Supplemental Material for "On the Convergence of a Bayesian Algorithm for Joint Dictionary Learning and Sparse Recovery"

Geethu Joseph and Chandra R. Murthy *Senior Member, IEEE*

I. Derivation of DL-SBL Algorithm

In this section, we provide the details of the EM-algorithm development, explaining how to obtain (3)-(8), and the γ_k update equations in Algorithm 1 and Algorithm 2. The EM algorithm computes the unknown parameter set $\boldsymbol{\Lambda}$ by minimizing the negative log likelihood $-\log p(\boldsymbol{y}^K;\boldsymbol{\Lambda})$. To compute the likelihood, we first note that the SBL framework imposes a Gaussian prior on the unknown vector $\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Gamma}_k)$, where $\boldsymbol{\Gamma}_k$ is an unknown diagonal matrix.. Thus, $\boldsymbol{y}_k$ also follows a Gaussian distribution: $\boldsymbol{y}_k \sim \mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}})$ because the noise term $\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I})$. Therefore, we have

$$p(\boldsymbol{y}^K;\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \frac{1}{\sqrt{(2\pi)^m \left|\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right|}} \times \exp\left(-\frac{1}{2}\boldsymbol{y}_k^{\mathsf{T}}\left(\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right)^{-1}\boldsymbol{y}_k\right). \quad (97)$$

Hence, the negative log likelihood is computed as follows:

$$-\log p(\boldsymbol{y}^K;\boldsymbol{\Lambda}) = \frac{1}{2}\sum_{k=1}^{K}\left[m\log(2\pi)+\log\left|\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right| + \frac{1}{2}\boldsymbol{y}_k^{\mathsf{T}}\left(\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right)^{-1}\boldsymbol{y}_k\right]. \quad (98)$$

Since the $\log(2\pi)$ term is a constant independent of $\boldsymbol{\Lambda}$, we omit that term and the scaling factor of $\frac{1}{2}$ to obtain the cost function $T(\boldsymbol{\Lambda})$ in (3).

The EM algorithm treats the unknowns $\boldsymbol{x}^K$ as the hidden data and the observations $\boldsymbol{y}^K$ as the known data. It is an iterative procedure which updates the estimate of the parameters $\boldsymbol{\Lambda}$ in every iteration using two steps: an expectation step (E-step) and a maximization step (M-step). Let $\boldsymbol{\Lambda}^{(r)}$ be the estimate of $\boldsymbol{\Lambda}$ at the r^{th} iteration. The E-step computes the marginal log-likelihood of the observed data $Q\left(\boldsymbol{\Lambda};\boldsymbol{\Lambda}^{(r-1)}\right)$, and the M-step computes the parameter tuple $\boldsymbol{\Lambda}$ that maximizes $Q\left(\boldsymbol{\Lambda};\boldsymbol{\Lambda}^{(r-1)}\right)$.

$$\textbf{E-step: } Q\left(\boldsymbol{\Lambda};\boldsymbol{\Lambda}^{(r-1)}\right) = \mathbb{E}_{\boldsymbol{x}^K|\boldsymbol{y}^K;\boldsymbol{\Lambda}^{(r-1)}}\left\{\log p\left(\boldsymbol{y}^K,\boldsymbol{x}^K;\boldsymbol{\Lambda}\right)\right\}$$
$$\textbf{M-step: } \boldsymbol{\Lambda}^{(r)} = \underset{\boldsymbol{\Lambda}\in\mathbb{O}\times\mathbb{R}_+^{NK}}{\arg\max}\; Q\left(\boldsymbol{\Lambda};\boldsymbol{\Lambda}^{(r-1)}\right). \quad (99)$$

The authors are with the Dept. of ECE at IISc, Bangalore, India, Emails:{geethu, cmurthy}@iisc.ac.in.

