Relaxed Majorization-Minimization for Non-smooth and Non-convex Optimization

Chen Xu[1], Zhouchen Lin[1,2,*], Zhenyu Zhao[3], Hongbin Zha[1]
[1] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, P. R. China
[2] Cooperative Medianet Innovation Center, Shanghai Jiaotong University, P. R. China
[3]Department of Mathematics, School of Science, National University of Defense Technology, P. R. China
xuen@pku.edu.cn, zlin@pku.edu.cn, dwightzzy@gmail.com, zha@cis.pku.edu.cn

Abstract

We propose a new majorization-minimization (MM) method for non-smooth and non-convex programs, which is general enough to include the existing MM methods. Besides the local majorization condition, we only require that the difference between the directional derivatives of the objective function and its surrogate function vanishes *when the number of iterations approaches infinity*, which is a very weak condition. So our method can use a surrogate function that directly approximates the non-smooth objective function. In comparison, all the existing MM methods construct the surrogate function by approximating the smooth component of the objective function. We apply our relaxed MM methods to the robust matrix factorization (RMF) problem with different regularizations, where our locally majorant algorithm shows great advantages over the state-of-the-art approaches for RMF. This is the first algorithm for RMF ensuring, *without extra assumptions*, that any limit point of the iterates is a stationary point.

Introduction

Consider the following optimization problem:

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}), \tag{1}$$

where $\mathcal{C}$ is a closed convex subset in $\mathbb{R}^n$ and $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is a continuous function bounded below, which could be non-smooth and non-convex. Often, $f(x)$ can be split as:

$$f(\mathbf{x}) = \tilde{f}(\mathbf{x}) + \hat{f}(\mathbf{x}), \tag{2}$$

where $\tilde{f}(\mathbf{x})$ is differentiable and $\hat{f}(\mathbf{x})$ is non-smooth[1]. Such an optimization problem is ubiquitous, e.g., in statistics (Chen, 2012), computer vision and image processing (Bruckstein, Donoho, and Elad, 2009; Ke and Kanade, 2005), data mining and machine learning (Pan et al., 2014; Kong, Ding, and Huang, 2011). There have been a variety of methods to tackle problem (1). Typical methods include subdifferential (Clarke, 1990), bundle methods (Mäkelä, 2002), gradient sampling (Burke, Lewis, and Overton, 2005), smoothing methods (Chen, 2012), and majorization-minimization (MM) (Hunter and Lange, 2004). In this paper, we focus on the MM methods.

*Corresponding author.
Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
[1] $\tilde{f}(\mathbf{x})$ and $\hat{f}(\mathbf{x})$ may vanish.

Algorithm 1 Sketch of MM

Input: $\mathbf{x}_0 \in \mathcal{C}$.
while not converged **do**
  Construct a surrogate function $g_k(\mathbf{x})$ of $f(\mathbf{x})$ at the current iterate $\mathbf{x}_k$.
  Minimize the surrogate to get the next iterate: $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}\in\mathcal{C}} g_k(\mathbf{x})$.
  $k \leftarrow k+1$.
end while
Output: The solution $\mathbf{x}_k$.

Table 1: Comparison of surrogate functions among existing MM methods. $\#\mathbf{1}$ represents globally majorant MM in (Mairal, 2013), $\#\mathbf{2}$ represents strongly convex MM in (Mairal, 2013), $\#\mathbf{3}$ represents successive MM in (Razaviyayn, Hong, and Luo, 2013) and $\#\mathbf{4}$ represents our relaxed MM. In the second to fourth rows, $\times$ means that a function is not necessary to have the feature. In the last three rows, $\surd$ means that a function has the ability.

Surrogate Functions	$\#\mathbf{1}$	$\#\mathbf{2}$	$\#\mathbf{3}$	$\#\mathbf{4}$
Globally Majorant	$\surd$	$\times$	$\surd$	$\times$
Smoothness of Difference	$\surd$	$\surd$	$\times$	$\times$
Equality of Directional Derivative	$\surd$	$\surd$	$\surd$	$\times$
Approximate $\tilde{f}(\mathbf{x})$ in (2)	$\surd$	$\surd$	$\surd$	$\surd$
Approximate $f(\mathbf{x})$ in (1) ($\hat{f}\neq 0$)	$\times$	$\times$	$\times$	$\surd$
Sufficient Descent	$\times$	$\surd$	$\times$	$\surd$

