Since *K* is positive semi-definite, it can be expressed in terms of a Cholesky decomposition $K = LL^t$. With reparameterization $\boldsymbol{a} = L^t\boldsymbol{\theta}$, the problem transforms to

$$\min_{b,\epsilon,\alpha} \quad \sum_{i=1}^{n}\epsilon_i + \frac{\lambda}{2}||\boldsymbol{\alpha}||_2^2$$

subject to

$$1 - y_i\left(b + \sum_{j=1}^{n}\boldsymbol{l}_{ij}\alpha_j\right) \leq \epsilon_i \quad and \quad \epsilon_i \geq 0 \quad for\ all\ i,$$

which is essentially the same as the original SVM. The Cholesky decomposition costs $n^3/6$ flops and might be a concern for data with huge number of observations. Some kernels used in genomics are naturally low rank with trivial Cholesky factors ℓ and L^t. Even for a full-rank kernel *K* , one can resort to the fast Lanczos algorithm [23] to extract its top *r* eigen-pairs $K \approx U_r D_r U_r^t$ and set $L = U_r D_r^{1/2}$, an $n \times r$ matrix.

3.3 The fire station problem

Finally, we give another example that distance majorization need not be fettered to Euclidean distances. Indeed, Euclidean distances may be inappropriate in some problems. Consider the problem of determining the optimal location of a new fire station in a city where the streets occur on a rectangular grid. The station should be situated to guarantee the shortest routes to several major buildings spread throughout the city. This is just the generalized Heron problem with the ℓ_1 norm substituting for the Euclidean norm [16]. More general treatment of the problem under arbitrary norms and infinite dimensions can be found in [34–36]. Here we are concerned with efficient computation with a particular norm. The projection operators $P_C^1\left(\boldsymbol{x}\right)$ are now harder to calculate. Indeed, they are often sets rather than points. Fortunately, when *C* is a rectangle [$\boldsymbol{a}, \boldsymbol{b}$] with sides parallel to the standard axes, $P_C^1\left(\boldsymbol{x}\right)$ is a point with components

$$P_C^1(\boldsymbol{x})_i \begin{cases} a_i & \boldsymbol{x}_i < a_i \\ \boldsymbol{x}_i & a_i \leq \boldsymbol{x}_i \leq b_i \\ b_i & \boldsymbol{x}_i > b_i. \end{cases}$$

To minimize the objective function, we minimize the surrogate function

$$g\left(\boldsymbol{x} \mid \boldsymbol{x}_n\right) = \sum_{i=1}^{m} \sum_{j=1}^{p} |\boldsymbol{x}_j - P_{C_i}^1(\boldsymbol{x}_n)_j|.$$

Because the ℓ_1 norm separates variables, we obtain a very simple update formula.

$$\boldsymbol{x}_{n+1,j} = median\left[P_{C_1}^1(\boldsymbol{x}_n)_j, \ldots, P_{C_m}^1(\boldsymbol{x}_n)_j\right].$$

Consider the example where the buildings have centers (−7, 0.5), (−5, −8), (4, 7), (5, 2), and (−4, 6) and half-side lengths of 0.5. Minimizing the sum of ℓ_1 and ℓ_2 distances yields the results shown in Fig. 5. The optimal position clearly depends on the underlying norm. For more general ℓ_1 problems, the solution may not be unique because the projection operator does not reduce to a single point.

4 Convergence analysis

We now prove convergence of the distance majorization algorithm under conditions pertinent to Euclidean distances. For broader impact, we relax the convexity requirement on $\ell(\boldsymbol{x})$. For example, in statistics, the objective function corresponding to many widely used robust estimators are often not-convex [26]. Some of our convergence results hold for such objective functions. When $\ell(\boldsymbol{x})$ is convex, it is possible to prove stronger results, and we comment on what changes when convexity is assumed. Let us first consider the convergence of the MM algorithm for solving subproblem (3). The convergence theory of MM algorithms hinges on the properties of the algorithm map $\psi(\boldsymbol{x}) \equiv$ arg min$_{\boldsymbol{y}} g(\boldsymbol{y} \mid \boldsymbol{x})$. For easy reference, we state a simple version of Meyer's monotone convergence theorem [32] instrumental in proving convergence in our setting.

Proposition 1

Let f($\boldsymbol{x}$) be a continuous function on a domain S and $\psi(\boldsymbol{x})$ be a continuous algorithm map from S into S satisfying f($\psi(\boldsymbol{x})$) < f($\boldsymbol{x}$) for all $\boldsymbol{x} \in$ S with $\psi(\boldsymbol{x})$ $\boldsymbol{x}$. Suppose for some initial point $\boldsymbol{x}_0$ that the set

$$\mathscr{L}_f(\boldsymbol{x}_0) \equiv \{\boldsymbol{x} \in S : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$$

is compact. Then (a) $\lim_{m \to \infty} \|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\| = 0$, *(b) all cluster points are fixed points of $\psi(\boldsymbol{x})$, and (c)* $\mathbf{x}_m$ *converges to one of the fixed points if they are finite in number.*

