Beyond Alternating Updates for Matrix Factorization with Inertial Bregman Proximal Gradient Algorithms

Mahesh Chandra Mukkamala* Peter Ochs†

Abstract

Matrix Factorization is a popular non-convex objective, for which alternating minimization schemes are mostly used. They usually suffer from the major drawback that the solution is biased towards one of the optimization variables. A remedy is non-alternating schemes. However, due to a lack of Lipschitz continuity of the gradient in matrix factorization problems, convergence cannot be guaranteed. A recently developed remedy relies on the concept of Bregman distances, which generalizes the standard Euclidean distance. We exploit this theory by proposing a novel Bregman distance for matrix factorization problems, which, at the same time, allows for simple/closed form update steps. Therefore, for non-alternating schemes, such as the recently introduced Bregman Proximal Gradient (BPG) method and an inertial variant Convex–Concave Inertial BPG (CoCaIn BPG), convergence of the whole sequence to a stationary point is proved for Matrix Factorization. In several experiments, we observe a superior performance of our non-alternating schemes in terms of speed and objective value at the limit point.

2010 Mathematics Subject Classification: Primary 90C26; Secondary 26B25, 90C30, 49M27, 47J25, 52A41, 65K05, 65F22.
Keywords: Composite nonconvex nonsmooth minimization, non Euclidean distances, Bregman distance, Bregman proximal gradient method, inertial methods, matrix factorization, matrix completion.

1 Introduction

Matrix factorization has numerous applications in Machine Learning [42, 55], Computer Vision [16, 56, 60, 27], Bio-informatics [54, 11] and many others. Given a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, one is interested in the factors $\mathbf{U} \in \mathbb{R}^{M\times K}$ and $\mathbf{Z} \in \mathbb{R}^{K\times M}$ such that $\mathbf{A} \approx \mathbf{UZ}$ holds. This is usually cast into the following non-convex optimization problem

$$\min_{\mathbf{U}\in\mathcal{U},\mathbf{Z}\in\mathcal{Z}} \left\{ \Psi \equiv \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \mathcal{R}_1(\mathbf{U}) + \mathcal{R}_2(\mathbf{Z}) \right\} , \tag{1.1}$$

where $\mathcal{U}, \mathcal{Z}$ are constraint sets and $\mathcal{R}_1, \mathcal{R}_2$ are regularization terms. The most frequently used techniques for solving matrix factorization problems involve Gauss-Seidel method [25] based alternating updates like PALM [7], iPALM [52], BCD [61], BC-VMFB [17], HALS [18] and many others. A common disadvantage of these schemes is their bias towards one of the variables. Such alternating schemes involve fixing a subset of variables to do the updates. Typically, Lipschitz continuity of gradients holds for remaining variables, which is used to obtain the convergence to a stationary point. However, in general Lipschitz continuity of gradients fails to hold for all variables. The same problem appears in various practical applications such as Quadratic Inverse Problems, Poisson Linear Inverse Problems, Cubic Regularized Non-convex Quadratic Problems and Robust Denoising Problems with Non-convex

*Faculty of Mathematics and Computer Science, Saarland University, 66123 Saarbrücken, Germany, E-mail: mukkamala@math.uni-sb.de
†Faculty of Mathematics and Computer Science, Saarland University, 66123 Saarbrücken, Germany, E-mail: ochs@math.uni-sb.de

Total Variation Regularization [45, 8, 4]. They belong to the following broad class of additive non-convex composite minimization problems

$$(\mathcal{P}) \qquad \inf\left\{\Psi \equiv f\left(x\right)+g\left(x\right):\ x\in\overline{C}\right\}, \tag{1.2}$$

where f and g are potentially non-convex extended real valued functions, g is a smooth function and $\overline{C}$ is a nonempty, closed, convex set in $\mathbb{R}^d$. The extensions of Lipschitz continuity of gradients was initiated by [4] in convex setting and for non-convex problems in [8]. Such extensions are based on a generalized proximity measure known as Bregman distance and have recently led to new algorithms to solve (1.2): Bregman Proximal Gradient (BPG) method [8] and its inertial variant Convex–Concave Inertial BPG (CoCaIn BPG) [45].

BPG uses proximal gradient method like step but with Bregman distance as the proximity measure. Its convergence theory relies on L-smad property, an extension of Lipschitz continuity of gradients for non-convex problems [8]. It involves an upper bound and a lower bound, where the upper bound involving a convex majorant controls the step-size of BPG. However, the significance of lower bounds for BPG was not clear. In non-convex optimization literature, the lower bounds which involve concave minorants were largely ignored. Recently, extending on [59, 49], CoCaIn BPG changed this trend by justifying the usage of lower bounds to incorporate inertia for faster convergence [45]. Moreover, the generated inertia is adaptive, in the sense that it changes according to the function behavior. Thus, CoCaIn BPG does not use an iteration number dependent inertial parameter, unlike Nesterov Accelerated Gradient (NAG) method [46] (also FISTA [5, 59]) in the convex setting.

In this paper we ask the question, *"Can we apply BPG and CoCaIn BPG efficiently for Matrix Factorization problems?"*. This question is significant, since convergence of the Bregman minimization variants BPG and CoCaIn BPG relies on the L-smad property, which is non-trivial and is an open problem for Matrix Factorization. Another crucial issue is the efficient computability of the algorithm's update steps, which is particularly hard due to the coupling between two subsets of variables. We successfully solve these challenges with our contributions, which we review below.

Contributions. We make recently introduced powerful Bregman minimization based algorithms BPG [8] and CoCaIn BPG [45] and the corresponding convergence results applicable to the matrix factorization problems. Experiments show a significant advantage of BPG and CoCaIn BPG which are non-alternating by construction, compared to popular alternating minimization schemes in particular PALM [7] and iPALM [52]. The proposed algorithms require the following non-trivial contributions:

- We propose a novel Bregman distance for Matrix Factorization with the following auxiliary function (called kernel generating distance) with certain $c_1, c_2 > 0$:

$$h(\mathbf{U},\mathbf{Z}) = c_1\left(\frac{\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2}{2}\right)^2 + c_2\left(\frac{\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2}{2}\right).$$

  The generated Bregman distance embeds the crucial coupling between the variables $\mathbf{U}$, $\mathbf{Z}$. We prove the L-smad property with such a kernel generating distance and infer convergence of BPG and CoCaIn BPG to a stationary point.

- We compute the analytic solution for subproblems of proposed variants of BPG and CoCaIn BPG, where the usual analytic solutions based on Euclidean distances cannot be used.

We illustrate below the proposed variant of BPG on a simple problem. Here, our Bregman distance with BPG incurs an update step like Gradient Descent, but with an additional scaling step based on a cubic equation. The scaling step is crucial for convergence and is an artifact of the Bregman distance.

Simple Illustration of BPG for Matrix Factorization. Consider the following simple matrix factorization optimization problem, where we set $\mathcal{R}_1 := 0$ and $\mathcal{R}_2 := 0$ in (1.1)

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2 \right\} . \tag{1.3}$$

For this problem, the update steps of **Bregman Proximal Gradient for Matrix Factorization (BPG-MF)** given in Section 2.1 (also see Section 2.4) with a chosen $\lambda \in (0,1)$ are the following:

In each iteration, compute $t_k = 3(\left\|\mathbf{U^k}\right\|_F^2 + \left\|\mathbf{Z^k}\right\|_F^2) + \left\|\mathbf{A}\right\|_F$ and perform the intermediary gradient descent steps (non-alternating) for $\mathbf{U}$ and $\mathbf{Z}$ independently with step-size $\frac{\lambda}{t_k}$:

$$\mathbf{P^k} = \mathbf{U^k} - \frac{\lambda}{t_k}\left[(\mathbf{U^k}\mathbf{Z^k}-\mathbf{A})(\mathbf{Z^k})^T\right] , \quad \mathbf{Q^k} = \mathbf{Z}^k - \frac{\lambda}{t_k}\left[(\mathbf{U^k})^T(\mathbf{U^k}\mathbf{Z^k}-\mathbf{A})\right] .$$

Then, the additional scaling steps $\mathbf{U^{k+1}} = rt_k\mathbf{P^k}$ and $\mathbf{Z^{k+1}} = rt_k\mathbf{Q^k}$ are required, where the scaling factor $r \geq 0$ satisfies a cubic equation: $3t_k^2\left(\left\|\mathbf{P^k}\right\|_F^2 + \left\|\mathbf{Q^k}\right\|_F^2\right) r^3 + \left\|\mathbf{A}\right\|_F r - 1 = 0$.

1.1 Related Work

Alternating Minimization is the go-to strategy for matrix factorization problems due to coupling between two subsets of variables [23, 1, 62]. In the context of non-convex and non-smooth optimization, recently PALM [7] was proposed and convergence to stationary point was proved. An inertial variant, iPALM was proposed in [52]. However, such methods require a subset of variables to be fixed. We remove such a restriction here and take the contrary view by proposing non-alternating schemes based on powerful Bregman proximal minimization framework, which we review below.

Bregman Proximal Minimization framework extends upon the standard proximal minimization, where Bregman distances are used as proximity measures. This work initiated by [4] in convex setting inspired various extensions to non-convex optimization [8]. Related inertial variants were proposed in [45, 65]. Related line-search methods were proposed in [51] based on [9, 10]. More related works in convex optimization include [48, 39, 41]. Recently, the symmetric non-negative matrix factorization problem was solved with a non-alternating Bregman proximal minimization scheme [20] with the following kernel generating distance

$$h(\mathbf{U}) = \frac{\left\|\mathbf{U}\right\|_F^4}{4} + \frac{\left\|\mathbf{U}\right\|_F^2}{2} .$$

However for the following applications, such a h is not suitable, unlike our Bregman distance.

Non-negative Matrix Factorization (NMF) is a variant of matrix factorization problem which requires the factors to have non-negative entries [24, 36]. Some applications are hyperspectral unmixing, clustering and others [23, 21]. The non-negativity constraints pose new challenges [36] and only convergence to a stationary point [23, 30] is guaranteed, as NMF is NP-hard in general. Under certain restrictions, NMF can be solved exactly [2, 43] but such methods are computationally infeasible. We give efficient algorithms for NMF and show the superior performance empirically.

Matrix Completion is another variant of matrix factorization problem arising in recommender systems [34] and Bio-informatics [38, 58], is an active research topic due to the hard non-convex optimization problem [14, 22]. The state-of-the art methods were proposed in [32, 63] and other recent methods include [64]. Here, our algorithms are either faster or competitive.

Our algorithms are also applicable to Graph Regularized NMF (GNMF) [12], Sparse NMF [7], Nuclear Norm Regularized problems [13, 31], Symmetric NMF via non-symmetric extension [66].

2 Matrix Factorization Problem Setting and Algorithms

Notation. We refer to [53] for standard notation, unless specified otherwise.

Formally, in a matrix factorization problem, given a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ we want to obtain the factors $\mathbf{U} \in \mathbb{R}^{M\times K}$ and $\mathbf{Z}^{K\times N}$ such that $\mathbf{A} \approx \mathbf{UZ}$, which is captured by the following non-convex problem

$$\min_{\mathbf{U}\in\mathcal{U},\mathbf{Z}\in\mathcal{Z}} \left\{\Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \mathcal{R}_1(\mathbf{U}) + \mathcal{R}_2(\mathbf{Z})\right\}, \tag{2.1}$$

where $\mathcal{R}_1(\mathbf{U}) + \mathcal{R}_2(\mathbf{Z})$ is the separable regularization term, $\frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2$ is the data-fitting term, and $\mathcal{U},\mathcal{Z}$ are the constraint sets for $\mathbf{U}$ and $\mathbf{Z}$ respectively. Here, $\mathcal{R}_1(\mathbf{U})$ and $\mathcal{R}_2(\mathbf{Z})$ can be potentially non-convex extended real valued functions and possibly non-smooth. In this paper, we propose to make use of BPG and its inertial variant CoCaIn BPG to solve (2.1). The introduction of these algorithms requires the following preliminary considerations.

Definition 2.1. (Kernel Generating Distance [8]) Let C be a nonempty, convex and open subset of $\mathbb{R}^d$. Associated with C, a function $h : \mathbb{R}^d \to (-\infty,+\infty]$ is called a *kernel generating distance* if it satisfies:

(i) h is proper, lower semicontinuous and convex, with $\operatorname{dom} h \subset \overline{C}$ and $\operatorname{dom}\partial h = C$.

(ii) h is C^1 on $\operatorname{int}\operatorname{dom} h \equiv C$.

We denote the class of kernel generating distances by $\mathcal{G}(C)$.

For every $h \in \mathcal{G}(C)$, the associated Bregman distance is given by $D_h : \operatorname{dom} h \times \operatorname{int}\operatorname{dom} h \to \mathbb{R}_+$:

$$D_h(x,y) := h(x) - \left[h(y) + \langle \nabla h(y), x-y\rangle\right].$$

For examples, consider the following kernel generating distances:

$$h_0(x) = \frac{1}{2}\|x\|^2, \quad h_1(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2 \quad \text{and} \quad h_2(x) = \frac{1}{3}\|x\|^3 + \frac{1}{2}\|x\|^2.$$

It is easy to see that the Bregman distance associated with $h_0(x)$ is the Euclidean distance. The Bregman distances associated with h_1 and h_2 appear in the context of non-convex quadratic inverse problems [8, 45] and non-convex cubic regularized problems [45] respectively. For a review on the recent literature, we refer the reader to [57] and for early work on Bregman distances to [15].
These distance measures are key for development of algorithms for the following class of non-convex additive composite problems

$$(\mathcal{P}) \qquad \inf\left\{\Psi \equiv f(x) + g(x) : \ x \in \overline{C}\right\}, \tag{2.2}$$

which is assumed to satisfy the following standard assumption [8].

Assumption A. (i) $h \in \mathcal{G}(C)$ with $\overline{C} = \overline{\operatorname{dom} h}$.

(ii) $f : \mathbb{R}^d \to (-\infty,+\infty]$ is a proper and lower semicontinuous function (potentially non-convex) with $\operatorname{dom} f \cap C \neq \emptyset$.

(iii) $g : \mathbb{R}^d \to (-\infty,+\infty]$ is a proper and lower semicontinuous function (potentially non-convex) with $\operatorname{dom} h \subset \operatorname{dom} g$, which is continuously differentiable on C.

(iv) $v(\mathcal{P}) := \inf\left\{\Psi(x) : \ x \in \overline{C}\right\} > -\infty$.

Matrix Factorization Example. A special case of (2.2) is the following problem,

$$\inf \left\{ \Psi(\mathbf{U}, \mathbf{Z}) := f_1(\mathbf{U}) + f_2(\mathbf{Z}) + g(\mathbf{U}, \mathbf{Z}) : (\mathbf{U}, \mathbf{Z}) \in \overline{C} \right\} . \tag{2.3}$$

We denote $f(\mathbf{U}, \mathbf{Z}) = f_1(\mathbf{U}) + f_2(\mathbf{Z})$. Many practical matrix factorization problems can be cast into the form of (2.1). The choice of f and g is dependent on the problem, for which we provide some examples in Section 3. Here f_1, f_2 satisfy the assumptions of f with dimensions chosen accordingly. Moreover by definition, f is separable in $\mathbf{U}$ and $\mathbf{Z}$, which we assume only for practical reasons. Also, the choice of f, g may not be unique. For example, in (2.1) when $\mathcal{R}_1(\mathbf{U}) = \frac{\lambda_0}{2} \|\mathbf{U}\|_F^2$ and $\mathcal{R}_2(\mathbf{Z}) = \frac{\lambda_0}{2} \|\mathbf{Z}\|_F^2$ the choice of f as in (2.3) can be $\mathcal{R}_1 + \mathcal{R}_2$ and $g = \frac{1}{2} \|\mathbf{A} - \mathbf{U}\mathbf{Z}\|_F^2$. However, the other choice is to set $g = \Psi$ and $f := 0$.

2.1 BPG-MF: Bregman Proximal Gradient for Matrix Factorization

We require the notion of Bregman Proximal Gradient Mapping [8, Section 3.1] given by

$$T_\lambda(x) \in \operatorname{argmin} \left\{ f(u) + \langle \nabla g(x), u - x \rangle + \frac{1}{\lambda} D_h(u, x) : \ u \in \overline{C} \right\} . \tag{2.4}$$

Then, the update step of Bregman Proximal Gradient (BPG) [8] for solving (2.2) is $x^{k+1} \in T_\lambda(x^k)$, for some $\lambda > 0$ and $h \in \mathcal{G}(C)$. Convergence of BPG relies on a generalized notion of Lipschitz continuity, the so-called L-smad property (Defintion 2.2).

Beyond Lipschitz continuity. BPG extends upon the popular proximal gradient methods, for which convergence relies on Lipschitz continuity of the smooth part of the objective in (2.2). However, such a notion of Lipschitz continuity is restrictive for many practical applications such as Poisson linear inverse problems [4], quadratic inverse problems [8, 45], cubic regularized problems [45] and robust denoising problems with non-convex total variation regularization [45]. The extensions for generalized notions of Lipschitz continuity of gradients is an active area of research [4, 39, 8]. We consider the following from [8].

Definition 2.2 (L-smad property)**.** The function g is said to be L-smooth adaptable (L-smad) on C with respect to h, if and only if $Lh - g$ and $Lh + g$ are convex on C.

Note that L-smad property is equivalent to having Lipschitz continuous gradients when $h(x) = \frac{1}{2} \|x\|^2$. Therefore, a non-trivial example is $f(x) = x^4$. It is L-smad with $h(x) = x^4$ and $L \geq 1$, however ∇f is not Lipschitz continuous.

Now, we are ready to present the BPG algorithm for Matrix Factorization.

BPG-MF: BPG for Matrix Factorization.
Input. Choose $h \in \mathcal{G}(C)$ with $C \equiv \operatorname{int} \operatorname{dom} h$ such that g satisfies L-smad with respect to h on C.
Initialization. $(\mathbf{U^1}, \mathbf{Z^1}) \in \operatorname{int} \operatorname{dom} h$ and let $\lambda > 0$.
General Step. For $k = 1, 2, \ldots$, compute

$$\mathbf{P^k} = \lambda \nabla_{\mathbf{U}} g\left(\mathbf{U^k}, \mathbf{Z^k}\right) - \nabla_{\mathbf{U}} h(\mathbf{U^k}, \mathbf{Z^k}), \quad \mathbf{Q^k} = \lambda \nabla_{\mathbf{Z}} g\left(\mathbf{U^k}, \mathbf{Z^k}\right) - \nabla_{\mathbf{Z}} h(\mathbf{U^k}, \mathbf{Z^k}),$$

$$(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}) \in \underset{(\mathbf{U},\mathbf{Z}) \in \overline{C}}{\operatorname{argmin}} \left\{ \lambda f(\mathbf{U}, \mathbf{Z}) + \left\langle \mathbf{P^k}, \mathbf{U} \right\rangle + \left\langle \mathbf{Q^k}, \mathbf{Z^k}), \mathbf{Z} \right\rangle + h(\mathbf{U}, \mathbf{Z}) \right\} . \tag{2.5}$$

Under Assumption A and the following one (mostly satisfied in practice), BPG is well-defined [8].

Assumption B. The range of T_λ lies in C and, for all $\lambda > 0$, the function $h + \lambda f$ is supercoercive.

The update step for BPG-MF is easy to derive from BPG, however convergence of BPG also relies on the "right" choice of kernel generating distance h and the L-smad condition. Finding h such that

L-smad holds (also see Section 2.2) and that the update step can be given in closed form (also see Section 2.4) is our main contribution and allows us to invoke the convergence results from [8]. The convergence result states that the whole sequence of iterates generated by BPG-MF converges to a stationary point, precisely given in Theorem 2.2. The result depends on the non-smooth KL-property (see [6, 3, 7]) which is a mild requirement and is satisfied by most practical objectives. We provide below the convergence result in [8, Theorem 4.1] adapted to BPG-MF.

Theorem 2.1 (Global Convergence of BPG-MF)**.** *Let Assumptions A and B hold and let g be L-smad with respect to h, where h is assumed to be σ-strongly convex with full domain. Assume $\nabla g, \nabla h$ to be Lipschitz continuous on any bounded subset. Let $\left\{(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}})\right\}_{k\in\mathbb{N}}$ be a bounded sequence generated by BPG-MF with $0 < \lambda L < 1$, and suppose Ψ satisfies the KL property, then, such a sequence has finite length, and converges to a critical point.*

2.2 New Bregman Distance for Matrix Factorization

We prove the L-smad property for the term $g(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A} - \mathbf{U}\mathbf{Z}\right\|_F^2$ of the matrix factorization problem in (2.1). The kernel generating distance is a linear combination of

$$h_1(\mathbf{U},\mathbf{Z}) := \left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2}\right)^2 \quad \text{and} \quad h_2(\mathbf{U},\mathbf{Z}) := \frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2}\,. \tag{2.6}$$

and it is designed to also allow for closed form updates (see Section 2.4).

Proposition 2.1. *Let g, h_1, h_2 be as defined above. Then, for a certain constant $L \geq 1$, the function g satisfies the L-smad property with respect to the following kernel generating distance*

$$h_a(\mathbf{U},\mathbf{Z}) = 3h_1(\mathbf{U},\mathbf{Z}) + \|\mathbf{A}\|_F\, h_2(\mathbf{U},\mathbf{Z})\,. \tag{2.7}$$

The proof is given in Section F.1 in the appendix. The Bregman distances considered in previous works [45, 8] are separable and not applicable for matrix factorization problems. The inherent coupling between two subsets of variables $\mathbf{U},\mathbf{Z}$ is the main source of non-convexity in the objective g. The kernel generating distance (in particular h_1 in (2.7)) contains the interaction/coupling terms between $\mathbf{U},\mathbf{Z}$ which makes it amenable for matrix factorization problems.