Existing MM for Non-smooth and Non-convex Optimization

MM has been successfully applied to a wide range of problems. Mairal (2013) has given a comprehensive review on MM. Conceptually, the MM methods consist of two steps (see Algorithm 1). First, construct a surrogate function $g_k(\mathbf{x})$ of $f(\mathbf{x})$ at the current iterate $\mathbf{x}_k$. Second, minimize the surrogate $g_k(\mathbf{x})$ to update $\mathbf{x}$. The choice of surrogate is critical for the efficiency of solving (1) and also the quality of solution. The most popular choice of surrogate is the class of "first order surrogates", whose difference from the objective function is differentiable with a Lipschitz continuous gradient (Mairal, 2013). For non-smooth and non-convex objectives, to the best of our knowledge, "first order surrogates" are only used to approximate the differentiable part $\tilde{f}(\mathbf{x})$ of the objective $f(\mathbf{x})$ in (2). More precisely, denoting $\tilde{g}_k(\mathbf{x})$ as an approximation of $\tilde{f}(\mathbf{x})$ at iteration k, the surrogate is

$$g_k(\mathbf{x}) = \tilde{g}_k(\mathbf{x}) + \hat{f}(\mathbf{x}). \tag{3}$$

Such a split approximation scheme has been successfully applied, e.g., to minimizing the difference of convex functions (Candes, Wakin, and Boyd, 2008) and in the proximal splitting algorithm (Attouch, Bolte, and Svaiter, 2013). In parallel to (Mairal, 2013), Razaviyayn, Hong, and Luo (2013) also showed that many popular methods for minimizing non-smooth functions could be regarded as MM methods. They proposed the block coordinate descent method, where the traditional MM could be regarded as a special case by gathering all variables in one block. Different from (Mairal, 2013), they suggested using the directional derivative to ensure the first order smoothness between the objective and the surrogate, which is weaker than the condition in (Mairal, 2013) that the difference between the objective and the surrogate should be smooth. However, Razaviyayn, Hong, and Luo (2013) only discussed the choice of surrogates by approximating $f(\mathbf{x})$ as (3).

Contributions

The contributions of this paper are as follows:

(a) We further relax the condition on the difference between the objective and the surrogate. We only require that the directional derivative of the difference vanishes *when the number of iterations approaches infinity* (see (8)). Our even weaker condition ensures that the non-smooth and non-convex objective can be approximated directly. Our relaxed MM is general enough to include the existing MM methods (Mairal, 2013; Razaviyayn, Hong, and Luo, 2013).

(b) We also propose the conditions ensuring that the iterates produced by our relaxed MM converge to stationary points[2], even for general non-smooth and non-convex objectives.

(c) As a concrete example, we apply our relaxed MM to the robust matrix factorization (RMF) problem with different regularizations. Experimental results testify to the robustness and effectiveness of our locally majorant algorithm over the state-of-the-art algorithms for RMF. To the best of our knowledge, this is the first work that ensures convergence to stationary points *without extra assumptions*, as the objective and the constructed surrogate naturally fulfills the convergence conditions of our relaxed MM.

Table 1 summarizes the differences between our relaxed MM and the existing MM. We can see that ours is general enough to include existing works (Mairal, 2013; Razaviyayn, Hong, and Luo, 2013). In addition, it requires less smoothness on the difference between the objective and the surrogate and has higher approximation ability and better convergence property.

[2]For convenience, we say that a sequence converges to stationary points meaning that any limit point of the sequence is a stationary point.

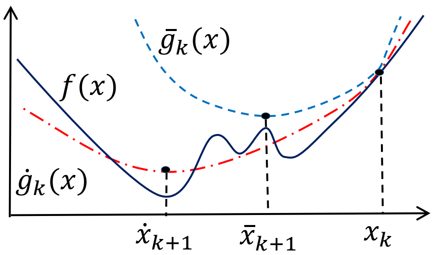


Figure 1: Illustration of a locally majorant surrogate $\dot{g}_k(\mathbf{x})$ and a globally majorant surrogate $\bar{g}_k(\mathbf{x})$. Quite often, a globally majorant surrogate cannot approximate the objective function well, thus giving worse solutions.

Our Relaxed MM

Before introducing our relaxed MM, we recall some definitions that will be used later.

Definition 1. ***(Sufficient Descent)*** *$\{f(\mathbf{x}_k)\}$ is said to have sufficient descent on the sequence $\{\mathbf{x}_k\}$ if there exists a constant $\alpha > 0$ such that:*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \alpha \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2, \quad \forall k. \tag{4}$$

Definition 2. ***(Directional Derivative*** *(Borwein and Lewis, 2010, Chapter 6.1)) The directional derivative of function $f(\mathbf{x})$ in the feasible direction $\mathbf{d}$ ($\mathbf{x} + \mathbf{d} \in \mathcal{C}$) is defined as:*

$$\nabla f(\mathbf{x}; \mathbf{d}) = \liminf_{\theta \downarrow 0} \frac{f(\mathbf{x} + \theta \mathbf{d}) - f(\mathbf{x})}{\theta}. \tag{5}$$

Definition 3. ***(Stationary Point*** *(Razaviyayn, Hong, and Luo, 2013)) A point $\mathbf{x}^*$ is a (minimizing) stationary point of $f(\mathbf{x})$ if $\nabla f(\mathbf{x}^*; \mathbf{d}) \geq 0$ for all $\mathbf{d}$ such that $\mathbf{x}^* + \mathbf{d} \in \mathcal{C}$.*