The function $f(\boldsymbol{x})$ is the objective to be minimized. In our context, the objective function is $f_\mu(\boldsymbol{x}) = \ell(\boldsymbol{x}) + \frac{\mu}{2} \sum_{i=1}^{m} dist(\boldsymbol{x}, C_i)^2$. We make the following assumptions: (a) $\ell(\boldsymbol{x})$ is continuously differentiable and $\ell(\boldsymbol{x}) + \frac{\kappa}{2}\|\boldsymbol{x}\|^2$ is convex for some constant $\kappa > 0$, (b) $f_\kappa(\boldsymbol{x})$ is coercive in the sense that $\lim_{\|\boldsymbol{x}\| \to \infty} f_\kappa(\boldsymbol{x}) = \infty$, and (c) $\mu > \kappa$. Note that $f_\mu(\boldsymbol{z})$ inherits coerciveness from $f_\kappa(\boldsymbol{z})$. Assumption (b) is met in several different scenarios, for example, if at least one of the C_i is bounded or if $\ell(\boldsymbol{x})$ itself is coercive. When $\ell(\boldsymbol{x})$ is convex,

$\ell(\boldsymbol{x}) + \frac{\kappa}{2}||\boldsymbol{x}||^2$ is convex for any $\kappa > 0$. Consequently, assumption (c) holds for any positive μ. If $f(\boldsymbol{x})$ is non-convex, but the smallest eigenvalue of the Hessian $d^2 f(\boldsymbol{x})$ is bounded below by λ, then one can take $\kappa = -\lambda$. As a rule, it can be challenging to identify κ in advance, and consequently in practice we would not know how large to choose μ to ensure the conditions for convergence when κ is unknown. Nonetheless, κ can be explicitly determined in many useful cases. In the "Appendix", we derive κ for the classic Tukey biweight of robust estimation.

Proposition 2

The cluster points of the MM iterates for solving subproblem (3) are stationary points of $f_\mu(\boldsymbol{x})$ under assumptions (a) through (c) above. If the number of stationary points is finite, then the MM iterates converge. Finally, if $f_\mu(\boldsymbol{x})$ has a unique stationary point, then the MM iterates converge to that stationary point, which globally minimizes $f_\mu(\boldsymbol{x})$.

Proof—We first argue that the surrogate function $g_\mu(\boldsymbol{y} \mid \boldsymbol{x})$ is strongly convex, a crucial fact invoked later. For all $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, Assumption (a) implies

$$\ell(y) + \frac{\kappa}{2}||\boldsymbol{y}||^2 \geq \ell(z) + \frac{\kappa}{2}||z||^2 + [\nabla \ell(z) + \kappa z]^t (y - z),$$

which in turn entails

$$\begin{aligned} \ell(y) &\geq \ell(z) + \nabla \ell(z)^t (y - z) + \kappa \left[\tfrac{1}{2}||z||^2 + z^t(y-z) - \tfrac{1}{2}||y||^2 \right] \\ &= \ell(z) + \nabla \ell(z)^t (y - z) - \tfrac{\kappa}{2}||z - y||^2. \end{aligned} \tag{13}$$

The quadratic expansion

$$\begin{aligned} \ell(y) &\geq \ell(z) + \nabla \ell(z)^t (y - z) + \kappa \left[\tfrac{1}{2}||z||^2 + z^t(y-z) - \tfrac{1}{2}||y||^2 \right] \\ &= \ell(z) + \nabla \ell(z)^t (y - z) - \tfrac{\kappa}{2}||z - y||^2. \end{aligned} \tag{14}$$

also holds. Combining inequality (13) with equality (14) leads to

$$g_\mu(\boldsymbol{y} \mid \boldsymbol{x}) \geq g_\mu(z \mid \boldsymbol{x}) + \nabla g_\mu(z \mid \boldsymbol{x})^t (y - z) + \frac{\mu - \kappa}{2}||z - y||_2^2.$$

which is equivalent to the strong convexity of $\boldsymbol{y} \mapsto g_\mu(\boldsymbol{y} \mid \boldsymbol{y})$. In view of this result, $\boldsymbol{y} \mapsto g_\mu(\boldsymbol{y} \mid \boldsymbol{x})$ has a single stationary point, which is also its unique global minimizer.

We now proceed to check the conditions given in Proposition 1. It is easy to verify that $f_\mu(\boldsymbol{x})$ is continuous. We must also show that the algorithm map $\psi(\boldsymbol{x})$ is continuous. Take an arbitrary convergent sequence $\boldsymbol{x}_n$ that tends to the limit $\boldsymbol{x}$. It suffices to prove that the sequence $\boldsymbol{y}_n = \psi(\boldsymbol{x}_n)$ tends to $\boldsymbol{y} = \psi(\boldsymbol{x})$. Now there exists a constant b such that

$$f_\mu(\boldsymbol{x}_n) \leq f_\mu(\boldsymbol{x}) + b$$

for all x_n. In view of the descent property, we have $f_\mu(y_n)$ $f_\mu(x) + b$ as well. Hence, coerciveness implies y_n is bounded. Consider any convergent subsequence y_{n_k} with limit z. The points y_{n_k} and x_{n_k} are related through the stationarity condition

$$0 = \nabla \ell \left(y_{n_k} \right) + \mu \sum_{i=1}^{m} \left[y_{n_k} - P_{C_i} \left(x_{n_k} \right) \right].$$

Since $\ell(x)$ is continuously differentiable and Euclidean projections are continuous functions, taking limits gives,

$$0 = \nabla \ell \left(z \right) + \mu \sum_{i=1}^{m} \left[z - P_{C_i} \left(x \right) \right] = \nabla g_\mu \left(z \mid x \right).$$

Because the surrogate function $z \mapsto g_\mu(z \mid x)$ possesses a unique stationary point y, the subsequence y_{n_k} converges to y. Given this conclusion for all subsequences of the bounded sequence y_n, the sequence y_n in fact converges to y.

