Supplementary

8.3. Proof of Lemma 1

From the optimality condition of the x problem (7a) we have

$$\nabla f(x^{r+1}) + A^T(\mu^r + \beta A x^{r+1}) + \beta B^T B(x^{r+1} - x^r) = 0.$$

Applying (7b), we have

$$A^T \mu^{r+1} = -\nabla f(x^{r+1}) - \beta B^T B(x^{r+1} - x^r). \tag{33}$$

From equation (7b) ($\mu^{r+1} = \mu^r + \beta A^T \mu^r$) it is clear the difference of the dual variables lies in the column space of A. Therefore the following is true

$$\sigma_{\min}^{1/2} \|\mu^{r+1} - \mu^r\| \le \|A^T(\mu^{r+1} - \mu^r)\|.$$

This inequality combined with (33) implies that

$$\begin{aligned}\|\mu^{r+1} - \mu^r\| &\le \frac{1}{\sigma_{\min}^{1/2}} \| - \nabla f(x^{r+1}) - \beta B^T B(x^{r+1} - x^r) - (-\nabla f(x^r) - \beta B^T B(x^r - x^{r-1}))\| \\ &= \frac{1}{\sigma_{\min}^{1/2}} \left\|\nabla f(x^r) - \nabla f(x^{r+1}) - \beta B^T B w^r\right\|.\end{aligned}$$

Squaring both sides and dividing by β, we obtain the desired result. **Q.E.D.**

8.4. Proof of Lemma 2

Since $f(x)$ has Lipschitz continuous gradient, and that $A^T A + B^T B \succeq I$ by Assumption [A1], it is known that if $\beta > L$, then the x-problem (7a) is strongly convex with modulus $\gamma := \beta - L > 0$; See (Zlobec, 2005) [Theorem 2.1]. That is, we have

$$\begin{aligned}&L_\beta(x, \mu^r) + \frac{\beta}{2}\|x - x^r\|^2_{B^T B} - (L_\beta(z, \mu^r) + \frac{\beta}{2}\|z - x^r\|^2_{B^T B}) \\ &\ge \langle \nabla_x L_\beta(z, \mu^r) + \beta(B^T B(z - x^r)), x - z\rangle + \frac{\gamma}{2}\|x - z\|^2, \ \forall\ x, z \in \mathbb{R}^N, \ \forall\ \mu^r.\end{aligned} \tag{34}$$

Using this property, we have

$$\begin{aligned}&L_\beta(x^{r+1}, \mu^{r+1}) - L_\beta(x^r, \mu^r) \\ &= L_\beta(x^{r+1}, \mu^{r+1}) - L_\beta(x^{r+1}, \mu^r) + L_\beta(x^{r+1}, \mu^r) - L_\beta(x^r, \mu^r) \\ &\le L_\beta(x^{r+1}, \mu^{r+1}) - L_\beta(x^{r+1}, \mu^r) + L_\beta(x^{r+1}, \mu^r) + \frac{\beta}{2}\|x^{r+1} - x^r\|^2_{B^T B} - L_\beta(x^r, \mu^r) \\ &\overset{\text{(i)}}{\le} \frac{\|\mu^{r+1} - \mu^r\|^2}{\beta} + \langle \nabla_x L_\beta(x^{r+1}, \mu^r) + \beta(B^T B(x^{r+1} - x^r)), x^{r+1} - x^r\rangle - \frac{\gamma}{2}\|x^{r+1} - x^r\|^2 \\ &\overset{\text{(ii)}}{\le} \frac{\|\mu^{r+1} - \mu^r\|^2}{\beta} - \frac{\gamma}{2}\|x^{r+1} - x^r\|^2 \\ &\le \frac{1}{\sigma_{\min}}\left(\frac{2L^2}{\beta}\left\|x^r - x^{r+1}\right\|^2 + 2\beta\left\|B^T B w^r\right\|^2\right) - \frac{\gamma}{2}\|x^{r+1} - x^r\|^2 \\ &= -\left(\frac{\beta - L}{2} - \frac{2L^2}{\beta\sigma_{\min}}\right)\|x^{r+1} - x^r\|^2 + \frac{2\beta}{\sigma_{\min}}\left\|B^T B w^r\right\|^2\end{aligned} \tag{35}$$

where in (i) we have used (34) with the identification $z = x^{r+1}$ and $x = x^r$ and the fact that

$$L_\beta(x^{r+1}, \mu^{r+1}) - L_\beta(x^{r+1}, \mu^r) = \langle \mu^{r+1} - \mu^r, A x^{r+1}\rangle = \frac{1}{\beta}\|\mu^{r+1} - \mu^r\|^2$$

; in (ii) we have used the optimality condition for the x-subproblem (7a). The claim is proved. **Q.E.D.**

8.5. Proof of Lemma 3

From the optimality condition of the x-subproblem (7a) we have

$$\langle \nabla f(x^{r+1}) + A^T\mu^r + \beta A^T A x^{r+1} + \beta B^T B(x^{r+1} - x^r), x^{r+1} - x\rangle \le 0,\ \forall\ x \in \mathbb{R}^Q. \tag{36}$$

If we shift r to $r-1$, we get

$$\langle \nabla f(x^r) + A^T\mu^{r-1} + \beta A^T A x^r + \beta B^T B(x^r - x^{r-1}), x^r - x\rangle \le 0,\ \forall\ x \in \mathbb{R}^Q.$$

Plugging $x = x^r$ into the first inequality and $x = x^{r+1}$ into the second, adding the resulting inequalities and utilizing the μ-update step (7b) we obtain

$$\langle \nabla f(x^{r+1}) - \nabla f(x^r) + A^T(\mu^{r+1} - \mu^r) + \beta B^T B w^r, x^{r+1} - x^r\rangle \le 0.$$

Rearranging, we have

$$\langle A^T(\mu^{r+1} - \mu^r), x^{r+1} - x^r\rangle \le -\langle \nabla f(x^{r+1}) - \nabla f(x^r) + \beta B^T B w^r, x^{r+1} - x^r\rangle. \tag{37}$$

Let us bound the lhs and the rhs of (37) separately.

First the lhs of (37) can be expressed as

$$\begin{aligned}\langle A^T(\mu^{r+1} - \mu^r), x^{r+1} - x^r\rangle &= \langle \beta A^T A x^{r+1}, x^{r+1} - x^r\rangle \\ &= \langle \beta A x^{r+1}, A x^{r+1} - A x^r\rangle \\ &= \beta\|Ax^{r+1}\|^2 - \beta\langle A x^{r+1}, A x^r\rangle \\ &= \frac{\beta}{2}\left(\|Ax^{r+1}\|^2 - \|Ax^r\|^2 + \|A(x^{r+1} - x^r)\|^2\right). \end{aligned} \tag{38}$$

Second we have the following bound for the rhs of (37)

$$\begin{aligned}&-\langle \nabla f(x^{r+1}) - \nabla f(x^r) + \beta B^T B w^r, x^{r+1} - x^r\rangle \\ &\le L\|x^{r+1} - x^r\|^2 - \beta\langle B^T B w^r, x^{r+1} - x^r\rangle \\ &= L\|x^{r+1} - x^r\|^2 + \frac{\beta}{2}\Big(\|x^r - x^{r-1}\|^2_{B^T B} - \|x^{r+1} - x^r\|^2_{B^T B} - \|w^r\|^2_{B^T B}\Big). \end{aligned} \tag{39}$$

Combining the above two bounds, we have

$$\begin{aligned}\frac{\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1} - x^r\|^2_{B^T B}\right) \le L\|x^{r+1} - x^r\|^2 + \frac{\beta}{2}\left(\|x^r - x^{r-1}\|^2_{B^T B} + \|Ax^r\|^2\right) \\ - \frac{\beta}{2}\left(\|w^r\|^2_{B^T B} + \|A(x^{r+1} - x^r)\|^2\right). \end{aligned}$$