The surrogate function in our relaxed MM should satisfy the following three conditions:

$$f(\mathbf{x}_k) = g_k(\mathbf{x}_k), \tag{6}$$

$$f(\mathbf{x}_{k+1}) \leq g_k(\mathbf{x}_{k+1}), \quad \text{(Locally Majorant)} \tag{7}$$

$$\begin{aligned} &\lim_{k \to \infty} \left(\nabla f(\mathbf{x}_k; \mathbf{d}) - \nabla g_k(\mathbf{x}_k; \mathbf{d}) \right) = 0, \\ &\forall \mathbf{x}_k + \mathbf{d} \in \mathcal{C}. \quad \text{(Asymptotic Smoothness)} \end{aligned} \tag{8}$$

By combining conditions (6) and (7), we have the non-increment property of MM:

$$f(\mathbf{x}_{k+1}) \leq g_k(\mathbf{x}_{k+1}) \leq g_k(\mathbf{x}_k) = f(\mathbf{x}_k). \tag{9}$$

However, we will show that with a careful choice of surrogate, $\{f(\mathbf{x}_k)\}$ can have sufficient descent, which is stronger than non-increment and is critical for proving the convergence of our MM.

In the traditional MM, the global majorization condition (Mairal, 2013; Razaviyayn, Hong, and Luo, 2013) is assumed:

$$f(\mathbf{x}) \leq g_k(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{C}, \quad \text{(Globally Majorant)} \tag{10}$$

which also results in the non-increment of the objective function, i.e., (9). However, a globally majorant surrogate cannot approximate the object well (Fig. 1). Moreover, the step length between successive iterates may be too small. So a globally majorant surrogate is likely to produce an inferior solution and converges slower than a locally majorant one ((Mairal, 2013) and our experiments).

Condition (8) requires very weak first order smoothness of the difference between $g_k(\mathbf{x})$ and $f(\mathbf{x})$. It is weaker than that in (Razaviyayn, Hong, and Luo, 2013, Assumption 1.(A3)), which requires that the directional derivatives of $g_k(\mathbf{x})$ and $f(\mathbf{x})$ are equal at every $\mathbf{x}_k$. Here we only require the equality when the number of iterations goes to infinity, which provides more flexibility in constructing the surrogate function. When $f(\cdot)$ and $g_k(\cdot)$ are both smooth, condition (8) can be satisfied when the two hold the same gradient at $\mathbf{x}_k$. For non-smooth functions, condition (8) can be enforced on the differentiable part $\tilde{f}(\mathbf{x})$ of $f(\mathbf{x})$. These two cases have been discussed in the literature (Mairal, 2013; Razaviyayn, Hong, and Luo, 2013). If $\tilde{f}(\mathbf{x})$ vanishes, the case not yet discussed in the literature, the condition can still be fulfilled by approximating the whole objective function $f(\mathbf{x})$ directly, as long as the resulted surrogate satisfies certain properties, as stated below[3].

Proposition 1. *Assume that* $\exists\ K > 0,\ +\infty > \gamma_u, \gamma_l > 0,$ *and* $\epsilon > 0$*, such that*

$$\hat{g}_k(\mathbf{x})+\gamma_u\|\mathbf{x}-\mathbf{x}_k\|_2^2 \geq f(\mathbf{x}) \geq \hat{g}_k(\mathbf{x})-\gamma_l\|\mathbf{x}-\mathbf{x}_k\|_2^2 \quad (11)$$

holds for all $k \geq K$ *and* $\mathbf{x} \in \mathcal{C}$ *such that* $\|\mathbf{x} - \mathbf{x}_k\| \leq \epsilon$*, where the equality holds if and only if* $\mathbf{x} = \mathbf{x}_k$*. Then condition* (8) *holds for* $g_k(\mathbf{x}) = \hat{g}_k(\mathbf{x}) + \gamma_u\|\mathbf{x} - \mathbf{x}_k\|_2^2$.

We make two remarks. First, for sufficiently large γ_u and γ_l the inequality (11) naturally holds. However, a larger γ_u leads to slower convergence. Fortunately, as we only require the inequality to hold for sufficiently large k, adaptively increasing γ_u from a small value is allowed and also beneficial. In many cases, the bounds of γ_u and γ_l may be deduced from the objective function. Second, the above proposition does not specify any smoothness property on either $\hat{g}_k(\mathbf{x})$ or the difference $g_k(\mathbf{x}) - f(\mathbf{x})$. It is not odd to add the proximal term $\|\mathbf{x} - \mathbf{x}_k\|_2^2$, which has been widely used, e.g. in proximal splitting algorithm (Attouch, Bolte, and Svaiter, 2013) and alternating direction method of multipliers (ADMM) (Lin, Liu, and Li, 2013).

For general non-convex and non-smooth problems, proving the convergence to a global (or local) minimum is out of reach, classical analysis focuses on converging to stationary points instead. For general MM methods, since only non-increment property is ensured, even the convergence to stationary points cannot be guaranteed. Mairal (Mairal, 2013) proposed using strongly convex surrogates and thus proved that the iterates of corresponding MM converge to stationary points. Here we have similar results, as stated below.