The strict descent property of $\psi(x)$ follows from the uniqueness of the global minimizer of $g_\mu(y \mid x)$. Because $f_\mu(x)$ is coercive and continuous (in fact, continuously differentiable), the set $\mathscr{L}_{f_\mu} \left(x_0 \right)$ is compact for any initial point x_0. Therefore, Proposition 1 implies that all cluster points of the sequence $x_{n+1} = \psi(x_n)$ are fixed points. Since $\triangledown f_\mu(x) = \triangledown g_\mu(x \mid x)$, fixed points coincide with stationary points of $f_\mu(x)$. If $f_\mu(x)$ has finitely many stationary points, conclusion (c) of Proposition 1 implies that the iterates converge to one of the stationary points. If the coercive function $f_\mu(x)$ possesses a single stationary point, then that point represents a global minimum, and the MM iterates x_n converge to it.

Observe that Proposition 2 does not explicitly require the loss function $\ell(x)$ to be convex. This is in sharp contrast to the strong convexity condition on $\ell(x)$ needed to ensure the global convergence of the dual ascent algorithm. The convergence of the dual ascent method is discussed further in the "Appendix". For a sequence of penalization parameters $\mu_k \uparrow \infty$, we intuitively expect the solutions to the penalized problems to approach a solution to the original problem. Indeed, this is the case. We restate Theorem 17.1 in [38] in our notation.

Proposition 3

Suppose each $x(\mu_k)$ exactly solves subproblem (3), *and that $\mu_k \uparrow \infty$. Then every cluster point of the sequence* $\mathbf{x}(\mu_k)$ *is a global solution to the original problem* (1).

When $\ell(x)$ is coercive and possesses a unique minimizer subject to the constraints, one can justify the stronger claim that the sequence $x(\mu_k)$ converges to the minimizer. Under these assumptions the sequence $x(\mu_k)$ is bounded and possesses exactly one cluster point. Therefore, the sequence $x(\mu_k)$ converges to that cluster point. Boundedness of $x(\mu_k)$ follows from the inequalities

$$\ell \left[x \left(\mu \right) \right] \leq f_\mu \left[x \left(\mu \right) \right] \leq f_\mu \left(y \right) = \ell \left(y \right)$$

for any feasible point y.

5 Discussion

The MM principle is a versatile tool. Here we demonstrate how majorizing a distance function can be leveraged to solve a variety of optimization problems with non-trivial convex constraints. The resulting MM algorithms have simple update formulas that open the door to straightforward parallelization and graceful handling of large data sets. In the case of projection onto an intersection of closed convex sets, we have demonstrated that accelerated variants of the MM algorithm are competitive with the current state-of-the-art algorithms for solving non-smooth convex programs.

Several of our examples rely on the classical penalty method. This raises the questions of how to select the ultimate penalty constant and how fast to increase it from a low starting value. The quality of our solutions and the rate of convergence of the MM algorithms depend on these choices. We have given some rough guidelines that work well in practice, but more theoretical and empirical insight would be helpful. We have not encountered disastrous numerical instabilities in using the penalty method, partially because all of our computations were carried out in double precision.

Distance majorization works best for Euclidean distance. This follows from the fact that explicit formulas are available for several important projection operators. For others, such as projection onto the unit simplex, fast algorithms have been devised. Nonetheless, as the feasible point and fire station examples show, distance majorization can be applied to non-Euclidean distances. Is it possible to devise fast MM algorithms for computing non-Euclidean distances? This is an problem area deserving more thorough study.

In the examples we considered here all constraint violations were equally weighted. In problems where some constraints are softer than others, employing nonuniform weights on the penalty terms in the objective function (6) may be advantageous. Introducing nonuniform weights does not change our qualitative conclusions about convergence but may improve the numerical performance of the algorithm if constraint violations are weighted differently.

Another intriguing issue is the application of distance majorization to minimization of non-convex loss functions $\ell(x)$ over the intersection of convex sets. In statistics, many useful robust parameter estimates employ non-convex $\ell(x)$, for example Tukey's biweight function and more generally M-estimators [26]. Although the strongest convergence guarantees require the uniqueness of a global solution, much of the convergence theory remains intact if convexity is no longer assumed. Our convergence theory shows that the convexity assumption on $\ell(x)$ can be relaxed. Extending these results and constructing new practical examples are worthy targets of future research.

Acknowledgments

Appendix

Dual ascent algorithm

We derive a modest generalization of an iterative algorithm for the dual of the projection problem [19]. Because constructing the dual program and the associated projected gradient algorithm are exercises in modern convex analysis, we first review a few key facts from this discipline. Readers can consult the references [6,8,25,41,43] for proofs and further background material.