To simplify $Q\left(\boldsymbol{\Lambda},\boldsymbol{\Lambda}^{(r-1)}\right)$, we note that

$$p\left(\boldsymbol{y}^K,\boldsymbol{x}^K;\boldsymbol{\Lambda}\right) = \prod_{k=1}^{K} p\left(\boldsymbol{y}_k|\boldsymbol{x}_k;\boldsymbol{\Lambda}\right)p\left(\boldsymbol{x}_k;\boldsymbol{\Lambda}\right). \quad (100)$$

Here, $p\left(\boldsymbol{y}_k|\boldsymbol{x}_k;\boldsymbol{\Lambda}\right) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_k,\sigma^2\boldsymbol{I})$, and $p\left(\boldsymbol{x}_k;\boldsymbol{\Lambda}\right) = \mathcal{N}(\boldsymbol{0},\boldsymbol{\Gamma}_k)$. Thus, we get,

$$\begin{aligned}\log p\left(\boldsymbol{y}^K,\boldsymbol{x}^K;\boldsymbol{\Lambda}\right) &= \log\left\{\prod_{k=1}^{K}\frac{1}{\sqrt{(2\pi\sigma)^{2m}}}\exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k\right\|^2\right) \times \frac{1}{\sqrt{(2\pi)^N\left|\boldsymbol{\Gamma}_k\right|}}\exp\left(-\frac{1}{2}\boldsymbol{x}_k^{\mathsf{T}}\boldsymbol{\Gamma}_k^{-1}\boldsymbol{x}_k\right)\right\} \quad (101)\\ &= -\frac{Km}{2}\log((2\pi)^{N+1}\sigma^2) - \frac{1}{2}\sum_{k=1}^{K}\left[\log\left|\boldsymbol{\Gamma}_k\right| + \mathsf{Tr}\left\{\boldsymbol{\Gamma}_k^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^{\mathsf{T}}\right\}\right]\\ &\quad - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\left(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k\right)^{\mathsf{T}}\left(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k\right). \quad (102)\end{aligned}$$

Therefore, eliminating the constant terms, we obtain (5) as follows:

$$\begin{aligned}Q\left(\boldsymbol{\Lambda};\boldsymbol{\Lambda}^{(r-1)}\right) = &-\frac{1}{2}\sum_{k=1}^{K}\left[\log\left|\boldsymbol{\Gamma}_k\right| + \mathsf{Tr}\left\{\boldsymbol{\Gamma}_k^{-1}\mathbb{E}\left\{\boldsymbol{x}_k\boldsymbol{x}_k^{\mathsf{T}}|\boldsymbol{y}^K;\boldsymbol{\Lambda}^{(r-1)}\right\}\right\}\right]\\ &-\frac{1}{2\sigma^2}\sum_{k=1}^{K}\mathbb{E}\left\{\left(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k\right)^{\mathsf{T}}\left(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k\right)|\boldsymbol{y}^K;\boldsymbol{\Lambda}^{(r-1)}\right\}.\end{aligned} \quad (103)$$

We notice that the expectation terms in the above expression depend only on $\boldsymbol{\Lambda}^{(r-1)}$, and are independent of $\boldsymbol{\Lambda}$. Thus, the dependence of $\boldsymbol{\Gamma}_k$ in $Q\left(\boldsymbol{\Lambda};\boldsymbol{\Lambda}^{(r-1)}\right)$ is only through the k^{th} term in the first summation, and the dependence on $\boldsymbol{A}$ is only through the last summation term. Therefore, the optimization in the M-step is separable in its variables $\boldsymbol{\Gamma}_k$ and $\boldsymbol{A}$. Hence, the M-step reduces as follows:

$$\gamma_k^{(r)} = \underset{\boldsymbol{\gamma}\in\mathbb{R}_+^N}{\arg\min}\log\left|\boldsymbol{\Gamma}_k\right| + \mathsf{Tr}\left\{\boldsymbol{\Gamma}_k^{-1}\mathbb{E}\left\{\boldsymbol{x}_k\boldsymbol{x}_k^{\mathsf{T}}|\boldsymbol{y}^k;\boldsymbol{\Lambda}^{(r-1)}\right\}\right\} \quad (104)$$

$$\boldsymbol{A}^{(r)} = \underset{\boldsymbol{A}\in\mathbb{O}}{\arg\min}\sum_{k=1}^{K}\mathbb{E}\left\{\left(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k\right)^{\mathsf{T}}\left(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k\right)|\boldsymbol{y}^k;\boldsymbol{\Lambda}^{(r-1)}\right\}. \quad (105)$$