2.3 CoCaIn BPG-MF: An Adaptive Inertial Bregman Proximal Gradient Method

The goal of this section is to introduce an inertial variant of BPG-MF, called CoCaIn BPG-MF. The effective step-size choice for BPG-MF can be restrictive due to large constant like $\|\mathbf{A}\|_F$ (see (2.7)), for which we present a practical example in the numerical experiments. In order to allow for larger step-sizes, one needs to adapt it locally, which is often done via a backtracking procedure. CoCaIn BPG-MF combines inertial steps with a novel backtracking procedure proposed in [45].

Inertial algorithms often lead to better convergence [50, 52, 45]. The classical Nesterov Accelerated Gradient (NAG) method [46] and the popular Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [5] employ an extrapolation based inertial strategy. However, the extrapolation is governed by a parameter which is typically scheduled to follow certain iteration-dependent scheme [46, 28]and is restricted to the convex setting. Recently with Convex–Concave Inertial Bregman Proximal Gradient (CoCaIn BPG) [45], it was shown that one could leverage the upper bound (convexity of $Lh - g$) and lower bound (convexity of $Lh + g$) to incorporate inertia in an adaptive manner.

We recall now the update steps of CoCaIn BPG [45] to solve (2.2). Let $h \in \mathcal{G}(C)$, $\lambda > 0$, and $x^0 = x^1 \in \mathbb{R}^d$ be an initalization, then in each iteration the extrapolated point $y^k = x^k + \gamma_k(x^k - x^{k-1})$ is computed followed by a BPG like update (at y^k) given by $x^{k+1} \in T_{\tau_k}(y^k)$, where γ_k is the inertial parameter and τ_k is the step-size parameter. Similar conditions to BPG are required for the convergence to a stationary point. We use CoCaIn BPG for Matrix Factorization (CoCaIn BPG-MF) and our proposed novel kernel generating distance h from (2.7) makes the convergence results of [45] applicable. Along with Assumption B, we require the following assumption.

Assumption C. (i) There exists $\alpha \in \mathbb{R}$ such that $f(\mathbf{U},\mathbf{Z}) - \frac{\alpha}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2\right)$ is convex.

(ii) The kernel generating distance h is σ-strongly convex on $\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}$.

The Assumption C(i) refers to notion of semi-convexity of the function f, (see [49, 45]) and seems to be closely connected to the inertial feature of an algorithm. For notational brevity, we use $D_g(x,y) := g(x) - [g(y) + \langle \nabla g(y), x-y\rangle]$ which may also be negative if g is not a kernel generating distance. Moreover, we use $D_h((\mathbf{X}_1,\mathbf{Y}_1),(\mathbf{X}_2,\mathbf{Y}_2))$ as $D_h(\mathbf{X}_1,\mathbf{Y}_1,\mathbf{X}_2,\mathbf{Y}_2)$. We provide CoCaIn BPG-MF below.

CoCaIn BPG-MF: Convex–Concave Inertial BPG for Matrix Factorization.
Input. Choose $\delta,\varepsilon > 0$ with $1 > \delta > \epsilon$, $h \in \mathcal{G}(C)$ with $C \equiv \operatorname{int}\operatorname{dom} h$ satisfying L-smad on C.
Initialization. $(\mathbf{U^1},\mathbf{Z^1}) = (\mathbf{U^0},\mathbf{Z^0}) \in \operatorname{int}\operatorname{dom} h \cap \operatorname{dom} f$, $\bar{L}_0 > \frac{-\alpha}{(1-\delta)\sigma}$ and $\tau_0 \le \bar{L}_0^{-1}$.
General Step. For $k = 1,2,\ldots$, compute extrapolated points

$$Y_{\mathbf{U}}^{\mathbf{k}} = \mathbf{U}^k + \gamma_k\left(\mathbf{U^k} - \mathbf{U^{k-1}}\right) \quad \text{and} \quad Y_{\mathbf{Z}}^{\mathbf{k}} = \mathbf{Z}^k + \gamma_k\left(\mathbf{Z^k} - \mathbf{Z^{k-1}}\right), \tag{2.8}$$

where $\gamma_k \ge 0$ such that

$$(\delta-\varepsilon)D_h\left(\mathbf{U^{k-1}},\mathbf{Z^{k-1}},\mathbf{U^k},\mathbf{Z^k}\right) \ge (1+\underline{L}_k\tau_{k-1})D_h\left(\mathbf{U^k},\mathbf{Z^k},Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}\right), \tag{2.9}$$

where $\underline{L}_k$ satisfies

$$D_g\left(\mathbf{U^k},\mathbf{Z^k},Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}\right) \ge -\underline{L}_k D_h\left(\mathbf{U^k},\mathbf{Z^k},Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}\right). \tag{2.10}$$

Choose $\bar{L}_k \ge \bar{L}_{k-1}$, and set $\tau_k \le \min\{\tau_{k-1},\bar{L}_k^{-1}\}$. Now, compute

$$\mathbf{P^k} = \tau_k\nabla_{\mathbf{U}}g\left(Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}\right) - \nabla_{\mathbf{U}}h(Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}), \quad \mathbf{Q^k} = \tau_k\nabla_{\mathbf{Z}}g\left(Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}\right) - \nabla_{\mathbf{Z}}h(Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}),$$

$$(\mathbf{U^{k+1}},\mathbf{Z^{k+1}}) \in \underset{(\mathbf{U},\mathbf{Z})\in\overline{C}}{\operatorname{argmin}}\left\{\tau_k f(\mathbf{U},\mathbf{Z}) + \left\langle\mathbf{P^k},\mathbf{U}\right\rangle + \left\langle\mathbf{Q^k},\mathbf{Z}\right\rangle + h(\mathbf{U},\mathbf{Z})\right\}, \tag{2.11}$$

such that $\bar{L}_k$ satisfies

$$D_g\left(\mathbf{U^{k+1}},\mathbf{Z^{k+1}},Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}\right) \le \bar{L}_k D_h\left(\mathbf{U^{k+1}},\mathbf{Z^{k+1}},Y_{\mathbf{U}}^{\mathbf{k}},Y_{\mathbf{Z}}^{\mathbf{k}}\right). \tag{2.12}$$

The extrapolation step is performed in (2.8), which is similar to NAG/FISTA. However, the inertia cannot be arbitrary and the analysis from [45] requires step (2.9) which is governed by the convexity of lower bound, $\underline{L}_k h + g$, however only locally as in (2.10). The update step (2.11) is similar to BPG-MF, however the step-size is controlled via the convexity of upper bound $\bar{L}_k h - g$, but only locally as in (2.12). The local adaptation of the steps (2.10) and (2.12) is performed via backtracking. Since, $\bar{L}_k$ can be potentially very small compared to L, hence potentially large steps can be taken. There is no restriction on $\underline{L}_k$ in each iteration, and smaller $\underline{L}_k$ can result in high value for the inertial parameter γ_k. Thus the algorithm in essence aims to detect "local convexity" of the objective. The update steps of CoCaIn BPG-MF can be executed sequentially without any nested loops for the backtracking. One can always find the inertial parameter γ_k in (2.9) due to [45, Lemma 4.1]. For certain cases, (2.9) yields an explicit condition on γ_k. For example, for $h(\mathbf{U},\mathbf{Z}) = \frac{1}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2)$, we have $0 \le \gamma_k \le \sqrt{\frac{\delta-\varepsilon}{1+\tau_{k-1}\underline{L}_k}}$. We now provide below the convergence result from [45, Theorem 5.2] adapted to CoCaIn BPG-MF.

Theorem 2.2 (Global Convergence of CoCaIn BPG-MF)**.** *Let Assumptions A, B and C hold, let g be L-smad with respect to h with full domain. Assume $\nabla g, \nabla h$ to be Lipschitz continuous on any bounded subset. Let $\left\{(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}})\right\}_{k\in\mathbb{N}}$ be a bounded sequence generated by CoCaIn BPG-MF, and suppose f, g satisfy the KL property, then, such a sequence has finite length, and converges to a critical point.*

2.4 Closed Form Solutions for Update Steps of BPG-MF and CoCaIn BPG-MF

Our second significant contribution is to make BPG-MF and CoCaIn BPG-MF an efficient choice for solving Matrix Factorization, namely closed form expressions for the main update steps (2.5), (2.11). For the derivation, we refer to the appendix, here we just state our results.

For the L2-regularized problem

$$g(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2\,,\quad f(\mathbf{U},\mathbf{Z}) = \frac{\lambda_0}{2}\left(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2\right),\quad h = h_a$$

with $c_1 = 3, c_2 = \left\|\mathbf{A}\right\|_F$ and $0 < \lambda < 1$ the BPG-MF updates are:

$$\mathbf{U^{k+1}} = -r\mathbf{P^k}\,,\ \mathbf{Z^{k+1}} = -r\mathbf{Q^k} \text{ with } r \geq 0\,,\ c_1\left(\left\|-\mathbf{P^k}\right\|_F^2 + \left\|-\mathbf{Q^k}\right\|_F^2\right)r^3 + (c_2+\lambda_0)r - 1 = 0\,.$$

For NMF with additional non-negativity constraints, we replace $-\mathbf{P^k}$ and $-\mathbf{Q^k}$ by $\Pi_+(-\mathbf{P^k})$ and $\Pi_+(-\mathbf{Q^k})$ respectively where $\Pi_+(.) = \max\{0,.\}$ and max is applied element wise.
Now consider the following L1-Regularized problem

$$g(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2\,,\quad f(\mathbf{U},\mathbf{Z}) = \lambda_1\left(\left\|\mathbf{U}\right\|_1 + \left\|\mathbf{Z}\right\|_1\right),\quad h = h_a\,. \tag{2.13}$$

The soft-thresholding operator is defined for any $y \in \mathbb{R}^d$ by $\mathcal{S}_\theta(y) = \max\{|y| - \theta, 0\}\,\mathrm{sgn}(y)$ where $\theta > 0$. Set $c_1 = 3, c_2 = \left\|\mathbf{A}\right\|_F$ and $0 < \lambda < 1$ the BPG-MF updates with the above given g, f, h are:

$\mathbf{U^{k+1}} = r\mathcal{S}_{\lambda_1\lambda}(-\mathbf{P^k})$, $\mathbf{Z^{k+1}} = r\mathcal{S}_{\lambda_1\lambda}(-\mathbf{Q^k})$ with $r \geq 0$ and

$$c_1\left(\left\|\mathcal{S}_{\lambda_1\lambda}(-\mathbf{P^k})\right\|_F^2 + \left\|\mathcal{S}_{\lambda_1\lambda}(-\mathbf{Q^k})\right\|_F^2\right)r^3 + c_2 r - 1 = 0\,.$$

We denote a vector of ones as $\mathbf{e}_D \in \mathbb{R}^D$. For additional non-negativity constraints we need to replace $\mathcal{S}_{\lambda_1\lambda}(-\mathbf{P^k})$ with $\Pi_+(-\left(\mathbf{P^k} + \lambda_1\lambda\mathbf{e}_M\mathbf{e}_K^T\right))$ and $\mathcal{S}_{\lambda_1\lambda}\left(-\mathbf{Q^k}\right)$ to $\Pi_+(-\left(\mathbf{Q^k} + \lambda_1\lambda\mathbf{e}_K\mathbf{e}_N^T\right))$. Excluding the gradient computation, the computational complexity of our updates is $O(MK + NK)$ only, thanks to linear operations. PALM and iPALM additionally involve calculating Lipschitz constants with at most $O(K^2\max\{M,N\}^2)$ computations. Examples like Graph Regularized NMF (GNMF) [12], Sparse NMF [7], Matrix Completion [34], Nuclear Norm Regularization [13, 31], Symmetric NMF [66] and proofs are given in the appendix.

3 Experiments

In this section, we show experiments for (2.1). Denote the regularization settings, **R1:** with $R_1 \equiv R_2 \equiv 0$, **R2:** with L2 regularization $\mathcal{R}_1(\mathbf{U}) = \frac{\lambda_0}{2}\left\|\mathbf{U}\right\|_F^2$ and $\mathcal{R}_2(\mathbf{Z}) = \frac{\lambda_0}{2}\left\|\mathbf{Z}\right\|_F^2$ for some $\lambda_0 > 0$, **R3:** with L1 Regularization $\mathcal{R}_1(\mathbf{U}) = \lambda_0\left\|\mathbf{U}\right\|_1$ and $\mathcal{R}_2(\mathbf{Z}) = \lambda_0\left\|\mathbf{Z}\right\|_1$ for some $\lambda_0 > 0$.

Algorithms. We compare our first order optimization algorithms, BPG-MF and CoCaIn BPG-MF, and recent state of the art optimization methods iPALM [52] and PALM [7]. We focus on algorithms that guarantee convergence to a stationary point. We also use BPG-MF-WB, where WB stands for "with backtracking", which is equivalent to CoCaIn BPG-MF with $\gamma_k \equiv 0$. We use two settings for iPALM, where all the extrapolation parameters are set to a single value β set to 0.2 and 0.4. PALM is equivalent to iPALM if $\beta = 0$. We use the same initialization for the algorithms and we set $\bar{L}_0 = 0.1$ for CoCaIn BPG-MF[1]. We aim to empirically illustrate the differences of alternating minimization based schemes and non-alternating schemes, BPG-MF, BPG-MF-WB and CoCaIn BPG-MF.

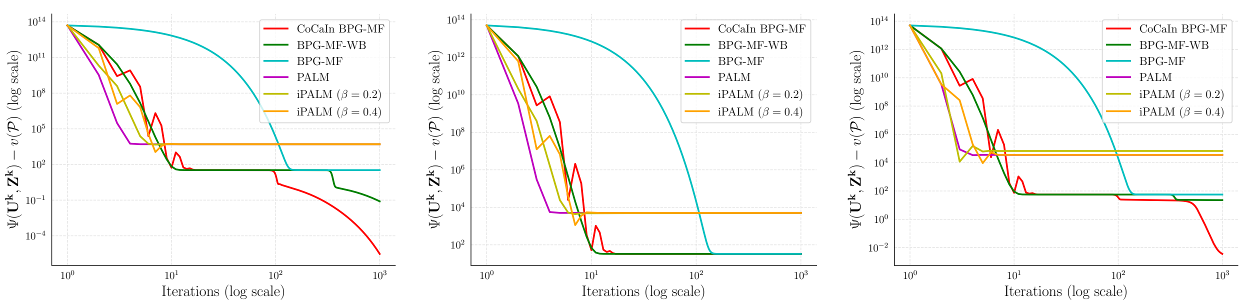
Simple Matrix Factorization. We set $\mathcal{U} = \mathbb{R}^{M\times K}$ and $\mathcal{Z} = \mathbb{R}^{K\times N}$. We use a randomly generated synthetic data matrix with $A \in \mathbb{R}^{200\times 200}$ and report performance in terms of function value for three regularization settings, **R1**, **R2** and **R3** with $K = 5$. For **R2** and **R3** we use $\lambda_0 = 0.1$. The results are given in Figure 1 where the performance of CoCaIn BPG-MF is superior [2].

Non-negative Matrix Factorization. We set $\mathcal{U} = \mathbb{R}_+^{M\times K}$ and $\mathcal{Z} = \mathbb{R}_+^{K\times N}$. We consider Medulloblastoma dataset [11] dataset with matrix $A \in \mathbb{R}^{5893\times 34}$. The other settings are same as Simple Matrix Factorization given above. As evident from Figure 2, BPG-MF, BPG-MF-WB, CoCaIn BPG-MF consistently chooses a better local minimum with lower objective value.

Matrix Completion. In recommender systems [34] given a matrix A with entries at few index pairs in set Ω, the goal is to obtain factors $\mathbf{U}$ and $\mathbf{Z}$ that generalize via following optimization problem

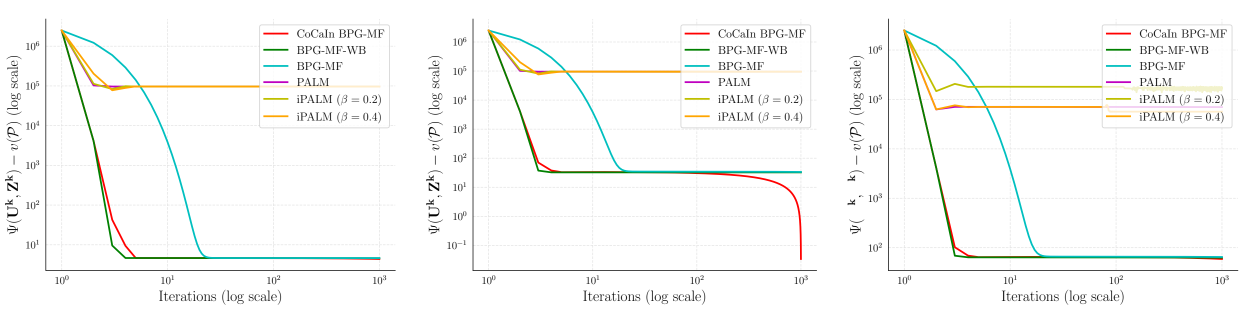
$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|P_\Omega\left(\mathbf{A} - \mathbf{UZ}\right)\right\|_F^2 + \frac{\lambda_0}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2\right)\right\}, \tag{3.1}$$

where P_Ω preserves the given matrix entries and sets others to zero. We use 80% data of MovieLens-100K, MovieLens-1M and MovieLens-10M [29] datasets and use other 20% to test (details in the appendix). CoCaIn BPG-MF is faster than all methods as given in Figure 3.



(a) No Regularization (b) L2-Regularization (c) L1-Regularization

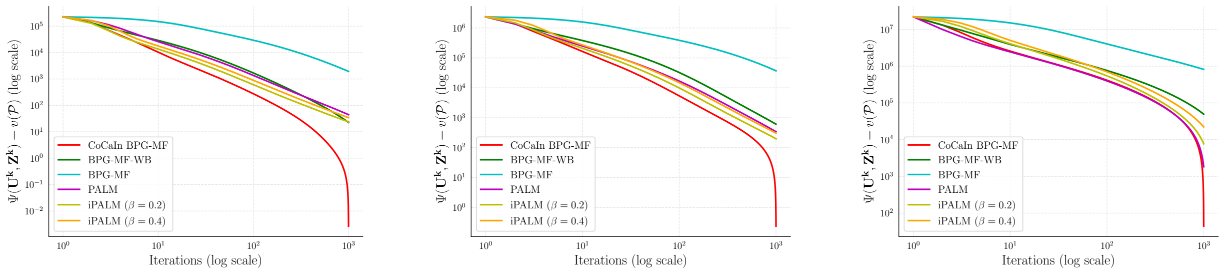
Figure 1: **Simple Matrix Factorization on Synthetic Dataset.**



(a) No-Regularization (b) L2-Regularization (c) L1-Regularization

Figure 2: **Non-negative Matrix Factorization on Medulloblastoma Dataset [11].**

[1] This value can be set even smaller, however it can become high in the first iteration itself via backtracking.
[2] Note that in the y-axis label $v(\mathcal{P})$ is the least objective value attained by any of the methods.



(a) MovieLens-100K (b) MovieLens-1M (c) MovieLens-10M

Figure 3: **Matrix Completion on MovieLens Datasets [29].**

As evident from Figures 1, 2, 3, CoCaIn BPG-MF, BPG-MF-WB and BPG-MF can result in better performance than well known alternating methods. BPG-MF on MovieLens dataset is not better than PALM and iPALM because of prohibitively small step-sizes (due to $\|\mathbf{A}\|_F$ in (2.7)), which is resolved by CoCaIn BPG-MF and BPG-MF-WB using backtracking. Time comparisons are provided in the appendix, where we show that our methods are competitive.

Conclusion and Extensions

We proposed non-alternating algorithms to solve matrix factorization problems, contrary to the typical alternating strategies. We use Bregman proximal minimization framework to propose BPG [8] and an inertial variant CoCaIn BPG [45] for matrix factorization problems. We developed a novel Bregman distance, crucial for convergence to a stationary point. Moreover, we also provide non-trivial efficient closed form update steps for many matrix factorization problems. This line of thinking raises new open questions, such as extensions to Tensor Factorization [33], to Robust Matrix Factorization [63], to stochastic variants [19, 26, 44, 47] and to state-of-the-art matrix factorization model [32].

Acknowledgments

Mahesh Chandra Mukkamala and Peter Ochs were supported by the German Research Foundation (DFG Grant OC 150/1-1).