The Fenchel conjugate $f\star(y)$ of a function $f(x)$ is defined as

$$f^\star(y) = \sup_{\boldsymbol{x}} \left[\boldsymbol{y}^t \boldsymbol{x} - f(\boldsymbol{x}) \right].$$

When $f(x)$ is convex and lower semicontinuous, it satisfies the biconjugate relation $f\star\star(x) = f(x)$. In particular, the conjugate of the indicator function $\delta_C(x)$ of a set C is the support function

$$\delta_C^\star(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in C} \boldsymbol{y}^t \boldsymbol{x}$$

of C. When C is closed and convex, $\delta_C^{\star\star}(\boldsymbol{x}) = \delta_C(\boldsymbol{x})$.

Recall that a function $f(x)$ is strongly convex with parameter $\eta > 0$ if the difference $f(\boldsymbol{x}) - \frac{\eta}{2}\|\boldsymbol{x}\|_2^2$ is convex. Thus, a strongly convex function has a curvature bounded away from zero. If $f(x)$ is strongly convex, then the value $f\star(y)$ is attained at a single point x. In this case, $f\star(y)$ is differentiable with gradient $\triangledown f\star(y) = x$. Furthermore, $\triangledown f\star(y)$ satisfies the Lipschitz inequality

$$\|\nabla f^\star(z) - \nabla f^\star(y)\|_2 \le \frac{1}{\eta}\|z - y\|_2.$$

Lipschitz continuity ensures global convergence of the proximal gradient algorithm for solving the dual problem. The proximity-operator $\text{prox}_h(z)$ associated with a function $h(x)$ is defined as

$$prox_h(z) = \arg\min_{\boldsymbol{x}} \left[h(\boldsymbol{x}) + \frac{1}{2}\|z - \boldsymbol{x}\|_2^2 \right].$$

Here the right hand side has a unique minimizer whenever $h(x)$ is convex and lower semicontinuous. The proximal gradient method [37] is guaranteed to minimize the function $f(x)+g(x)$ when $f(x)$ is differentiable, convex, and has a Lipschitz continuous gradient, and

$g(\boldsymbol{x})$ is lower-semicontinuous and convex. The proximal gradient method iterates according to

$$\boldsymbol{x}^{n+1} = prox_{\sigma g}\left[\boldsymbol{x}^n - \sigma\nabla f\left(\boldsymbol{x}^n\right)\right],$$

where σ denotes a step size and $\boldsymbol{x}^n$ the nth iterate. We recover the classic gradient descent method when $g(\boldsymbol{x})$ is the zero function, and we recover the projected gradient algorithm when $g(\boldsymbol{x}) = \delta_C(\boldsymbol{x})$ is the indicator of a closed convex set C. Thus, the proximal gradient algorithm generalizes two important algorithm classes.

We are now ready to derive an iterative algorithm for solving the dual program of interest. Consider the slightly more general problem of minimizing a strongly convex function $f(\boldsymbol{x})$ over the intersection of a finite collection of closed convex sets $C_1, \ldots, C_m$. This problem can be reposed as minimizing the function

$$f\left(\boldsymbol{x}\right) + \sum_{i=1}^{m}\delta_{C_i}\left(\boldsymbol{x}_i\right)$$

subject to the constraints $\boldsymbol{x}_1 = \boldsymbol{x}, \ldots, \boldsymbol{x}_m = \boldsymbol{x}$. The Lagrangian for this problem

$$L\left(\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, z_1, \ldots, z_m\right) = f\left(\boldsymbol{x}\right) - \left(\sum_{i=1}^{m}z_i\right)^t\boldsymbol{x} + \sum_{i=1}^{m}\left[\delta_{C_i}\left(\boldsymbol{x}_i\right) + z_i^t\boldsymbol{x}_i\right].$$

gives rise to the dual function

$$\mathscr{D}\left(z\right) = -f^\star\left(z_1 + \cdots + z_m\right) - \sum_{i=1}^{m}h_i\left(z_i\right),$$

where $z = (z_1, \ldots, z_m)$ denotes the concatenation of the dual variables z_i and $h_i(z_i)$ is the support function of the set C_i at $-z_i$. Thus, the dual problem of maximizing $\mathscr{D}\left(z\right)$ is equivalent to minimizing $f^\star\left(z_1 + \cdots + z_m\right) + \sum_{i=1}^{m}h_i\left(z_i\right)$. Given that $f(\boldsymbol{x})$ is strongly convex, $f\star(z)$ is differentiable and in fact $\triangledown f\star(z)$ is Lipschitz continuous. Therefore, the dual is a prime candidate to be solved via the proximal gradient method. Since $\sum_i h_i\left(z_i\right)$ separates the variable z_i, the dual proximal gradient step can be computed blockwise as

$$z_i^{n+1} = prox_{\sigma h_i}\left[z_i^n - \sigma\nabla f^\star\left(z_1^n + \cdots + z_m^n\right)\right], \quad (15)$$

where σ denotes a step size. The algorithm simplifies further by applying the Moreau decomposition [18, Lemma 2.10].

$$\boldsymbol{u} = prox_{\sigma h_i}\left(\boldsymbol{u}\right) + \sigma prox_{h_i^\star/\sigma}\left(\boldsymbol{u}/\sigma\right). \quad (16)$$

Note that $prox_{h_i^\star/\sigma}(u/\sigma) = -P_{C_i}(-u/\sigma)$ and $\triangledown f\star(s)$ is a minimizer of the convex function $f(x)-s^t x$. Combining these identities with (15) and (16) gives the algorithm

$$
\begin{aligned}
x^n &= arg\,\min_x \left[f(x) - (z_1^n + \cdots + z_m^n)^t x \right] \\
z_i^{n+1} &= z_i^n + \sigma \left[P_{C_i}(x^n - \sigma^{-1} z_i^n) - x^n \right].
\end{aligned}
$$

Convergence is assured by setting the step length $\sigma = \eta$, where $1/\eta$ is the Lipschitz constant of $\triangledown f\star(z)$. Thus, the strong convexity condition on $f(x)$ is actually required for convergence, since a closed, convex function f is Lipschitz continuous if and only if its conjugate function is strongly convex [27].