Here, we note that (105) is same as (7). Further, differentiating the objective function, we get the update equation (6):

$$\gamma_k^{(r)} = \mathsf{Diag}\left\{\mathbb{E}\left\{\boldsymbol{x}_k\boldsymbol{x}_k^{\mathsf{T}}|\boldsymbol{y}^k;\boldsymbol{\Lambda}^{(r-1)}\right\}\right\} \tag{106}$$

$$= \mathsf{Diag}\left\{\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathsf{T}} + \boldsymbol{\Sigma}_{(k)}\right\}, \tag{107}$$

where we use the following facts:

$$\boldsymbol{\mu}_k \triangleq \mathbb{E}\left\{\boldsymbol{x}_k|\boldsymbol{y}_k;\boldsymbol{\Lambda}^{(r-1)}\right\} \tag{108}$$

$$\boldsymbol{\Sigma}_{(k)} \triangleq \mathbb{E}\left\{(\boldsymbol{x}_k-\boldsymbol{\mu}_k)(\boldsymbol{x}_k-\boldsymbol{\mu}_k)^{\mathsf{T}}|\boldsymbol{y}_k;\boldsymbol{\Lambda}^{(r-1)}\right\} \tag{109}$$

$$= \mathrm{cov}\left\{\boldsymbol{x}_k|\boldsymbol{y}^K;\boldsymbol{\Lambda}^{(r-1)}\right\}. \tag{110}$$

Next, we compute the conditional expectations terms needed to find $\gamma_k^{(r)}$. We start with the following cross-covariance matrix:

$$\begin{aligned}\mathbb{E}\left\{\boldsymbol{y}_k\boldsymbol{x}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\} &= \mathbb{E}\left\{(\boldsymbol{A}\boldsymbol{x}_k+\boldsymbol{w}_k)\boldsymbol{x}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\}\\ &= \mathbb{E}\left\{\boldsymbol{A}\boldsymbol{x}_k\boldsymbol{x}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\}\\ &= \boldsymbol{A}\boldsymbol{\Gamma}_k.\end{aligned} \tag{111}$$

Thus, the conditional mean and covariance are given as follows:

$$\begin{aligned}&\mathrm{cov}\left\{\boldsymbol{x}_k|\boldsymbol{y}^K;\boldsymbol{\Lambda}\right\}\\ &= \mathbb{E}\left\{\boldsymbol{x}_k\boldsymbol{x}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\} - \mathbb{E}\left\{\boldsymbol{x}_k\boldsymbol{y}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\}\\ &\qquad\times\mathbb{E}\left\{\boldsymbol{y}_k\boldsymbol{y}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\}^{-1}\mathbb{E}\left\{\boldsymbol{y}_k\boldsymbol{x}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\}\\ &= \boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\left(\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right)^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_k \end{aligned} \tag{112}$$

$$\begin{aligned}&\mathbb{E}\left\{\boldsymbol{x}_k|\boldsymbol{y}^K;\boldsymbol{\Lambda}\right\}\\ &= \mathbb{E}\left\{\boldsymbol{x}_k|\gamma_k,\sigma^2\right\} + \mathbb{E}\left\{\boldsymbol{x}_k\boldsymbol{y}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\}\\ &\quad\times\mathbb{E}\left\{\boldsymbol{y}_k\boldsymbol{y}_k^{\mathsf{T}}|\gamma_k,\sigma^2\right\}^{-1}\left(\boldsymbol{y}_k-\mathbb{E}\left\{\boldsymbol{y}_k|\gamma_k,\sigma^2\right\}\right)\\ &= \boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\left(\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right)^{-1}\boldsymbol{y}_k\\ &= \sigma^{-2}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\left(\boldsymbol{I}-\left(\sigma^2\boldsymbol{I}+\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right)^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_k\boldsymbol{A}^{\mathsf{T}}\right)\boldsymbol{y}_k\\ &= \sigma^{-2}\mathrm{cov}\left\{\boldsymbol{x}_k|\boldsymbol{y}^K;\boldsymbol{\Lambda}\right\}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{y}_k.\end{aligned} \tag{113}$$

Therefore, (106), (116) and (121) together gives the update step for $\boldsymbol{\gamma}_k$ used in Algorithm 1 and Algorithm 2.