The desired claim is proved. **Q.E.D.**

8.6. Proof of Lemma 4

Multiplying both sides of (10) by the constant c and then add them to (9), we obtain

$$\begin{aligned}&L_\beta(x^{r+1}, \mu^{r+1}) + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1} - x^r\|^2_{B^T B}\right) \\ &\le L_\beta(x^r, \mu^r) + cL\|x^{r+1} - x^r\|^2 + \frac{c\beta}{2}\left(\|x^r - x^{r-1}\|^2_{B^T B} + \|Ax^r\|^2\right) \\ &- \left(\frac{\beta - L}{2} - \frac{2L^2}{\beta\sigma_{\min}}\right)\|x^{r+1} - x^r\|^2 + \frac{2\beta}{\sigma_{\min}}\left\|B^T B w^r\right\|^2 \\ &- \frac{c\beta}{2}\left(\|w^r\|^2_{B^T B} + \|A(x^{r+1} - x^r)\|^2\right) \\ &\le L_\beta(x^r, \mu^r) + \frac{c\beta}{2}\left(\|x^r - x^{r-1}\|^2_{B^T B} + \|Ax^r\|^2\right) \\ &\quad - \left(\frac{\beta - L}{2} - \frac{2L^2}{\beta\sigma_{\min}} - cL\right)\|x^{r+1} - x^r\|^2 - \left(\frac{c\beta}{2} - \frac{2\beta\|B^T B\|_F}{\sigma_{\min}}\right)\|w^r\|^2_{B^T B}. \end{aligned}$$

The desired result is proved. **Q.E.D.**

8.7. Proof of Lemma 5

To prove this we need to utilize the boundedness assumption in [A2].

First, we can express the augmented Lagrangian function as following

$$\begin{aligned} L_\beta(x^{r+1},\mu^{r+1}) &= f(x^{r+1}) + \langle \mu^{r+1}, Ax^{r+1}\rangle + \frac{\beta}{2}\|Ax^{r+1}\|^2 \\ &= f(x^{r+1}) + \frac{1}{\beta}\langle \mu^{r+1}, \mu^{r+1}-\mu^r\rangle + \frac{\beta}{2}\|Ax^{r+1}\|^2 \\ &= f(x^{r+1}) + \frac{1}{2\beta}\left(\|\mu^{r+1}\|^2 - \|\mu^r\|^2 + \|\mu^{r+1}-\mu^r\|^2\right) + \frac{\beta}{2}\|Ax^{r+1}\|^2. \end{aligned}$$

Therefore, summing over $r = 1\cdots, T$, we obtain

$$\sum_{r=1}^{T} L_\beta(x^{r+1},\mu^{r+1}) = \sum_{r=1}^{T}\left(f(x^{r+1}) + \frac{\beta}{2}\|Ax^{r+1}\|^2 + \frac{1}{2\beta}\|\mu^{r+1}-\mu^r\|^2\right) + \frac{1}{2\beta}\left(\|\mu^{T+1}\|^2 - \|\mu^1\|^2\right).$$

Suppose Assumption [A2] is satisfied and β is chosen according to (13) and (14), then clearly the above sum is lower bounded since

$$f(x) + \frac{\beta}{2}\|Ax\|^2 \ge f(x) + \frac{\delta}{2}\|Ax\|^2 \ge 0,\ \forall\, x\in\mathbb{R}^Q.$$

This fact implies that the sum of the potential function is also lower bounded (note, the remaining terms in the potential function are all nonnegative), that is

$$\sum_{r=1}^{T} P_{c,\beta}(x^{r+1},x^r,\mu^{r+1}) > -\infty, \quad \forall\, T>0.$$

Note that if c and β are chosen according to (13) and (14), then $P_{c,\beta}(x^{r+1},x^r,\mu^{r+1})$ is nonincreasing. Combined with the lower boundedness of the sum of the potential function, we can conclude that the following is true

$$P_{c,\beta}(x^{r+1},x^r,\mu^{r+1}) > -\infty, \quad \forall\, r>0. \tag{40}$$

This completes the proof. **Q.E.D.**

8.8. Proof of Theorm 1

First we prove part (1). Combining Lemmas 4 and 5, we conclude that $\|x^{r+1}-x^r\|^2\to 0$. Then according to (8), in the limit we have $\mu^{r+1}\to\mu^r$, or equivalently $Ax^r\to 0$. That is, the constraint violation will be satisfied in the limit.

Then we prove part (2). From the optimality condition of x-update step (7a) we have

$$\nabla f(x^{r+1}) + A^T\mu^r + \beta A^T(Ax^{r+1}) + \beta B^T B(x^{r+1}-x^r) = 0.$$

Then we argue that $\{\mu^r\}$ is a bounded sequence if $\nabla f(x^{r+1})$ is bounded. Indeed the fact that $\|x^{r+1}-x^r\|^2\to 0$ and $Ax^{r+1}\to 0$ imply that both $(x^{r+1}-x^r)$ and Ax^{r+1} are bounded. Then the boundedness of μ^r follows from the assumption that $\nabla f(x)$ is bounded for any $x\in\mathbb{R}^Q$, and that μ^r lies in the column space of A.

Then we argue that $\{x^r\}$ is bounded if $f(x)+\frac{\beta}{2}\|Ax\|^2$ is coercive. Note that the potential function can be expressed as

$$\begin{aligned} P_{c,\beta}(x^{r+1},x^r,\mu^{r+1}) &= f(x^{r+1}) + \langle\mu^{r+1},Ax^{r+1}\rangle + \frac{\beta}{2}\|Ax^{r+1}\|^2 + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1}-x^r\|^2_{B^TB}\right) \\ &= f(x^{r+1}) + \frac{1}{2\beta}(\|\mu^{r+1}\|^2 - \|\mu^r\|^2 + \|\mu^{r+1}-\mu^r\|^2) + \frac{\beta}{2}\|Ax^{r+1}\|^2 \\ &\quad + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1}-x^r\|^2_{B^TB}\right) \end{aligned}$$

and by our analysis in Lemma 5 we know that it is decreasing thus *upper bounded*. Suppose that $\{x^r\}$ is unbounded and let $\mathcal{K}$ denote an infinite subset of iteration index in which $\lim_{r\in\mathcal{K}} x^r = \infty$. Passing limit to $P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1})$ over $\mathcal{K}$, and using the fact that $x^{r+1} \to x^r$, $\mu^{r+1} \to \mu^r$, we have

$$\lim_{r\in\mathcal{K}} P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) = \lim_{r\in\mathcal{K}} f(x^{r+1}) + \frac{c\beta+\beta}{2}\|Ax^{r+1}\| = \infty$$

where the last equality comes from the coerciveness assumption. This is a contradiction to the fact that the potential function $P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1})$ is upper bounded. This concludes the proof for the second part of the result.

Then we prove part (3). Let $\mathcal{K}$ denote any converging infinite iteration index such that $\{(\mu^r, x^r)\}_{r\in\mathcal{K}}$ converges to the limit point (μ^*, x^*). Passing limit in $\mathcal{K}$, and using the fact that $\|x^{r+1} - x^r\| \to 0$, we have

$$\nabla f(x^*) + A^T\mu^* + \beta A^T Ax^* = 0.$$

Combined with the fact that $Ax^* = 0$, we conclude that (μ^*, x^*) is indeed a stationary point of the original problem (5), satisfying (16).

Additionally, even if the sequence $\{x^{r+1}, \mu^{r+1}\}$ does not have a limit point, from part (1) we still have $\|\mu^{r+1} - \mu^r\| \to 0$ and $\|x^r - x^{r+1}\| \to 0$. Hence

$$\lim_{r\to\infty} \nabla_x L_\beta(x^{r+1}, \mu^r) = \lim_{r\to\infty} \nabla f(x^{r+1}) \stackrel{\text{(i)}}{=} \lim_{r\to\infty} -\beta B^T B(x^{r+1} - x^r) = 0$$

where (i) is from the optimality condition of the x-subproblem (7a). Therefore we have $Q(x^{r+1}, \mu^r) \to 0$.