The Nesterov acceleration mentioned earlier requires just a minor adjustment. The first two iterates are computed as above; subsequent updates use the following extrapolation steps.

$$
\begin{aligned}
x^n &= arg\,\min_x \left[f(x) - (z_1^n + \cdots + z_m^n)^t x \right] \\
s^n &= z_i^n + \tfrac{n-2}{n+1} \left[z^n - z^{n-1} \right] \\
z_i^{n+1} &= s_i^n + \sigma \left[P_{C_i}(x^n - t^{-1} s_i^n) - x^n \right].
\end{aligned}
$$

An MM formulation that fails the SUMMA condition

Consider minimizing the univariate function $f(x) = \cos(x)$. According to the quadratic upper bound principle [7], the function

$$
g(y \mid x) = cos(x) - sin(x)(y-x) + \frac{1}{2}(y-x)^2
$$

majorizes $f(y)$. The MM algorithm $x_n = \psi(x_{n-1})$ employs the iteration map

$$
\psi(x) = x + sin(x). \quad (17)
$$

Figure 6b depicts the first two majorizations starting $x_0 = 1$. The global SUMMA condition requires that

$$
g(x \mid x_0) - g(x_1 \mid x_0) \geq (x \mid x_1) - f(x) \quad (18)
$$

for all x, but Fig. 6b shows that this inequality fails for some x. Restricted to the interval $C = \left[\frac{\pi}{2}, \frac{3\pi}{2} \right]$, however, the MM algorithm does belong to the SUMMA class. It is geometrically obvious that all iterates reside in C when iteration commences there. Furthermore, the objective function is convex, and the MM algorithm reduces to gradient descent with a fixed step size that is no greater than twice the inverse of the Lipschitz constant of $f'(x)$. In these circumstances Byrne [13] verifies the SUMMA condition.

On the other hand, one can prove convergence without invoking the SUMMA machinery. The MM algorithm has fixed points at integer multiples of π. Even multiples correspond to maxima and odd multiples to minima. The maxima are repelling, and the minima are attracting. If the algorithm starts on the interval $[2k\pi, 2(k + 1)\pi]$, then it remains there. Hence, the hypotheses of Proposition 1 are met. Although the SUMMA condition is helpful in forcing convergence and understanding the rate of convergence, there is no need to compel majorization to satisfy it.

Tukey's Biweight

In robust linear regression, outliers are a major concern. In standard regression one minimizes the squared error loss

$$f\left(\beta\right) = \sum_{i=1}^{n} \phi\left(y_i - x_i^t \beta\right),$$

with $\phi(t) = \frac{1}{2}t^2$, where y_i is the response of case i, x_i is the predictor vector for case i, and β is the vector of regression coefficients. One can moderate the influence of outliers by substituting Tukey's biweight [26]

$$\phi(t) = \begin{cases} \frac{c^2}{6}\left\{1 - \left[1 - \left(\frac{t}{c}\right)^2\right]^3\right\} & if \quad |t| \le c \\ c^2/6 & if \quad >c. \end{cases} \tag{19}$$

for $\phi(t) = \frac{1}{2}t^2$. The new loss determined by the function (19) discounts the contribution of residuals $y_i - x_i^t \beta$ whose absolute value exceeds c. To calculate a global lower bound λ on the eigenvalues of $d^2 f(x)$, we note that

$$d^2 f\left(\beta\right) = \sum_{i=1}^{n} \phi''\left(y_i - x_i^t \beta\right) x_i x_i^t,$$

where

$$\phi''(t) = \begin{cases} 1 - 6\left(\frac{t}{c}\right)^2 + 5\left(\frac{t}{c}\right)^4 & if \quad |t| \le c \\ 0 & if \quad |t| > c. \end{cases}$$

The function $\varphi''(t)$ achieves a minimum of $-\frac{4}{5}$ at $t = \pm c\sqrt{\frac{3}{5}}$. It follows that we can take

$$\kappa \ge \frac{4}{5}\rho\left(X^t X\right),$$

where X is the matrix with columns x_i and $\rho(M)$ denotes the largest eigenvalue of the symmetric matrix M. Interestingly, κ does not depend on c. Similar calculations can be carried out for other robust choices of $\varphi(t)$.

References

1. Barlow, RE.; Bartholomew, D.; Bremner, JM.; Brunk, HD. Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression. Wiley; New York: 1972.

2. Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2009; 2(1):183–202.

3. Becker MP, Yang I, Lange K. EM algorithms without missing data. Stat. Methods Med. Res. 1997; 6:38–54. [PubMed: 9185289]

4. Bertsekas, DP. Convex Analysis and Optimization. With Angelia Nedi and Asuman E. Ozdaglar; Athena Scientific, Belmont: 2003.

