Block MM algorithm on the Grassmann manifold and its applications

Carlos Alejandro Lopez[1] and **Jaume Riba Sagarra**[1]

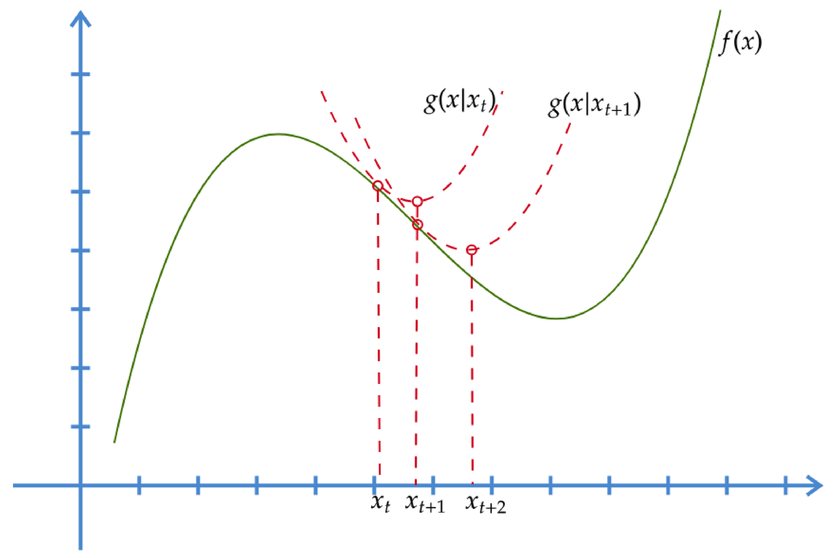
[1]Signal Processing and Communications Group (SPCOM), Universitat Politècnica de Catalunya (UPC)

This work was supported by project MAYTE (PID2022-136512OB-C21 financed by MICIU/AEI/10.13039/501100011033 and by ERDF/EU

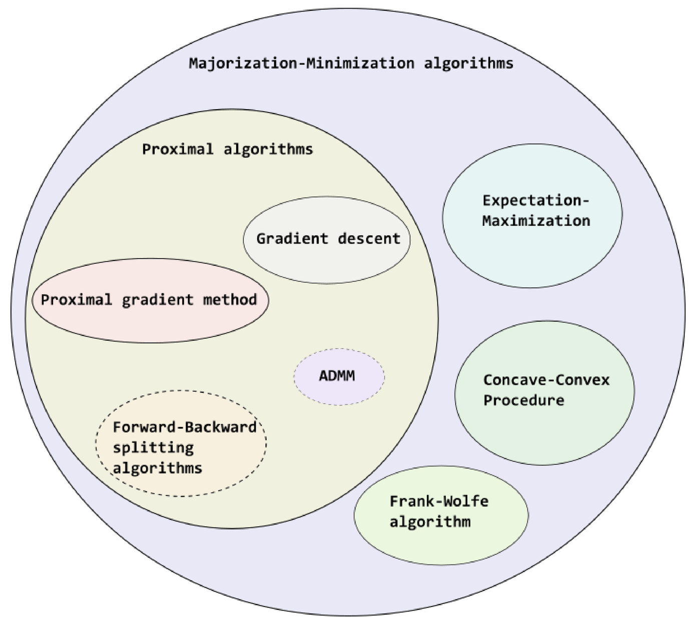
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

The Majorization-Minimization algorithm

The Majorization-Minimization (MM) is a widely used framework that finds the solution of (almost) any optimization problem. It consists on the following principle [1]:



It is the generalization of known optimization frameworks (among others).



Issue: Classical proofs of convergence of the MM algorithm assume that the underlying constraint sets are **convex**.