A Experiments and Implementation Details

A.1 Double Backtracking Implementation

This subsection where we provide certain crucial implementation details of CoCaIn BPG-MF algorithm, is largely based on [45, Section 5.4]. Note that CoCaIn BPG-MF is a sequential algorithm in the sense one can compute $Y_{\mathbf{U}}^{\mathbf{k}}, Y_{\mathbf{Z}}^{\mathbf{k}}$ first via the steps (2.8), (2.9) and (2.10). Then, the updates can be done exactly like BPG-MF, where step-size depends on the parameter $\bar{L}_k$ obtained via (2.12). In (2.10) it is required to find $\underline{L}_k$ such that the following holds

$$D_g\left(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}, Y_{\mathbf{U}}^{\mathbf{k}}, Y_{\mathbf{Z}}^{\mathbf{k}}\right) \geq -\underline{L}_k D_h\left(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}, Y_{\mathbf{U}}^{\mathbf{k}}, Y_{\mathbf{Z}}^{\mathbf{k}}\right), \tag{A.1}$$

similarly in (2.12) it is required to find $\bar{L}_k$ such that

$$D_g\left(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}, Y_{\mathbf{U}}^{\mathbf{k}}, Y_{\mathbf{Z}}^{\mathbf{k}}\right) \leq \bar{L}_k D_h\left(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}, Y_{\mathbf{U}}^{\mathbf{k}}, Y_{\mathbf{Z}}^{\mathbf{k}}\right). \tag{A.2}$$

The above mentioned steps can be solved via the classical backtracking strategy for $\underline{L}_k$ and $\bar{L}_k$ individually, hence the name "double backtracking". We describe the backtracking procedure for $\underline{L}_k$

and it is easy to extend to $\bar{L}_k$. The backtracking strategy involves a scaling parameter $\nu \geq 1$ and an initialization point $\underline{L}_{k,0} > 0$ (preferably small) both chosen by the user and the parameter $\underline{L}_k$ is set to the smallest element from the set $\{\underline{L}_{k,0}, \nu\underline{L}_{k,0}, \nu^2\underline{L}_{k,0}, \ldots\}$ such that (2.10) holds. For $\bar{L}_k$ one requires to use (2.12) and also due to the additional restriction that $\bar{L}_k \geq \bar{L}_{k-1}$ in CoCaIn BPG-MF it is required to start the initialization $\bar{L}_{k,0} = \bar{L}_{k-1}$.

A.2 Matrix Completion

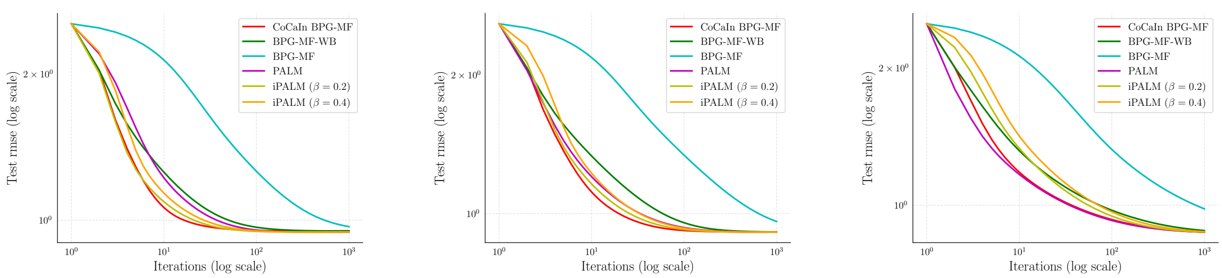
The MovieLens datasets are essentially a matrix $A \in \mathbb{R}^{M\times N}$, where M denotes the number of users and N denotes the number of movies. Only a few non-zero entries are given and the entries denote the ratings which the user has provided for a particular movie. The ratings can take the value between 1 and 5, which we refer to as scale. The exact statistics of all the MovieLens datasets are given below.

Dataset	**Users**	**Movies**	**Non-zero entries**	**Scale**
MovieLens100K	943	1682	100000	1-5
MovieLens1M	6040	3952	1000209	1-5
MovieLens10M	71567	10681	10000054	1-5

The plots provided for the matrix completion problem in Section 3 uses only 80% of the data and we use the remaining 20% as test data in order to obtain the generalization performance to unseen matrix entries with the resulting factors $\mathbf{U} \in \mathbb{R}^{M\times K}$ and $\mathbf{Z} \in \mathbb{R}^{K\times N}$ where we use $K = 5$. The predicted rating to a particular $i \in \{1, 2, \ldots, M\}$ and $j \in \{1, 2, \ldots, N\}$ is given by $(\mathbf{UZ})_{ij}$. The test data is comprised of matrix indices with unseen entries and we denote this set of indices as Ω_T. A popular measure for the test data is the Test RMSE, which is given by the following entity

$$\text{Test RMSE} = \sqrt{\frac{1}{|\Omega_T|}\sum_{i=1}^{M}\sum_{j=1}^{N}\mathbf{I}_{(i,j)\in\Omega_T}\left(\mathbf{A}_{ij} - (\mathbf{UZ})_{ij}\right)^2}$$

where $|\Omega_T|$ denotes the cardinality of the set Ω_T and $\mathbf{I}_{(i,j)\in\Omega_T} = 1$ if the index pair (i, j) lies in the set Ω_T else it is zero. The Test RMSE comparisons for the MovieLens Dataset are given below in Figure 4.



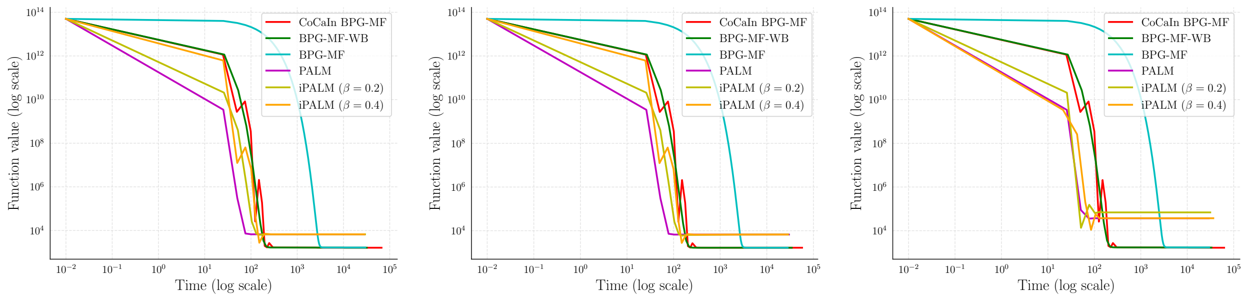
(a) MovieLens-100K (b) MovieLens-1M (c) MovieLens-10M

Figure 4: **Test RMSE plot on MovieLens Datasets [29].**

The above given figures show that the proposed methods BPG-MF-WB and CoCaIn BPG-MF are competitive to PALM and iPALM. BPG-MF is slow in the beginning, however it is competitive to other methods towards the end.

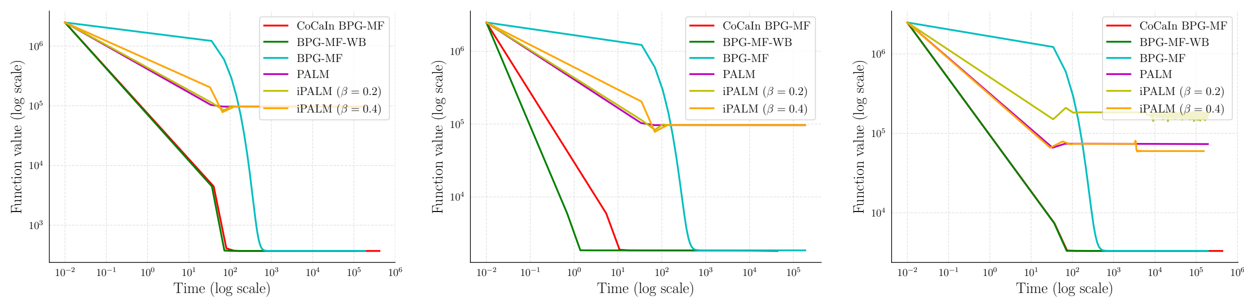
A.3 Time Comparisons

We provide time comparisons in Figures 5,6,7 for all the experimental settings mentioned in Section 3, where we mention the dataset in the caption. Since, we used logarithmic scaling, we used an offset of 10^{-2} for all algorithms for better visualization.



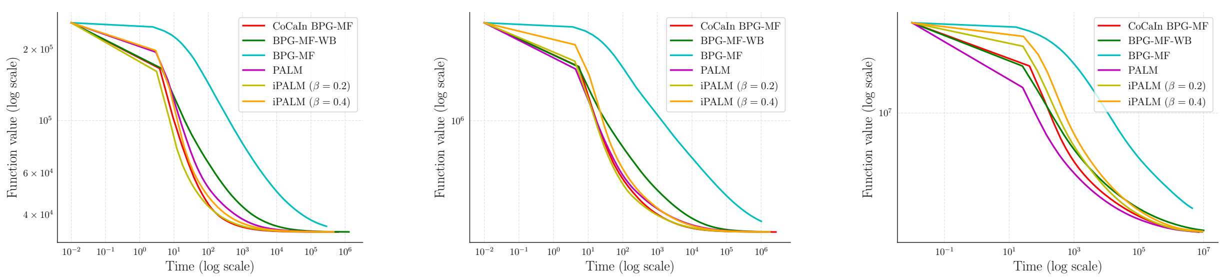
(a) No Regularization (b) L2-Regularization (c) L1-Regularization

Figure 5: **Time plots for Simple Matrix Factorization on Synthetic Dataset.**



(a) No-Regularization (b) L2-Regularization (c) L1-Regularization

Figure 6: **Time plots for Non-negative Matrix Factorization on Medulloblastoma dataset [11].**



(a) MovieLens-100K (b) MovieLens-1M (c) MovieLens-10M

Figure 7: **Time plots for Matrix Completion on MovieLens Datasets [29].**

As evident from the plots, the proposed variants BPG-MF-WB and CoCaIn BPG-MF are faster or competitive that PALM and iPALM. And, BPG-MF is mostly slow, however this behavior is compensated with a lower function value to the end.

B Closed Form Solutions Part I for Matrix Factorization

Since, the update steps of BPG-MF and CoCaIn BPG-MF have same structure, we provide the closed form expressions to just BPG-MF. We start with the following technical lemma.

Lemma B.1. *Let* $\mathbf{Q} \in \mathbb{R}^{A\times B}$ *for some positive integers* A *and* B. *Let* $t \geq 0$ *and* $\|\mathbf{Q}\|_F \neq 0$ *then*

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X} \rangle : \|\mathbf{X}\|_F^2 = t^2 \right\} \equiv \min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X} \rangle : \|\mathbf{X}\|_F^2 \leq t^2 \right\} = -t\,\|\mathbf{Q}\|_F \ ,$$

with the minimizer at $\mathbf{X}^* = -t\mathbf{Q}/\left\|\mathbf{Q}\right\|_F$ *.*

Proof. The proof is inspired from [40, Lemma 9]. On rewriting we have the following equivalence

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \le t^2 \right\} \equiv - \max_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle -\mathbf{Q}, \mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \le t^2 \right\} .$$

The expression $\langle -\mathbf{Q}, \mathbf{X}\rangle$ is maximized at $\mathbf{X}^* = c(-\mathbf{Q})$ for certain constant c. On substituting we have

$$\langle -\mathbf{Q}, \mathbf{X}^*\rangle = c\left\|\mathbf{Q}\right\|_F^2 .$$

Since, the dependence on c is linear and we additionally require $\left\|\mathbf{X}\right\|_F^2 \le t^2$, we can set $c = \frac{t}{\left\|\mathbf{Q}\right\|_F}$ if $\left\|\mathbf{Q}\right\|_F \neq 0$ else $c = 0$. Hence, the minimizer to

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \le t^2 \right\}$$

is attained at $\mathbf{X}^* = -t\frac{\mathbf{Q}}{\left\|\mathbf{Q}\right\|_F}$ for $\left\|\mathbf{Q}\right\|_F \neq 0$ else $\mathbf{X}^* = 0$. The equivalence in the statement follows as $\left\|\mathbf{X}^*\right\|_F^2 = t^2$. □

Consider the following non-convex matrix factorization problem

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A} - \mathbf{U}\mathbf{Z}\right\|_F^2 \right\} . \tag{B.1}$$

Denote $g = \Psi$, $f := 0$, $h = h_a$.

Proposition B.1. *In BPG-MF, with above defined* g, f, h *the update steps in each iteration are given by* $\mathbf{U^{k+1}} = -r\,\mathbf{P^k}$, $\mathbf{Z^{k+1}} = -r\,\mathbf{Q^k}$ *where* r *is the non-negative real root of*

$$c_1\left(\left\|\mathbf{Q^k}\right\|_F^2 + \left\|\mathbf{P^k}\right\|_F^2\right) r^3 + c_2 r - 1 = 0 , \tag{B.2}$$

with $c_1 = 3$ *and* $c_2 = \left\|\mathbf{A}\right\|_F$*.*

Proof. Consider the following subproblem

$$(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}) \in \underset{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}}{\operatorname{argmin}} \left\{ \left\langle \mathbf{P^k}, \mathbf{U}\right\rangle + \left\langle \mathbf{Q^k}, \mathbf{Z}\right\rangle \right. \\ \left. + c_1\left(\frac{\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2}{2}\right)^2 + c_2\left(\frac{\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2}{2}\right) \right\} .$$

Denote the objective in the above minimization problem as $\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})$. Now, the following holds

$$\min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}} \left(\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})\right)$$

$$\equiv \min_{t_1\ge 0, t_2\ge 0} \left\{ \min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}, \left\|\mathbf{U}\right\|_F = t_1, \left\|\mathbf{Z}\right\|_F = t_2} \left(\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})\right) \right\} , \tag{B.3}$$

$$\equiv \min_{t_1\ge 0, t_2\ge 0} \left\{ \min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}, \left\|\mathbf{U}\right\|_F \le t_1, \left\|\mathbf{Z}\right\|_F \le t_2} \left(\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})\right) \right\} , \tag{B.4}$$

where the first step is a simple rewriting of the objective. The second step is non-trivial. In order to prove (B.4) we rewrite (B.3) as

$$\min_{t_1\ge 0, t_2\ge 0} \left\{ \min_{\mathbf{U_1}\in\mathbb{R}^{M\times K}} \left\{ \left\langle \mathbf{P^k}, \mathbf{U_1}\right\rangle : \left\|\mathbf{U_1}\right\|_F^2 = t_1 \right\} \right. \\ + \min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}} \left\{ \left\langle \mathbf{Q^k}, \mathbf{Z_1}\right\rangle : \left\|\mathbf{Z_1}\right\|_F^2 = t_2 \right\} \\ \left. + c_1\left(\frac{t_1^2 + t_2^2}{2}\right)^2 + c_2\left(\frac{t_1^2 + t_2^2}{2}\right) \right\} .$$

Now, note the following equivalence due to Lemma B.1

$$\min_{\mathbf{U_1} \in \mathbb{R}^{M\times K}} \left\{ \left\langle \mathbf{P^k}, \mathbf{U_1} \right\rangle : \|\mathbf{U_1}\|_F^2 = t_1 \right\} \equiv \min_{\mathbf{U_1} \in \mathbb{R}^{M\times K}} \left\{ \left\langle \mathbf{P^k}, \mathbf{U_1} \right\rangle : \|\mathbf{U_1}\|_F^2 \le t_1 \right\},$$
$$\min_{\mathbf{Z_1} \in \mathbb{R}^{K\times N}} \left\{ \left\langle \mathbf{Q^k}, \mathbf{Z_1} \right\rangle : \|\mathbf{Z_1}\|_F^2 = t_2 \right\} \equiv \min_{\mathbf{Z_1} \in \mathbb{R}^{K\times N}} \left\{ \left\langle \mathbf{Q^k}, \mathbf{Z_1} \right\rangle : \|\mathbf{Z_1}\|_F^2 \le t_2 \right\}.$$

This proves (B.4). Now, we solve for $(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}})$ via the following strategy. Denote

$$\mathbf{U}_1^*(t_1) \in \operatorname{argmin} \left\{ \left\langle \mathbf{P^k}, \mathbf{U_1} \right\rangle : \mathbf{U_1} \in \mathbb{R}^{M\times K}, \|\mathbf{U_1}\|_F^2 \le t_1 \right\},$$
$$\mathbf{Z}_1^*(t_2) \in \operatorname{argmin} \left\{ \left\langle \mathbf{Q^k}, \mathbf{Z_1} \right\rangle : \mathbf{Z_1} \in \mathbb{R}^{K\times N}, \|\mathbf{Z_1}\|_F^2 \le t_2 \right\}.$$

Then we obtain $(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}) = (\mathbf{U}_1^*(t_1^*), \mathbf{Z}_1^*(t_2^*))$, where t_1^* and t_2^* are obtained by solving the following two dimensional subproblem

$$\begin{aligned}(t_1^*, t_2^*) \in \operatorname*{argmin}_{t_1\ge 0, t_2 \ge 0} \Bigg\{ & \min_{\mathbf{U_1} \in \mathbb{R}^{M\times K}} \left\{ \left\langle \mathbf{P^k}, \mathbf{U_1} \right\rangle : \|\mathbf{U_1}\|_F^2 \le t_1 \right\} \\ & + \min_{\mathbf{Z_1} \in \mathbb{R}^{K\times N}} \left\{ \left\langle \mathbf{Q^k}, \mathbf{Z_1} \right\rangle : \|\mathbf{Z_1}\|_F^2 \le t_2 \right\} \\ & + c_1 \left(\frac{t_1^2 + t_2^2}{2} \right)^2 + c_2 \left(\frac{t_1^2 + t_2^2}{2} \right) \Bigg\}.\end{aligned}$$

Note that inner minimization subproblems can be trivially solved once we obtain $\mathbf{U}_1^*(t_1)$ and $\mathbf{Z}_1^*(t_2)$ via Lemma B.1. Then the solution to the subproblem in each iteration is as follows:

$$\mathbf{U^{k+1}} = \begin{cases} t_1^* \frac{-\mathbf{P^k}}{\left\|\mathbf{P^k}\right\|_F}, & \text{for } \left\|\mathbf{P^k}\right\|_F \neq 0, \\ \mathbf{0} & \textit{otherwise}. \end{cases}$$
$$\mathbf{Z^{k+1}} = \begin{cases} t_2^* \frac{-\mathbf{Q^k}}{\left\|\mathbf{Q^k}\right\|_F}, & \text{for } \left\|\mathbf{Q^k}\right\|_F \neq 0, \\ \mathbf{0} & \textit{otherwise}. \end{cases}$$

We solve for t_1^* and t_2^* with the following two dimensional minimization problem

$$\operatorname*{argmin}_{t_1\ge 0, t_2\ge 0} \left\{ -t_1 \left\|\mathbf{P^k}\right\|_F - t_2 \left\|\mathbf{Q^k}\right\|_F + c_1 \left(\frac{t_1^2 + t_2^2}{2} \right)^2 + c_2 \left(\frac{t_1^2 + t_2^2}{2} \right) \right\}.$$

Thus, the solutions t_1^* and t_2^* are the non-negative real roots of the following equations

$$-\left\|\mathbf{P^k}\right\|_F + c_1(t_1^2 + t_2^2)t_1 + c_2 t_1 = 0$$
$$-\left\|\mathbf{Q^k}\right\|_F + c_1(t_1^2 + t_2^2)t_2 + c_2 t_2 = 0$$

Further simplifications lead to $t_1 = r\left\|\mathbf{P^k}\right\|_F$ and $t_2 = r\left\|\mathbf{Q^k}\right\|_F$ for some $r \ge 0$ such that r satisfies the following cubic equation

$$c_1 \left(\left\|\mathbf{Q^k}\right\|_F^2 + \left\|\mathbf{P^k}\right\|_F^2 \right) r^3 + c_2 r - 1 = 0.$$

□

B.1 Extensions to L2-Regularized Matrix Factorization

We consider the following L2-Regularized Matrix Factorization problem [37].

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \frac{\lambda_0}{2}\left(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2\right)\right\} . \tag{B.5}$$

Denote $g := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2$, $f := \frac{\lambda_0}{2}\left(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2\right)$ and $h = h_a$.

Proposition B.2. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = -r\,\mathbf{P^k}$, $\mathbf{Z^{k+1}} = -r\,\mathbf{Q^k}$ where r is the non-negative real root of*

$$c_1\left(\left\|\mathbf{Q^k}\right\|_F^2 + \left\|\mathbf{P^k}\right\|_F^2\right) r^3 + (c_2+\lambda_0)r - 1 = 0\,, \tag{B.6}$$

with $c_1 = 3$ and $c_2 = \left\|\mathbf{A}\right\|_F$.

We skip the proof as it is very similar to Proposition B.1 and only change is in c_2.

B.2 Extensions to Graph Regularized Matrix Factorization

Graph Regularized Matrix Factorization was proposed in [12]. However, they used non-negativity constraints. We simplify the problem here by not considering the non-negativity constraints. We later show in Section C.3, how the non-negativity constraints are handled. Here, given $\mathcal{L}\in\mathbb{R}^{M\times M}$ we are interested to solve

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \frac{\mu_0}{2}\mathbf{tr}(\mathbf{U}^T\mathcal{L}\mathbf{U}) + \frac{\lambda_0}{2}\left(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2\right)\right\} .$$

In such a case, it is easy to extend the following ideas to Graph Regularized Non-negative Matrix Factorization. We show here L-smad property. We first need the following technical lemma.