Similarly, the optimization problem corresponding the dictionary update (105) reduces as follows:

$$\underset{\boldsymbol{A}\in\mathbb{O}}{\arg\min}\ \sum_{k=1}^{K}\mathbb{E}\left\{(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k)^{\mathsf{T}}(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k)\middle|\,\boldsymbol{y}_k;\boldsymbol{\Lambda}^{(r-1)}\right\} \tag{114}$$

$$\begin{aligned}&= \underset{\boldsymbol{A}\in\mathbb{O}}{\arg\min}\ \sum_{k=1}^{K}\mathbb{E}\left\{-\boldsymbol{y}_k^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}_k+\frac{1}{2}\boldsymbol{x}_k^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}_k\middle|\,\boldsymbol{y}_k;\boldsymbol{\Lambda}^{(r-1)}\right\}\\ &= \underset{\boldsymbol{A}\in\mathbb{O}}{\arg\min}\ -\mathsf{Tr}\left\{\left(\sum_{k=1}^{K}\boldsymbol{\mu}_k\boldsymbol{y}_k^{\mathsf{T}}\right)\boldsymbol{A}+\frac{1}{2}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}\right\}\\ &= \underset{\boldsymbol{A}\in\mathbb{O}}{\arg\min}\ \mathsf{Tr}\left\{-\boldsymbol{M}\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{A}+\frac{1}{2}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}\right\}.\end{aligned} \tag{115}$$

Since $\boldsymbol{A}\in\mathbb{O}$, we can further simplify the second term here as follows:

$$\mathsf{Tr}\left\{\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}\right\} = \sum_{i,j=1;i\neq j}^{N}\boldsymbol{\Sigma}[i,j]\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{A}_j + \sum_{i=1}^{N}\boldsymbol{\Sigma}[i,i]\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{A}_i \tag{116}$$

$$= \mathsf{Tr}\left\{\boldsymbol{A}\left(\boldsymbol{\Sigma}-\mathcal{D}\left\{\boldsymbol{\Sigma}\right\}\right)\boldsymbol{A}^{\mathsf{T}}\right\} + \sum_{i=1}^{N}\boldsymbol{\Sigma}[i,i]. \tag{117}$$

Here, the second term does not depend on $\boldsymbol{A}$, and hence, we remove the term from the objective function to get an equivalent optimization objective function as in (8). Thus, the derivation of algorithm development given by (3)-(8), and the update equations for $\boldsymbol{\gamma}_k$ in Algorithm 1 and Algorithm 2 are completed.

Learning the noise variance

Following a similar approach as the above, we can learn the noise variance σ^2 along with the dictionary $\boldsymbol{A}$ and covariance matrices $\boldsymbol{\Gamma}_k$. If σ^2 is unknown, we have to incorporate its update to the M-step by maximizing the Q function defined in (103). Thus, considering the terms that depend on $\sigma 62$, we get

$$\begin{aligned}\left(\sigma^2\right)^{(r)} &= \underset{\sigma^2\in\mathbb{R}_+}{\arg\min}\ Km\log(\sigma^2)\\ &\quad+\frac{1}{\sigma^2}\sum_{k=1}^{K}\mathbb{E}\left\{(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k)^{\mathsf{T}}(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k)\,|\boldsymbol{y}^K;\boldsymbol{\Lambda}^{(r-1)}\right\}\\ &= \frac{1}{Km}\sum_{k=1}^{K}\mathbb{E}\left\{(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k)^{\mathsf{T}}(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{x}_k)\,|\boldsymbol{y}^K;\boldsymbol{\Lambda}^{(r-1)}\right\}\\ &= \frac{1}{Km}\mathsf{Tr}\left\{\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y}-2\boldsymbol{M}\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{A}+\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}\right\},\end{aligned} \tag{118}$$

where the last step follows because of the same arguments used to derive (125) from (122).

II. Proof of Kurdyka-Łojasiewicz property based Convergence Result

Theorem 6. *A bounded sequence of iterates $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$ generated by the ALS algorithm converges to a stationary point of $\tilde{g}$ if the following four conditions hold:*

(i) The objective function $\tilde{g}(\boldsymbol{A})$ satisfies

$$\inf_{\boldsymbol{A}\in\mathbb{R}^{m\times N}}\tilde{g}(\boldsymbol{A}) > -\infty. \tag{119}$$