Theorem 1. ***(Convergence)*** *Assume that the surrogate* $g_k(\mathbf{x})$ *satisfies* (6) *and* (7)*, and further is strongly convex, then the sequence* $\{f(\mathbf{x}_k)\}$ *has sufficient descent. If* $g_k(\mathbf{x})$ *further satisfies* (8) *and* $\{\mathbf{x}_k\}$ *is bounded, then the sequence* $\{\mathbf{x}_k\}$ *converges to stationary points.*

Remark 1. *If* $g_k(\mathbf{x}) = \dot{g}_k(\mathbf{x}) + \rho/2\|\mathbf{x}-\mathbf{x}_k\|_2^2$*, where* $\dot{g}_k(\mathbf{x})$ *is locally majorant (not necessarily convex) as* (7) *and* $\rho > 0$*, then the strongly convex condition can be removed and the same convergence result holds.*

In the next section, we will give a concrete example on how to construct appropriate surrogates for the RMF problem.

Solving Robust Matrix Factorization by Relaxed MM

Matrix factorization is to factorize a matrix, which usually has missing values and noises, into two matrices. It is widely used for structure from motion (Tomasi and Kanade, 1992), clustering (Kong, Ding, and Huang, 2011), dictionary learning (Mairal et al., 2010), etc. Normally, people aim at minimizing the error between the given matrix and the product of two matrices at the observed entries, measured in squared ℓ_2 norm. Such models are fragile to outliers. Recently, ℓ_1-norm has been suggested to measure the error for enhancing robustness. Such models are thus called robust matrix factorization (RMF). Their formulation is as follows:

$$\min_{U\in\mathcal{C}_u, V\in\mathcal{C}_v} \|W \odot (M - UV^T)\|_1 + R_u(U) + R_v(V), \quad (12)$$

where $M \in \mathbb{R}^{m\times n}$ is the observed matrix and $\|\cdot\|_1$ is the ℓ_1-norm, namely the sum of absolute values in a matrix. $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ are the unknown factor matrices. W is the 0-1 binary mask with the same size as M. The entry value 0 means that the corresponding entry in M is missing, and 1 otherwise. The operator $\odot$ is the Hadamard entry-wise product. $\mathcal{C}_u \subseteq \mathbb{R}^{m\times r}$ and $\mathcal{C}_v \subseteq \mathbb{R}^{n\times r}$ are some closed convex sets, e.g., non-negative cones or balls in some norm. $R_u(U)$ and $R_v(V)$ represent some convex regularizations, e.g., ℓ_1-norm, squared Frobenius norm, or elastic net. By combining different constraints and regularizations, we can get variants of RMF, e.g., low-rank matrix recovery (Ke and Kanade, 2005), non-negative matrix factorization (NMF) (Lee and Seung, 2001), and dictionary learning (Mairal et al., 2010).

Suppose that we have obtained (U_k, V_k) at the k-th iteration. We split (U, V) as the sum of (U_k, V_k) and the unknown increment $(\Delta U, \Delta V)$:

$$(U, V) = (U_k, V_k) + (\Delta U, \Delta V). \quad (13)$$

Then (12) can be rewritten as:

$$\min_{\Delta U+U_k\in\mathcal{C}_u, \Delta V+V_k\in\mathcal{C}_v} F_k(\Delta U, \Delta V) = \|W \odot (M - (U_k + \Delta U)(V_k^T + \Delta V)^T)\|_1 + R_u(U_k + \Delta U) + R_v(V_k + \Delta V). \quad (14)$$

Now we aim at finding an increment $(\Delta U, \Delta V)$ such that the objective function decreases properly. However, problem (14) is not easier than the original problem (12). Inspired by MM, we try to approximate (14) with a convex surrogate.

[3] The proofs and more details can be found at `http://arxiv.org/pdf/1511.08062v1.pdf`

By the triangular inequality of norms, we have the following inequality:

$$F_k(\Delta U, \Delta V) \le \|W \odot (M - U_k V_k^T - \Delta U V_k^T - U_k \Delta V^T)\|_1 + \|W \odot (\Delta U \Delta V^T)\|_1 + R_u(U_k + \Delta U) + R_v(V_k + \Delta V), \tag{15}$$

where the term $\|W \odot (\Delta U \Delta V^T)\|_1$ can be further approximated by $\rho_u/2\|U\|_F^2 + \rho_v/2\|V\|_F^2$, in which ρ_u and ρ_v are some positive constants. Denoting

$$\hat{G}_k(\Delta U, \Delta V) = \|W \odot (M - U_k Y_k^T - \Delta U V_k^T - U_k \Delta V^T)\|_1 + R_u(U_k + \Delta U) + R_v(V_k + \Delta V), \tag{16}$$

we have a surrogate function of $F_k(\Delta U, \Delta V)$ as follows:

$$G_k(\Delta U, \Delta V) = \hat{G}_k(\Delta U, \Delta V) + \frac{\rho_u}{2}\|\Delta U\|_F^2 + \frac{\rho_v}{2}\|\Delta V\|_F^2, \quad \text{s.t. } \Delta U + U_k \in \mathcal{C}_u, \Delta V + V_k \in \mathcal{C}_v. \tag{17}$$

Denoting $\#W_{(i,.)}$ and $\#W_{(.,j)}$ as the number of observed entries in the corresponding column and row of M, respectively, and $\epsilon > 0$ as any positive scalar, we have the following proposition.