Finally we prove part (4). Our first step is to bound the size of the gradient of the augmented Lagrangian. From the optimality condition of the x-problem (7a), we have

$$\begin{aligned}\|\nabla_x L_\beta(x^r, \mu^{r-1})\|^2 &= \|\nabla_x L_\beta(x^{r+1}, \mu^r) + \beta B^T B(x^{r+1} - x^r) - \nabla_x L_\beta(x^r, \mu^{r-1})\|^2 \\ &= \|\nabla f(x^{r+1}) - \nabla f(x^r) + A^T(\mu^{r+1} - \mu^r) + \beta B^T B(x^{r+1} - x^r)\|^2 \\ &\le 3L^2\|x^{r+1} - x^r\|^2 + 3\|\mu^{r+1} - \mu^r\|^2\|A^T A\| + 3\beta^2\|B^T B(x^{r+1} - x^r)\|^2.\end{aligned}$$

By utilizing the estimate (8), we see that there must exist a constant $\xi > 0$ such that the following is true

$$Q(x^r, \mu^{r-1}) = \|\nabla_x L_\beta(x^r, \mu^{r-1})\|^2 + \beta\|Ax^r\|^2 \le \xi\left\|x^r - x^{r+1}\right\|^2 + \xi\left\|B^T Bw^r\right\|^2.$$

From the descent estimate (9) we see that there must exist a constant $\nu > 0$ such that

$$P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) - P_{c,\beta}(x^r, x^{r-1}, \mu^r) \le -\nu\|x^{r+1} - x^r\|^2 - \nu\left\|B^T Bw^r\right\|^2.$$

Matching the above two bounds, we have

$$Q(x^r, \mu^{r-1}) \le \frac{\nu}{\xi}\left(P_{c,\beta}(x^r, x^{r-1}, \mu^r) - P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1})\right).$$

Summing over r, and let T denote the first time that $Q(x^r, \mu^{r-1})$ reaches below φ, we obtain

$$\begin{aligned}\varphi \le \frac{1}{T-1}\sum_{r=1}^{T-1} Q(x^r, \mu^{r-1}) &\le \frac{1}{T-1}\frac{\nu}{\xi}\left(P_{c,\beta}(x^1, x^0, \mu^1) - P_{c,\beta}(x^T, x^{T-1}, \mu^T)\right) \\ &\le \frac{1}{T-1}\frac{\nu}{\xi}\left(P_{c,\beta}(x^1, x^0, \mu^1) - \underline{P}\right) := \frac{\nu}{T-1}.\end{aligned}$$

We conclude that the convergence in term of the optimality gap function $Q(x^{r+1}, \mu^r)$ is sublinear. **Q.E.D.**

8.9. The Analysis Outline for Prox-GPDA

First, following the derivation leading to (8) we obtain

$$\frac{1}{\beta}\|\mu^{r+1}-\mu^r\|^2 \le \frac{2L^2}{\beta\sigma_{\min}}\left\|x^r - x^{r-1}\right\|^2 + \frac{2\beta}{\sigma_{\min}}\left\|B^T B w^r\right\|^2. \tag{41}$$

Note that the first term is now related to the square of the difference between the *previous* two iterations.

Following the proof steps in Lemma 2, the descent of the augmented Lagrangian is given by

$$\begin{aligned} & L_\beta(x^{r+1},\mu^{r+1}) - L_\beta(x^r,\mu^r) \\ & \le -\frac{\beta-L}{2}\|x^{r+1}-x^r\|^2 + \frac{2\beta}{\sigma_{\min}}\left\|B^T B w^r\right\|^2 + \frac{2L^2}{\beta\sigma_{\min}}\left\|x^r - x^{r-1}\right\|^2. \end{aligned} \tag{42}$$

In the third step we have the following estimate

$$\begin{aligned} & \frac{\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1}-x^r\|^2_{B^T B}\right) \\ & \le \frac{L}{2}\|x^{r-1}-x^r\|^2 + \frac{L}{2}\|x^{r+1}-x^r\|^2 + \frac{\beta}{2}\left(\|x^r - x^{r-1}\|^2_{B^T B} + \|Ax^r\|^2\right) \\ & \quad - \frac{\beta}{2}\left(\|w^r\|^2_{B^T B} + \|A(x^{r+1}-x^r)\|^2\right). \end{aligned} \tag{43}$$

Note that the first two terms come from the following estimate

$$\begin{aligned} -\langle x^{r+1}-x^r, \nabla f(x^r) - \nabla f(x^{r-1})\rangle & \le \frac{L}{2}\|x^{r+1}-x^r\|^2 + \frac{1}{2L}\|\nabla f(x^r) - \nabla f(x^{r-1})\|^2 \\ & \le \frac{L}{2}\|x^{r+1}-x^r\|^2 + \frac{L}{2}\|x^r - x^{r-1}\|^2, \end{aligned}$$

where the first inequality is the application of Young's inequality.

In the fourth step we have the following overall descent estimate

$$\begin{aligned} & L_\beta(x^{r+1},\mu^{r+1}) + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1}-x^r\|^2_{B^T B}\right) \\ & \le L_\beta(x^r,\mu^r) + \frac{c\beta}{2}\left(\|x^r - x^{r-1}\|^2_{B^T B} + \|Ax^r\|^2\right) - \left(\frac{\beta-L}{2} - \frac{cL}{2}\right)\|x^{r+1}-x^r\|^2 \\ & \quad + \left(\frac{2L^2}{\beta\sigma_{\min}} + \frac{cL}{2}\right)\|x^{r-1}-x^r\|^2 - \left(\frac{c\beta}{2} - \frac{2\beta\|B^T B\|}{\sigma_{\min}}\right)\|w^r\|^2_{B^T B}. \end{aligned} \tag{44}$$

Note that there is a slight difference between this descent estimate and our previous estimate (12), because now there is a positive term in the rhs, which involves $\|x^r - x^{r-1}\|^2$. Therefore the potential function is difficult to decrease by itself. Fortunately, such extra term can be bounded by the descent of the *previous* iteration. We can take the summation over all the iterations and obtain

$$\begin{aligned} & L_\beta(x^{T+1},\mu^{T+1}) + \frac{c\beta}{2}\left(\|Ax^{T+1}\|^2 + \|x^{T+1}-x^T\|^2_{B^T B}\right) \\ & \le L_\beta(x^1,\mu^1) + \frac{c\beta}{2}\left(\|x^1 - x^0\|^2_{B^T B} + \|Ax^1\|^2\right) + \left(\frac{2L^2}{\beta\sigma_{\min}} + cL\right)\|x^0 - x^1\|^2 \\ & \quad - \sum_{r=1}^{T-1}\left(\frac{\beta-L}{2} - \frac{2L^2}{\beta\sigma_{\min}} - cL\right)\|x^{r+1}-x^r\|^2 - \sum_{r=1}^{T}\left(\frac{c\beta}{2} - \frac{2\beta\|B^T B\|}{\sigma_{\min}}\right)\|w^r\|^2_{B^T B}. \end{aligned}$$

Clearly as long as the potential function is lower bounded, we have $x^{r+1}\to x^r$ and $x^{r+1}-x^r \to x^r - x^{r-1}$. The rest of the proof follows similar steps leading to Theorem 1, hence is omitted.

9. Proof of Convergence for Prox-PDA-IP

In this part we present the convergence analysis for Prox-PDA-IP algorithm which main steps are given in (19) and (20). Our analysis consists of a series of steps.

Step 1. Our first step is again to bound the size of the successive difference of $\{\mu^r\}$. To this end, write down the optimality condition for the x-update (19) as

$$A^T\mu^{r+1} = -\nabla f(x^{r+1}) - \beta^{r+1}B^TB(x^{r+1}-x^r). \tag{45}$$

Subtracting the previous iteration, we obtain

$$A^T(\mu^{r+1}-\mu^r) = -(\nabla f(x^{r+1}) - \nabla f(x^r)) - \beta^r B^TB\,(w^r) - (\beta^{r+1}-\beta^r)B^TB(x^{r+1}-x^r). \tag{46}$$

Therefore, using the fact that $\mu^{r+1}-\mu^r \in \mathrm{col}(A)$, we have

$$\frac{1}{\beta^{r+1}}\|\mu^{r+1}-\mu^r\|^2 \le \frac{3}{\beta^{r+1}\sigma_{\min}}\left(L^2 + (\beta^{r+1}-\beta^r)^2\|B^TB\|\right)\|x^{r+1}-x^r\|^2 + \frac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}}\left\|B^TB\,(w^r)\right\|^2. \tag{47}$$