(ii) There exist constants $\theta\in[0,1)$, $C,\epsilon>0$ such that

$$|\tilde{g}(\boldsymbol{A})-\tilde{g}(\boldsymbol{A}^*)|^{\theta} \leq C\,\|\boldsymbol{Z}\| \tag{120}$$

for any stationary point $\boldsymbol{A}^*$ of $\tilde{g}$, any $\boldsymbol{A}$ such that $\|\boldsymbol{A}-\boldsymbol{A}^*\|\leq\epsilon$, and any $\boldsymbol{Z}$ such that $\boldsymbol{Z}\in\partial g(\boldsymbol{A})$. The constant θ is called the Łojasiewicz exponent *of the Łojasiewicz gradient inequality.*

(iii) There exists $C_1>0$ such that

$$\tilde{g}\left(\boldsymbol{A}^{(r,u-1)}\right)-\tilde{g}\left(\boldsymbol{A}^{(r,u)}\right) \geq C_1\left\|\boldsymbol{A}^{(r,u-1)}-\boldsymbol{A}^{(r,u)}\right\|^2 \tag{121}$$

(iv) There exist $u_0 > 1$, $C_2 > 0$ and $\boldsymbol{Z} \in \partial g\left(\boldsymbol{A}^{(r,u)}\right)$ such that for all $u > u_0$

$$\|\boldsymbol{Z}\| \leq C_2 \left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u)}\right\|. \tag{122}$$

The proof is adapted from the proof of [40, Theorem 2]. At a high level, there are four steps to the proof:

A We first prove that the sequence $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$ converges to a bounded connected set $\mathbb{G} \subseteq \text{crit}(\tilde{g}) \subseteq \mathbb{O}$, where $\text{crit}(\tilde{g})$ is the set of stationary points of $\tilde{g}$. Moreover, $\tilde{g}$ is constant over the set $\mathbb{G}$.

B Next, we connect the above result to Condition (ii). To establish the connection, we define a new function $\bar{g} : \mathbb{O} \to \mathbb{R}_+$ as $\bar{g}(\boldsymbol{A}) \triangleq \tilde{g}(\boldsymbol{A}) - \tilde{g}(\boldsymbol{A}^{(r)})$, where $\boldsymbol{A}^{(r)}$ is a limit point of the sequence $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$, and $\boldsymbol{A}$ is any point in the set $\mathbb{O}$. We note that the definition of $\bar{g}$ is unambiguous because Step A shows that $\tilde{g}$ is constant over the set $\mathbb{G}$. We then show that there exists a positive integer $U_0 \in \mathbb{N}$ and $\tilde{C} > 0$ such that for all $u \geq U_0$,

$$\left(\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right)^{\theta} \geq \tilde{C}\,\|\boldsymbol{Z}\|, \tag{123}$$

for any $\boldsymbol{Z}$ such that $\boldsymbol{Z} \in \partial\tilde{g}\left(\boldsymbol{A}^{(r,u)}\right)$.

C Finally, using the above relation and other conditions of the theorem, we show that the desired result follows.

Next, we present the details of the above steps:

A. Characterization of $\mathbb{G}$

From Condition (iii), we get that

$$\sum_{u=1}^{\infty}\left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u)}\right\|^2 \leq \frac{1}{C_1}\left[\lim_{u\to\infty}\tilde{g}\left(\boldsymbol{A}^{(r,u-1)}\right) - \tilde{g}\left(\boldsymbol{A}^{(r,0)}\right)\right] < \infty, \tag{124}$$

where the last step follows because $\lim_{u\to\infty}\tilde{g}\left(\boldsymbol{A}^{(r,u-1)}\right) < \infty$ due to Proposition 1. Further, [45, Theorem 1] states that the set of subsequential limit points of a sequence $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$ in a compact metric space is a connected set if it satisfies the following:

$$\sum_{u=1}^{\infty}\left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u)}\right\|^2 < \infty. \tag{125}$$

Consequently, the result applies to any bounded sequence satisfying (137). Since the sequence $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$ generated by the AM procedure belongs to the bounded set $\mathbb{O}$, it converges to a bounded connected set $\mathbb{G} \subseteq \mathbb{O}$. Also, since the set of subsequential limits is closed, $\mathbb{G}$ is a connected compact set.