Proposition 2. $\hat{G}_k(\Delta U, \Delta V) + \bar{\rho}_u/2\|\Delta U\|_F^2 + \bar{\rho}_v/2\|\Delta V\|_F^2 \ge F_k(\Delta U, \Delta V) \ge \hat{G}_k(\Delta U, \Delta V) - \bar{\rho}_u/2\|\Delta U\|_F^2 - \bar{\rho}_v/2\|\Delta V\|_F^2$ *holds for all possible* $(\Delta U, \Delta V)$ *and the equality holds if and only if* $(\Delta U, \Delta V) = (\mathbf{0}, \mathbf{0})$*, where* $\bar{\rho}_u = \max\{\#W_{(i,.)}, i = 1, \ldots, m\} + \epsilon$*, and* $\bar{\rho}_v = \max\{\#W_{(.,j)}, j = 1, \ldots, n\} + \epsilon$.

By choosing ρ_u and ρ_v in different ways, we have two versions of relaxed MM for RMF: RMF by globally majorant MM (RMF-GMMM for short) and RMF by locally majorant MM (RMF-LMMM for short). In RMF-GMMM, ρ_u and ρ_v are fixed to be $\bar{\rho}_u$ and $\bar{\rho}_v$ in Proposition 2, respectively, throughout the iterations. In RMF-LMMM, ρ_u and ρ_v are instead initialized with relatively small values and then increase gradually, using the line search technique in (Beck and Teboulle, 2009) to ensure the locally majorant condition (7). ρ_u and ρ_v eventually reach the upper bounds $\bar{\rho}_u$ and $\bar{\rho}_v$ in Proposition 2, respectively. As we will show, RMF-LMMM significantly outperforms RMF-GMMM in all our experiments, in both convergence speed and quality of solution.

Since the chosen surrogate G_k naturally fulfills the conditions in Theorem 1, we have the following convergence result for RMF solved by relax MM.

Theorem 2. *By minimizing* (17) *and updating* (U, V) *according to* (13)*, the sequence* $\{F(U_k, V_k)\}$ *has sufficient descent and the sequence* $\{(U_k, V_k)\}$ *converges to stationary points.*

To the best of our knowledge, this is the first convergence guarantee for variants of RMF without extra assumptions. In the following, we will give two examples of RMF in (12).

Two Variants of RMF

Low Rank Matrix Recovery exploits the fact $r \ll \min(m, n)$ to recover the intrinsic low rank data from the measurement matrix with missing data. When the error is measured by the squared Frobenius norm, many algorithms have been proposed (Buchanan and Fitzgibbon, 2005; Mitra, Sheorey, and Chellappa, 2010). For robustness, Ke and Kanade (2005) proposed to adopt the ℓ_1-norm. They minimized U and V alternatively, which could easily get stuck at non-stationary points (Bertsekas, 1999). So Eriksson and van den Hengel (2012) represented V implicitly with U and extended the Wiberg Algorithm to ℓ_1-norm. They only proved the convergence of the objective function value, not the sequence $\{(U_k, V_k)\}$ itself. Moreover, they had to assume that the dependence of V on U is differentiable, which is unlikely to hold everywhere. Additionally, as it unfolds matrix U into a vector and adopt an () its memory requirement is very high, which prevents it from large scale computation. Recently, ADMM was used for matrix recovery. By assuming that the variables are bounded and convergent, Shen, Wen, and Zhang (2014) proved that any accumulation point of their algorithm is the Karush-Kuhn-Tucker (KKT) point. However, the method was only able to handle outliers with magnitudes comparable to the low rank matrix (Shen, Wen, and Zhang, 2014). Moreover, the penalty parameter was fixed, which was not easy to tune for fast convergence. Some researches (Zheng et al., 2012; Cabral et al., 2013) further extended that by adding different regularizations on (U, V) and achieved state-of-the-art performance. As the convergence analysis in (Shen, Wen, and Zhang, 2014) cannot be directly extended, it remains unknown whether the iterates converge to KKT points. In this paper, we adopt the same formulation as (Cabral et al., 2013):

$$\min_{U,V} \|W \odot (M - UV^T)\|_1 + \frac{\lambda_u}{2}\|U\|_F^2 + \frac{\lambda_v}{2}\|V\|_F^2, \tag{18}$$

where the regularizers $\|U\|_F^2$ and $\|V\|_F^2$ are for reducing the solution space (Buchanan and Fitzgibbon, 2005).