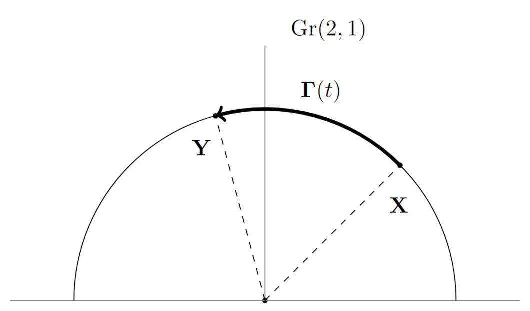
Grassmann manifold and G-convex optimization

Relevant concepts of geometry for optimization:

- The Grassmann manifold, $\text{Gr}(N, D)$, is the set that contains the subspaces of dimension D in $\mathbb{R}^D$ (it is a **non-convex** set):

$$[\mathbf{X}] = \{\mathbf{X}\mathbf{R} \in \mathbb{R}^{N\times D} : \mathbf{X}^T\mathbf{X} = \mathbf{I}_D, \mathbf{R} \in \text{O}(D)\}. \tag{1}$$

- Grassmann geodesics:



- Geodesic quasiconvexity:

$$f(\mathbf{\Gamma}(t)) \leq \max(f(\mathbf{X}), f(\mathbf{Y}))\ \forall \mathbf{X}, \mathbf{Y} \in \mathcal{G}. \tag{2}$$

- Geodesic convexity (it is restrictive in practice [2]):

$$f(\mathbf{\Gamma}(t)) \leq (1-t)f(\mathbf{X}) + tf(\mathbf{Y})\ \forall t \in [0, 1]. \tag{3}$$

Block MM algorithm on the Grassmannian

The block MM algorithm on $\text{Gr}(N, D)$ aims to solve an optimization problem of the following form:

$$\hat{\mathbf{G}}, \hat{\mathbf{c}} = \arg\min_{\mathbf{G},\mathbf{c}} f(\mathbf{G}, \mathbf{c}) \quad \text{s.t. } \mathbf{G} \in \mathcal{G} \subseteq \text{Gr}(N, D), \mathbf{c} \in \mathcal{C} \subseteq \mathbb{R}^M, \tag{4}$$

using update equations that are given by:

$$\mathbf{G}_{i+1} = \arg\min_{\mathbf{G}} g_G(\mathbf{G}|\mathbf{G}_i, \mathbf{c}_i) \quad \mathbf{G} \in \mathcal{G}, \tag{5a}$$

$$\mathbf{c}_{i+1} = \arg\min_{\mathbf{c}} g_c(\mathbf{c}|\mathbf{G}_{i+1}, \mathbf{c}_i) \quad \mathbf{c} \in \mathcal{C}. \tag{5b}$$

The previous setting must satisfy the following assumptions so that the aforementioned update equations converges to a stationary point of the original problem [3]:

(A1) The surrogates must have the same value as the original cost at the current iterate:

$$g_G(\mathbf{G}|\mathbf{G}, \mathbf{c}) = f(\mathbf{G}, \mathbf{c})\ \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C},$$
$$g_c(\mathbf{c}|\mathbf{G}, \mathbf{c}) = f(\mathbf{G}, \mathbf{c})\ \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C}.$$

(A2) The surrogates must majorize the original cost function:

$$g_G(\mathbf{H}|\mathbf{G}, \mathbf{c}) \geq f(\mathbf{G}, \mathbf{c})\ \forall \mathbf{G}, \mathbf{H} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C},$$
$$g_c(\mathbf{d}|\mathbf{G}, \mathbf{c}) \geq f(\mathbf{G}, \mathbf{c})\ \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c}, \mathbf{d} \in \mathcal{C}.$$

(A3) The first (lower) directional derivatives of the surrogates and of the original cost must agree:

$$g'_G(\mathbf{G}|\mathbf{G}, \mathbf{c}; \mathbf{\Delta}) = f'(\mathbf{G}, \mathbf{c}; \mathbf{\Delta}, \mathbf{0}),$$

for all tangent directions $\mathbf{\Delta} \in \mathcal{T}_{\mathbf{G}}\text{Gr}(N, D)$ whose resulting geodesic remains on $\mathcal{G}$ and:

$$g'_c(\mathbf{c}|\mathbf{G}, \mathbf{c}; \boldsymbol{\delta}) = f'(\mathbf{G}, \mathbf{c}; \mathbf{0}, \boldsymbol{\delta}) \quad \text{s.t. } \mathbf{c} + \boldsymbol{\delta} \in \mathcal{C}.$$

(A4) $g_G(\cdot|\cdot)$ and $g_c(\cdot|\cdot)$ must be continuous on its input arguments and must have unique minimizers.