5. Bertsekas D. Projected Newton methods for optimization problems with simple constraints. SIAM J. Control Optim. 1982; 20(2):221–246.

6. Bertsekas, DP. Convex Optimization Theory. Athena Scientific, Belmont: 2009.

7. Böhning D, Lindsay BG. Monotonicity of quadratic-approximation algorithms. Ann. Inst. Stat. Math. 1988; 40:641–663.

8. Borwein, JM.; Lewis, AS. Convex Analysis and Nonlinear Optimization: Theory and Examples. Springer; New York: 2000.

9. Boyd, S.; Vandenberghe, L. Convex Optimization. Cambridge University Press; Cambridge: 2004.

10. Byrne C. An elementary proof of convergence of the forward-backward splitting algorithm. 2013 submitted for publication.

11. Byrne, C. Applied Iterative Methods. Ak Peters Series, AK Peters; Wellesly: 2008.

12. Byrne C, Censor Y. Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback–Leibler distance minimization. Ann. Oper. Res. 2001; 105(1–4):77–98.

13. Byrne C. Sequential unconstrained minimization algorithms for constrained optimization. Inverse Probl. 2008; 24(1):015,013.

14. Byrne C. Alternating minimization as sequential unconstrained minimization: a survey. J. Optim. Theory Appl. 2013; 156(3):554–566.

15. Censor Y, Chen W, Combettes P, Davidi R, Herman G. On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints. Comput. Optim. Appl. 2012; 51:1065–1088.

16. Chi EC, Lange K. A look at the generalized Heron problem through the lens of majorization-minimization. Am. Math. Mon. 2013 appear.

17. Cimmino G. Calcolo approssimato per soluzioni dei sistemi di equazioni lineari. La Ricerca Scientifica XVI Series II Anno IX(1). 1938:326–333.

18. Combettes P, Wajs V. Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. 2005; 4(4):1168–1200.

19. Combettes, PL.; Pesquet, JC. Proximal splitting methods in signal processing. In: Bauschke, HH.; Burachik, RS.; Combettes, PL.; Elser, V.; Luke, DR.; Wolkowicz, H., editors. Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer Optimization and Its Applications. Vol. 49. Springer; New York: 2011. p. 185-212.

20. Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. Proceedings of the International Conference on Machine Learning; 2008.

21. Dykstra RL. An algorithm for restricted least squares regression. J. Am. Stat. Assoc. 1983; 78(384):837–842.

22. Fiacco, AV.; McCormick, GP. Classics in Applied Mathematics. SIAM; Philadelphia: 1990. Nonlinear Programming: Sequential Unconstrained Minimization Techniques.

23. Golub, GH.; Van Loan, CF. Johns Hopkins Studies in the Mathematical Sciences. 3rd edn. Johns Hopkins University Press; Baltimore: 1996. Matrix Computations.

24. Gould N. How good are projection methods for convex feasibility problems? Comput. Optim. Appl. 2008; 40:1–12.

25. Hiriart-Urruty, JB.; Lemaréchal, C. Fundamentals of Convex Analysis. Springer; Berlin: 2004.

26. Huber, PJ.; Ronchetti, EM. Robust Statistics. 2nd edn. Wiley; New York: 2009.

27. Kakade SM, Shalev-Shwartz S, Tewari A. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization, Technical report. Toyota Technological Institute. 2009

28. Kim D, Sra S, Dhillon I. Tackling box-constrained optimization via a new projected quasi-newton approach. SIAM J. Sci. Comput. 2010; 32(6):3548–3563.

29. Lange, K. Statistics and Computing. 2nd edn. Springer; New York: 2010. Numerical Analysis for Statisticians.

30. Lange, K. Springer Texts in Statistics. 2nd edn. Springer; New York: 2012. Optimization.

31. Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions (with discussion). J. Comput. Graph. Stat. 2000; 9:1–20.

32. Meyer R. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. J. Comput. Syst. Sci. 1976; 12(1):108–121.

33. Michelot C. A finite algorithm for finding the projection of a point onto the canonical simplex of Rn. J. Optim. Theory Appl. 1986; 50:195–200.

34. Mordukhovich BS, Nam NM, Salinas J. Applications of variational analysis to a generalized Heron problems. Appl. Anal. 2012; 91(10):1915–1942.

35. Mordukhovich B, Nam NM. Applications of variational analysis to a generalized Fermat-Torricelli problem. J. Optim. Theory Appl. 2011; 148:431–454.

36. Mordukhovich B, Nam NM, Salinas J. Solving a generalized Heron problem by means of convex analysis. Am. Math. Mon. 2012; 119(2):87–99.

37. Nesterov, Y. Gradient methods for minimizing composite objective function. CORE Discussion Papers. 2007.

38. Nocedal, J.; Wright, SJ. Springer Series in Operations Research and Financial Engineering. 2nd edn. Springer; New York: 2006. Numerical Optimization.

39. Ortega, JM.; Rheinboldt, WC. Iterative Solutions of Nonlinear Equations in Several Variables. Academic; New York: 1970.

40. Robertson, T.; Wright, FT.; Dykstra, RL. Order Restricted Statistical Inference. Probability and Mathematical Statistics: Wiley Series in Probability and Mathematical Statistics. Wiley; Chichester: 1988.