Lemma B.2. *Let $g_1(\mathbf{U}) = \mathbf{tr}(\mathbf{U}^T\mathcal{L}\mathbf{U})$, then for any $\mathbf{H}\in\mathbb{R}^{M\times K}$ we have $\nabla g_1(\mathbf{U}) = \mathcal{L}\mathbf{U} + \mathcal{L}^T\mathbf{U}$,*

$$\left\langle \mathbf{H}, \nabla^2 g_1(\mathbf{U})\mathbf{H}\right\rangle = 2\left\langle \mathcal{L}\mathbf{H},\mathbf{H}\right\rangle .$$

Proof. Note that $\mathbf{tr}(\mathbf{U}^T\mathcal{L}\mathbf{U}) = \langle \mathcal{L}\mathbf{U},\mathbf{U}\rangle$, now we obtain for $\mathbf{H}\in\mathbb{R}^{M\times K}$ the following

$$\begin{aligned}\langle \mathcal{L}(\mathbf{U}+\mathbf{H}),\mathbf{U}+\mathbf{H}\rangle &= \langle \mathcal{L}(\mathbf{U}+\mathbf{H}),\mathbf{U}+\mathbf{H}\rangle\\ &= \langle \mathcal{L}\mathbf{U},\mathbf{U}\rangle + \langle \mathcal{L}\mathbf{U},\mathbf{H}\rangle + \langle \mathcal{L}\mathbf{H},\mathbf{U}\rangle + \langle \mathcal{L}\mathbf{H},\mathbf{H}\rangle\,,\\ &= \langle \mathcal{L}\mathbf{U},\mathbf{U}\rangle + \langle \mathcal{L}\mathbf{U},\mathbf{H}\rangle + \left\langle \mathcal{L}^T\mathbf{U},\mathbf{H}\right\rangle + \langle \mathcal{L}\mathbf{H},\mathbf{H}\rangle\,.\end{aligned}$$

Thus the statement holds, by collecting the first and second order terms. □

Now, we prove the L-smad property.

Proposition B.3. *Let $g(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \frac{\mu_0}{2}\mathbf{tr}(\mathbf{U}^T\mathcal{L}\mathbf{U})$. Then, for a certain constant $L\geq 1$, the function g satisfies L-smad property with respect to the following kernel generating distance,*

$$h_c(\mathbf{U},\mathbf{Z}) = 3h_1(\mathbf{U},\mathbf{Z}) + \left(\left\|\mathbf{A}\right\|_F + \mu_0\left\|\mathcal{L}\right\|_F\right)h_2(\mathbf{U},\mathbf{Z})\,.$$

Proof. The proof is similar to Proposition 2.1 and Lemma B.2 must be applied for the result. □

Denote $g := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \frac{\mu_0}{2}\mathbf{tr}(\mathbf{U}^T\mathcal{L}\mathbf{U})$, $f := \frac{\lambda_0}{2}\left(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2\right)$ and $h = h_c$.

Proposition B.4. *In BPG-MF, with the above defined f, g, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = -r\,\mathbf{P^k}$, $\mathbf{Z^{k+1}} = -r\,\mathbf{Q^k}$ where $r\geq 0$ and satisfies*

$$c_1\left(\left\|\mathbf{Q^k}\right\|_F^2 + \left\|\mathbf{P^k}\right\|_F^2\right) r^3 + (c_2 + \mu_0\left\|\mathcal{L}\right\|_F + \lambda_0)r - 1 = 0\,, \tag{B.7}$$

with $c_1 = 3$ and $c_2 = \left\|\mathbf{A}\right\|_F$.

The proof is similar to Proposition B.1 and only c_2 changes.

B.3 Extensions to L1-Regularized Matrix Factorization

Now consider the following matrix factorization problem with L1-Regularization

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2 + \lambda_1\left(\left\|\mathbf{U}\right\|_1 + \left\|\mathbf{Z}\right\|_1\right)\right\} . \tag{B.8}$$

Recall that soft-thresholding operator is defined for any $y \in \mathbb{R}^d$ by

$$\mathcal{S}_\theta\left(y\right) = \operatorname{argmin}_{x\in\mathbb{R}^d}\left\{\theta\left\|x\right\|_1 + \frac{1}{2}\left\|x-y\right\|^2\right\} = \max\left\{|y|-\theta,0\right\}\operatorname{sgn}\left(y\right) , \tag{B.9}$$

where $\theta > 0$ and the operations are applied element-wise. We require the following technical result.

Lemma B.3. *Let $\mathbf{Q} \in \mathbb{R}^{A\times B}$ for some positive integers A and B. Let $t_0 > 0$ and let $t \geq 0$ then*

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle + t_0\left\|\mathbf{X}\right\|_1 : \left\|\mathbf{X}\right\|_F^2 \leq t^2\right\} = -t\left\|S_{t_0}(-\mathbf{Q})\right\| .$$

with the minimizer at $\mathbf{X}^* = t\frac{S_{t_0}(-\mathbf{Q})}{\|S_{t_0}(-\mathbf{Q})\|}$ for $\|S_{t_0}(-\mathbf{Q})\| \neq 0$ and otherwise all $\mathbf{X}$ such that $\|\mathbf{X}\|_F^2 \leq t^2$ are minimizers. Moreover we have the following equivalence,

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle + t_0\left\|\mathbf{X}\right\|_1 : \left\|\mathbf{X}\right\|_F^2 \leq t^2\right\} \equiv \min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle + t_0\left\|\mathbf{X}\right\|_1 : \left\|\mathbf{X}\right\|_F^2 = t^2\right\} . \tag{B.10}$$

Proof. We have the following equivalence

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle + t_0\left\|\mathbf{X}\right\|_1 : \left\|\mathbf{X}\right\|_F^2 \leq t^2\right\} \equiv -\max_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle-\mathbf{Q},\mathbf{X}\rangle - t_0\left\|\mathbf{X}\right\|_1 : \left\|\mathbf{X}\right\|_F^2 \leq t^2\right\} .$$

Then the result follows due to [40, Proposition 14] with the minimizer at $\mathbf{X}^* = t\frac{S_{t_0}(-\mathbf{Q})}{\|S_{t_0}(-\mathbf{Q})\|}$ for $\|S_{t_0}(-\mathbf{Q})\| \neq 0$ and $\mathbf{0}$ otherwise. The equivalence statement in (B.10) follows as $\|\mathbf{X}^*\|_F^2 = t^2$ for $\|S_{t_0}(-\mathbf{Q})\| \neq 0$ and otherwise all the points satisfying $\|\mathbf{X}\|_F^2 = t^2$ are minimizers. ☐

Denote $g := \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2$, $f := \lambda_1\left(\left\|\mathbf{U}\right\|_1 + \left\|\mathbf{Z}\right\|_1\right)$ and $h = h_a$.

Proposition B.5. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = r\mathcal{S}_{\lambda_1\lambda}(-\mathbf{P^k})$, $\mathbf{Z^{k+1}} = r\mathcal{S}_{\lambda_1\lambda}(-\mathbf{Q^k})$ where $r \geq 0$ and satisfies*

$$c_1\left(\left\|\mathcal{S}_{\lambda_1\lambda}\left(-\mathbf{Q^k}\right)\right\|_F^2 + \left\|\mathcal{S}_{\lambda_1\lambda}\left(-\mathbf{P^k}\right)\right\|_F^2\right)r^3 + c_2 r - 1 = 0\,, \tag{B.11}$$

with $c_1 = 3$ and $c_2 = \|\mathbf{A}\|_F$.

Proof. The proof is similar to that of Proposition B.1, but with certain changes due to the L1 norm in the objective. Consider the following subproblem

$$\begin{aligned}(\mathbf{U^{k+1}},\mathbf{Z^{k+1}}) \in \operatorname*{argmin}_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}} &\left\{\lambda\lambda_1\left(\left\|\mathbf{U}\right\|_1 + \left\|\mathbf{Z}\right\|_1\right) + \left\langle\mathbf{P^k},\mathbf{U}\right\rangle + \left\langle\mathbf{Q^k},\mathbf{Z}\right\rangle\right.\\ &\left. + c_1\left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2}\right)^2 + c_2\left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2}\right)\right\} ,\end{aligned}$$

Denote the objective in the above minimization problem as $\mathcal{O}(\mathbf{U^k},\mathbf{Z^k})$. Now, we show that the following holds

$$\begin{aligned}&\min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}}\left(\mathcal{O}(\mathbf{U^k},\mathbf{Z^k})\right)\\ &\equiv \min_{t_1\geq 0,t_2\geq 0}\left\{\min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N},\|\mathbf{U}\|_F=t_1,\|\mathbf{Z}\|_F=t_2}\left(\mathcal{O}(\mathbf{U^k},\mathbf{Z^k})\right)\right\} , &\text{(B.12)}\\ &\equiv \min_{t_1\geq 0,t_2\geq 0}\left\{\min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N},\|\mathbf{U}\|_F\leq t_1,\|\mathbf{Z}\|_F\leq t_2}\left(\mathcal{O}(\mathbf{U^k},\mathbf{Z^k})\right)\right\} . &\text{(B.13)}\end{aligned}$$

where the first step is a simple rewriting of the objective. The second step is non-trivial. In order to prove (B.13) we rewrite (B.12) as

$$\min_{t_1\geq 0, t_2\geq 0}\left\{\min_{\mathbf{U_1}\in\mathbb{R}^{M\times K}}\left\{\left\langle\mathbf{P^k},\mathbf{U_1}\right\rangle+\lambda\lambda_1\|\mathbf{U}\|_1:\|\mathbf{U_1}\|_F^2=t_1\right\}\right.$$
$$+\min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}}\left\{\left\langle\mathbf{Q^k},\mathbf{Z_1}\right\rangle+\lambda\lambda_1\|\mathbf{Z}\|_1:\|\mathbf{Z_1}\|_F^2=t_2\right\}$$
$$\left.+c_1\left(\frac{t_1^2+t_2^2}{2}\right)^2+c_2\left(\frac{t_1^2+t_2^2}{2}\right)\right\}.$$

where the second step (B.13) uses Lemma B.3 and strong convexity of h. Now, note the following equivalence due to Lemma B.3

$$\begin{aligned}&\min_{\mathbf{U_1}\in\mathbb{R}^{M\times K}}\left\{\left\langle\mathbf{P^k},\mathbf{U_1}\right\rangle+\lambda\lambda_1\|\mathbf{U}\|_1:\|\mathbf{U_1}\|_F^2=t_1\right\}\\ \equiv&\min_{\mathbf{U_1}\in\mathbb{R}^{M\times K}}\left\{\left\langle\mathbf{P^k},\mathbf{U_1}\right\rangle+\lambda\lambda_1\|\mathbf{U}\|_1:\|\mathbf{U_1}\|_F^2\leq t_1\right\},\end{aligned}\tag{B.14}$$

and

$$\begin{aligned}&\min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}}\left\{\left\langle\mathbf{Q^k},\mathbf{Z_1}\right\rangle+\lambda\lambda_1\|\mathbf{Z}\|_1:\|\mathbf{Z_1}\|_F^2=t_2\right\}\\ \equiv&\min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}}\left\{\left\langle\mathbf{Q^k},\mathbf{Z_1}\right\rangle+\lambda\lambda_1\|\mathbf{Z}\|_1:\|\mathbf{Z_1}\|_F^2\leq t_2\right\}.\end{aligned}\tag{B.15}$$

We solve the subproblems via the following strategy. Denote

$$\mathbf{U}_1^*(t_1)\in\operatorname{argmin}\left\{\left\langle\mathbf{P^k},\mathbf{U_1}\right\rangle+\lambda\lambda_1\|\mathbf{U}\|_1:\mathbf{U_1}\in\mathbb{R}^{M\times K},\|\mathbf{U_1}\|_F^2\leq t_1\right\}$$

$$\mathbf{Z}_1^*(t_2)\in\operatorname{argmin}\left\{\left\langle\mathbf{Q^k},\mathbf{Z_1}\right\rangle+\lambda\lambda_1\|\mathbf{Z}\|_1:\mathbf{Z_1}\in\mathbb{R}^{K\times N},\|\mathbf{Z_1}\|_F^2\leq t_2\right\}$$

Then we obtain $(\mathbf{U^{k+1}},\mathbf{Z^{k+1}})=(\mathbf{U}_1^*(t_1^*),\mathbf{Z}_1^*(t_2^*))$, where t_1^* and t_2^* are obtained by solving the following two dimensional subproblem

$$(t_1^*,t_2^*)\in\operatorname*{argmin}_{t_1\geq 0, t_2\geq 0}\left\{\min_{\mathbf{U_1}\in\mathbb{R}^{M\times K}}\left\{\left\langle\mathbf{P^k},\mathbf{U_1}\right\rangle+\lambda\lambda_1\|\mathbf{U}\|_1:\|\mathbf{U_1}\|_F^2\leq t_1\right\}\right.$$
$$+\min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}}\left\{\left\langle\mathbf{Q^k},\mathbf{Z_1}\right\rangle+\lambda\lambda_1\|\mathbf{Z}\|_1:\|\mathbf{Z_1}\|_F^2\leq t_2\right\}$$
$$\left.+c_1\left(\frac{t_1+t_2}{2}\right)^2+c_2\left(\frac{t_1+t_2}{2}\right)\right\}.$$

Note that inner minimization subproblems can be trivially solved once we obtain $\mathbf{U}_1^*(t_1)$ and $\mathbf{Z}_1^*(t_2)$. Due to Lemma B.3 we obtain the solution to the subproblem in each iteration as follows

$$\mathbf{U^{k+1}}=\begin{cases}t_1^*\frac{S_{\lambda\lambda_1}(-\mathbf{P^k})}{\left\|S_{\lambda\lambda_1}(-\mathbf{P^k})\right\|_F}, & \text{for }\left\|S_{\lambda\lambda_1}(-\mathbf{P^k})\right\|_F\neq 0,\\ \mathbf{0} & \textit{otherwise}.\end{cases}$$
$$\mathbf{Z^{k+1}}=\begin{cases}t_2^*\frac{S_{\lambda\lambda_1}(-\mathbf{Q^k})}{\left\|S_{\lambda\lambda_1}(-\mathbf{Q^k})\right\|_F}, & \text{for }\left\|S_{\lambda\lambda_1}(-\mathbf{Q^k})\right\|_F\neq 0,\\ \mathbf{0} & \textit{otherwise}.\end{cases}$$

We solve for t_1^* and t_2^* with the following two dimensional minimization problem

$$\operatorname*{argmin}_{t_1\geq 0,t_2\geq 0}\left\{-t_1\left\|S_{\lambda\lambda_1}(-\mathbf{P^k})\right\|_F-t_2\left\|S_{\lambda\lambda_1}(-\mathbf{Q^k})\right\|_F+c_1\left(\frac{t_1^2+t_2^2}{2}\right)^2+c_2\left(\frac{t_1^2+t_2^2}{2}\right)\right\}.$$

Thus, the solutions t_1^* and t_2^* are the non-negative real roots of the following equations

$$
\begin{aligned}
-\left\|S_{\lambda\lambda_1}(-\mathbf{P^k})\right\|_F + c_1(t_1^2+t_2^2)t_1 + c_2 t_1 = 0 \\
-\left\|S_{\lambda\lambda_1}(-\mathbf{Q^k})\right\|_F + c_1(t_1^2+t_2^2)t_2 + c_2 t_2 = 0\,.
\end{aligned}
$$

Set $t_1 = r\left\|S_{\lambda\lambda_1}(-\mathbf{P^k})\right\|_F$ and $t_2 = r\left\|S_{\lambda\lambda_1}(-\mathbf{Q^k})\right\|_F$ for some $r \geq 0$. This results in the following cubic equation,

$$
c_1\left(\left\|S_{\lambda\lambda_1}(-\mathbf{Q^k})\right\|_F^2 + \left\|S_{\lambda\lambda_1}(-\mathbf{P^k})\right\|_F^2\right) r^3 + c_2 r - 1 = 0\,,
$$

where the solution is the non-negative real root. □

B.4 Extensions with nuclear norm regularization

We start with the notion of Singular Value Shrinkage Operator [13], where given a matrix $\mathbf{Q} \in \mathbb{R}^{A\times B}$ of rank K with Singular Value Decomposition given by $\mathbf{U\Sigma V}^T$ with $\mathbf{U} \in \mathbb{R}^{A\times K}$, $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$ and $\mathbf{V} \in \mathbb{R}^{K\times N}$ for $t \geq 0$ the output is

$$
\mathcal{D}_t(\mathbf{Q}) = \mathbf{U}\mathcal{S}_t(\mathbf{\Sigma})\mathbf{V}^T\,, \tag{B.16}
$$

where the soft-thresholding operator is applied only to the singular values. Before we proceed, we require the following technical lemma.

Lemma B.4. *Let $\mathbf{Q} \in \mathbb{R}^{A\times B}$ of rank K with Singular Value Decomposition given by $\mathbf{U\Sigma V}^T$ with $\mathbf{U} \in \mathbb{R}^{A\times K}$, $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$ and $\mathbf{Z} \in \mathbb{R}^{K\times N}$. Let $t \geq 0$ and $\|\mathbf{Q}\|_F \neq 0$ then*

$$
\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{\langle \mathbf{Q},\mathbf{X}\rangle + t_0\|\mathbf{X}\|_* : \|\mathbf{X}\|_F^2 \leq t^2\right\} = -t\left\|\mathcal{S}_{t_0}(-\mathbf{\Sigma})\right\|\,.
$$

with $\mathbf{X}^* = t\frac{\mathcal{D}_{t_0}(-\mathbf{Q})}{\|\mathcal{D}_t(-\mathbf{Q})\|_F}$ if $\|\mathcal{D}_{t_0}(-\mathbf{Q})\| \neq 0$ else any $\mathbf{X}$ such that $\|\mathbf{X}\|_F^2 \leq t^2$ is a minimizer. Moreover we have the following equivalence

$$
\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{\langle \mathbf{Q},\mathbf{X}\rangle + t_0\|\mathbf{X}\|_* : \|\mathbf{X}\|_F^2 \leq t^2\right\} = \min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{\langle \mathbf{Q},\mathbf{X}\rangle + t_0\|\mathbf{X}\|_* : \|\mathbf{X}\|_F^2 = t^2\right\}. \tag{B.17}
$$

Proof. The sub-differential of the nuclear norm [13] is given by

$$
\partial\|\mathbf{X}\|_* = \left\{\mathbf{UV^T} + \mathbf{W} : \mathbf{W} \in \mathbb{R}^{A\times B}, \mathbf{U^T W} = 0, \mathbf{WV} = 0, \|\mathbf{W}\|_2 \leq 1\right\}. \tag{B.18}
$$

The normal cone for the set $C_1 = \left\{\mathbf{X} : \|\mathbf{X}\|_F^2 \leq t^2\right\}$ is given by

$$
\mathcal{N}_{C_1}(\bar{\mathbf{X}}) = \left\{\mathbf{V} \in \mathbb{R}^{A\times B} : \left\langle \mathbf{V}, \mathbf{X} - \bar{\mathbf{X}}\right\rangle \leq 0 \text{ for all } \mathbf{X} \in C_1\right\} \equiv \left\{\theta\bar{\mathbf{X}} : \theta \geq 0\right\}.
$$

We consider the following problem

$$
\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{\langle \mathbf{Q},\mathbf{X}\rangle + t_0\|\mathbf{X}\|_* : \|\mathbf{X}\|_F^2 \leq t^2\right\}.
$$

and the optimality condition [53, Theorem 10.1, p. 422] results in

$$
\mathbf{0} \in \mathbf{Q} + t_0\partial\|\mathbf{X}\|_* + \mathcal{N}_{C_1}(\mathbf{X})\,.
$$

We follow the strategy from [13, Theorem 2.1]. One can decompose $-\mathbf{Q}$ as

$$
-\mathbf{Q} = \mathbf{U_0\Sigma_0V_0^T} + \mathbf{U_1\Sigma_1V_1^T}\,.
$$

where $\mathbf{U_0}, \mathbf{V_0}$ contain the singular vectors for singular values greater than t_0 and $\mathbf{U_1}, \mathbf{V_1}$ for less than equal to t_0. Then with $\mathbf{X} = \mathbf{U_0\Sigma V_0^T}$, the optimality condition becomes

$$
\mathbf{0} = \mathbf{Q} + t_0(\mathbf{U_0V_0^T} + \mathbf{W}) + \theta\mathbf{U_0\Sigma V_0^T}\,, \tag{B.19}
$$

and thus we obtain

$$\mathbf{U_0\Sigma_0V_0^T} + \mathbf{U_1\Sigma_1V_1^T} = t_0\left(\mathbf{U_0V_0^T} + \mathbf{W}\right) + \theta\mathbf{U_0\Sigma V_0^T}\,.$$

With $\mathbf{W} = t_0^{-1}\mathbf{U_1\Sigma_1V_1^T}$ all the conditions in (B.18) are satisfied. For some unknown $\theta \geq 0$ we have

$$\theta\mathbf{\Sigma} = \mathbf{\Sigma_0} - t_0\mathbf{I}\,.$$

The objective $\langle\mathbf{Q},\mathbf{X}\rangle + t_0\left\|\mathbf{X}\right\|_*$ is now monotonically decreasing with θ after substituting. Thus, we obtain the solution $\mathbf{X} = \frac{t}{\|\mathbf{\Sigma_0} - t_0\mathbf{I}\|}\mathbf{U_0}\left(\mathbf{\Sigma_0} - t_0\mathbf{I}\right)\mathbf{V}_0^T$ for $\|\mathbf{\Sigma_0} - t_0\mathbf{I}\| \neq 0$ else the solution is $\mathbf{0}$. The equivalence statement in (B.17) follows trivially because if $\|\mathbf{\Sigma_0} - t_0\mathbf{I}\| \neq 0$ we have $\left\|\mathbf{X}\right\|_F^2 = t^2$ otherwise all the points satisfying $\left\|\mathbf{X}\right\|_F^2 \leq t^2$ are minimizers. □

Here, we want to solve matrix factorization problem with nuclear norm regularization, where for certain constant $\lambda_2 > 0$ we want to solve

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}}\left\{\Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A} - \mathbf{UZ}\right\|_F^2 + \lambda_2\left(\left\|\mathbf{U}\right\|_* + \left\|\mathbf{Z}\right\|_*\right)\right\}. \tag{B.20}$$

Denote $g := \frac{1}{2}\left\|\mathbf{A} - \mathbf{UZ}\right\|_F^2$, $f := \lambda_2\left(\left\|\mathbf{U}\right\|_* + \left\|\mathbf{Z}\right\|_*\right)$ and $h = h_a$.