Also from the optimality condition we have the following relation

$$x^{r+1} = x^r - \frac{1}{\beta^{r+1}}(B^TB)^{-1}\left(\nabla f(x^{r+1}) + A^T\mu^{r+1}\right) := x^r - \frac{1}{\beta^{r+1}}v^{r+1}, \tag{48}$$

where we have defined the primal update direction v^{r+1} as

$$v^{r+1} = (B^TB)^{-1}\left(\nabla f(x^{r+1}) + A^T\mu^{r+1}\right).$$

Step 2. In the second step we analyze the descent of the augmented Lagrangian. We have the following estimate

$$\begin{aligned}
&L_{\beta^{r+1}}(x^{r+1},\mu^{r+1}) - L_{\beta^r}(x^r,\mu^r)\\
&= L_{\beta^{r+1}}(x^{r+1},\mu^{r+1}) - L_{\beta^{r+1}}(x^{r+1},\mu^r) + L_{\beta^{r+1}}(x^{r+1},\mu^r) - L_{\beta^{r+1}}(x^r,\mu^r) + L_{\beta^{r+1}}(x^r,\mu^r) - L_{\beta^r}(x^r,\mu^r)\\
&\overset{\text{(i)}}{\le} \frac{1}{\beta^{r+1}}\|\mu^{r+1}-\mu^r\|^2 + \frac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\|\mu^r-\mu^{r-1}\|^2 - \frac{\beta^{r+1}-L}{2}\|x^{r+1}-x^r\|^2\\
&\overset{\text{(ii)}}{\le} -\left(\frac{\beta^{r+1}-L}{2} - \frac{3}{\beta^{r+1}\sigma_{\min}}\left(L^2 + (\beta^{r+1}-\beta^r)^2\|B^TB\|\right)\right)\|x^{r+1}-x^r\|^2 + \frac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\|\mu^r-\mu^{r-1}\|^2\\
&\quad + \frac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}}\left\|B^TB\,(w^r)\right\|^2
\end{aligned} \tag{49}$$

where in (i) we have used the optimality of the x-subproblem (cf. the derivation in (35)), and the fact that

$$L_{\beta^{r+1}}(x^r,\mu^r) - L_{\beta^r}(x^r,\mu^r) = \frac{\beta^{r+1}-\beta^r}{2}\|Ax^r\|^2 = \frac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\|\mu^r-\mu^{r-1}\|^2; \tag{50}$$

in (ii) we have applied (47).

Step 3. In the third step, we construct the remaining part of the potential function. We have the following two inequalities from the optimality condition of the x-update (19)

$$\begin{aligned}
\left\langle \nabla f(x^{r+1}) + A^T\mu^{r+1} + \beta^{r+1}B^TB(x^{r+1}-x^r), x^{r+1}-x\right\rangle \le 0,\ \forall\, x\in\mathbb{R}^Q\\
\left\langle \nabla f(x^r) + A^T\mu^r + \beta^r B^TB(x^r-x^{r-1}), x^r-x\right\rangle \le 0,\ \forall\, x\in\mathbb{R}^Q.
\end{aligned}$$

Plugging $x=x^r$ and $x=x^{r+1}$ to these two equations and adding them together, we obtain

$$\begin{aligned}
&\langle A^T(\mu^{r+1}-\mu^r), x^{r+1}-x^r\rangle\\
&\le -\langle \nabla f(x^{r+1}) - \nabla f(x^r), x^{r+1}-x^r\rangle - \langle B^TB(\beta^{r+1}(x^{r+1}-x^r) - \beta^r(x^r-x^{r-1})), x^{r+1}-x^r\rangle.
\end{aligned}$$

The lhs of the above inequality can be expressed as

$$\begin{aligned}
&\langle A^T(\mu^{r+1}-\mu^r), x^{r+1}-x^r\rangle \\
&= \frac{\beta^{r+1}}{2}\left(\|Ax^{r+1}\|^2 - \|Ax^r\|^2 + \|A(x^{r+1}-x^r)\|^2\right) \\
&= \frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2 - \frac{\beta^r}{2}\|Ax^r\|^2 + \frac{\beta^{r+1}}{2}\|A(x^{r+1}-x^r)\|^2 + \frac{\beta^r-\beta^{r+1}}{2}\|Ax^r\|^2,
\end{aligned}$$

while its rhs can be bounded as

$$\begin{aligned}
&-\langle \nabla f(x^{r+1}) - \nabla f(x^r), x^{r+1}-x^r\rangle - \langle B^TB(\beta^{r+1}(x^{r+1}-x^r) - \beta^r(x^r-x^{r-1})), x^{r+1}-x^r\rangle \\
&\le L\|x^{r+1}-x^r\|^2 - (\beta^{r+1}-\beta^r)\|x^{r+1}-x^r\|^2_{B^TB} \\
&\quad + \frac{\beta^r}{2}\left(\|x^r-x^{r-1}\|^2_{B^TB} - \|x^r-x^{r+1}\|^2_{B^TB} - \|w^r\|^2_{B^TB}\right) \\
&= L\|x^{r+1}-x^r\|^2 - \frac{\beta^{r+1}-\beta^r}{2}\|x^{r+1}-x^r\|^2_{B^TB} \\
&\quad + \frac{\beta^r}{2}\|x^r-x^{r-1}\|^2_{B^TB} - \frac{\beta^{r+1}}{2}\|x^r-x^{r+1}\|^2_{B^TB} - \frac{\beta^r}{2}\|w^r\|^2_{B^TB} \\
&\overset{(21)}{\le} L\|x^{r+1}-x^r\|^2 + \frac{\beta^r}{2}\|x^r-x^{r-1}\|^2_{B^TB} - \frac{\beta^{r+1}}{2}\|x^r-x^{r+1}\|^2_{B^TB} - \frac{\beta^r}{2}\|w^r\|^2_{B^TB}.
\end{aligned}$$

Therefore, combining the above three inequalities we obtain

$$\begin{aligned}
&\frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2 + \frac{\beta^{r+1}}{2}\|x^r-x^{r+1}\|^2_{B^TB} \\
&\le \frac{\beta^r}{2}\|Ax^r\|^2 + \frac{\beta^r}{2}\|x^r-x^{r-1}\|^2_{B^TB} + \frac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\|\mu^{r-1}-\mu^r\|^2 + L\|x^{r+1}-x^r\|^2 - \frac{\beta^r}{2}\|w^r\|^2_{B^TB}.
\end{aligned}$$

Multiplying both sides by β^r, we obtain

$$\begin{aligned}
&\frac{\beta^{r+1}\beta^r}{2}\|Ax^{r+1}\|^2 + \frac{\beta^{r+1}\beta^r}{2}\|x^r-x^{r+1}\|^2_{B^TB} \\
&\le \frac{\beta^r\beta^{r-1}}{2}\|Ax^r\|^2 + \frac{\beta^r\beta^{r-1}}{2}\|x^r-x^{r-1}\|^2_{B^TB} + \frac{\beta^{r+1}-\beta^r}{2\beta^r}\|\mu^{r-1}-\mu^r\|^2 + \beta^rL\|x^{r+1}-x^r\|^2 \\
&\quad - \frac{(\beta^r)^2}{2}\|w^r\|^2_{B^TB} + \frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|Ax^r\|^2 + \frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|x^r-x^{r-1}\|^2_{B^TB} \\
&= \frac{\beta^r\beta^{r-1}}{2}\|Ax^r\|^2 + \frac{\beta^r\beta^{r-1}}{2}\|x^r-x^{r-1}\|^2_{B^TB} + \frac{\beta^{r+1}-\beta^{r-1}}{2\beta^r}\|\mu^{r-1}-\mu^r\|^2 + \beta^rL\|x^{r+1}-x^r\|^2 \\
&\quad - \frac{(\beta^r)^2}{2}\|w^r\|^2_{B^TB} + \frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|x^r-x^{r-1}\|^2_{B^TB}
\end{aligned} \tag{51}$$

where in the last equality we have merged the terms $\frac{\beta^{r+1}-\beta^r}{2\beta^r}\|\mu^{r-1}-\mu^r\|^2$ and $\frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|Ax^r\|^2$.