Non-negative Matrix Factorization (NMF) has been popular since the seminal work of Lee and Seung (2001). Kong, Ding, and Huang (2011) extended the squared ℓ_2 norm to the ℓ_{21}-norm for robustness. Recently, Pan et al. (2014) further introduced the ℓ_1-norm to handle outliers in non-negative dictionary learning, resulting in the following model:

$$\min_{U \ge 0, V \ge 0} \|M - UV^T\|_1 + \frac{\lambda_u}{2}\|U\|_F^2 + \lambda_v\|V\|_1, \tag{19}$$

where $\|U\|_F^2$ is added in to avoid the trivial solution and $\|V\|_1$ is to induce sparsity. All the three NMF models use multiplicative updating schemes, which only differ in the weights used. The multiplicative updating scheme is intrinsically a globally majorant MM. Assuming that the iterates converge, they proved that the limit of sequence is a stationary point. However, Gonzalez and Zhang (2005) pointed out that with such a multiplicative updating scheme is hard to reach the convergence condition even on toy data.

Minimizing the Surrogate Function

Now we show how to find the minimizer of the convex $G_k(\Delta U, \Delta V)$ in (17). This can be easily done by using the linearized alternating direction method with parallel splitting and adaptive penalty (LADMPSAP) (Lin, Liu, and Li, 2013). LADMPSAP fits for solving the following linearly constrained separable convex programs:

$$\min_{\mathbf{x}_1,\cdots,\mathbf{x}_n} \sum_{j=1}^{n} f_j(\mathbf{x}_j), \quad s.t. \quad \sum_{j=1}^{n} \mathcal{A}_j(\mathbf{x}_j) = \mathbf{b}, \tag{20}$$

where $\mathbf{x}_j$ and $\mathbf{b}$ could be either vectors or matrices, f_j is a proper convex function, and $\mathcal{A}_j$ is a linear mapping. To apply LADMPSAP, we first introduce an auxiliary matrix E such that $E = M - U_k Y_k^T - \Delta U V_k^T - U_k \Delta V^T$. Then minimizing $G_k(\Delta U, \Delta V)$ in (17) can be transformed into:

$$\begin{aligned}
&\min_{E,\Delta U,\Delta V} \|W \odot E\|_1 \\
&+ \left(\frac{\rho_u}{2}\|\Delta U\|_F^2 + R_u(U_k + \Delta U) + \delta_{\mathcal{C}_u}(U_k + \Delta U)\right) \\
&+ \left(\frac{\rho_v}{2}\|\Delta V\|_F^2 + R_v(V_k + \Delta V) + \delta_{\mathcal{C}_v}(V_k + \Delta V)\right), \\
&\text{s.t.} \quad E + \Delta U V_k^T + U_k \Delta V^T = M - U_k Y_k^T,
\end{aligned} \tag{21}$$

where the indicator function $\delta_{\mathcal{C}}(\mathbf{x}) : \mathbb{R}^p \rightarrow \mathbb{R}$ is defined as:

$$\delta_{\mathcal{C}}(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{C}, \\ +\infty, & \text{otherwise.} \end{cases} \tag{22}$$

Then problem (21) naturally fits into the model problem (20). For more details, please refer to Supplementary Materials.

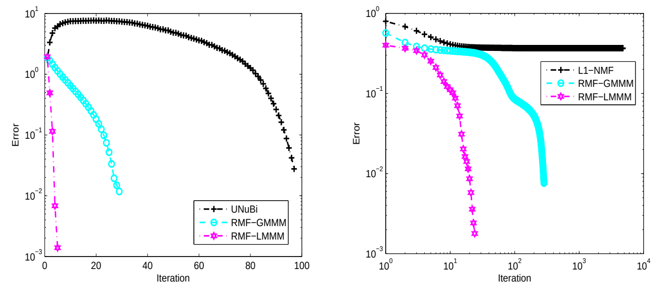
Experiments

In this section, we compare our relaxed MM algorithms with state-of-the-art RMF algorithms: UNuBi (Cabral et al., 2013) for low rank matrix recovery and ℓ_1-NMF (Pan et al., 2014) for robust NMF. The code of UNuBi (Cabral et al., 2013) was kindly provided by the authors. We implemented the code of ℓ_1-NMF (Pan et al., 2014) ourselves.

Synthetic Data

We first conduct experiments on synthetic data. Here we set the regularization parameters $\lambda_u = \lambda_v = 20/(m+n)$ and stop our relaxed MM algorithms when the relative change in the objective function is less than 10^{-4}.

Low Rank Matrix Recovery: We generate a data matrix $M = U_0 V_0^T$, where $U_0 \in \mathbb{R}^{500\times 10}$ and $V_0 \in \mathbb{R}^{500\times 10}$ are sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0,1)$. We additionally corrupt 40% entries of M with outliers uniformly distributed in $[-10, 10]$ and choose W with 80% data missing. The positions of both outliers and missing data are chosen uniformly at random. We initialize all compared algorithms with the rank-r truncation of the singular-value decomposition of $W \odot M$. The performance is evaluated by measuring relative error with the ground truth:



(a) Low-rank Matrix Recovery (b) Non-negative Matrix Factorization

Figure 2: Iteration number versus relative error in log-10 scale on synthetic data. (a) The locally majorant MM, RMF-LMMM, gets better solution in much less iterations than the globally majorant, RMF-GMMM, and the state-of-the-art algorithm, UNuBi (Cabral et al., 2013). (b) RMF-LMMM outperforms the globally majorant MM, RMF-GMMM, and the state-of-the-art algorithm (also globally majorant), ℓ_1-NMF (Pan et al., 2014). The iteration number in (b) is in log-10 scale.