Proposition B.6. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = r\mathcal{D}_{\lambda_1\lambda}(-\mathbf{P^k})$, $\mathbf{Z^{k+1}} = r\mathcal{D}_{\lambda_1\lambda}(-\mathbf{Q^k})$ where $r \geq 0$ and satisfies*

$$c_1\left(\left\|\mathcal{D}_{\lambda_1\lambda}\left(-\mathbf{Q^k}\right)\right\|_F^2 + \left\|\mathcal{D}_{\lambda_1\lambda}\left(-\mathbf{P^k}\right)\right\|_F^2\right)r^3 + c_2 r - 1 = 0\,, \tag{B.21}$$

with $c_1 = 3$ and $c_2 = \left\|\mathbf{A}\right\|_F$.

The proof is similar to Proposition B.5 but Lemma B.4 must be used instead of Lemma B.3.

B.5 Extensions with non-convex sparsity constraints

We want to solve the matrix factorization problem with non-convex sparsity constraints [7]

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}}\left\{\Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A} - \mathbf{UZ}\right\|_F^2 : \left\|\mathbf{U}\right\|_0 \leq s_1, \left\|\mathbf{Z}\right\|_0 \leq s_2,\right\}. \tag{B.22}$$

The problem with additional non-negativity constraints, the so called Sparse NMF is considered in Section C.5. Now, denote $g := \frac{1}{2}\left\|\mathbf{A} - \mathbf{UZ}\right\|_F^2$, $f := \mathbf{I}_{\|\mathbf{U}\|_0 \leq s_1} + \mathbf{I}_{\|\mathbf{Z}\|_0 \leq s_2}$ and $h = h_a$. Note that the Assumption C is not valid here, hence CoCaIn BPG-MF theory does not hold and hints at possible extensions of CoCaIn BPG-MF, which is an interesting open question. Before, we proceed, we require the following concept. Let $y \in \mathbb{R}^d$ and without loss of generality we can assume that $|y_1| \geq |y_2| \geq \ldots \geq |y_d|$, then the hard-thresholding operator [40] is given by

$$\mathcal{H}_s\left(y\right) = \mathrm{argmin}_{x\in\mathbb{R}^d}\left\{\left\|x - y\right\|^2 : \left\|x\right\|_0 \leq s\right\} = \begin{cases} y_i, & i \leq s, \\ 0, & \text{otherwise}, \end{cases} \tag{B.23}$$

where $s > 0$ and the operations are applied element-wise. We require the following technical lemma.

Lemma B.5. *Let $\mathbf{Q} \in \mathbb{R}^{A\times B}$ for some positive integers A and B. Let $t \geq 0$ and $\left\|\mathbf{Q}\right\|_F \neq 0$ then*

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2, \left\|\mathbf{X}\right\|_0 \leq s\right\} = -t\left\|\mathcal{H}_s(-\mathbf{Q})\right\|\,.$$

with the minimizer $\mathbf{X}^* = \frac{t\mathcal{H}_s(-\mathbf{Q})}{\|\mathcal{H}_s(-\mathbf{Q})\|}$ if $\|\mathcal{H}_s(-\mathbf{Q})\| \neq 0$ else $\mathbf{X}^* = \mathbf{0}$. Moreover we have the following equivalence

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2, \left\|\mathbf{X}\right\|_0 \leq s\right\} = \min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 = t^2, \left\|\mathbf{X}\right\|_0 \leq s\right\}.$$

Proof. The proof is similar to [40, Proposition 11]. We have

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 \le t^2, \|\mathbf{X}\|_0 \le s \right\} = -\max_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle -\mathbf{Q}, \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 \le t^2, \|\mathbf{X}\|_0 \le s \right\},$$
$$= -\max_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathcal{H}_s(-\mathbf{Q}), \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 \le t^2 \right\}.$$

The first equality is a simple rewriting of the objective. Then, the corresponding objective $\langle -\mathbf{Q}, \mathbf{X}\rangle$ can be maximized with $\sum_{i=1}^{A}\sum_{j=1}^{B} \mathbf{I}_{(i,j)\in\Omega_0}(-\mathbf{Q}_{ij}\mathbf{X}_{ij})$ where Ω_0 is set of index pairs and $\mathbf{I}_{(i,j)\in\Omega_0}$ is 1 if the index pair if $(i,j)\in\Omega_0$ and zero otherwise. Note that the objective $\langle -\mathbf{Q}, \mathbf{X}\rangle$ is maximized if Ω_0 contains all the index pairs corresponding to the elements of $-\mathbf{Q}$ with highest absolute value which is captured by Hard-thresholding operator. Thus, the second equality follows and the solution follows due to Lemma B.1. The equivalence statement follows as $\|\mathbf{X}^*\|_F^2 = t^2$ for $\|\mathcal{H}_s(-\mathbf{Q})\| \neq 0$ else the function value is zero and is attained by all the points in the set $\left\{\mathbf{X} : \|\mathbf{X}\|_F^2 \le t^2\right\}$ are minimizers, hence the equivalence. □

Proposition B.7. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = r\mathcal{H}_{s_1}(-\mathbf{P^k})$, $\mathbf{Z^{k+1}} = r\mathcal{H}_{s_2}(-\mathbf{Q^k})$ where $r \ge 0$ and satisfies*

$$c_1\left(\left\| \mathcal{H}_{s_1}\left(-\mathbf{Q^k}\right)\right\|_F^2 + \left\| \mathcal{H}_{s_2}\left(-\mathbf{P^k}\right)\right\|_F^2 \right) r^3 + c_2 r - 1 = 0\,, \tag{B.24}$$

with $c_1 = 3$ and $c_2 = \|\mathbf{A}\|_F$.

The proof is similar to Proposition B.5 but Lemma B.5 must be used instead of Lemma B.3.

C Closed Form Solutions Part II for NMF variants

For simplicity we consider the following problem [35, 36]

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\|\mathbf{A}-\mathbf{U}\mathbf{Z}\|_F^2 + \mathbf{I}_{\mathbf{U}\ge\mathbf{0}} + \mathbf{I}_{\mathbf{Z}\ge\mathbf{0}} \right\}. \tag{C.1}$$

We set $\mathcal{R}_1(\mathbf{U}) = 0$, $\mathcal{R}_2(\mathbf{Z}) = 0$, $g = \Psi$ and $f = \mathbf{I}_{\mathbf{U}\ge\mathbf{0}} + \mathbf{I}_{\mathbf{Z}\ge\mathbf{0}}$ where $\mathbf{I}$ is the indicator operator. We start with the following technical lemma.

Lemma C.1. *Let $\mathbf{Q}\in\mathbb{R}^{A\times B}$ for some positive integers A and B. Let $t \ge 0$ and $\|\mathbf{Q}\|_F \neq 0$ then*

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 \le t^2, \mathbf{X} \ge \mathbf{0} \right\} = -t\,\|\Pi_+(-\mathbf{Q})\|_F\,,$$

with the minimizer $\mathbf{X}^* = t\frac{\Pi_+(-\mathbf{Q})}{\|\Pi_+(-\mathbf{Q})\|_F}$ if $\|\Pi_+(-\mathbf{Q})\|_F \neq 0$ else $\mathbf{X}^* = \mathbf{0}$. For $\|\Pi_+(-\mathbf{Q})\|_F \neq 0$, we have the following equivalence

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 \le t^2, \mathbf{X} \ge \mathbf{0} \right\} \equiv \min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 = t^2, \mathbf{X} \ge \mathbf{0} \right\}. \tag{C.2}$$

Proof. On rewriting we have the following equivalence

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle \mathbf{Q}, \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 \le t^2, \mathbf{X} \ge \mathbf{0} \right\} \equiv -\max_{\mathbf{X}\in\mathbb{R}^{A\times B}} \left\{ \langle -\mathbf{Q}, \mathbf{X}\rangle : \|\mathbf{X}\|_F^2 \le t^2, \mathbf{X} \ge \mathbf{0} \right\}.$$

The expression $\langle -\mathbf{Q}, \mathbf{X}\rangle$ is maximized at $\mathbf{X}^* = c\Pi_+(-\mathbf{Q})$ for certain constant c. On substituting we have

$$\langle -\mathbf{Q}, \mathbf{X}^*\rangle = c\,\|\Pi_+(-\mathbf{Q})\|_F^2\,.$$

Since, the dependence on c is linear and we additionally require $\|\mathbf{X}\|_F^2 \le t^2$, we can set $c = \frac{t}{\|\Pi_+(-\mathbf{Q})\|_F}$ if $\|\Pi_+(-\mathbf{Q})\|_F \neq 0$ else $c = 0$. Hence, the minimizer to

$$\min_{\mathbf{X} \in \mathbb{R}^{A \times B}} \left\{ \langle \mathbf{Q}, \mathbf{X} \rangle : \|\mathbf{X}\|_F^2 \le t^2 \right\}$$

is attained at $\mathbf{X}^* = -t \frac{\Pi_+(-\mathbf{Q})}{\|\Pi_+(-\mathbf{Q})\|_F}$ for $\|\Pi_+(-\mathbf{Q})\|_F \neq 0$ else $\mathbf{X}^* = 0$. The equivalence in the statement follows as $\|\mathbf{X}^*\|_F^2 = t^2$. □

Denote $g = \Psi$, $f = \mathbf{I}_{\mathbf{U} \ge \mathbf{0}} + \mathbf{I}_{\mathbf{Z} \ge \mathbf{0}}$ and $h = h_a$.

Proposition C.1. *In BPG-MF, when* $g = \Psi$ *in* (C.1) *the update step in each iteration are given by* $\mathbf{U}^{\mathbf{k+1}} = \Pi_+(-\mathbf{P}^{\mathbf{k}})$, $\mathbf{Z}^{\mathbf{k+1}} = \Pi_+(-\mathbf{Q}^{\mathbf{k}})$ *where* $r \ge 0$ *and satisfies*

$$c_1 \left(\left\| \Pi_+(-\mathbf{Q}^{\mathbf{k}}) \right\|_F^2 + \left\| \Pi_+(-\mathbf{P}^{\mathbf{k}}) \right\|_F^2 \right) r^3 + c_2 r - 1 = 0\,., \tag{C.3}$$

with $c_1 = 3$ *and* $c_2 = \|\mathbf{A}\|_F$.

Proof. The proof is similar to that of Proposition B.1, but with certain changes due to the involved non-negativity constraints for the objective. Consider the following subproblem

$$\begin{aligned}(\mathbf{U}^{\mathbf{k+1}}, \mathbf{Z}^{\mathbf{k+1}}) \in \underset{(\mathbf{U},\mathbf{Z}) \in \mathbb{R}_+^{M \times K} \times \mathbb{R}_+^{K \times N}}{\operatorname{argmin}} & \left\{ \left\langle \mathbf{P}^{\mathbf{k}}, \mathbf{U} \right\rangle + \left\langle \mathbf{Q}^{\mathbf{k}}, \mathbf{Z} \right\rangle \right. \\ & \left. + c_1 \left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2} \right)^2 + c_2 \left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2} \right) \right\}.\end{aligned}$$

Denote the objective in the above minimization problem as $\mathcal{O}(\mathbf{U}^{\mathbf{k}}, \mathbf{Z}^{\mathbf{k}})$. Now, we show that the following holds

$$\begin{aligned}&\min_{(\mathbf{U},\mathbf{Z}) \in \mathbb{R}^{M \times K} \times \mathbb{R}^{K \times N}} \left(\mathcal{O}(\mathbf{U}^{\mathbf{k}}, \mathbf{Z}^{\mathbf{k}}) \right) \\ &\equiv \min_{t_1 \ge 0, t_2 \ge 0} \left\{ \min_{(\mathbf{U},\mathbf{Z}) \in \mathbb{R}^{M \times K} \times \mathbb{R}^{K \times N}, \|\mathbf{U}\|_F = t_1, \|\mathbf{Z}\|_F = t_2} \left(\mathcal{O}(\mathbf{U}^{\mathbf{k}}, \mathbf{Z}^{\mathbf{k}}) \right) \right\}, \qquad (C.4) \\ &\equiv \min_{t_1 \ge 0, t_2 \ge 0} \left\{ \min_{(\mathbf{U},\mathbf{Z}) \in \mathbb{R}^{M \times K} \times \mathbb{R}^{K \times N}, \|\mathbf{U}\|_F \le t_1, \|\mathbf{Z}\|_F \le t_2} \left(\mathcal{O}(\mathbf{U}^{\mathbf{k}}, \mathbf{Z}^{\mathbf{k}}) \right) \right\}, \qquad (C.5)\end{aligned}$$

where the first step is a simple rewriting of the objective and involved variables and the second equivalence proof is similar to that equivalence of (B.13) and (B.12) in Proposition B.5, which we describe now. The second step is non-trivial. In order to prove (C.5) we rewrite (C.4) as

$$\begin{aligned}\min_{t_1 \ge 0, t_2 \ge 0} & \left\{ \min_{\mathbf{U_1} \in \mathbb{R}^{M \times K}} \left\{ \left\langle \mathbf{P}^{\mathbf{k}}, \mathbf{U_1} \right\rangle + \lambda\lambda_1 \|\mathbf{U}\|_1 : \|\mathbf{U_1}\|_F^2 = t_1 \right\} \right. \\ & + \min_{\mathbf{Z_1} \in \mathbb{R}^{K \times N}} \left\{ \left\langle \mathbf{Q}^{\mathbf{k}}, \mathbf{Z_1} \right\rangle + \lambda\lambda_1 \|\mathbf{Z}\|_1 : \|\mathbf{Z_1}\|_F^2 = t_2 \right\} \\ & \left. + c_1 \left(\frac{t_1^2 + t_2^2}{2} \right)^2 + c_2 \left(\frac{t_1^2 + t_2^2}{2} \right) \right\}.\end{aligned}$$

where the second step uses Lemma B.3 and strong convexity of h. Now, due to Lemma B.3, if $\left\| \Pi_+(-\mathbf{P}^{\mathbf{k}}) \right\|_F \neq 0$ we have

$$\min_{\mathbf{U_1} \in \mathbb{R}^{M \times K}} \left\{ \left\langle \mathbf{P}^{\mathbf{k}}, \mathbf{U_1} \right\rangle : \|\mathbf{U_1}\|_F^2 = t_1\,, \mathbf{U_1} \ge 0 \right\} \equiv \min_{\mathbf{U_1} \in \mathbb{R}^{M \times K}} \left\{ \left\langle \mathbf{P}^{\mathbf{k}}, \mathbf{U_1} \right\rangle : \|\mathbf{U_1}\|_F^2 \le t_1\,, \mathbf{U_1} \ge 0 \right\}, \tag{C.6}$$

and similarly if $\left\|\Pi_+(-\mathbf{Q^k})\right\|_F \neq 0$ we have

$$\min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}} \left\{\left\langle \mathbf{Q^k}, \mathbf{Z_1}\right\rangle : \|\mathbf{Z_1}\|_F^2 = t_2\,, \mathbf{Z_1} \geq 0\right\} \equiv \min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}} \left\{\left\langle \mathbf{Q^k}, \mathbf{Z_1}\right\rangle : \|\mathbf{Z_1}\|_F^2 \leq t_2\,, \mathbf{Z_1} \geq 0\right\}\,. \quad \text{(C.7)}$$

Note that if $\left\|\Pi_+(-\mathbf{P^k})\right\|_F = 0$ and $\left\|\mathbf{P^k}\right\|_F \neq 0$ then the objective

$$\min_{\mathbf{U_1}\in\mathbb{R}^{M\times K}} \left\{\left\langle \mathbf{P^k}, \mathbf{U_1}\right\rangle : \|\mathbf{U_1}\|_F^2 = t_1\,, \mathbf{U_1} \geq 0\right\}$$

with minimum function value of a positive value $t_1 \min_{i\in[M],\, j\in[K]} \{(\mathbf{P^k})_{i,j}\}$ where we have $[A] = \{1, 2, \ldots, A\}$ for a positive integer A. Similarly if $\left\|\Pi_+(-\mathbf{Q^k})\right\|_F = 0$ and $\left\|\mathbf{Q^k}\right\|_F \neq 0$ the minimum function value for

$$\min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}} \left\{\left\langle \mathbf{Q^k}, \mathbf{Z_1}\right\rangle : \|\mathbf{Z_1}\|_F^2 = t_2\,, \mathbf{Z_1} \geq 0\right\}$$

is a positive value $t_2 \min_{i\in[K],\, j\in[N]} \{(\mathbf{Q^k})_{i,j}\}$. Thus for $\left\|\mathbf{P^k}\right\|_F \neq 0$ with $\left\|\Pi_+(-\mathbf{P^k})\right\|_F = 0$ (or $\left\|\mathbf{Q^k}\right\|_F \neq 0$ with $\left\|\Pi_+(-\mathbf{Q^k})\right\|_F = 0$) the final objective (C.4) is monotonically increasing in t_1 (or t_2) which will drive t_1 (or t_2) to 0 due to the constraint $t_1 \geq 0$ (or $t_2 \geq 0$). So, without loss of generality we can consider $\left\|\Pi_+(-\mathbf{Q^k})\right\|_F \neq 0$ and $\left\|\Pi_+(-\mathbf{Q^k})\right\|_F = 0$. Now, we obtain the solutions via the following strategy. Denote

$$\mathbf{U}_1^*(t_1) \in \operatorname{argmin}\left\{\left\langle \mathbf{P^k}, \mathbf{U_1}\right\rangle : \mathbf{U_1} \in \mathbb{R}_+^{M\times K}\,, \|\mathbf{U_1}\|_F^2 \leq t_1\right\}\,,$$
$$\mathbf{Z}_1^*(t_2) \in \operatorname{argmin}\left\{\left\langle \mathbf{Q^k}, \mathbf{Z_1}\right\rangle : \mathbf{Z_1} \in \mathbb{R}_+^{K\times N}\,, \|\mathbf{Z_1}\|_F^2 \leq t_2\right\}\,.$$

Then we obtain $(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}) = (\mathbf{U}_1^*(t_1^*), \mathbf{Z}_1^*(t_2^*))$, where t_1^* and t_2^* are obtained by solving the following two dimensional subproblem

$$\begin{aligned}(t_1^*, t_2^*) \in \operatorname*{argmin}_{t_1\geq 0, t_2\geq 0} \Bigg\{ &\min_{\mathbf{U_1}\in\mathbb{R}_+^{M\times K}} \left\{\left\langle \mathbf{P^k}, \mathbf{U_1}\right\rangle : \|\mathbf{U_1}\|_F^2 \leq t_1\right\} \\ &+ \min_{\mathbf{Z_1}\in\mathbb{R}_+^{K\times N}} \left\{\left\langle \mathbf{Q^k}, \mathbf{Z_1}\right\rangle : \|\mathbf{Z_1}\|_F^2 \leq t_2\right\} \\ &+ c_1\left(\frac{t_1+t_2}{2}\right)^2 + c_2\left(\frac{t_1+t_2}{2}\right)\Bigg\}\,.\end{aligned}$$

Note that inner minimization subproblems can be trivially solved once we obtain $\mathbf{U}_1^*(t_1)$ and $\mathbf{Z}_1^*(t_2)$. Due to Lemma C.1 we obtain the solution to the subproblem in each iteration as follows

$$\mathbf{U^{k+1}} = \begin{cases} t_1^* \frac{\Pi_+(-\mathbf{P^k})}{\left\|\Pi_+(-\mathbf{P^k})\right\|_F}, & \text{for } \left\|\Pi_+(-\mathbf{P^k})\right\|_F \neq 0\,, \\ \mathbf{0}, & \textit{otherwise}\,. \end{cases}$$
$$\mathbf{Z^{k+1}} = \begin{cases} t_2^* \frac{\Pi_+(-\mathbf{Q^k})}{\left\|\Pi_+(-\mathbf{Q^k})\right\|_F}, & \text{for } \left\|\Pi_+(-\mathbf{Q^k})\right\|_F \neq 0\,, \\ \mathbf{0}, & \textit{otherwise}\,. \end{cases}$$

We solve for t_1^* and t_2^* with the following two dimensional minimization problem

$$\operatorname*{argmin}_{t_1\geq 0, t_2\geq 0} \left\{ -t_1 \left\|\Pi_+(-\mathbf{P^k})\right\|_F - t_2\left\|\Pi_+(-\mathbf{Q^k})\right\|_F + c_1\left(\frac{t_1^2+t_2^2}{2}\right)^2 + c_2\left(\frac{t_1^2+t_2^2}{2}\right)\right\}\,.$$

Thus, the solutions t_1^* and t_2^* are the non-negative real roots of the following equations

$$-\left\|\Pi_+(-\mathbf{P^k})\right\|_F + c_1(t_1^2+t_2^2)t_1 + c_2 t_1 = 0\,,$$
$$-\left\|\Pi_+(-\mathbf{Q^k})\right\|_F + c_1(t_1^2+t_2^2)t_2 + c_2 t_2 = 0\,.$$

Further simplifications lead to $t_1 = r \left\| \Pi_+(-\mathbf{P^k}) \right\|_F$ and $t_2 = r \left\| \Pi_+(-\mathbf{Q^k}) \right\|_F$ for some $r \geq 0$. This results in the following cubic equation,

$$c_1 \left(\left\| \Pi_+(-\mathbf{Q^k}) \right\|_F^2 + \left\| \Pi_+(-\mathbf{P^k}) \right\|_F^2 \right) r^3 + c_2 r - 1 = 0 \,,$$

where the solution is the non-negative real root. □

C.1 Extensions to L2-regularized NMF

Here, the goal is solve the following minimization problem

$$\min_{\mathbf{U} \in \mathbb{R}^{M \times K}, \mathbf{Z} \in \mathbb{R}^{K \times N}} \left\{ \Psi(\mathbf{U}, \mathbf{Z}) := \frac{1}{2} \left\| \mathbf{A} - \mathbf{UZ} \right\|_F^2 + \frac{\lambda_0}{2} \left(\left\| \mathbf{U} \right\|_F^2 + \left\| \mathbf{Z} \right\|_F^2 \right) + \mathbf{I}_{\mathbf{U} \geq \mathbf{0}} + \mathbf{I}_{\mathbf{Z} \geq \mathbf{0}} \right\} .$$

Denote $g := \frac{1}{2} \left\| \mathbf{A} - \mathbf{UZ} \right\|_F^2 + \frac{\lambda_0}{2} \left(\left\| \mathbf{U} \right\|_F^2 + \left\| \mathbf{Z} \right\|_F^2 \right)$, $f := \mathbf{I}_{\mathbf{U} \geq \mathbf{0}} + \mathbf{I}_{\mathbf{Z} \geq \mathbf{0}}$ and $h = h_b$.