Step 4. In this step we construct and estimate the descent of the potential function. For some given $c>0$, let us define the potential function as

$$P_{\beta^{r+1},c}(x^{r+1},x^r,\mu^{r+1}) = L_{\beta^{r+1}}(x^{r+1},\mu^{r+1}) + \frac{c\beta^{r+1}\beta^r}{2}\|Ax^{r+1}\|^2 + \frac{c\beta^{r+1}\beta^r}{2}\|x^r-x^{r+1}\|^2_{B^TB}.$$

Note that this potential function has some major differences compared with the one we used before; cf. (11). In particular, the second and the third terms are now quadratic, rather than linear, in the penalty parameters. This new construction is the key to our following analysis.

Then combining the estimate in (51) and (49), we obtain

$$\begin{aligned}
&P_{\beta^{r+1},c}(x^{r+1},x^r,\mu^{r+1}) - P_{\beta^r,c}(x^r,x^{r-1},\mu^r)\\
&\le -\left(\frac{\beta^{r+1}-L}{2} - \frac{3}{\beta^{r+1}\sigma_{\min}}\left(L^2+(\beta^{r+1}-\beta^r)^2\|B^TB\|\right) - c\beta^r L\right)\|x^{r+1}-x^r\|^2\\
&\quad + \frac{\beta^{r+1}-\beta^{r-1}}{2\beta^r}(\frac{1}{\beta^r}+c)\|\mu^r-\mu^{r-1}\|^2 + \frac{c\beta^r(\beta^r-\beta^{r-1})}{2}\|x^r-x^{r-1}\|^2_{B^TB}\\
&\quad - \left(\frac{c(\beta^r)^2}{2} - \frac{3(\beta^r)^2\|B^TB\|}{\beta^{r+1}\sigma_{\min}}\right)\|w^r\|^2_{B^TB} \qquad (52)
\end{aligned}$$

where in the inequality we have also used the fact that $\beta^r \ge \beta^{r-1}$.

Taking the sum of r from t to T (for some $T > t > 1$) and utilize again the estimate in (47), we have

$$\begin{aligned}
&P_{\beta^{T+1},c}(x^{T+1},x^T,\mu^{T+1}) - P_{\beta^t,c}(x^t,x^{t-1},\mu^t)\\
&\le \sum_{r=t}^{T} -\Bigg(\frac{\beta^{r+1}-L}{2} - \frac{3+3(1/\beta^r+c)(\beta^{r+1}-\beta^{r-1})/2\beta^r}{\beta^{r+1}\sigma_{\min}}\left(L^2+(\beta^{r+1}-\beta^{r-1})^2\|B^TB\|\right)\\
&\quad - c\beta^r L - \frac{c\beta^{r+1}(\beta^{r+1}-\beta^r)\|B^TB\|}{2}\Bigg)\|x^{r+1}-x^r\|^2\\
&\quad - \left(\frac{c(\beta^r)^2}{2} - \frac{(3+3(1/\beta^r+c)(\beta^{r+1}-\beta^{r-1})/2\beta^r)(\beta^r)^2\|B^TB\|}{\beta^{r+1}\sigma_{\min}}\right)\|w^r\|^2_{B^TB}\\
&\quad + \frac{c\beta^t(\beta^t-\beta^{t-1})}{2}\|x^t-x^{t-1}\|^2_{B^TB} + \frac{\beta^{t+1}-\beta^{t-1}}{2\beta^t}(1/\beta^t+c)\|\mu^t-\mu^{t-1}\|^2. \qquad (53)
\end{aligned}$$

First, note that for any $c \in (0,1)$, the coefficient in front of $\|w^r\|^2_{B^TB}$ becomes negative for sufficiently large (but finite) t. This is because $\{\beta^r\} \to \infty$, and that the first term in the parenthesis scales in $\mathcal{O}((\beta^r)^2)$ while the second term scales in $\mathcal{O}(\beta^r)$. For the first term to be negative, we need $c > 0$ to be *small enough* such that the following is true for large enough r

$$\frac{\beta^{r+1}-L}{2} - c\beta^r L - \frac{c\beta^{r+1}(\beta^{r+1}-\beta^r)\|B^TB\|}{2} > \frac{\beta^{r+1}}{24}.$$

Suppose that r is large enough such that $(\beta^{r+1}-L)/2 > \beta^{r+1}/3$, or equivalently $\beta^{r+1} > 3L$. Also choose $c = \min\{1/(4L), 1/(12\kappa\|B^TB\|)\}$, where κ is given in (21). Then we have

$$\frac{\beta^{r+1}-L}{2} - c\beta^r L - \frac{c\beta^{r+1}(\beta^{r+1}-\beta^r)\|B^TB\|}{2} > \frac{\beta^{r+1}}{3} - \frac{\beta^{r+1}}{4} - \frac{\beta^{r+1}}{24} = \frac{\beta^{r+1}}{24}. \qquad (54)$$

For this given c, we can also show that the following is true for sufficiently large r

$$\begin{aligned}
&\frac{3+3(1/\beta^r+c)(\beta^{r+1}-\beta^{r-1})/2\beta^r}{\beta^{r+1}\sigma_{\min}}\left(L^2+(\beta^{r+1}-\beta^r)^2\|B^TB\|\right) \le \frac{\beta^{r+1}}{48}\\
&\left(\frac{c(\beta^r)^2}{2} - \frac{(3+3(1/\beta^r+c)(\beta^{r+1}-\beta^{r-1})/2\beta^r)(\beta^r)^2\|B^TB\|}{\beta^{r+1}\sigma_{\min}}\right) \ge \frac{c(\beta^r)^2}{48}.
\end{aligned}$$

In conclusion we have that for sufficiently large but finite t_0, we have

$$\begin{aligned}
&P_{\beta^{T+1},c}(x^{T+1},x^T,\mu^{T+1}) - P_{\beta^{t_0},c}(x^{t_0-1},x^{t_0},\mu^{t_0})\\
&\le \sum_{r=t_0}^{T}\left(-\frac{\beta^{r+1}}{48}\|x^{r+1}-x^r\|^2 - \frac{c(\beta^r)^2}{48}\|w^r\|^2_{B^TB}\right)\\
&\quad + \frac{c\beta^{t_0}(\beta^{t_0}-\beta^{t_0-1})}{2}\|x^{t_0}-x^{t_0-1}\|^2_{B^TB} + \frac{\beta^{t_0+1}-\beta^{t_0-1}}{2\beta^{t_0}}(1/\beta^{t_0}+c)\|\mu^{t_0}-\mu^{t_0-1}\|^2. \qquad (55)
\end{aligned}$$

Therefore we conclude that if $\{\beta^{r+1}\}$ satisfies (21), and for $c > 0$ sufficiently small, there exits a finite $t_0 > 0$ such that for all $T > t_0$, the first two terms of the rhs of (53) are negative.

Step 5. Next we show that the potential function must be lower bounded. Observe that the augmented Lagrangian is given by

$$\begin{aligned}
&L_{\beta^{r+1}}(x^{r+1}, \mu^{r+1}) \\
&= f(x^{r+1}) + \langle \mu^{r+1}, Ax^{r+1} \rangle + \frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2 \\
&= f(x^{r+1}) + \frac{1}{2\beta^{r+1}}\left(\|\mu^{r+1}\|^2 - \|\mu^r\|^2 + \|\mu^{r+1} - \mu^r\|^2\right) + \frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2 \\
&= f(x^{r+1}) + \frac{1}{2\beta^{r+1}}\|\mu^{r+1}\|^2 - \frac{1}{2\beta^r}\|\mu^r\|^2 + \frac{1}{2\beta^{r+1}}\|\mu^{r+1} - \mu^r\|^2 + \left(\frac{1}{2\beta^r} - \frac{1}{2\beta^{r+1}}\right)\|\mu^r\|^2 + \frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2 \\
&\geq f(x^{r+1}) + \frac{1}{2\beta^{r+1}}\|\mu^{r+1}\|^2 - \frac{1}{2\beta^r}\|\mu^r\|^2 + \frac{1}{2\beta^{r+1}}\|\mu^{r+1} - \mu^r\|^2 + \frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2
\end{aligned}$$

where we have used the fact that $\beta^{r+1} \geq \beta^r$. Note that t_0 in (55) is a finite number hence $\frac{1}{2\beta^{t_0}}\|\mu^{t_0}\|^2$ is finite, and utilize Assumption [A2], we conclude that

$$\sum_{r=t_0}^{\infty} L_{\beta^{r+1}}(x^{r+1}, \mu^{r+1}) > -\infty. \tag{56}$$

By noting that the remaining terms of the potential function are all nonnegative, we have

$$\sum_{r=1}^{\infty} P_{\beta^{r+1},c}(x^{r+1}, x^r, \mu^{r+1}) > -\infty. \tag{57}$$

Combining (57) and the bound (55) (which is true for a finite $t_0 > 0$), we conclude that the potential function $P_{\beta^{r+1},c}(x^{r+1}, x^r, \mu^{r+1})$ is lower bounded for all r.