41. Rockafellar, RT. Convex Analysis. Princeton University Press; Princeton: 1996.

42. Rosen JB. The gradient projection method for nonlinear programming. Part I. Linear constraints. J. Soc. Ind. Appl. Math. 1960; 8(1):181–217.

43. Ruszczy ski, A. Nonlinear Optimization. Princeton University Press; Princeton: 2006.

44. Schmidt, M.; van den Berg, E.; Friedlander, MP.; Murphy, K. Optimizing costly functions with simple constraints: a limited-memory projected quasi-newton algorithm. In: van Dyk, D.; Welling, M., editors. Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS); Florida: Clearwater Beach; 2009. p. 456-463.2009

45. Scholkopf, B.; Smola, AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press; Cambridge: 2001.

46. Silvapulle, MJ.; Sen, PK. Wiley Series in Probability and Statistics. Wiley; Hoboken: 2005. Constrained Statistical Inference: Inequality, Order, and Shape Restrictions.

47. Vapnik, VN. Statistics for Engineering and Information Science. 2nd edn. Springer; New York: 2000. The Nature of Statistical Learning Theory.

48. Wu TT, Lange K. The MM alternative to EM. Stat. Sci. 2010; 25(4):492–505.

49. Zhou H, Alexander D, Lange K. A quasi-Newton acceleration for high-dimensional optimization algorithms. Stat. Comput. 2011; 21:261–273. [PubMed: 21359052]
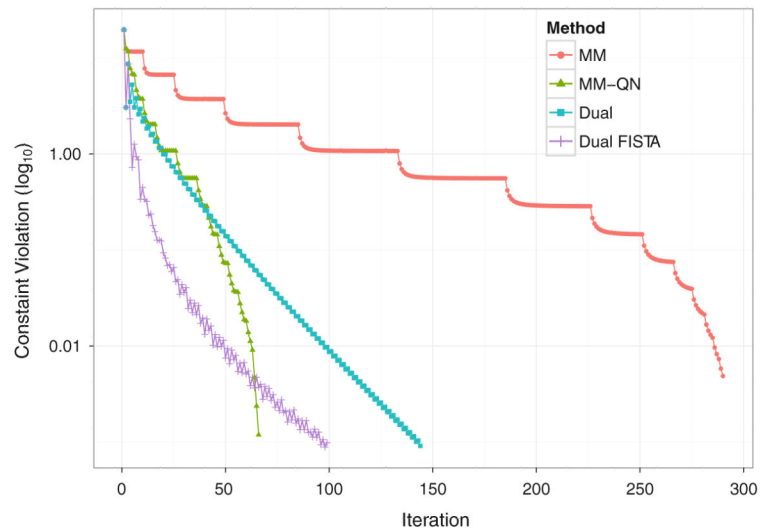
Fig. 1.
A comparison of the MM algorithm, its quasi-Newton acceleration, the dual proximal gradient algorithm, and its FISTA acceleration applied to the problem of projecting a 200×200 matrix onto the set of doubly nonnegative matrices
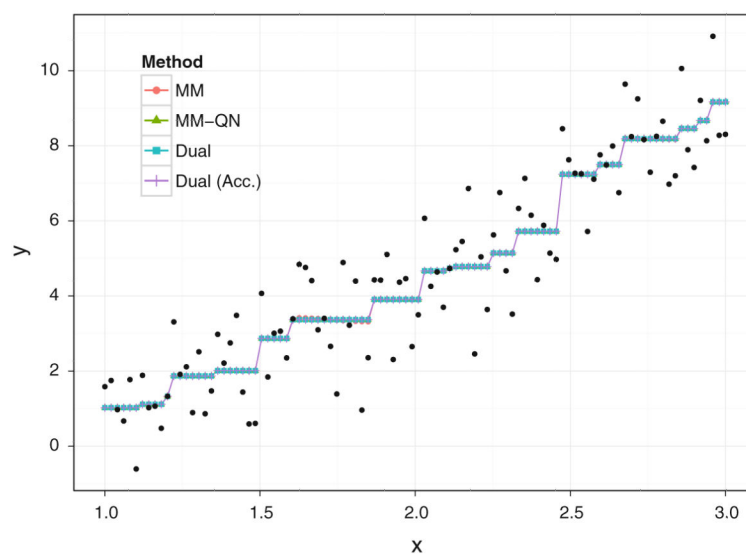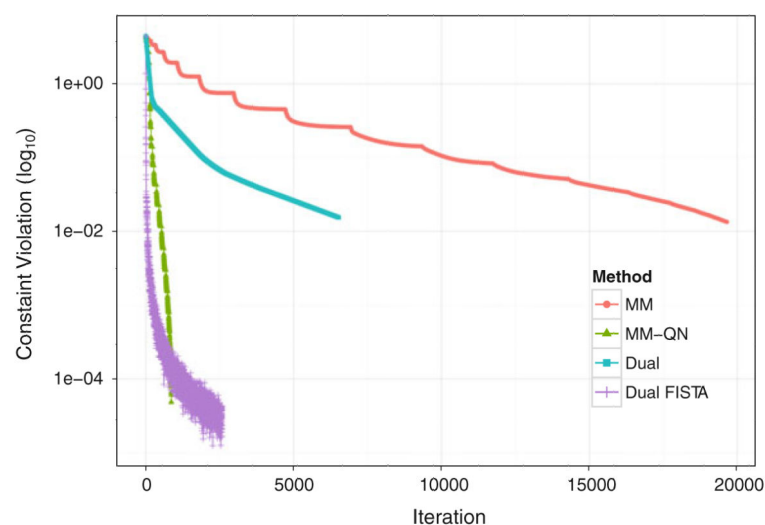
Fig. 2.
Fitted data for isotonic regression