$\|U_{est}V_{est}^T - U_0 V_0^T\|_1/(mn)$, where U_{est} and V_{est} are the estimated matrices. The results are shown in Fig. 2(a), where RMF-LMMM reaches the lowest relative error in much less iterations.

Non-negative Matrix Factorization: We generate a data matrix $M = U_0 V_0^T$, where $U_0 \in \mathbb{R}^{500\times 10}$ and $V_0 \in \mathbb{R}^{500\times 10}$ are sampled i.i.d. from a uniform distribution $\mathcal{U}(0,1)$. For sparsity, we further randomly set 30% entries of V as 0. We further corrupt 40% entries of M with outliers uniformly distributed in $[0, 10]$. All the algorithms are initialized with the same non-negative random matrix. The results are shown in Fig. 2(b), where RMF-LMMM also gets the best result with much less iterations. ℓ_1-NMF tends to be stagnant and cannot approach a high precision solution even after 5000 iterations.

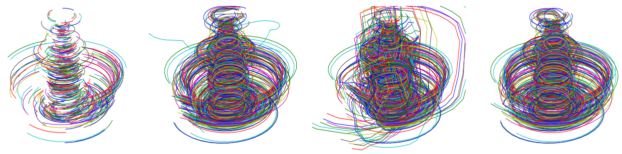
Real Data

In this subsection, we conduct experiment on real data. Since there is no ground truth, we measure the relative error by $\|W \odot (M_{est} - M)\|_1/\#W$, where $\#W$ is the number of observed entries. For NMF, W becomes an all-one matrix.

Low Rank Matrix Recovery: Tomasi and Kanade (1992) first modelled the affine rigid structure from motion as a rank-4 matrix recovery problem. Here we use the famous Oxford Dinosaur sequence [4], which consists of 36 images with a resolution of 720×576 pixels. We pick out a portion of the raw feature points which are observed by at least 6 views (Fig. 3(a)). The observed matrix is of size 72×557 with a missing data ratio 79.5% and shows a band diagonal pattern. We register the image origin to the image center, $(360, 288)$. We adopt the same initialization and parameter setting as the synthetic data above.

Figures 3(b)-(d) show the full tracks reconstructed by all algorithms. As the dinosaur sequence is taken on a turntable,

[4] http://www.robots.ox.ac.uk/~vgg/data1.html



(a) Raw Data (b) UNuBi Error=0.336 (c) RMF-GMMM Error=0.488 (d) RMF-LMMM Error=**0.322**

Figure 3: Original incomplete and recovered data of the Dinosaur sequence. (a) Raw input tracks. (b-d) Full tracks reconstructed by UNuBi (Cabral et al., 2013), RMF-GMMM, and RMF-LMMM, respectively. The relative error is presented below the tracks.

all the tracks are supposed to be circular. Among them, the tracks reconstructed by RMF-GMMM are the most inferior. UNuBi gives reasonably good results. However, most of the reconstructed tracks in large radii do not appear closed. Some tracks in the upper part are not reconstructed well either, including one obvious failure. In contrast, almost all the tracks reconstructed by RMF-LMMM are circular and appear closed, which are the most visually plausible. The lowest relative error also confirms the effectiveness of RMF-LMMM.

Non-negative Matrix Factorization: We test the performance of robust NMF by clustering (Kong, Ding, and Huang, 2011; Pan et al., 2014). The experiments are conducted on four benchmark datasets of face images, which includes: AT&T, UMIST, a subset of PIE [5] and a subset of AR [6]. We use the first 10 images in each class for PIE and the first 13 images for AR. The descriptions of the datasets are summarized in the second row of Table 2. The evaluation metrics we use here are accuracy (ACC), normalized mutual information (NMI) and purity (PUR) (Kong, Ding, and Huang, 2011; Pan et al., 2014). We change the regularization parameter λ_u to $2000/(m+n)$ and maintain λ_v as $20/(m+n)$. The number r of clusters is equal to the number of classes in each dataset. We adopt the initializations in (Kong, Ding, and Huang, 2011). Firstly, we use the principal component analysis (PCA) to get a subspace with dimension r. Then we employ k-means on the PCA-reduced data to get the clustering results V'. Finally, V is initialized as $V = V' + 0.3$ and U is by computing the clustering centroid for each class. We empirically terminate ℓ_1-NMF, RMF-GMMM, and RMF-LMMM after 5000, 500, and 20 iterations, respectively. The clustering results are shown in the last three rows of Table 2. We can see that RMF-LMMM achieves tremendous improvements over the two majorant algorithms across all datasets. RMF-GMMM is also better than ℓ_1-NMF. The lowest relative error in the third row shows that RMF-LMMM can always approximate the measurement matrix much better than the other two.

[5] `http://www.zjucadcg.cn/dengcai/Data/data.html`

[6] `http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html`

Table 2: Dataset descriptions, relative errors, and clustering results. For all the three metrics, larger values are better.