Step 6. In this step we show that the successive differences of various quantities converge.

The lower boundedness of the potential function combined with the bound (55) (which is true for a finite $t_0 > 0$) implies that

$$\sum_{r=1}^{\infty} \beta^{r+1}\|x^{r+1} - x^r\|^2 < \infty, \tag{58a}$$

$$\sum_{r=1}^{\infty} (\beta^r)^2 \|w^r\|^2_{B^TB} < \infty. \tag{58b}$$

Therefore, we have

$$\beta^{r+1}\|x^{r+1} - x^r\|^2 \to 0, \tag{59a}$$

$$(\beta^r)^2\|w^r\|^2_{B^TB} \to 0. \tag{59b}$$

These two facts applied to (46), combined with $\mu^{r+1} - \mu^r \in \mathrm{col}(A)$, indicate that the following is true

$$\mu^{r+1} - \mu^r \to 0. \tag{60}$$

Also (55) implies that the potential function is *upper bounded* as well, and this indicates that

$$\frac{c\beta^{r+1}\beta^r}{2}\|Ax^{r+1}\|^2 \text{ is bounded}, \quad \frac{c\beta^{r+1}\beta^r}{2}\|x^r - x^{r+1}\|^2 \text{ is bounded}. \tag{61}$$

The second of the above inequality implies that $\beta^{r+1}B^TB(x^{r+1} - x^r)$ is bounded. If we further assume that $\nabla f(x)$ is bounded, and use (45), we can conclude that $\{\mu^r\}$ is bounded.

Step 7. Next we show that every limit point of (x^r, μ^r) converges to a stationary solution of problem (5). Let us pass a subsequence $\mathcal{K}$ to (x^r, μ^r) and denote (x^*, μ^*) as its limit point. For notational simplicity, in the following the index r all belongs to the set $\mathcal{K}$.

From relation (58a) we have that any given $\epsilon > 0$, there exists t large enough such that the following is true

$$\sum_{r=t-1}^{\infty} \beta^{r+1}\|x^{r+1} - x^r\|^2 \le \frac{\epsilon}{c\kappa 16}. \tag{62}$$

Utilizing (48), we have that the following is true

$$\sum_{r=1}^{\infty} \frac{1}{\beta^{r+1}}\|v^{r+1}\|^2 < \infty, \quad \lim_{t\to\infty}\sum_{r=t}^{\infty}(\beta^r)^2 \|w^r\|^2_{B^TB} = 0. \tag{63}$$

The first relation implies that $\liminf_{r\to\infty} \|v^{r+1}\| = 0$. Applying these relations to (47), we have

$$\sum_{r=1}^{\infty} \frac{1}{\beta^{r+1}}\|\mu^{r+1} - \mu^r\|^2 < \infty.$$

This implies that for any given $\epsilon > 0$, $c > 0$, there exists an index t sufficiently large such that

$$\sum_{r=t-1}^{\infty} \frac{1}{\beta^{r+1}}\|\mu^{r+1} - \mu^r\|^2 < \frac{\epsilon^2}{4096L\|B^TB\|_F\kappa(1+c)}. \tag{64}$$

Applying this inequality and (62) to (55), we have that for large enough t and for any $T > t$ the following is true

$$P_{\beta^{T+1},c}(x^{T+1}, x^T, \mu^{T+1}) - P_{\beta^t,c}(x^t, x^{t-1}, \mu^t) \le -\sum_{r=t}^{T}\left(\frac{\beta^{r+1}}{48}\|x^{r+1} - x^r\|^2\right) + \frac{\epsilon^2}{4096L\|B^TB\|}. \tag{65}$$

Next we modify a classical argument in (Bertsekas & Tsitsiklis, 1996)[Proposition 3.5] to show that

$$\lim_{r\to\infty} \|v^{r+1}\| \to 0.$$

We already know from the first relation in (63) that $\liminf_{r\to\infty} \|v^{r+1}\| = 0$. Suppose that $\|v^{r+1}\|$ does not converge to 0, then we must have $\limsup_{r\to\infty} \|v^{r+1}\| > 0$. Hence there exists an $\epsilon > 0$ such that $\|v^{r+1}\| < \epsilon/2$ for infinitely many r, and $\|v^{r+1}\| > \epsilon$ for infinitely many r. Then there exists an infinite subset of iteration indices $\mathcal{R}$ such that for each $r \in \mathcal{R}$, there exits a $t(r)$ such that

$$\|v^r\| < \epsilon/2, \quad , \|v^{t(r)}\| > \epsilon,$$
$$\epsilon/2 < \|v^t\| \le \epsilon, \quad \forall\ r < t < t(r).$$

Using the fact that $\lim_{r\in\mathcal{K}} \mu^r = \mu^*$, we have that for r large enough, the following is true for all $t \ge 0$

$$\|\mu^r - \mu^{r+t}\| \le \frac{\epsilon}{8}\frac{1}{\|(B^TB)^{-1}\|\|A^TA\|}. \tag{66}$$

Without loss of generality we can assume that this relation holds for all $r \in \mathcal{R}$. Note that the following is true

$$\begin{aligned}
\frac{\epsilon}{2} \le \|v^{t(r)}\| - \|v^r\| \le \|v^{t(r)} - v^r\| &= \left\|(B^TB)^{-1}\sum_{t=r}^{t(r)-1}\left(\nabla f(x^{t+1}) - \nabla f(x^t) + A^T(\mu^{t+1} - \mu^t)\right)\right\| \\
&\le \|(B^TB)^{-1}\|\left(\sum_{t=r}^{t(r)-1}\|\nabla f(x^{t+1}) - \nabla f(x^t)\| + \|A^TA\|\|\mu^{t(r)} - \mu^r\|\right) \\
&\overset{(48)}{\le} \|(B^TB)^{-1}\|\left(\sum_{t=r}^{t(r)-1}\frac{L}{\beta^{t+1}}\|v^{t+1}\| + \|A^TA\|\|\mu^{t(r)} - \mu^r\|\right) \\
&\le \epsilon L\|(B^TB)^{-1}\|\sum_{t=r}^{t(r)-1}\frac{1}{\beta^{t+1}} + \frac{\epsilon}{8}
\end{aligned} \tag{67}$$

where in the last inequality we have used (66) and the fact that for all $t \in (r+1, t(r))$, we have $\|v^t\| < \epsilon$. This implies that

$$\frac{3}{8L\|(B^TB)^{-1}\|} \leq \sum_{t=r}^{t(r)-1} \frac{1}{\beta^{t+1}}. \tag{68}$$

Using the descent of the potential function (65) we have, for $r \in \mathcal{R}$ and r large enough

$$\begin{aligned}
&P_{\beta^{t(r)},c}(x^{t(r)}, x^{t(r)-1}, \mu^{t(r)}) - P_{\beta^r,c}(x^r, x^{r-1}, \mu^r)\\
&\leq -\sum_{t=r}^{t(r)-1} \frac{1}{48\beta^{t+1}}\|v^{t+1}\|^2 + \frac{\epsilon^2}{4096L\|B^TB\|}\\
&\overset{\text{(i)}}{\leq} -\left(\frac{\epsilon}{4}\right)^2 \sum_{t=r}^{t(r)-1} \frac{1}{48\beta^{t+1}} + \frac{\epsilon^2}{4096L\|B^TB\|}\\
&\overset{\text{(ii)}}{\leq} -\frac{\epsilon^2}{2048L\|B^TB\|} + \frac{\epsilon^2}{4096L\|B^TB\|}\\
&\leq -\frac{\epsilon^2}{4096L\|B^TB\|}
\end{aligned} \tag{69}$$

where in (i) we have used the fact that for all $r \in \mathcal{R}$, $\|v^{r+i}\| \geq \frac{\epsilon}{2}$ for $i = 1, \cdots, t(r)$; in (ii) we have used (68). However we know that the potential function is convergent, i.e.,

$$\lim_{r\to\infty} P_{\beta^{t(r)},c}(x^{t(r)}, x^{t(r)-1}, \mu^{t(r)}) \to P_{\beta^r,c}(x^r, x^{r-1}, \mu^r) = 0$$

which contradicts to (69). Therefore we conclude that $\|v^{r+1}\| \to 0$.