Fig. 3.
A comparison of the MM algorithm, its quasi-Newton acceleration, the dual proximal gradient algorithm, and its FISTA acceleration applied to a univariate isotonic regression problem
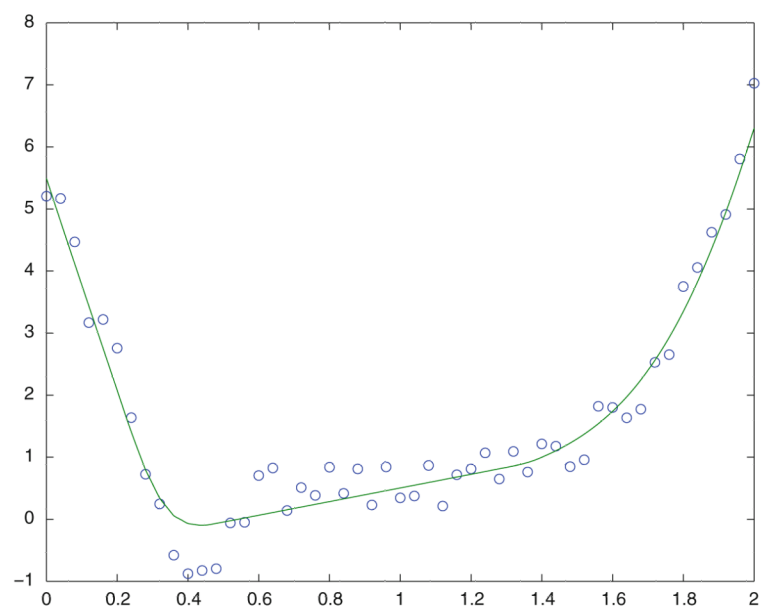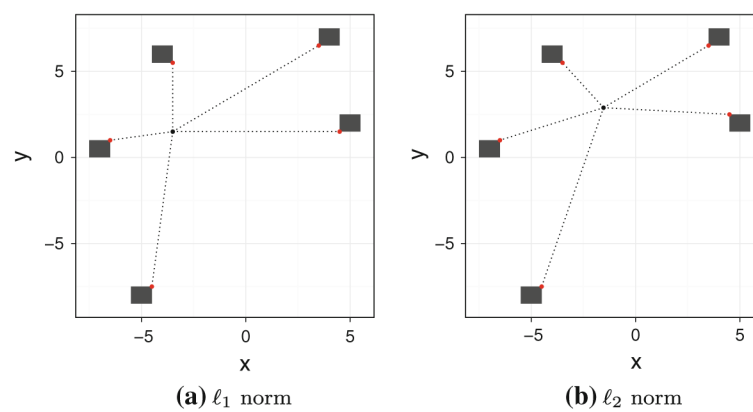
Fig. 4.
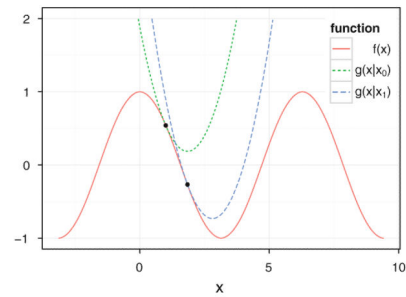Fitted data for convex regression
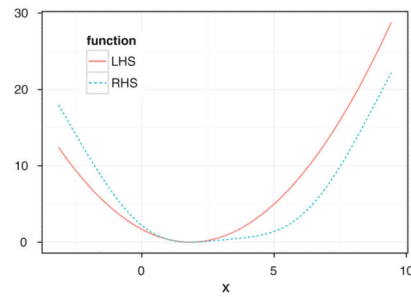
(a) ℓ_1 norm **(b)** ℓ_2 norm

Fig. 5.
Optimal location for the fire station

(a) An MM algorithm for minimizing $f(x)=\cos(x)$. The first two majorizations are shown when $x_0 = 1$.

(b) Violation of the SUMMA inequality. The solid line plots the left hand side (LHS) of the SUMMA inequality (18). The dashed line plots the right hand side (RHS) of the SUMMA inequality (18). The SUMMA inequality requires the dashed line to never cross above the solid line.

Fig. 6.

An example of an MM algorithm whose auxiliary functions fail the SUMMA condition

Table 1

Timing comparisons and constraint violations for projecting onto the set of doubly nonnegative matrices

Method	Time (s)	Iterations	Distance	Constraint violation
MM	16.526	290	120.9110	−0.0048710012
MM-QN	11.098	98	120.9131	−0.0007433297
Dual	19.882	144	120.9131	−0.0009122053
Dual (Acc.)	13.926	99	120.9136	−0.0009862162

Table 2

Timing comparisons and constraint violations for the isotonic regression example

Method	Time (s)	Iterations	Distance	Constraint violation
MM	24.45	19,651	9.633144	−1.330351e−02
MM-QN	3.27	863	9.677731	−4.869077e−05
Dual	12.70	6,526	9.637778	−1.525531e−02
Dual (Acc.)	5.46	2,578	9.677847	−4.104088e−05