Dataset		AT&T	UMINT	CMUPIE	AR
Desc	# Size	400	360	680	1300
	# Dim	644	625	576	540
	# Class	40	20	68	100
Error	L_1-NMF	13.10	14.52	15.33	16.96
	RMF-GMMM	13.10	11.91	14.64	17.16
	RMF-LMMM	**9.69**	**9.26**	**4.67**	**7.53**
ACC	L_1-NMF	0.5250	0.7550	0.2176	0.1238
	RMF-GMMM	0.5325	0.7889	0.2691	0.1085
	RMF-LMMM	**0.7250**	**0.8333**	**0.4250**	**0.2015**
NMI	L_1-NMF	0.7304	0.8789	0.5189	0.4441
	RMF-GMMM	0.7499	0.8744	0.5433	0.4289
	RMF-LMMM	**0.8655**	**0.9012**	**0.6654**	**0.4946**
PUR	L_1-NMF	0.5700	0.8139	0.2368	0.1138
	RMF-GMMM	0.5725	0.8222	0.2897	0.1354
	RMF-LMMM	**0.7500**	**0.8778**	**0.4456**	**0.2192**

Conclusions

In this paper, we propose a weaker condition on surrogates in MM, which enables better approximation of the objective function. Our relaxed MM is general enough to include the existing MM methods. In particular, the non-smooth and non-convex objective function can be approximated directly, which is never done before. Using the RMF problems as examples, our locally majorant relaxed MM beats the state-of-the-art methods with margin, in both solution quality and convergence speed. We prove that the iterates converge to stationary points. To our best knowledge, this is the first convergence guarantee for variants of RMF without extra assumptions.

Acknowledgements

Zhouchen Lin is supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502), National Natural Science Foundation of China (NSFC) (grant no. 61272341 and 61231002), and Microsoft Research Asia Collaborative Research Program. Zhenyu Zhao is supported by NSFC (Grant no. 61473302). Hongbin Zha is supported by 973 Program (grant no. 2011CB302202).

References

Attouch, H.; Bolte, J.; and Svaiter, B. F. 2013. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming* 137(1-2):91–129.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1):183–202.

Bertsekas, D. 1999. *Nonlinear Programming*. Athena Scientific, 2nd edition.

Borwein, J. M., and Lewis, A. S. 2010. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media.

Bruckstein, A. M.; Donoho, D. L.; and Elad, M. 2009. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review* 51(1):34–81.

Buchanan, A. M., and Fitzgibbon, A. W. 2005. Damped Newton algorithms for matrix factorization with missing data. In *CVPR*, volume 2, 316–322. IEEE.

Burke, J. V.; Lewis, A. S.; and Overton, M. L. 2005. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization* 15(3):751–779.

Cabral, R.; Torre, F. D. L.; Costeira, J. P.; and Bernardino, A. 2013. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *ICCV*, 2488–2495. IEEE.

Candes, E. J.; Wakin, M. B.; and Boyd, S. P. 2008. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications* 14(5-6):877–905.

Chen, X. 2012. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical Programming* 134(1):71–99.

Clarke, F. H. 1990. *Optimization and nonsmooth analysis*, volume 5. SIAM.

Eriksson, A., and van den Hengel, A. 2012. Efficient computation of robust weighted low-rank matrix approximations using the L_1 norm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34(9):1681–1690.

Gonzalez, E. F., and Zhang, Y. 2005. Accelerating the Lee-Seung algorithm for non-negative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*.

Hunter, D. R., and Lange, K. 2004. A tutorial on MM algorithms. *The American Statistician* 58(1):30–37.

Ke, Q., and Kanade, T. 2005. Robust L_1-norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR*, volume 1, 739–746. IEEE.

Kong, D.; Ding, C.; and Huang, H. 2011. Robust nonnegative matrix factorization using L_{21}-norm. In *CIKM*, 673–682. ACM.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *NIPS*, 556–562.

Lin, Z.; Liu, R.; and Li, H. 2013. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. *Machine Learning* 1–39.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research* 11:19–60.

Mairal, J. 2013. Optimization with first-order surrogate functions. *arXiv preprint arXiv:1305.3120*.

Mäkelä, M. 2002. Survey of bundle methods for nonsmooth optimization. *Optimization Methods and Software* 17(1):1–29.

Mitra, K.; Sheorey, S.; and Chellappa, R. 2010. Large-scale matrix factorization with missing data under additional constraints. In *NIPS*, 1651–1659.

Pan, Q.; Kong, D.; Ding, C.; and Luo, B. 2014. Robust nonnegative dictionary learning. In *AAAI*.

Razaviyayn, M.; Hong, M.; and Luo, Z.-Q. 2013. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization* 23(2):1126–1153.

Shen, Y.; Wen, Z.; and Zhang, Y. 2014. Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software* 29(2):239–263.

Tomasi, C., and Kanade, T. 1992. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision* 9(2):137–154.

Zheng, Y.; Liu, G.; Sugimoto, S.; Yan, S.; and Okutomi, M. 2012. Practical low-rank matrix approximation under robust L_1-norm. In *CVPR*, 1410–1417. IEEE.