Finally, combining $\|v^{r+1}\| \to 0$ with the convergence of $\mu^{r+1} - \mu^r$ (cf. (60)), we conclude that every limit point of $\{x^r, \mu^r\}$ satisfies

$$\nabla f(x^*) + A^T\mu^* = 0, \quad Ax^* = 0.$$

Therefore it is a stationary solution for problem (5). This completes the proof.

10. Proof of Convergence for Algorithm 2

To make the derivation compact, define the following matrix

$$\begin{aligned}
\boldsymbol{M}^{r+1} &:= \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^{r+1}, Y^{r+1})\\
&= \left[((X_1^{r+1}y_1^{r+1}) - z_1)(y_1^{r+1})^T + 2\gamma X_1^{r+1}; \cdots; ((X_N^{r+1}y_N^{r+1}) - z_N)(y_N^{r+1})^T + 2\gamma X_N^{r+1}\right].
\end{aligned} \tag{70}$$

The proof consists of six steps.

Step 1. First we note that the optimality condition for the $\boldsymbol{X}$-subproblem (30c) is given by

$$\boldsymbol{A}^T\boldsymbol{\Omega}^{r+1} = -\boldsymbol{M}^{r+1} - \beta\langle \boldsymbol{B}^T\boldsymbol{B}, (\boldsymbol{X}^{r+1} - \boldsymbol{X}^r)\rangle. \tag{71}$$

By utilizing the fact that $\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r$ lies in the column space of $\boldsymbol{A}$, and the eigenvalues of A^TA equal to those of $\boldsymbol{A}^T\boldsymbol{A}$, we have the following bound

$$\|\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r\|_F^2 \leq \frac{2}{\sigma_{\min}}\left(\|\boldsymbol{M}^{r+1} - \boldsymbol{M}^r\|_F^2 + \beta^2\|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2\right).$$

Next let us analyze the first term in the rhs of the above inequality. The following identity holds true

$$\begin{aligned}
&\|\boldsymbol{M}^{r+1} - \boldsymbol{M}^r\|_F^2 \\
&= \sum_{i=1}^N \|(X_i^{r+1}y_i^{r+1} - z_i)(y_i^{r+1})^T - (X_i^r y_i^r - z_i)(y_i^r)^T + 2\gamma(X_i^{r+1} - X_i^r)\|_F^2 \\
&\le \sum_{i=1}^N 4\|X_i^{r+1} - X_i^r\|_F^2 \|y_i^{r+1}(y_i^{r+1})^T\|^2 + 4\|X_i^r y_i^r - z_i\|^2 \|y_i^{r+1} - y_i^r\|^2 + 4\|X_i^r(y_i^{r+1} - y_i^r)\|^2 \|y_i^{r+1}\|^2 \\
&\quad + 16\gamma^2 \|X_i^{r+1} - X_i^r\|_F^2 \\
&\le \sum_{i=1}^N 4(\tau^2 + 4\gamma^2)\|X_i^{r+1} - X_i^r\|_F^2 + 4\theta_i^r \|y_i^{r+1} - y_i^r\|^2 + 4\tau\|X_i^r(y_i^{r+1} - y_i^r)\|^2
\end{aligned} \tag{72}$$

where in the last inequality we have defined the constant θ_i^r as

$$\theta_i^r := \|X_i^r y_i^r - z_i\|^2. \tag{73}$$

Therefore, combining the above two inequalities, we obtain

$$\begin{aligned}
\frac{1}{\beta}\|\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r\|_F^2 \le \frac{8}{\beta\sigma_{\min}} &\sum_{i=1}^N \left((\tau^2 + 4\gamma^2)\|X_i^{r+1} - X_i^r\|_F^2 + \theta_i^r\|y_i^{r+1} - y_i^r\|^2 + \tau\|X_i^r(y_i^{r+1} - y_i^r)\|^2\right) \\
&+ \frac{2\beta}{\sigma_{\min}} \|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2
\end{aligned} \tag{74}$$

Step 2. Next let us analyze the descent of the augmented Lagrangian. First we have

$$\begin{aligned}
&L_\beta(\boldsymbol{X}^r, Y^{r+1}, \boldsymbol{\Omega}^r) - L_\beta(\boldsymbol{X}^r, Y^r, \boldsymbol{\Omega}^r) \\
&= \sum_{i=1}^N \left(\frac{1}{2}\|X_i^r y_i^{r+1} - z_i\|^2 + h_i(y_i^{r+1}) - \frac{1}{2}\|X_i^r y_i^r - z_i\|^2 - h_i(y_i^r)\right) \\
&\le \sum_{i=1}^N \left(\frac{1}{2}\|X_i^r y_i^{r+1} - z_i\|^2 + h_i(y_i^{r+1}) + \frac{\theta_i^r}{2}\|y_i^{r+1} - y_i^r\|^2 - \frac{1}{2}\|X_i^r y_i^r - z_i\|^2 - h_i(y_i^r)\right) \\
&\le \sum_{i=1}^N \Bigg(\left\langle (X_i^r)^T(X_i^r y_i^{r+1} - z_i) + \theta_i^r(y_i^{r+1} - y_i^r), y_i^{r+1} - y_i^r \right\rangle - \frac{1}{2}\|X_i^r(y_i^{r+1} - y_i^r)\|^2 - \frac{\theta_i^r}{2}\|y_i^{r+1} - y_i^r\|^2 \\
&\quad + \langle \zeta_i^{r+1}, y_i^{r+1} - y_i^r \rangle \Bigg) \\
&\le -\sum_{i=1}^N \left(\frac{1}{2}\|X_i^r(y_i^{r+1} - y_i^r)\|^2 + \frac{\theta_i^r}{2}\|y_i^{r+1} - y_i^r\|^2\right)
\end{aligned} \tag{75}$$

where in the second to the last equality we have used the convexity of h_i, and $\zeta_i^{r+1} \in \partial h_i(y_i^{r+1})$; the last inequality uses the optimality condition of the y-step (30b). Similarly, we can show that

$$L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^r) - L_\beta(\boldsymbol{X}^r, Y^{r+1}, \boldsymbol{\Omega}^r) \le -\frac{\beta + 2\gamma}{2}\|\boldsymbol{X}^{r+1} - \boldsymbol{X}^r\|_F^2 \tag{76}$$

where we have utilized the fact that $\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{B}^T\boldsymbol{B} = 2\mathbf{D} \succeq \mathbf{I}_{NM}$. Therefore, combining the estimate (74), we obtain

$$\begin{aligned}
&L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^{r+1}) - L_\beta(\boldsymbol{X}^r, Y^r, \boldsymbol{\Omega}^r) \\
&\le -\left(\frac{\beta+2\gamma}{2} - \frac{8(\tau^2+4\gamma^2)}{\beta\sigma_{\min}}\right)\sum_{i=1}^N \|X_i^{r+1} - X_i^r\|_F^2 - \sum_{i=1}^N\left(\frac{\theta_i^r}{2} - \frac{8\theta_i^r}{\beta\sigma_{\min}}\right)\|y_i^{r+1} - y_i^r\|^2 \\
&\quad -\left(\frac{1}{2} - \frac{8\tau}{\sigma_{\min}\beta}\right)\sum_{i=1}^N \|X_i^r(y_i^{r+1} - y_i^r)\|^2 \\
&\quad + \frac{2\beta}{\sigma_{\min}}\|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2. \qquad (77)
\end{aligned}$$

Step 3. This step follows Lemma 3 in the analysis of Algorithm 1. In particular, after writing down the optimality condition of the $\boldsymbol{X}^{r+1}$ and $\boldsymbol{X}^r$ step, we can obtain

$$\begin{aligned}
&\langle \boldsymbol{A}^T(\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r), \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle \\
&\le -\left\langle \boldsymbol{M}^{r+1} - \boldsymbol{M}^r + \beta\boldsymbol{B}^T\boldsymbol{B}\left[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})\right], \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\right\rangle.
\end{aligned}$$