Now, for any limit point $\boldsymbol{A}^{(r)} \in \mathbb{G}$ of the sequence $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$, there exists a sequence $\{u_j\}_{j\in\mathbb{N}}$ of natural numbers such that $\left\{\left(\boldsymbol{A}^{(r,u_j)}, \boldsymbol{Z}^{(r,u_j)}, \tilde{g}\left(\boldsymbol{A}^{(r,u_j)}\right)\right)\right\}_{j\in\mathbb{N}}$ converges to the tuple $\left(\boldsymbol{A}^{(r)}, \boldsymbol{0}, \tilde{g}\left(\boldsymbol{A}^{(r)}\right)\right)$. This is because the subsequence $\left\{\left(\boldsymbol{Z}^{(r,u_j)}, \tilde{g}\left(\boldsymbol{A}^{(r,u_j)}\right)\right)\right\}_{j\in\mathbb{N}}$ converges to the same limit point as that of the sequence $\left\{\left(\boldsymbol{Z}^{(r,u)}, \tilde{g}\left(\boldsymbol{A}^{(r,u)}\right)\right)\right\}_{u\in\mathbb{N}}$ which is $\left(\boldsymbol{0}, \tilde{g}\left(\boldsymbol{A}^{(r)}\right)\right)$ due to (13) and Proposition 1. Therefore, we conclude that $\mathbb{G} \subset \text{crit}(\tilde{g})$ and $\tilde{g}$ is constant over the set $\mathbb{G}$, completing Step A.

B. Connection to Kurdyka-Łojasiewicz property

The compact set $\mathbb{G}$ can be covered with finite number of closed balls $\mathcal{B}_j = \left\{\boldsymbol{A} \in \mathbb{O} : \left\|\boldsymbol{A} - \boldsymbol{A}^{*(j)}\right\| \leq \epsilon_j\right\}$ such that Condition (ii) is satisfied by $\boldsymbol{A}^{(r,j)}$ with constants $C^{(j)}$ and $\epsilon_j > 0$. Therefore, we have the following relation for $\boldsymbol{A} \in \mathcal{B}_j$:

$$\left|\tilde{g}(\boldsymbol{A}) - \tilde{g}\left(\boldsymbol{A}^{*(j)}\right)\right|^{\theta_j} \leq C^{(j)}\,\|\boldsymbol{Z}\|, \tag{126}$$

for some θ_j and any $\boldsymbol{Z}$ such that $\boldsymbol{Z} \in \partial\tilde{g}(\boldsymbol{A})$. Setting $\epsilon = \min_j \epsilon_j$, $\tilde{C} = \max_j C^{(j)}$, and $\theta = \max_j \theta_j$ we get the following:

$$\left|\tilde{g}(\boldsymbol{A}) - \tilde{g}(\boldsymbol{A}^*)\right|^{\theta} \leq \tilde{C}\,\|\boldsymbol{Z}\|, \tag{127}$$

for any $\boldsymbol{A}^* \in \mathbb{G}$ of $\tilde{g}$, any $\boldsymbol{A}$ such that $\|\boldsymbol{A} - \mathbb{G}\| \leq \epsilon$, and any $\boldsymbol{Z}$ such that $\boldsymbol{Z} \in \partial\tilde{g}(\boldsymbol{A})$. Further, since $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$ converges to $\mathbb{G}$, for any $\epsilon > 0$, there exists a positive integer U_0 such that for all $u \geq U_0$, we have $\left\|\boldsymbol{A}^{(r,u)} - \mathbb{G}\right\| \leq \epsilon$. Therefore, for all $u \geq U_0$,

$$\left|\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right|^{\theta} = \left|\tilde{g}\left(\boldsymbol{A}^{(r,u)}\right) - \tilde{g}\left(\boldsymbol{A}^{(r)}\right)\right|^{\theta} \leq \tilde{C}\,\|\boldsymbol{Z}\|. \tag{128}$$

Thus, Step B is completed.

C. Convergence to a single point

Since $\left\{\tilde{g}\left(\boldsymbol{A}^{(r,u)}\right)\right\}_{u\in\mathbb{N}}$ is a non-increasing sequence, we have $\bar{g}\left(\boldsymbol{A}^{(r,u)}\right) \geq 0$, and the following relation holds.