Proposition C.2. *In BPG-MF, with above defined g, f, h the update step in each iteration are given by $\mathbf{U^{k+1}} = \Pi_+(-\mathbf{P^k})$, $\mathbf{Z^{k+1}} = \Pi_+(-\mathbf{Q^k})$ where $r \geq 0$ and satisfies*

$$c_1 \left(\left\| \Pi_+(-\mathbf{Q^k}) \right\|_F^2 + \left\| \Pi_+(-\mathbf{P^k}) \right\|_F^2 \right) r^3 + (c_2 + \lambda_0) r - 1 = 0 \,,$$

with $c_1 = 3$ and $c_2 = \left\| \mathbf{A} \right\|_F$.

The proof is similar to Proposition C.1 with only change in c_2.

C.2 Extensions to L1-regularized NMF

Here, the goal is solve the following minimization problem

$$\min_{\mathbf{U} \in \mathbb{R}^{M \times K}, \mathbf{Z} \in \mathbb{R}^{K \times N}} \left\{ \Psi(\mathbf{U}, \mathbf{Z}) := \frac{1}{2} \left\| \mathbf{A} - \mathbf{UZ} \right\|_F^2 + \lambda_1 \left(\left\| \mathbf{U} \right\|_1 + \left\| \mathbf{Z} \right\|_1 \right) + \mathbf{I}_{\mathbf{U} \geq \mathbf{0}} + \mathbf{I}_{\mathbf{Z} \geq \mathbf{0}} \right\} .$$

We denote $\mathbf{e_D}$ to be a vector of dimension $\mathbf{D}$ with all its elements set to 1.

Lemma C.2. *Let $\mathbf{Q} \in \mathbb{R}^{A \times B}$ for some positive integers A and B. Let $t \geq 0$ and $\left\| \mathbf{Q} \right\|_F \neq 0$ then*

$$\min_{\mathbf{X} \in \mathbb{R}^{A \times B}} \left\{ \left\langle \mathbf{Q}, \mathbf{X} \right\rangle + t_0 \left\| \mathbf{X} \right\|_1 : \left\| \mathbf{X} \right\|_F^2 \leq t^2, \mathbf{X} \geq 0 \right\} = -t \left\| \Pi_+(-\left(\mathbf{Q} + t_0 \mathbf{e_A} \mathbf{e_B}^T \right)) \right\|_F$$

with the minimizer $\mathbf{X}^* = t \frac{\Pi_+(-\left(\mathbf{Q} + t_0 \mathbf{e_A} \mathbf{e_B}^T \right))}{\left\| \Pi_+(-\left(\mathbf{Q} + t_0 \mathbf{e_A} \mathbf{e_B}^T \right)) \right\|_F}$ if the condition $\left\| \Pi_+(-\left(\mathbf{Q} + t_0 \mathbf{e_A} \mathbf{e_B}^T \right)) \right\|_F \neq 0$ holds.

Proof. By using $\mathbf{X} \geq 0$ and the basic trace properties we have the following equivalence

$$\left\| \mathbf{X} \right\|_1 = \sum_{i,j} \mathbf{X}_{ij} = \mathbf{e_A}^T \mathbf{X} \mathbf{e_B} = \mathbf{tr} \left(\mathbf{e_A}^T \mathbf{X} \mathbf{e_B} \right) = \mathbf{tr} \left(\mathbf{e_B} \mathbf{e_A}^T \mathbf{X} \right) = \left\langle \mathbf{e_A} \mathbf{e_B}^T, \mathbf{X} \right\rangle \,,$$

hence we have the following equivalence

$$\begin{aligned} &\min_{\mathbf{X} \in \mathbb{R}^{A \times B}} \left\{ \left\langle \mathbf{Q}, \mathbf{X} \right\rangle + t_0 \left\| \mathbf{X} \right\|_1 : \left\| \mathbf{X} \right\|_F^2 \leq t^2, \mathbf{X} \geq 0 \right\} \\ \equiv &\min_{\mathbf{X} \in \mathbb{R}^{A \times B}} \left\{ \left\langle \mathbf{Q} + t_0 \mathbf{e_A} \mathbf{e_B}^T, \mathbf{X} \right\rangle : \left\| \mathbf{X} \right\|_F^2 \leq t^2, \mathbf{X} \geq 0 \right\} \end{aligned}$$

Now, the solution follows due to Lemma C.1. □

Denote $g := \frac{1}{2} \left\| \mathbf{A} - \mathbf{UZ} \right\|_F^2$, $f := \lambda_1 \left(\left\| \mathbf{U} \right\|_1 + \left\| \mathbf{Z} \right\|_1 \right) + \mathbf{I}_{\mathbf{U} \geq \mathbf{0}} + \mathbf{I}_{\mathbf{Z} \geq \mathbf{0}}$ and $h = h_a$.

Proposition C.3. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = r \Pi_+(-\left(\mathbf{P^k} + t_0 \mathbf{e}_M \mathbf{e}_K^T \right))$, $\mathbf{Z^{k+1}} = r \Pi_+(-\left(\mathbf{Q^k} + t_0 \mathbf{e}_K \mathbf{e}_N^T \right))$ where $r \geq 0$ and satisfies*

$$c_1 \left(\left\| \Pi_+(-\left(\mathbf{P^k} + t_0 \mathbf{e}_M \mathbf{e}_K^T \right)) \right\|_F^2 + \left\| \Pi_+(-\left(\mathbf{Q^k} + t_0 \mathbf{e}_K \mathbf{e}_N^T \right)) \right\|_F^2 \right) r^3 + c_2 r - 1 = 0 \,,$$

with $c_1 = 3$ and $c_2 = \left\| \mathbf{A} \right\|_F$.

We skip the proof as it is similar to Proposition C.1.

C.3 Extensions to Graph Regularized Non-negative Matrix Factorization

Graph Regularized Non-negative Matrix Factorization was proposed in [12]. Here, given $\mathcal{L} \in \mathbb{R}^{M\times M}$ we are interested to solve

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \{\Psi(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2 + \frac{\mu_0}{2}\mathbf{tr}(\mathbf{U}^T\mathcal{L}\mathbf{U}) + \frac{\lambda_0}{2}\left(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2\right) + \mathbf{I}_{\mathbf{U}\geq\mathbf{0}} + \mathbf{I}_{\mathbf{Z}\geq\mathbf{0}}\bigg\}\,.$$

Recall that

$$h_c(\mathbf{U},\mathbf{Z}) = 3h_1(\mathbf{U},\mathbf{Z}) + \left(\left\|\mathbf{A}\right\|_F + \mu_0\left\|\mathcal{L}\right\|_F\right)h_2(\mathbf{U},\mathbf{Z})\,.$$

Denote $g := \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2 + \frac{\mu_0}{2}\mathbf{tr}(\mathbf{U}^T\mathcal{L}\mathbf{U})$, $f := \frac{\lambda_0}{2}\left(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{Z}\right\|_F^2\right) + \mathbf{I}_{\mathbf{U}\geq\mathbf{0}} + \mathbf{I}_{\mathbf{Z}\geq\mathbf{0}}$ and $h = h_c$.

Proposition C.4. *In BPG-MF, with the above defined f, g, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = r\Pi_+(-\mathbf{P^k})$, $\mathbf{Z^{k+1}} = r\Pi_+(-\mathbf{Q^k})$ where $r \geq 0$ and satisfies*

$$c_1\left(\left\|\Pi_+(-\mathbf{Q^k})\right\|_F^2 + \left\|\Pi_+(-\mathbf{P^k})\right\|_F^2\right)r^3 + (c_2 + \mu_0\left\|\mathcal{L}\right\|_F + \lambda_0)r - 1 = 0\,, \tag{C.8}$$

with $c_1 = 3$ and $c_2 = \left\|\mathbf{A}\right\|_F$.

The proof is similar to Proposition C.1 and only c_2 changes.

C.4 Extensions to Symmetric NMF via Non-symmetric relaxation.

In [66], the following optimization problem was proposed in the context of Symmetric NMF where the factors $\mathbf{U}$ and $\mathbf{Z}^T$ are equal. The symmetricity of the factors was lifted via a quadratic penalty terms resulting in the following problem

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{\Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2 + \frac{\lambda_0}{2}\left\|\mathbf{U}-\mathbf{Z}^T\right\|_F^2 + \mathbf{I}_{\mathbf{U}\geq\mathbf{0}} + \mathbf{I}_{\mathbf{Z}\geq\mathbf{0}}\right\}\,.$$

Now, we prove the L-smad property. We need the following technical lemma.

Lemma C.3. *Let $g(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2 + \frac{\lambda_0}{2}\left\|\mathbf{U}-\mathbf{Z}^T\right\|_F^2$ be as defined above, we have the following*

$$\nabla_{\mathbf{U}}g(\mathbf{A},\mathbf{U}\mathbf{Z}) = \lambda_0\left(\mathbf{U}-\mathbf{Z}^T\right) - (\mathbf{A}-\mathbf{U}\mathbf{Z})\mathbf{Z}^T$$
$$\nabla_{\mathbf{Z}}g(\mathbf{A},\mathbf{U}\mathbf{Z}) = \lambda_0\left(\mathbf{U}-\mathbf{Z}^T\right) + \mathbf{U}^T(\mathbf{A}-\mathbf{U}\mathbf{Z})$$

and

$$\begin{aligned}&\left\langle(\mathbf{H_1},\mathbf{H_2}),\nabla^2 g(\mathbf{A},\mathbf{U}\mathbf{Z})(\mathbf{H_1},\mathbf{H_2})\right\rangle\\ &= -2\left\langle\mathbf{A}-\mathbf{U}\mathbf{Z},\mathbf{H_1}\mathbf{H_2}\right\rangle + \left\|\mathbf{U}\mathbf{H_2}+\mathbf{H_1}\mathbf{Z}\right\|_F^2 + \lambda_0\left\|\mathbf{H_1}-\mathbf{H_2}^T\right\|_F^2\,.\end{aligned}$$

Proof. The first part of proof for function $\frac{1}{2}\left\|\mathbf{A}-\mathbf{U}\mathbf{Z}\right\|_F^2$ follows from Proposition 2.1. For the other term, with the Forbenius dot product, we obtain

$$\begin{aligned}&\frac{\lambda_0}{2}\left\|\mathbf{U}+\mathbf{H_1}-\mathbf{Z}^T-\mathbf{H_2}^T\right\|_F^2\\ &= \frac{\lambda_0}{2}\left(\left\|\mathbf{U}-\mathbf{Z}^T\right\|_F^2 + 2\left\langle\mathbf{U}-\mathbf{Z}^T,\mathbf{H_1}-\mathbf{H_2}^T\right\rangle + \left\|\mathbf{H_1}-\mathbf{H_2}^T\right\|_F^2\right)\,.\end{aligned}$$

Combining with Lemma F.1, the statement follows from the collecting the first order and second order terms. ∎

Proposition C.5. *Let $g(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \frac{\lambda_0}{2}\left\|\mathbf{U}-\mathbf{Z}\right\|_F^2$. Then, for a certain constant $L \geq 1$, the function g satisfies L-smad property with respect to the following kernel generating distance,*

$$h_d(\mathbf{U},\mathbf{Z}) = 3h_1(\mathbf{U},\mathbf{Z}) + \left(\left\|\mathbf{A}\right\|_F + 2\lambda_0\right)h_2(\mathbf{U},\mathbf{Z})\,.$$

Proof. The proof is similar to Proposition 2.1 and Lemma C.3 must be applied for the result. □

Denote $g := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 + \frac{\lambda_0}{2}\left\|\mathbf{U}-\mathbf{Z}\right\|_F^2$, $f := \mathbf{I}_{\mathbf{U}\geq\mathbf{0}} + \mathbf{I}_{\mathbf{Z}\geq\mathbf{0}}$ and $h = h_d$.

Proposition C.6. *In BPG-MF, with the above defined update steps in each iteration are given by $\mathbf{U^{k+1}} = r\Pi_+\left(-\mathbf{P^k}\right)$, $\mathbf{Z^{k+1}} = r\Pi_+\left(-\mathbf{Q^k}\right)$ where $r \geq 0$ and satisfies*

$$c_1\left(\left\|\Pi_+\left(-\mathbf{P^k}\right)\right\|_F^2 + \left\|\Pi_+\left(-\mathbf{Q^k}\right)\right\|_F^2\right)r^3 + (c_2+2\lambda_0)r - 1 = 0\,, \tag{C.9}$$

with $c_1 = 3$ and $c_2 = \left\|\mathbf{A}\right\|_F$.

The proof is similar to Proposition C.1 and only c_2 changes.

C.5 Extensions to NMF with non-convex sparsity constraints (Sparse NMF)

Consider the following problem from [7]

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}}\left\{\Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 : \mathbf{U}\geq\mathbf{0}, \left\|\mathbf{U}\right\|_0 \leq s_1, \mathbf{Z}\geq\mathbf{0}, \left\|\mathbf{Z}\right\|_0 \leq s_2,\right\}\,,$$

where s_1 and s_2 are two known positive integers. Denote $g := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2$, $f := \mathbf{I}_{\mathbf{U}\geq\mathbf{0}} + \mathbf{I}_{\left\|\mathbf{U}\right\|_0\leq s_1} + \mathbf{I}_{\mathbf{Z}\geq\mathbf{0}} + \mathbf{I}_{\left\|\mathbf{Z}\right\|_0\leq s_2}$ and $h = h_a$. Note that the Assumption C is not valid here, hence CoCaIn BPG-MF theory does not hold and hints at possible extensions of CoCaIn BPG-MF, which is an interesting open question. We start with the following technical lemma.

Proposition C.7. *Let $\mathbf{Q}\in\mathbb{R}^{A\times B}$ for some positive integers A and B. Let $t \geq 0$ and $\left\|\mathbf{Q}\right\|_F \neq 0$ then*

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2, \left\|\mathbf{X}\right\|_0 \leq s, \mathbf{X}\geq\mathbf{0}\right\} = -t\left\|\mathcal{H}_s(\Pi_+(-\mathbf{Q}))\right\|_F\,.$$

with the minimizer $\mathbf{X}^* = t\frac{\mathcal{H}_s(\Pi_+(-\mathbf{Q}))}{\left\|\mathcal{H}_s(\Pi_+(-\mathbf{Q}))\right\|_F}$ if $\left\|\mathcal{H}_s(\Pi_+(-\mathbf{Q}))\right\|_F \neq 0$ else $\mathbf{X}^* = \mathbf{0}$. If $\left\|\mathcal{H}_s(\Pi_+(-\mathbf{Q}))\right\|_F \neq 0$ we have the following equivalence

$$\min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2, \left\|\mathbf{X}\right\|_0 \leq s, \mathbf{X}\geq\mathbf{0}\right\} \tag{C.10}$$

$$\equiv \min_{\mathbf{X}\in\mathbb{R}^{A\times B}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 = t^2, \left\|\mathbf{X}\right\|_0 \leq s, \mathbf{X}\geq\mathbf{0}\right\} \tag{C.11}$$

Proof. We have

$$\begin{aligned}
&\min_{\mathbf{X}}\left\{\langle\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2, \left\|\mathbf{X}\right\|_0 \leq s, \mathbf{X}\geq\mathbf{0}\right\}\\
&= -\max_{\mathbf{X}}\left\{\langle-\mathbf{Q},\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2, \left\|\mathbf{X}\right\|_0 \leq s, \mathbf{X}\geq\mathbf{0}\right\}\,,\\
&= -\max_{\mathbf{X}}\left\{\langle\Pi_+(-\mathbf{Q}),\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2, \left\|\mathbf{X}\right\|_0 \leq s\right\}\,,\\
&= -\max_{\mathbf{X}}\left\{\langle\mathcal{H}_s(\Pi_+(-\mathbf{Q})),\mathbf{X}\rangle : \left\|\mathbf{X}\right\|_F^2 \leq t^2\right\}\,.
\end{aligned}$$

The first equality is a simple rewriting of the objective. Then, the corresponding objective $\langle-\mathbf{Q},\mathbf{X}\rangle$ can be maximized with $\sum_{i=1}^{A}\sum_{j=1}^{B}\mathbf{I}_{(i,j)\in\Omega_0}(-\mathbf{Q}_{ij}\mathbf{X}_{ij})$ where Ω_0 is set of index pairs and $\mathbf{I}_{(i,j)\in\Omega_0}$ is 1 if the index pair if $(i,j)\in\Omega_0$ and zero otherwise. It is easy to see that the objective $\langle-\mathbf{Q},\mathbf{X}\rangle$ is maximized if Ω_0 contains all the index pairs corresponding to the elements of $-\mathbf{Q}$ with highest absolute value which is captured by Hard-thresholding operator. However due to the non-negativity constraint if there is any $-\mathbf{Q}_{ij}$ such that it is negative, then since $\mathbf{X}_{ij}$ will be driven to zero. So, before we use the Hard-thresholding operator, we need to use $\Pi_+(.) = \max\{0, .\}$ in second equality. The third equality follows as a consequence of hard sparsity constraint similar to Lemma B.5 and the solution follows due to Lemma B.1. The equivalence statement follows as $\|\mathbf{X}^*\|_F^2 = t^2$. □

Proposition C.8. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are* $\mathbf{U^{k+1}} = r\mathcal{H}_{s_1}(\Pi_+(-\mathbf{P^k}))$, $\mathbf{Z^{k+1}} = r\mathcal{H}_{s_2}(\Pi_+(-\mathbf{Q^k}))$ *where $r \geq 0$ and satisfies*

$$c_1 \left(\left\| \mathcal{H}_{s_1} \left(\Pi_+(-\mathbf{Q^k}) \right) \right\|_F^2 + \left\| \mathcal{H}_{s_2} \left(\Pi_+(-\mathbf{P^k}) \right) \right\|_F^2 \right) r^3 + c_2 r - 1 = 0 \,,$$

with $c_1 = 3$ and $c_2 = \|\mathbf{A}\|_F$.

The proof is similar to Proposition C.1.

D Matrix Completion Problem

Matrix Completion is an important non-convex optimization problem, which arises in practical real world applications, such as recommender systems [34, 14, 22]. Give a matrix $\mathbf{A}$ where only the values at the index set given by Ω are given. The goal is obtain the rest of the values. One of the popular strategy is to obtain the factors $\mathbf{U} \in \mathbb{R}^{M\times K}$ and $\mathbf{Z} \in \mathbb{R}^{K\times N}$ for a small positive integer K. This is cast into the following problem,

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K}, \mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{ \Psi(\mathbf{U}, \mathbf{Z}) := \frac{1}{2} \|P_\Omega (\mathbf{A} - \mathbf{UZ})\|_F^2 + \frac{\lambda_0}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2 \right) \right\} , \tag{D.1}$$

where P_Ω is an masking operator over index set Ω which preserves the given matrix entries and sets others to zero.. We require the following technical lemma.