(A5) $g_G(\mathbf{G}|\mathbf{G}, \mathbf{c})$ must be geodesically quasiconvex on $\mathcal{G}$ and $g_c(\mathbf{c}|\mathbf{G}, \mathbf{c})$ must be quasiconvex on $\mathcal{C}$.

(A6) $f(\mathbf{G}, \mathbf{c})$ must have a set of stationary points and must be regular in $\mathcal{G} \times \mathcal{C}$.

The kind of convergence (local or global optimum) depends on the cardinality of the set of stationary points of (4). In practice, one can only expect **local** optimums in the block MM algorithm on the Grassmannian.

Example 1: Blind sparse deconvolution

The original blind sparse deconvolution problem is based on the following optimization problem [4]:

$$\min_{\mathbf{a},\mathbf{x}} ||\mathbf{y} - \mathbf{a} \circledast \mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1 \quad \text{s.t. } ||\mathbf{a}||_2^2 = 1. \tag{6}$$

Note that the previous expression has a **sign ambiguity**. This means that it can be further rewritten as:

$$\min_{\mathbf{a},\mathbf{x}} ||\mathbf{y} - \mathbf{a} \circledast \mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1 \quad \text{s.t. } \mathbf{a} \in \text{Gr}(N, 1), \tag{7}$$

which does not have the sign ambiguity. One way to solve the previous optimization problem is by means of the following update equations [4] ($\psi(\mathbf{a}, \mathbf{x}) = ||\mathbf{y} - \mathbf{a} \circledast \mathbf{x}||_2^2$):

$$\mathbf{x}_{i+1} = \text{prox}_{\ell_1,\lambda\eta}(\mathbf{x}_i - \eta\nabla_{\mathbf{x}}\psi(\mathbf{a}_i, \mathbf{x}_i)), \tag{8a}$$

$$\mathbf{a}_{i+1} = \exp_{\mathbf{a}}\left(\text{grad}_{\mathbf{a}}\psi(\mathbf{a}_i, \mathbf{x}_{i+1})\right), \tag{8b}$$

where:

$$\exp_{\mathbf{a}}(\mathbf{g}) = \boldsymbol{\gamma}(1), \tag{9}$$

is the exponential mapping at $\mathbf{a}$ of the Grassmann manifold. The update equations in (8) is the block extension of two optimization frameworks: the Proximal gradient method for $\mathbf{x}$ and the Riemannian gradient descent for $\mathbf{a}$.

Example 2: Minimum error entropy criterion for the blind fusion and regression problem

Consider K realizations of a block model that gathers N measurements of N sensors:

$$\mathbf{Y}_k = \mathbf{x}_k\mathbf{1}^T + \mathbf{W}_k \quad k = 1, ..., K, \tag{10}$$

where:

$$\mathbf{x} = \mathbf{B}\mathbf{u}. \tag{11}$$

After doing some magic (!) using the Conditional Maximum Likelihood principle and introducing the Arithmetic Average fusion of the measurements, $\mathbf{g}_k = \mathbf{Y}_k\mathbf{f}$, the previous problem can be solved by minimizing:

$$\hat{\mathbf{f}}, \hat{\mathbf{H}} = \arg\min_{\mathbf{f},\mathbf{H}} \log(\det(\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}))) \quad \text{s.t. } \mathbf{f}^T\mathbf{1} = 1,\ \mathbf{H} \in \text{Gr}(N, D), \tag{12}$$

where $\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}) = \frac{1}{K}\sum_{k=1}^K \hat{\mathbf{C}}_k(\mathbf{H}, \mathbf{f})$ and:

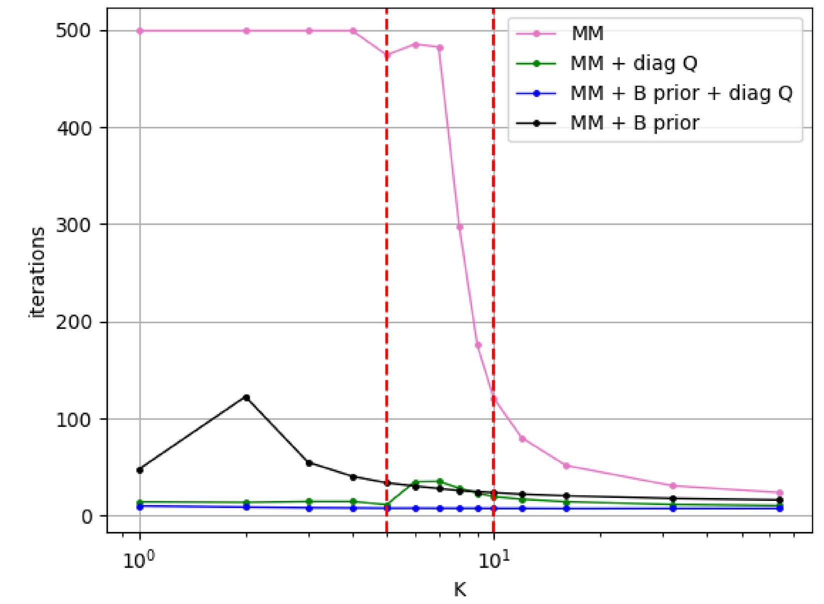
$$\hat{\mathbf{C}}_k(\mathbf{B}, \mathbf{f}) = \frac{1}{N}(\mathbf{Y}_k - \mathbf{P}_H\mathbf{Y}_k\mathbf{f}\mathbf{1}^T)^T(\mathbf{Y}_k - \mathbf{P}_H\mathbf{Y}_k\mathbf{f}\mathbf{1}^T). \tag{13}$$

Notice that the log-determinant function is majorized by its Taylor series expansion:

$$\ell(\mathbf{H}, \mathbf{f}) \leq \log(\det(\mathbf{Z}_k^{-1})) + \text{tr}\left(\mathbf{Z}_k\left(\hat{\mathbf{Q}}_{ML}(\mathbf{H}, \mathbf{f}) - \hat{\mathbf{Q}}_{ML}(\mathbf{H}_k, \mathbf{f}_k)\right)\right), \tag{14}$$

where $\mathbf{Z}_k = \left(\hat{\mathbf{Q}}_{ML}(\mathbf{H}_k, \mathbf{f}_k)\right)^{-1}$.

Numerical results of the blind fusion and regression problem



References

[1] Y. Sun, P. Babu and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," in IEEE Transactions on Signal Processing, vol. 65, no. 3, pp. 794-816, 1 Feb.1, 2017, doi: 10.1109/TSP.2016.2601299.

[2] Chen, Shixiang, Ma, Shiqian, So, Anthony, Zhang, Tong. (2020). Proximal Gradient Method for Nonsmooth Optimization over the Stiefel Manifold. SIAM Journal on Optimization. 30. 210-239. 10.1137/18M122457X.

[3] C. A. Lopez and J. Riba, "On the Convergence of Block Majorization-Minimization Algorithms on the Grassmann Manifold," in IEEE Signal Processing Letters, vol. 31, pp. 1314-1318, 2024, doi: 10.1109/LSP.2024.3396660.

[4] Y. Lau, Q. Qu, H. Kuo, P. Zhou, Y. Zhang, and J. Wright, "Short and sparse deconvolution - A geometric approach," in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020.