$$\lim_{u\to\infty}\bar{g}\left(\boldsymbol{A}^{(r,u)}\right) = 0. \tag{129}$$

We first note that the function $h : \mathbb{R}_+ \to \mathbb{R}$ defined as $h(s) = -s^{1-\theta}$ is convex for all $0 \leq \theta \leq 1$. Thus, for all $u \in \mathbb{N}$ and for θ in Condition (ii), it holds that

$$\begin{aligned}
&\left[\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right)\right]^{1-\theta} - \left[\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right]^{1-\theta} \\
&\quad = h\left(\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right)\right) - h\left(\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right) && (130)\\
&\quad \geq \left.\frac{dh(s)}{ds}\right|_{s=\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right)}\left[\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right) - \bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right] && (131)\\
&\quad = (1-\theta)\left[\bar{g}(\boldsymbol{A}^{(r,u-1)}\right]^{-\theta}\left[\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right) - \bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right] && (132)\\
&\quad \geq C_1(1-\theta)\left[\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right]^{-\theta}\left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u)}\right\|^2, && (133)
\end{aligned}$$

where we use Condition (iii) to obtain the last relation. Further, from Step B, we get that

$$\left[\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right)\right]^{1-\theta} - \left[\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right]^{1-\theta} \geq \frac{C_1(1-\theta)}{C}\frac{\left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\|^2}{\|\boldsymbol{Z}\|} \tag{134}$$

$$\geq \frac{C_1(1-\theta)}{CC_2} \frac{\left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\|^2}{\left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u-2)}\right\|}, \quad (135)$$

where we use Condition (iv).

Next, we fix a constant $0 < \tau < 1$. For some $u \geq U_0$, if $\left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| \geq \tau \left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u-2)}\right\|$, from (147), we get the following:

$$\frac{CC_2}{rC_1(1-\theta)} \left\{ \left[\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right)\right]^{1-\theta} - \left[\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right]^{1-\theta} \right\} \geq \left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\|. \quad (136)$$

For all other values of $u \geq U_0$, we have the following relation:

$$\left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| \leq \tau \left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u-2)}\right\|. \quad (137)$$

Combining (148) and (149), for all $u \geq U_0$, we get the upper bound as given below:

$$\left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| \leq \tau \left\|\boldsymbol{A}^{(r,u-1)} - \boldsymbol{A}^{(r,u-2)}\right\| + \frac{CC_2}{rC_1(1-\theta)} \left\{ \left[\bar{g}\left(\boldsymbol{A}^{(r,u-1)}\right)\right]^{1-\theta} - \left[\bar{g}\left(\boldsymbol{A}^{(r,u)}\right)\right]^{1-\theta} \right\}. \quad (138)$$

Summing both sides, and using (141), we can simplify the expression as follows:

$$\sum_{u=U_0}^{\infty} \left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| \leq \frac{\tau}{1-\tau} \left\|\boldsymbol{A}^{(r,U_0-1)} - \boldsymbol{A}^{(r,U_0-2)}\right\| + \frac{CC_2}{rC_1(1-\theta)} \left[\bar{g}\left(\boldsymbol{A}^{(r,U_0)}\right)\right]^{1-\theta}. \quad (139)$$

Thus, we conclude that the series converges, and there exists a finite constant $\kappa < \infty$ such that the following holds:

$$\sum_{u=1}^{\infty} \left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| = \kappa. \quad (140)$$

Hence, for any $\epsilon > 0$, there exists a positive integer U_1 such that for all $U \geq U_1$, we have

$$\kappa - \epsilon/2 \leq \sum_{u=1}^{U} \left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| \leq \kappa + \epsilon/2 \quad (141)$$

Thus, for any $U_1 \leq u_1 < u_2$, we have

$$\begin{aligned} &\left|\left\|\boldsymbol{A}^{(r,u_2)}\right\| - \left\|\boldsymbol{A}^{(r,u_1)}\right\|\right| \\ &\leq \sum_{u=u_1+1}^{u_2} \left|\left\|\boldsymbol{A}^{(r,u)}\right\| - \left\|\boldsymbol{A}^{(r,u-1)}\right\|\right| && (142) \\ &\leq \sum_{u=u_1+1}^{u_2} \left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| && (143) \\ &= \sum_{u=1}^{u_2} \left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| - \sum_{u=1}^{u_1} \left\|\boldsymbol{A}^{(r,u)} - \boldsymbol{A}^{(r,u-1)}\right\| && (144) \\ &\leq \epsilon. && (145) \end{aligned}$$

Therefore, the sequence $\left\{\boldsymbol{A}^{(r,u)}\right\}_{u\in\mathbb{N}}$ is Cauchy, hence it converges.