Lemma D.1. *Let $g := \frac{1}{2} \|P_\Omega (\mathbf{A} - \mathbf{UZ})\|_F^2$ be as defined above, we have the following*

$$\nabla_{\mathbf{U}} g(\mathbf{A}, \mathbf{UZ}) = -P_\Omega(\mathbf{A} - \mathbf{UZ})\mathbf{Z}^T, \quad \nabla_{\mathbf{Z}} g(\mathbf{A}, \mathbf{UZ}) = -\mathbf{U}^T P_\Omega(\mathbf{A} - \mathbf{UZ})$$

$$\left\langle (\mathbf{H_1}, \mathbf{H_2}), \nabla^2 g(\mathbf{A}, \mathbf{UZ})(\mathbf{H_1}, \mathbf{H_2}) \right\rangle = \|P_\Omega(\mathbf{UH_2} + \mathbf{H_1Z})\|_F^2 - 2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), \mathbf{H_1H_2} \right\rangle \,.$$

Proof. With the Forbenius dot product, we have

$$\|P_\Omega(\mathbf{A} - \mathbf{UZ})\|_F^2 = \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), P_\Omega(\mathbf{A} - \mathbf{UZ}) \right\rangle \,.$$

In the above expression by substituting $\mathbf{U}$ with $\mathbf{U} + \mathbf{H_1}$ and $\mathbf{Z}$ with $\mathbf{Z} + \mathbf{H_2}$, we obtain

$$\begin{aligned}
&\left\langle P_\Omega(\mathbf{A} - (\mathbf{U} + \mathbf{H_1})(\mathbf{Z} + \mathbf{H_2})), P_\Omega(\mathbf{A} - (\mathbf{U} + \mathbf{H_1})(\mathbf{Z} + \mathbf{H_2})) \right\rangle \,, \\
&= \|P_\Omega(\mathbf{A} - \mathbf{UZ})\|_F^2 + \|P_\Omega(\mathbf{UH_2} + \mathbf{H_1Z})\|_F^2 \\
&- 2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), P_\Omega(\mathbf{UH_2} + \mathbf{H_1Z}) \right\rangle - 2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), P_\Omega(\mathbf{H_1H_2}) \right\rangle
\end{aligned}$$

where in the last term we ignored the terms higher than second order. Collecting all the first order terms we have

$$\begin{aligned}
&- 2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), P_\Omega(\mathbf{UH_2} + \mathbf{H_1Z}) \right\rangle \\
&= -2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), \mathbf{UH_2} + \mathbf{H_1Z} \right\rangle \\
&= -2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ})\mathbf{Z}^T, \mathbf{H_1} \right\rangle - 2 \left\langle \mathbf{U}^T P_\Omega(\mathbf{A} - \mathbf{UZ}), \mathbf{H_2} \right\rangle
\end{aligned}$$

and similarly collecting all the second order terms we have

$$\begin{aligned}
&\|P_\Omega(\mathbf{UH_2} + \mathbf{H_1Z})\|_F^2 - 2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), P_\Omega(\mathbf{H_1H_2}) \right\rangle \\
&= \|P_\Omega(\mathbf{UH_2} + \mathbf{H_1Z})\|_F^2 - 2 \left\langle P_\Omega(\mathbf{A} - \mathbf{UZ}), \mathbf{H_1H_2} \right\rangle
\end{aligned}$$

Thus the statement follows using the second order Taylor expansion. □

Proposition D.1. *Let* $g := \frac{1}{2}\|P_\Omega(\mathbf{A}-\mathbf{UZ})\|_F^2$ *and* h_1, h_2 *be as defined as in* (2.6)*. Then, for a certain constant* $L \geq 1$*, the function* g *satisfies* L*-smad property with respect to the following kernel generating distance,*

$$h_a(\mathbf{U},\mathbf{Z}) = 3h_1(\mathbf{U},\mathbf{Z}) + \|P_\Omega(\mathbf{A})\|_F\, h_2(\mathbf{U},\mathbf{Z})\,.$$

Proof. With Lemma F.1 we obtain

$$\begin{aligned}
&\left\langle(\mathbf{H}_1,\mathbf{H}_2), \nabla^2 g(A,\mathbf{UZ})(\mathbf{H}_1,\mathbf{H}_2)\right\rangle \\
&= \|P_\Omega(\mathbf{UH_2}+\mathbf{H_1Z})\|_F^2 - 2\left\langle P_\Omega(\mathbf{A}-\mathbf{UZ}), \mathbf{H_1H_2}\right\rangle \\
&\leq \|\mathbf{H_1Z}+\mathbf{UH_2}\|_F^2 - 2\left\langle P_\Omega(\mathbf{A}-\mathbf{UZ}), \mathbf{H_1H_2}\right\rangle \\
&\leq 2\|\mathbf{H}_1\mathbf{Z}\|_F^2 + 2\|\mathbf{UH}_2\|_F^2 + 2\|P_\Omega(\mathbf{A})\|_F\|\mathbf{H}_1\mathbf{H}_2\|_F + 2\|P_\Omega(\mathbf{UZ})\|_F\|\mathbf{H}_1\mathbf{H}_2\|_F\,, \\
&\leq 2\|\mathbf{H}_1\mathbf{Z}\|_F^2 + 2\|\mathbf{UH}_2\|_F^2 + 2\|P_\Omega(\mathbf{A})\|_F\|\mathbf{H}_1\mathbf{H}_2\|_F + 2\|\mathbf{UZ}\|_F\|\mathbf{H}_1\mathbf{H}_2\|_F\,.
\end{aligned}$$

The rest of the proof is similar to Proposition 2.1. □

Proposition D.2. *Let* $g := \frac{1}{2}\|P_\Omega(\mathbf{A}-\mathbf{UZ})\|_F^2 + \frac{\lambda_0}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2\right)$ *and* h_1, h_2 *be as defined as in* (2.6)*. Then, for a certain constant* $L \geq 1$*, the function* g *satisfies* L*-smad property with respect to the following kernel generating distance,*

$$h_a(\mathbf{U},\mathbf{Z}) = 3h_1(\mathbf{U},\mathbf{Z}) + (\|P_\Omega(\mathbf{A})\|_F + \lambda_0)\, h_2(\mathbf{U},\mathbf{Z})\,.$$

The update steps are very similar as what we described earlier in Section B and C.

E Closed Form Solution with 5th-order Polynomial

The goal of this section is to show a case, where while obtaining the update step of BPG-MF we obtain a 5th order polynomial equation, for which Newton based method solvers can be used. We later show that we can obtain a cubic equation by slightly modifying the kernel generating distance. Let $\lambda_0 > 0$ and we consider the following problem

$$\min_{\mathbf{U}\in\mathbb{R}^{M\times K},\mathbf{Z}\in\mathbb{R}^{K\times N}} \left\{\Psi(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\|\mathbf{A}-\mathbf{UZ}\|_F^2 + \frac{\lambda_0}{2}\|\mathbf{U}\|_F^2\right\}. \tag{E.1}$$

We set $\mathcal{R}_1(\mathbf{U}) = \frac{\lambda_0}{2}\|\mathbf{U}\|_F^2$, $\mathcal{R}_2(\mathbf{Z}) = 0$, $g = \frac{1}{2}\|\mathbf{A}-\mathbf{UZ}\|_F^2$, $f(\mathbf{U},\mathbf{Z}) = \frac{\lambda_0}{2}\|\mathbf{U}\|_F^2$ and $h = h_a$.

Proposition E.1. *In BPG-MF, with above defined* g, f, h *the update steps in each iteration are given by* $\mathbf{U^{k+1}} = -\frac{\mathbf{P^k}}{r_1+\lambda_0}$, $\mathbf{Z^{k+1}} = -\frac{\mathbf{Q^k}}{r_1}$ *where* $r_1 \geq 0$ *and satisfies*

$$c_1\left(\left\|\mathbf{Q^k}\right\|_F^2 (r_1+\lambda_0)^2 + \left\|\mathbf{P^k}\right\|_F^2 r_1^2\right) + c_2 r_1^2 (r_1+\lambda_0)^2 - r_1^3(r_1+\lambda_0)^2 = 0\,, \tag{E.2}$$

with $c_1 = 3$ *and* $c_2 = \|\mathbf{A}\|_F$*.*

Proof. The proof is similar to that of Proposition B.1. Consider the following subproblem

$$\begin{aligned}
(\mathbf{U^{k+1}},\mathbf{Z^{k+1}}) \in \operatorname*{argmin}_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}} &\left\{\frac{\lambda_0}{2}\|\mathbf{U}\|_F^2 + \left\langle\mathbf{P^k},\mathbf{U}\right\rangle + \left\langle\mathbf{Q^k},\mathbf{Z}\right\rangle\right. \\
&\left. + c_1\left(\frac{\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2}{2}\right)^2 + c_2\left(\frac{\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2}{2}\right)\right\},
\end{aligned}$$

Denote the objective in the above minimization problem as $\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})$. Now, we show that the following holds

$$
\begin{aligned}
&\min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}} \Big(\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})\Big) \\
&\equiv \min_{t_1\geq 0, t_2\geq 0} \left\{ \min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}, \|\mathbf{U}\|_F = t_1, \|\mathbf{Z}\|_F = t_2} \Big(\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})\Big) \right\}, \\
&\equiv \min_{t_1\geq 0, t_2\geq 0} \left\{ \min_{(\mathbf{U},\mathbf{Z})\in\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}, \|\mathbf{U}\|_F \leq t_1, \|\mathbf{Z}\|_F \leq t_2} \Big(\mathcal{O}(\mathbf{U^k}, \mathbf{Z^k})\Big) \right\}.
\end{aligned}
$$

where the first step is a simple rewriting of the objective and the second step follows as there is no change in the constraint set and due to Lemma B.1, which is given precisely in Proposition B.1 where the equivalence argument used for (B.4) and (B.3) holds here. Note that in the first step, we used $\|\mathbf{U}\|_F = t_1$ this results in deviation of value of c_2 to $c_2 + \lambda_0$, corresponding to $\mathbf{U}$ (see below). We solve for $(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}})$ via the following strategy. Denote

$$
\mathbf{U}_1^*(t_1) \in \operatorname{argmin}\left\{ \left\langle \mathbf{P^k}, \mathbf{U_1} \right\rangle : \mathbf{U_1} \in \mathbb{R}^{M\times K}, \|\mathbf{U_1}\|_F^2 \leq t_1 \right\},
$$

$$
\mathbf{Z}_1^*(t_2) \in \operatorname{argmin}\left\{ \left\langle \mathbf{Q^k}, \mathbf{Z_1} \right\rangle : \mathbf{Z_1} \in \mathbb{R}^{K\times N}, \|\mathbf{Z_1}\|_F^2 \leq t_2 \right\}.
$$

Then we obtain $(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}) = (\mathbf{U}_1^*(t_1^*), \mathbf{Z}_1^*(t_2^*))$, where t_1^* and t_2^* are obtained by solving the following two dimensional subproblem

$$
\begin{aligned}
(t_1^*, t_2^*) \in \operatorname*{argmin}_{t_1\geq 0, t_2\geq 0} \Bigg\{ &\min_{\mathbf{U_1}\in\mathbb{R}^{M\times K}} \left\{ \left\langle \mathbf{P^k}, \mathbf{U_1} \right\rangle : \|\mathbf{U_1}\|_F^2 \leq t_1 \right\} \\
&+ \min_{\mathbf{Z_1}\in\mathbb{R}^{K\times N}} \left\{ \left\langle \mathbf{Q^k}, \mathbf{Z_1} \right\rangle : \|\mathbf{Z_1}\|_F^2 \leq t_2 \right\} \\
&+ c_1 \left(\frac{t_1^2 + t_2^2}{2}\right)^2 + c_2 \frac{t_2^2}{2} + (c_2 + \lambda_0)\frac{t_1^2}{2} \Bigg\}.
\end{aligned}
$$

Note that inner minimization subproblems can be trivially solved once we obtain $\mathbf{U}_1^*(t_1)$ and $\mathbf{Z}_1^*(t_2)$ via Lemma B.1. Then the solution to the subproblem in each iteration as follows:

$$
\mathbf{U^{k+1}} = \begin{cases} t_1^* \frac{-\mathbf{P^k}}{\|\mathbf{P^k}\|_F}, & \text{for } \left\|\mathbf{P^k}\right\|_F \neq 0, \\ \mathbf{0} & \textit{otherwise}. \end{cases}
$$

$$
\mathbf{Z^{k+1}} = \begin{cases} t_2^* \frac{-\mathbf{Q^k}}{\|\mathbf{Q^k}\|_F}, & \text{for } \left\|\mathbf{Q^k}\right\|_F \neq 0, \\ \mathbf{0} & \textit{otherwise}. \end{cases}
$$

We solve for t_1^* and t_2^* with the following two dimensional minimization problem

$$
\operatorname*{argmin}_{t_1\geq 0, t_2\geq 0} \left\{ -t_1 \left\|\mathbf{P^k}\right\|_F - t_2 \left\|\mathbf{Q^k}\right\|_F + c_1 \left(\frac{t_1^2 + t_2^2}{2}\right)^2 + c_2 \frac{t_2^2}{2} + (c_2 + \lambda_0)\frac{t_1^2}{2} \right\}.
$$

Thus, the solutions t_1^* and t_2^* are the non-negative real roots of the following equations

$$
-\left\|\mathbf{P^k}\right\|_F + c_1(t_1^2 + t_2^2)t_1 + (c_2 + \lambda_0)t_1 = 0, \tag{E.3}
$$

$$
-\left\|\mathbf{Q^k}\right\|_F + c_1(t_1^2 + t_2^2)t_2 + c_2 t_2 = 0. \tag{E.4}
$$

Further simplifications with $t_1 = \frac{\|\mathbf{P^k}\|_F}{r_1 + \lambda_0}$ and $t_2 = \frac{\|\mathbf{Q^k}\|_F}{r_1}$ denoting $r_1 = c_1(t_1^2 + t_2^2) + c_2$, then we have

$$
r_1 = c_1 \left(\left(\frac{\|\mathbf{P^k}\|_F}{r_1 + \lambda_0}\right)^2 + \left(\frac{\|\mathbf{Q^k}\|_F}{r_1}\right)^2 \right) + c_2
$$

This will result in following 5^{th} order equation,

$$c_1 \left(\left\| \mathbf{P^k} \right\|_F^2 r_1^2 + \left\| \mathbf{Q^k} \right\|_F^2 (r_1 + \lambda_0)^2 \right) + c_2 r_1^2 (r_1 + \lambda_0)^2 - r_1^3 (r_1 + \lambda_0)^2 = 0 \,.$$

□

E.1 Conversion to Cubic Equation

We set $\mathcal{R}_1(\mathbf{U}) = \frac{\lambda_0}{2} \|\mathbf{U}\|_F^2$, $\mathcal{R}_2(\mathbf{Z}) = 0$ and $g = \frac{1}{2} \|\mathbf{A} - \mathbf{UZ}\|_F^2$. Denote $f(\mathbf{U}, \mathbf{Z}) = \frac{\lambda_0}{2} \|\mathbf{U}\|_F^2$, $h(\mathbf{U}, \mathbf{Z}) = h_a(\mathbf{U}, \mathbf{Z}) + \frac{\lambda_0}{2} \|\mathbf{Z}\|_F^2$. Note that such a g satisfies L-smad property with respect to h satisfies L-smad trivially since only a quadratic term is added to h_a.

Proposition E.2. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = -r\,\mathbf{P^k}$, $\mathbf{Z^{k+1}} = -r\,\mathbf{Q^k}$ where r is the non-negative real root of*

$$c_1 \left(\left\| \mathbf{Q^k} \right\|_F^2 + \left\| \mathbf{P^k} \right\|_F^2 \right) r^3 + (c_2 + \lambda_0) r - 1 = 0 \,, \tag{E.5}$$

with $c_1 = 3$ and $c_2 = \|\mathbf{A}\|_F$.

Proof. The resulting subproblem is

$$(\mathbf{U^{k+1}}, \mathbf{Z^{k+1}}) \in \underset{(\mathbf{U},\mathbf{Z}) \in \mathbb{R}^{M\times K} \times \mathbb{R}^{K \times N}}{\operatorname{argmin}} \left\{ \left\langle \mathbf{P^k}, \mathbf{U} \right\rangle + \left\langle \mathbf{Q^k}, \mathbf{Z} \right\rangle \right.$$
$$\left. + c_1 \left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2} \right)^2 + (c_2 + \lambda_0) \left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2} \right) \right\} .$$

The rest of the proof is similar to Proposition B.1. □

E.2 Extensions to Mixed Regularization Terms

Let $\lambda_0 > 0$ and we consider the following problem

$$\min_{\mathbf{U} \in \mathbb{R}^{M \times K}, \mathbf{Z} \in \mathbb{R}^{K \times N}} \left\{ \Psi(\mathbf{U}, \mathbf{Z}) := \frac{1}{2} \|\mathbf{A} - \mathbf{UZ}\|_F^2 + \frac{\lambda_0}{2} \|\mathbf{U}\|_F^2 + \lambda_1 \|\mathbf{Z}\|_1 \right\} . \tag{E.6}$$

Note that the regularizer is a mixture of L1 and L2 regularization. The usual strategy with $h = h_a$ would result in a Fist order polynomial. In order to generate a cubic equation, we use the same strategy as given Section E.1. We set $h(\mathbf{U}, \mathbf{Z}) = h_a(\mathbf{U}, \mathbf{Z}) + \frac{\lambda_0}{2} \|\mathbf{Z}\|_F^2$, $g = \frac{1}{2} \|\mathbf{A} - \mathbf{UZ}\|_F^2$ and $f(\mathbf{U}, \mathbf{Z}) = \frac{\lambda_0}{2} \|\mathbf{U}\|_F^2 + \lambda_1 \|\mathbf{Z}\|_1$.

Proposition E.3. *In BPG-MF, with the above defined g, f, h the update steps in each iteration are given by $\mathbf{U^{k+1}} = -r\,\mathbf{P^k}$, $\mathbf{Z^{k+1}} = r\mathcal{S}_{\lambda\lambda_1}\left(-\mathbf{Q^k}\right)$ where r is the non-negative real root of*

$$c_1 \left(\left\| \mathbf{P^k} \right\|_F^2 + \left\| \mathcal{S}_{\lambda\lambda_1}\left(-\mathbf{Q^k}\right) \right\|_F^2 \right) r^3 + (c_2 + \lambda_0) r - 1 = 0 \,, \tag{E.7}$$

with $c_1 = 3$ and $c_2 = \|\mathbf{A}\|_F$.

The proof is similar to Proposition B.1 and Proposition B.5.

F Technical Lemmas and Proofs

Before we proceed to the proof of Proposition 2.1 we require the following technical lemma.

Lemma F.1. *Let* $g := \frac{1}{2}\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2$*, then we have the following*

$$\nabla g(\mathbf{A},\mathbf{UZ}) = \left(-(\mathbf{A}-\mathbf{UZ})\mathbf{Z}^T, -\mathbf{U}^T(\mathbf{A}-\mathbf{UZ})\right)$$

$$\left\langle(\mathbf{H_1},\mathbf{H_2}), \nabla^2 g(\mathbf{A},\mathbf{UZ})(\mathbf{H_1},\mathbf{H_2})\right\rangle = -2\left\langle\mathbf{A}-\mathbf{UZ},\mathbf{H_1H_2}\right\rangle + \left\langle\mathbf{UH_2}+\mathbf{H_1Z},\mathbf{UH_2}+\mathbf{H_1Z}\right\rangle .$$

Proof. With the Forbenius dot product, we have

$$\left\|\mathbf{A}-\mathbf{UZ}\right\|_F^2 = \left\langle\mathbf{A}-\mathbf{UZ},\mathbf{A}-\mathbf{UZ}\right\rangle .$$

In the above expression by substituting $\mathbf{U}$ with $\mathbf{U}+\mathbf{H_1}$ and $\mathbf{Z}$ with $\mathbf{Z}+\mathbf{H_2}$, we obtain

$$\begin{aligned}
&\left\langle\mathbf{A}-(\mathbf{U}+\mathbf{H_1})(\mathbf{Z}+\mathbf{H_2}),\mathbf{A}-(\mathbf{U}+\mathbf{H_1})(\mathbf{Z}+\mathbf{H_2})\right\rangle ,\\
&= \left\langle\mathbf{A}-\mathbf{UZ}-\mathbf{UH_2}-\mathbf{H_1Z}-\mathbf{H_1H_2},\mathbf{A}-\mathbf{UZ}-\mathbf{UH_2}-\mathbf{H_1Z}-\mathbf{H_1H_2}\right\rangle ,\\
&= \left\langle\mathbf{A},\mathbf{A}\right\rangle - \left\langle\mathbf{A},\mathbf{UZ}\right\rangle - \left\langle\mathbf{A},\mathbf{UH_2}\right\rangle - \left\langle\mathbf{A},\mathbf{H_1Z}\right\rangle - \left\langle\mathbf{A},\mathbf{H_1H_2}\right\rangle ,\\
&- \left\langle\mathbf{UZ},\mathbf{A}\right\rangle + \left\langle\mathbf{UZ},\mathbf{UZ}\right\rangle + \left\langle\mathbf{UZ},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{UZ},\mathbf{H_1Z}\right\rangle + \left\langle\mathbf{UZ},\mathbf{H_1H_2}\right\rangle\\
&- \left\langle\mathbf{UH_2},\mathbf{A}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{UZ}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{H_1Z}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{H_1H_2}\right\rangle\\
&- \left\langle\mathbf{H_1Z},\mathbf{A}\right\rangle + \left\langle\mathbf{H_1Z},\mathbf{UZ}\right\rangle + \left\langle\mathbf{H_1Z},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{H_1Z},\mathbf{H_1Z}\right\rangle + \left\langle\mathbf{H_1Z},\mathbf{H_1H_2}\right\rangle\\
&- \left\langle\mathbf{H_1H_2},\mathbf{A}\right\rangle + \left\langle\mathbf{H_1H_2},\mathbf{UZ}\right\rangle + \left\langle\mathbf{H_1H_2},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{H_1H_2},\mathbf{H_1Z}\right\rangle + \left\langle\mathbf{H_1H_2},\mathbf{H_1H_2}\right\rangle .
\end{aligned}$$

