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

Mahesh Chandra Mukkamala[1], Peter Ochs[1].

[1] Mathematical Optimization Group,
Saarland University, Germany.

Contributions

Motivation:

- Use **Bregman Distances** to go beyond Alternating methods.
- To obtain efficient closed form update steps with practical applicability.

Main Contributions:

- **Bregman Proximal Algorithms** for matrix factorization problems.
- **L-smad property is shown** via new Bregman distances.
- Efficient closed form updates for **BPG-MF** and **CoCaIn BPG-MF**, enabling practical applicability on nonconvex nonsmooth problems.
- Empirical comparisons of **BPG methods vs PALM methods**.

Matrix Factorization Problem

For $\mathbf{A} \in \mathbb{R}^{M \times N}$ obtain $\mathbf{U} \in \mathbb{R}^{M \times K}$ and $\mathbf{Z}^{K \times N}$ such that $\mathbf{A} \approx \mathbf{UZ}$?

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

- $\mathcal{R}_1(\mathbf{U}) + \mathcal{R}_2(\mathbf{Z})$ 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.

Standard way: Alternating methods.
Our way: Non-Alternating methods.

Kernel Generating Distance

Definition 1
(Bolte et al. [2018]) 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 $\mathcal{G}(C)$.

Bregman Distance

Let $\mathcal{G}(C)$ be class of standard kernel generating distances. For every $h \in \mathcal{G}(C)$, the associated Bregman distance is given by $D_h : \operatorname{dom} h \times \operatorname{int} \operatorname{dom} h \to \mathbb{R}_+$:

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

- $h_0(x) = \frac{1}{2} \|x\|^2$, (Proximal Gradient)
- $h_1(x) = \frac{1}{4} \|x\|^4 + \frac{1}{2} \|x\|^2$, (Phase retrieval)
- $h_2(x) = \frac{1}{3} \|x\|^3 + \frac{1}{2} \|x\|^2$. (Cubic regularization)

Beyond Lipschitz continuity

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

Definition 2 (L-smad property)

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.

Novel Bregman Distances

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

$$h_2(\mathbf{U}, \mathbf{Z}) := \left(\frac{\|\mathbf{U}\|_F^2 + \|\mathbf{Z}\|_F^2}{2} \right) .$$

Proposition 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

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

Bregman Proximal Mapping

Bregman Proximal Gradient Mapping (Bolte et al. [2018])

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

BPG update, for some $\lambda > 0$ and $h \in \mathcal{G}(C)$

$$\boxed{x^{k+1} \in T_\lambda(x^k)}$$

BPG-MF

Now, let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper lsc function (potentially non-convex) with $\operatorname{dom} f \cap C \neq \emptyset$.

BPG-MF: BPG for Matrix Factorization.

Input. $h \in \mathcal{G}(C)$ with $C \equiv \operatorname{int}\operatorname{dom} h$, g is L-smad w.r.t. 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

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

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

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

CoCaIn BPG: Outline

In Mukkamala et al. [2019] based on Nesterov's Momentum.

CoCaIn BPG: Convex-Concave Inertial BPG.

Step 0: Choose appropriate constants.

Step 1: Compute extrapolated points based on lower minorant.

Step 2: Update similar to BPG based on upper majorant.

Double backtracking for extrapolation and step-size.

CoCaIn BPG incorporates Adaptive Inertia.

CoCaIn BPG-MF: Inertial Step

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

Input. Choose $\delta, \epsilon > 0$ such that $1 > \delta > \epsilon > 0$, $h \in \mathcal{G}(C)$ with $C \equiv \operatorname{int}\operatorname{dom} h$, g is L-smad on C w.r.t. h.

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 $\tau_0 \leq \bar{L}_0^{-1}$.

General Step. For $k = 1, 2, \ldots$, compute extrapolated points

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

where $\gamma_k \geq 0$ such that

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

where $\underline{L}_k$ satisfies $D_g\left(\mathbf{U^k}, \mathbf{Z^k}, Y_{\mathbf{U}}^{\mathbf{k}}, Y_{\mathbf{Z}}^{\mathbf{k}}\right) \geq -\underline{L}_k D_h\left(\mathbf{U^k}, \mathbf{Z^k}, Y_{\mathbf{U}}^{\mathbf{k}}, Y_{\mathbf{Z}}^{\mathbf{k}}\right).$

CoCaIn BPG-MF: Update Step

CoCaIn BPG-MF ...

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

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

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

such that $\bar{L}_k$ satisfies

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

Closed Form Updates for L2-regularization

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

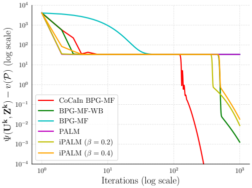
Step 0: Set $h = h_a$ with $c_1 = 3, c_2 = \|\mathbf{A}\|_F$ and $0 < \lambda < 1$.

Step 1: $\mathbf{U^{k+1}} = -r\mathbf{P^k}$, $\mathbf{Z^{k+1}} = -r\mathbf{Q^k}$.

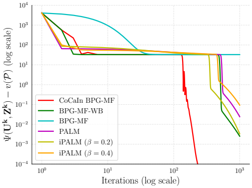
Step 2: $r \geq 0$ with $c_1\left(\left\|-\mathbf{P^k}\right\|_F^2 + \left\|-\mathbf{Q^k}\right\|_F^2\right)r^3 + (c_2+\lambda_0)r - 1 = 0\,.$

Extensions: L1-regularization, Nonnegative constraints etc

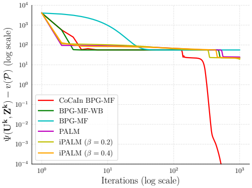
Simple Matrix Factorization Experiments



(a) CIFAR10



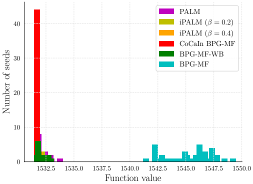
(b) CIFAR10



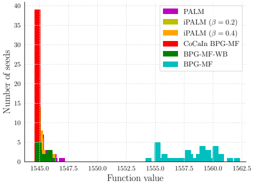
(c) CIFAR10

Simple Matrix Factorization on Synthetic Dataset.

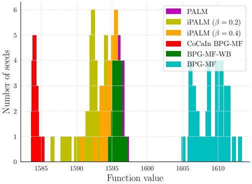
Statistical Evaluation Experiments



(d) CIFAR10



(e) CIFAR10



(f) CIFAR10

Statistical Evaluation on Simple Matrix Factorization.

Other settings

- Nonnegative Matrix Factorization.
- Matrix Completion.
- Graph Regularized NMF.
- Sparse NMF.
- and many more.

Conclusion

CoCaIn BPG-MF is competitive on
various matrix factorization problems.

Open question: Optimal Bregman Distances?

CODE: `github.com/mmahesh`

Thank you . . .

References I

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.

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.