Collecting all the first order terms we have

$$\begin{aligned}
&- \left\langle\mathbf{A},\mathbf{UH_2}\right\rangle - \left\langle\mathbf{A},\mathbf{H_1Z}\right\rangle + \left\langle\mathbf{UZ},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{UZ},\mathbf{H_1Z}\right\rangle\\
&- \left\langle\mathbf{UH_2},\mathbf{A}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{UZ}\right\rangle - \left\langle\mathbf{H_1Z},\mathbf{A}\right\rangle + \left\langle\mathbf{H_1Z},\mathbf{UZ}\right\rangle\\
&= - \left\langle\mathbf{A},\mathbf{H_1Z}\right\rangle + \left\langle\mathbf{UZ},\mathbf{H_1Z}\right\rangle - \left\langle\mathbf{H_1Z},\mathbf{A}\right\rangle + \left\langle\mathbf{H_1Z},\mathbf{UZ}\right\rangle\\
&- \left\langle\mathbf{A},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{UZ},\mathbf{UH_2}\right\rangle - \left\langle\mathbf{UH_2},\mathbf{A}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{UZ}\right\rangle ,\\
&= -2\left\langle\mathbf{A},\mathbf{H_1Z}\right\rangle - 2\left\langle\mathbf{A},\mathbf{UH_2}\right\rangle + 2\left\langle\mathbf{UZ},\mathbf{H_1Z}\right\rangle + 2\left\langle\mathbf{UZ},\mathbf{UH_2}\right\rangle ,\\
&= -2\mathbf{tr}((\mathbf{A}-\mathbf{UZ})\mathbf{Z}^T\mathbf{H_1}^T) - 2\mathbf{tr}((\mathbf{A}-\mathbf{UZ})\mathbf{H_2}^T\mathbf{U}^T),\\
&= -2\mathbf{tr}((\mathbf{A}-\mathbf{UZ})\mathbf{Z}^T\mathbf{H_1}^T) - 2\mathbf{tr}(\mathbf{U}^T(\mathbf{A}-\mathbf{UZ})\mathbf{H_2}^T),
\end{aligned}$$

and similarly collecting all the second order terms we have

$$\begin{aligned}
&- \left\langle\mathbf{A},\mathbf{H_1H_2}\right\rangle + \left\langle\mathbf{UZ},\mathbf{H_1H_2}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{UH_2},\mathbf{H_1Z}\right\rangle\\
&+ \left\langle\mathbf{H_1Z},\mathbf{UH_2}\right\rangle + \left\langle\mathbf{H_1Z},\mathbf{H_1Z}\right\rangle - \left\langle\mathbf{H_1H_2},\mathbf{A}\right\rangle + \left\langle\mathbf{H_1H_2},\mathbf{UZ}\right\rangle\\
&= -2\left\langle\mathbf{A}-\mathbf{UZ},\mathbf{H_1H_2}\right\rangle + \left\langle\mathbf{UH_2}+\mathbf{H_1Z},\mathbf{UH_2}+\mathbf{H_1Z}\right\rangle .
\end{aligned}$$

Thus the statement follows using the second order Taylor expansion. □

Lemma F.2. *Given* $h_1 := \left(\frac{\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2}{2}\right)^2$*, then we have the following*

$$\nabla h_1(\mathbf{U},\mathbf{Z}) = \left(\left(\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2\right)\mathbf{U}, \left(\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2\right)\mathbf{Z}\right) ,$$

$$\left\langle(\mathbf{H_1},\mathbf{H_2}), \nabla^2 h_1(\mathbf{U},\mathbf{Z})(\mathbf{H_1},\mathbf{H_2})\right\rangle = (\|\mathbf{H_1}\|_F^2+\|\mathbf{H_2}\|_F^2)(\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2) + 2\left\|\mathbf{H_1}\mathbf{U}^T+\mathbf{Z}\mathbf{H_2}^T\right\|_F^2$$

Proof. By the definition of Forbenius dot product, we have

$$\frac{1}{4}\|\mathbf{U}\|_F^4 + \frac{1}{4}\|\mathbf{Z}\|_F^4 + \frac{1}{2}\|\mathbf{U}\|_F^2\|\mathbf{Z}\|_F^2 = \frac{1}{4}\left\langle\mathbf{U},\mathbf{U}\right\rangle^2 + \frac{1}{4}\left\langle\mathbf{Z},\mathbf{Z}\right\rangle^2 + \frac{1}{2}\left\langle\mathbf{U},\mathbf{U}\right\rangle\left\langle\mathbf{Z},\mathbf{Z}\right\rangle$$

Now, considering $h_1(\mathbf{U}+\mathbf{H_1},\mathbf{Z}+\mathbf{H_2})$ we have

$$\begin{aligned}
&\frac{1}{4}\left\langle \mathbf{U}+\mathbf{H_1},\mathbf{U}+\mathbf{H_1}\right\rangle^2+\frac{1}{4}\left\langle \mathbf{Z}+\mathbf{H_2},\mathbf{Z}+\mathbf{H_2}\right\rangle^2+\frac{1}{2}\left\langle \mathbf{U}+\mathbf{H_1},\mathbf{U}+\mathbf{H_1}\right\rangle\left\langle \mathbf{Z}+\mathbf{H_2},\mathbf{Z}+\mathbf{H_2}\right\rangle\\
&=\frac{1}{4}\left(\left\langle \mathbf{U},\mathbf{U}\right\rangle+2\left\langle \mathbf{H_1},\mathbf{U}\right\rangle+\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\right)^2+\frac{1}{4}\left(\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+2\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\right)^2\\
&+\frac{1}{2}\left(\left\langle \mathbf{U},\mathbf{U}\right\rangle+2\left\langle \mathbf{H_1},\mathbf{U}\right\rangle+\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\right)\left(\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+2\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\right)\\
&=\frac{1}{4}\Big(\left\langle \mathbf{U},\mathbf{U}\right\rangle^2+4\left\langle \mathbf{H_1},\mathbf{U}\right\rangle^2+\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle^2+2\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\left\langle \mathbf{U},\mathbf{U}\right\rangle\\
&+4\left\langle \mathbf{U},\mathbf{U}\right\rangle\left\langle \mathbf{H_1},\mathbf{U}\right\rangle+4\left\langle \mathbf{H_1},\mathbf{U}\right\rangle\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle)\\
&+\frac{1}{4}\Big(\left\langle \mathbf{Z},\mathbf{Z}\right\rangle^2+4\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle^2+\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle^2+2\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle\\
&+4\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+4\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle)\\
&+\frac{1}{2}\left(\left\langle \mathbf{U},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+2\left\langle \mathbf{U},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+\left\langle \mathbf{U},\mathbf{U}\right\rangle\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\right)\\
&+\frac{1}{2}\left(2\left\langle \mathbf{H_1},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+4\left\langle \mathbf{H_1},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+2\left\langle \mathbf{H_1},\mathbf{U}\right\rangle\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\right)\\
&+\frac{1}{2}\left(\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+2\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\right)
\end{aligned}$$

Collecting all the first order terms, we have

$$\left\langle \mathbf{U},\mathbf{U}\right\rangle\left\langle \mathbf{H_1},\mathbf{U}\right\rangle+\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+\left\langle \mathbf{U},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+\left\langle \mathbf{H_1},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle\ ,$$

and similarly collecting all the second order terms we have

$$\begin{aligned}
&\frac{1}{4}\Big(4\left\langle \mathbf{H_1},\mathbf{U}\right\rangle^2+2\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\left\langle \mathbf{U},\mathbf{U}\right\rangle+4\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle^2+2\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle\Big)\\
&+\frac{1}{2}\left(\left\langle \mathbf{U},\mathbf{U}\right\rangle\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle+4\left\langle \mathbf{H_1},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\left\langle \mathbf{Z},\mathbf{Z}\right\rangle\right)\ ,\\
&=\frac{1}{2}\Big(2\left\langle \mathbf{H_1},\mathbf{U}\right\rangle^2+(\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle+\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle)(\left\langle \mathbf{U},\mathbf{U}\right\rangle+\left\langle \mathbf{Z},\mathbf{Z}\right\rangle)\\
&+2\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle^2+4\left\langle \mathbf{H_1},\mathbf{U}\right\rangle\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle\Big)\ ,\\
&=\frac{1}{2}\Big((\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle+\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle)(\left\langle \mathbf{U},\mathbf{U}\right\rangle+\left\langle \mathbf{Z},\mathbf{Z}\right\rangle)+2(\left\langle \mathbf{H_1},\mathbf{U}\right\rangle+\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle)^2\Big)\ .
\end{aligned}$$

Thus the statement follows. ☐

Lemma F.3. *Given* $h_2(\mathbf{U},\mathbf{Z}):=\frac{\|\mathbf{U}\|_F^2+\|\mathbf{Z}\|_F^2}{2}$, *then we have the following*

$$\nabla h_2(\mathbf{U},\mathbf{Z})=(\mathbf{U},\mathbf{Z})\,,$$

$$\left\langle (\mathbf{H_1},\mathbf{H_2}),\nabla^2 h_2(\mathbf{U},\mathbf{Z})(\mathbf{H_1},\mathbf{H_2})\right\rangle=\|\mathbf{H_1}\|_F^2+\|\mathbf{H_2}\|_F^2\ .$$

Proof. Considering $h_2(\mathbf{U}+\mathbf{H_1},\mathbf{Z}+\mathbf{H_2})$, we have

$$\begin{aligned}
&\frac{1}{2}\left\langle \mathbf{U}+\mathbf{H_1},\mathbf{U}+\mathbf{H_1}\right\rangle+\frac{1}{2}\left\langle \mathbf{Z}+\mathbf{H_2},\mathbf{Z}+\mathbf{H_2}\right\rangle\\
&=\frac{1}{2}\left(\left\langle \mathbf{U},\mathbf{U}\right\rangle+2\left\langle \mathbf{U},\mathbf{H_1}\right\rangle+\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle\right)+\frac{1}{2}\left(\left\langle \mathbf{Z},\mathbf{Z}\right\rangle+2\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle+\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\right)\ .
\end{aligned}$$

Collecting all the first order terms we have

$$\left\langle \mathbf{U},\mathbf{H_1}\right\rangle+\left\langle \mathbf{Z},\mathbf{H_2}\right\rangle\ ,$$

and similarly collecting all the second order terms we have

$$\frac{1}{2}\left(\left\langle \mathbf{H_1},\mathbf{H_1}\right\rangle+\left\langle \mathbf{H_2},\mathbf{H_2}\right\rangle\right)\ .$$

Thus the statement holds. ☐

F.1 Proof of Proposition 2.1

Proof. We prove here the convexity of $Lh_a - g$ for a certain constant $L \geq 1$. With Lemma F.1 we obtain

$$\begin{aligned}
&\left\langle (\mathbf{H}_1, \mathbf{H}_2), \nabla^2 g(\mathbf{A}, \mathbf{UZ})(\mathbf{H}_1, \mathbf{H}_2) \right\rangle \\
&= \|\mathbf{H}_1\mathbf{Z} + \mathbf{U}\mathbf{H}_2\|_F^2 - 2\left\langle \mathbf{A} - \mathbf{UZ}, \mathbf{H}_1\mathbf{H}_2 \right\rangle \,, \\
&\leq 2\|\mathbf{H}_1\mathbf{Z}\|_F^2 + 2\|\mathbf{U}\mathbf{H}_2\|_F^2 + 2\|\mathbf{A}\|_F\|\mathbf{H}_1\mathbf{H}_2\|_F + 2\|\mathbf{UZ}\|_F\|\mathbf{H}_1\mathbf{H}_2\|_F \,, \\
&\leq 2\|\mathbf{H}_1\|_F^2\|\mathbf{Z}\|_F^2 + 2\|\mathbf{U}\|_F^2\|\mathbf{H}_2\|_F^2 + 2\|\mathbf{A}\|_F\|\mathbf{H}_1\|_F\|\mathbf{H}_2\|_F + 2\|\mathbf{U}\|_F\|\mathbf{Z}\|_F\|\mathbf{H}_1\|_F\|\mathbf{H}_2\|_F \,.
\end{aligned}$$

With AM-GM inequality, for non-negative real numbers a, b we have $2\sqrt{ab} \leq a + b$, we have

$$2\|\mathbf{U}\|_F\|\mathbf{Z}\|_F\|\mathbf{H}_1\|_F\|\mathbf{H}_2\|_F \leq \|\mathbf{H}_1\|_F^2\|\mathbf{Z}\|_F^2 + \|\mathbf{U}\|_F^2\|\mathbf{H}_2\|_F^2 \,,$$

and similarly we have

$$2\|\mathbf{A}\|_F\|\mathbf{H}_1\|_F\|\mathbf{H}_2\|_F \leq \|\mathbf{A}\|_F\|\mathbf{H}_1\|_F^2 + \|\mathbf{A}\|_F\|\mathbf{H}_2\|_F^2 \,.$$

Using the above two inequalities, we obtain

$$\left\langle (\mathbf{H}_1, \mathbf{H}_2), \nabla^2 g(A, \mathbf{UZ})(\mathbf{H}_1, \mathbf{H}_2) \right\rangle \leq (3\|\mathbf{Z}\|_F^2 + \|\mathbf{A}\|_F)\|\mathbf{H}_1\|_F^2 + (3\|\mathbf{U}\|_F^2 + \|\mathbf{A}\|_F)\|\mathbf{H}_2\|_F^2 \,. \quad \text{(F.1)}$$

Now, considering the kernel generating distances, via Lemma F.2 and F.3 we obtain

$$\begin{aligned}
&\left\langle (\mathbf{H}_1, \mathbf{H}_2), \nabla^2 h_1(\mathbf{U}, \mathbf{Z})(\mathbf{H}_1, \mathbf{H}_2) \right\rangle \\
&= 2\|\mathbf{H}_1\mathbf{U} + \mathbf{H}_2\mathbf{Z}\|_F^2 + (\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2)\|\mathbf{H}_1\|_F^2 + (\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2)\|\mathbf{H}_2\|_F^2 \\
&\geq \|\mathbf{Z}\|_F^2\|\mathbf{H}_1\|_F^2 + \|\mathbf{U}\|_F^2\|\mathbf{H}_2\|_F^2 \,,
\end{aligned}$$

and

$$\left\langle (\mathbf{H}_1, \mathbf{H}_2), \nabla^2 h_2(\mathbf{U}, \mathbf{Z})(\mathbf{H}_1, \mathbf{H}_2) \right\rangle = \|\mathbf{H}_1\|_F^2 + \|\mathbf{H}_2\|_F^2 \,.$$

Now, it is easy to see that

$$\left\langle (\mathbf{H}_1, \mathbf{H}_2), \nabla^2 h_a(\mathbf{U}, \mathbf{Z})(\mathbf{H}_1, \mathbf{H}_2) \right\rangle \geq \left\langle (\mathbf{H}_1, \mathbf{H}_2), \nabla^2 g(\mathbf{A}, \mathbf{UZ})(\mathbf{H}_1, \mathbf{H}_2) \right\rangle \,.$$

A similar proof holds for the convexity of $Lh_a + g$, however the choice of L here need not be the same as it is for $Lh_a - g$ (see [8, Remark 2.1]). ☐

References

[1] P. Ablin, D. Fagot, H. Wendt, A. Gramfort, and C. Févotte. A Quasi-Newton algorithm on the orthogonal manifold for NMF with transform learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 700–704. IEEE, 2019.

[2] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.

[3] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

[4] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

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

[6] J. Bolte, A. Daniilidis, A.S. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[7] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[8] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[9] S. Bonettini, I. Loris, F. Porta, and M. Prato. Variable metric inexact line-search-based methods for nonsmooth optimization. *SIAM journal on optimization*, 26(2):891–921, 2016.

[10] S. Bonettini, I. Loris, F. Porta, M. Prato, and S. Rebegoldi. On the convergence of a linesearch based proximal-gradient method for nonconvex optimization. *Inverse Problems*, 33(5), 2017.

[11] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[12] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1548–1560, 2011.

[13] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20(4):1956–1982, mar 2010.

[14] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[15] Y. Censor and A. Lent. An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34(3):321–353, 1981.

[16] S. Chaudhuri, R. Velmurugan, and R. M. Rameshan. *Blind image deconvolution*. Springer, 2016.

[17] E. Chouzenoux, J. C. Pesquet, and A. Repetti. A block coordinate variable metric forward–backward algorithm. *Journal of Global Optimization*, 66(3):457–485, 2016.

[18] A. Cichocki, R. Zdunek, and S. Amari. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pages 169–176. Springer, 2007.

[19] D. Davis, D. Drusvyatskiy, and K. J. MacPhee. Stochastic model-based minimization under high-order growth. *ArXiv preprint arXiv:1807.00255*, 2018.

[20] R. A. Dragomir, J. Bolte, and A. d'Aspremont. Fast gradient methods for symmetric nonnegative matrix factorization. *ArXiv preprint arXiv:1901.10791*, 2019.

[21] F. Esposito, N. Gillis, and N. D. Buono. Orthogonal joint sparse NMF for microarray data analysis. *Journal of Mathematical Biology*, pages 1–25, 2019.

[22] H. Fang, Z. Zhang, Y. Shao, and C. J. Hsieh. Improved bounded matrix completion for large-scale recommender systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 1654–1660. AAAI Press, 2017.

[23] N. Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257), 2014.

[24] N. Gillis and S. A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):698–714, 2014.

[25] G. H. Golub and C. F.V. Loan. *Matrix computations*, volume 3. JHU press, 2012.

[26] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtarik. SGD: General analysis and improved rates. *ArXiv preprint arXiv:1901.09401*, 2019.

[27] Benjamin David Haeffele and René Vidal. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[28] F. Hanzely, P. Richtarik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *ArXiv preprint arXiv:1808.03045*, 2018.

[29] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19, 2016.

[30] C. J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1064–1072. ACM, 2011.

[31] C. J. Hsieh and P. Olsen. Nuclear norm minimization via active subspace selection. In *International Conference on Machine Learning*, pages 575–583, 2014.

[32] P. Jawanpuria and B. Mishra. A unified framework for structured low-rank matrix learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2254–2263. PMLR, 2018.

[33] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[34] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[35] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[36] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[37] W. Li and D.-Y. Yeung. Relation regularized matrix factorization. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, IJCAI'09, pages 1126–1131, 2009.

[38] C. Lu, M. Yang, F. Luo, F. X. Wu, M. Li, Y. Pan, Y. Li, and J. Wang. Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics*, 34(19):3357–3364, 2018.

[39] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[40] R. Luss and M. Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Review*, 55(1):65–98, 2013.

[41] C. J. Maddison, D. Paulin, Y. W. Teh, and A. Doucet. Dual space preconditioning for gradient descent. *ArXiv preprint arXiv:1902.02257*, 2019.

[42] Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

[43] A. Moitra. An almost optimal algorithm for computing nonnegative rank. *SIAM Journal on Computing*, 45(1):156–173, 2016.

[44] M. C. Mukkamala and M. Hein. Variants of RMSProp and Adagrad with Logarithmic Regret Bounds. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2545–2553, 2017.

[45] M. C. Mukkamala, P. Ochs, T. Pock, and S. Sabach. Convex-Concave Backtracking for Inertial Bregman Proximal Gradient Algorithms in Non-Convex Optimization. *ArXiv preprint arXiv:1904.03537*, 2019.

[46] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

[47] L. M. Nguyen, P. H. Nguyen, M. V. Dijk, P. Richtárik, K. Scheinberg, and M. Takáč. SGD and Hogwild! convergence without the bounded gradients assumption. *ArXiv preprint arXiv:1802.03801*, 2018.

[48] Q. V. Nguyen. Forward–Backward Splitting with Bregman Distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.

[49] P. Ochs. Local convergence of the heavy-ball method and ipiano for non-convex optimization. *Journal of Optimization Theory and Applications*, 177(1):153–180, 2018.

[50] P. Ochs, Y. Chen, T. Brox, and T. Pock. iPiano: inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[51] P. Ochs, J. Fadili, and T. Brox. Non-smooth non-convex Bregman minimization: Unification and new algorithms. *Journal of Optimization Theory and Applications*, 181(1):244–278, 2019.

[52] T. Pock and S. Sabach. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787, 2016.

[53] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Fundamental Principles of Mathematical Sciences*. Springer-Verlag, Berlin, 1998.

[54] S. Sra and I. S. Dhillon. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in neural information processing systems*, pages 283–290, 2006.

[55] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2005.

[56] J.-L. Starck, F. Murtagh, and J. Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge University Press, 2010.

[57] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[58] K. Thung, P. T. Yap, E. Adeli, S. W. Lee, D. Shen, and Alzheimer's Disease Neuroimaging Initiative. Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. *Medical image analysis*, 45:68–82, 2018.

[59] B. Wen, X. Chen, and T. K. Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27(1):124–145, 2017.

[60] Y. Xu, Z. Li, J. Yang, and D. Zhang. A survey of dictionary learning algorithms for face recognition. *IEEE access*, 5:8502–8514, 2017.

[61] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[62] Lei Yang, Ting Kei Pong, and Xiaojun Chen. A nonmonotone alternating updating method for a class of matrix factorization problems. *SIAM Journal on Optimization*, 28(4):3402–3430, 2018.

[63] Q. Yao and J. Kwok. Scalable robust matrix factorization with nonconvex loss. In *Advances in Neural Information Processing Systems*, pages 5061–5070, 2018.

[64] A. W. Yu, W. Ma, Y. Yu, J. Carbonell, and S. Sra. Efficient structured matrix rank minimization. In *Advances in neural information processing systems*, pages 1350–1358, 2014.

[65] X. Zhang, R. Barrio, M. Martinez, H. Jiang, and L. Cheng. Bregman proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *ArXiv preprint arXiv:1904.11295*, 2019.

[66] Z. Zhu, X. Li, K. Liu, and Q. Li. Dropping symmetry for fast symmetric nonnegative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 5154–5164, 2018.