Then it is easy to show that the above inequality implies the following

$$\begin{aligned}
&\frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r), \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle\right) \\
&\le \frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^r, \boldsymbol{A}\boldsymbol{X}^r\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^r - \boldsymbol{X}^{r-1}), \boldsymbol{X}^r - \boldsymbol{X}^{r-1}\rangle\right) - \frac{\beta}{2}\langle \boldsymbol{A}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r), \boldsymbol{A}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r)\rangle \\
&\quad - \langle \boldsymbol{M}^{r+1} - \boldsymbol{M}^r, \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle - \frac{\beta}{2}\|\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2.
\end{aligned}$$

Note the following fact

$$\begin{aligned}
&-\langle \boldsymbol{M}^{r+1} - \boldsymbol{M}^r, \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle \\
&= -\langle \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^{r+1}, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r), \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle \\
&= -\langle \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^{r+1}, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) + \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r), \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle \\
&\overset{\text{(i)}}{\le} -\langle \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r), \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle \\
&\overset{\text{(ii)}}{\le} \frac{1}{2d}\|\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r)\|_F^2 + \frac{d}{2}\|\boldsymbol{X}^{r+1} - \boldsymbol{X}^r\|_F^2 \\
&\overset{\text{(iii)}}{\le} \frac{1}{d}\sum_{i=1}^N\left(\theta_i^r\|y_i^{r+1} - y_i^r\|^2 + \tau\|X_i^r(y_i^{r+1} - y_i^r)\|^2\right) + \frac{d}{2}\|\boldsymbol{X}^{r+1} - \boldsymbol{X}^r\|_F^2 \qquad (78)
\end{aligned}$$

where in (i) we utilize the convexity of $f(\boldsymbol{X}, Y)$ wrt $\boldsymbol{X}$ for any fixed y; in (ii) we use the Cauchy-Swartz inequality, where $d > 0$ is a constant (to be determined later); (iii) is true due to a similar calculation as in (72).

Overall we have

$$\begin{aligned}
&\frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r), \boldsymbol{X}^{r+1} - \boldsymbol{X}^r\rangle\right) \\
&\le \frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^r, \boldsymbol{A}\boldsymbol{X}^r\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^r - \boldsymbol{X}^{r-1}), \boldsymbol{X}^r - \boldsymbol{X}^{r-1}\rangle\right) - \frac{\beta}{2}\langle \boldsymbol{A}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r), \boldsymbol{A}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r)\rangle \\
&\quad + \frac{1}{d}\sum_{i=1}^N\left(\theta_i^r\|y_i^{r+1} - y_i^r\|^2 + \tau\|X_i^r(y_i^{r+1} - y_i^r)\|^2\right) + \frac{d}{2}\|\boldsymbol{X}^{r+1} - \boldsymbol{X}^r\|_F^2 \\
&\quad - \frac{\beta}{2}\|\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2 \qquad (79)
\end{aligned}$$

Step 4. Let us define the potential function as

$$
\begin{aligned}
&P_{\beta,c}(\boldsymbol{X}^{r+1}, \boldsymbol{X}^r, Y^{r+1}, \boldsymbol{\Omega}^{r+1}) \\
&:= L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^{r+1}) + \frac{c\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r), \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\right).
\end{aligned} \tag{80}
$$

Then utilize the bounds (77) and (79), we obtain

$$
\begin{aligned}
&P_{\beta,c}(\boldsymbol{X}^{r+1}, \boldsymbol{X}^r, Y^{r+1}, \boldsymbol{\Omega}^{r+1}) - P_{\beta,c}(\boldsymbol{X}^r, \boldsymbol{X}^{r-1}, Y^r, \boldsymbol{\Omega}^r) \\
\leq\quad & -\left(\frac{\beta+2\gamma}{2} - \frac{8(\tau^2+4\gamma^2)}{\beta\sigma_{\min}} - \frac{cd}{2}\right)\sum_{i=1}^N \|X_i^{r+1}-X_i^r\|_F^2 \\
& -\sum_{i=1}^N\left(\frac{\theta_i^r}{2} - \frac{8\theta_i^r}{\beta\sigma_{\min}} - \frac{c\theta_i^r}{d}\right)\|y_i^{r+1}-y_i^r\|^2 - \left(\frac{1}{2} - \frac{8\tau}{\sigma_{\min}\beta} - \frac{c\tau}{d}\right)\sum_{i=1}^N \|X_i^r(y_i^{r+1}-y_i^r)\|^2 \\
& -\left(\frac{c\beta}{2} - \frac{2\beta\|B^TB\|}{\sigma_{\min}}\right)\|\boldsymbol{B}[(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r)-(\boldsymbol{X}^r-\boldsymbol{X}^{r-1})]\|_F^2.
\end{aligned}
$$

Therefore the following are the condition that guarantees the descent of the potential function

$$
\begin{aligned}
\frac{\beta+2\gamma}{2} - \frac{8(\tau^2+4\gamma^2)}{\beta\sigma_{\min}} - \frac{cd}{2} > 0, &\quad \frac{1}{2} - \frac{8}{\sigma_{\min}\beta} - \frac{c}{d} > 0 \\
\frac{1}{2} - \frac{8\tau}{\sigma_{\min}\beta} - \frac{c\tau}{d} > 0, &\quad \frac{c\beta}{2} - \frac{2\beta\|B^TB\|}{\sigma_{\min}} > 0.
\end{aligned} \tag{81}
$$

To see that it is always possible to find the tuple (β, c, d), first let us set c such that the last inequality is satisfied

$$
c > \frac{4\|B^TB\|}{\sigma_{\min}}. \tag{82}
$$

Second, let us pick any d such that the following is true

$$
d > \max\{2c\tau, 2c\}.
$$

Then clearly it is possible to make β large enough such that all the four conditions in (81) are satisfied.

Step 5. We need to prove that the potential function is lower bounded. We lower bound the augmented Lagrangian as follows

$$
\begin{aligned}
&L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^{r+1}) \\
&= \sum_{i=1}^N\left(\frac{1}{2}\|X_i^{r+1}y_i^{r+1}-z_i\|^2 + \gamma\|X_i^{r+1}\|_F^2 + h_i(y_i^{r+1})\right) + \langle \boldsymbol{\Omega}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle + \frac{\beta}{2}\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle \\
&= \sum_{i=1}^N\left(\frac{1}{2}\|X_i^{r+1}y_i^{r+1}-z_i\|^2 + \gamma\|X_i^{r+1}\|_F^2 + h_i(y_i^{r+1})\right) + \frac{\beta}{2}\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle \\
&\quad + \frac{1}{2\beta}\left(\|\boldsymbol{\Omega}^{r+1}-\boldsymbol{\Omega}^r\|_F^2 + \|\boldsymbol{\Omega}^{r+1}\|_F^2 - \|\boldsymbol{\Omega}^r\|_F^2\right).
\end{aligned} \tag{83}
$$

Then by the same argument leading to (40), we conclude that as long as h_i is lower bounded over its domain, then the potential function will be lower bounded.

Step 6. Combining the results in Step 5 and Step 4, we conclude the following

$$
\sum_{i=1}^N \|X_i^{r+1}-X_i^r\|_F^2 \to 0, \quad \sum_{i=1}^N \|y_i^{r+1}-y_i^r\|^2 \to 0 \tag{84a}
$$

$$
\sum_{i=1}^N \|X_i^r(y_i^{r+1}-y_i^r)\|^2 \to 0, \quad \left\|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r)-(\boldsymbol{X}^r-\boldsymbol{X}^{r-1})]\right\|_F \to 0. \tag{84b}
$$

Then utilizing (74), we have

$$\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r \to \mathbf{0}, \text{ or equivalently } \boldsymbol{A}\boldsymbol{X}^{r+1} \to 0.$$

That is, in the limit the network-wide consensus is achieved. Next we show that the primal and dual iterates are bounded.

Note that the potential function is both lower and upper bounded. Combined with (84) we must have that the augmented Lagrangian is both upper and lower bounded. Using the expression (83), the assumption that $h_i(y_i)$ is lower bounded, and the fact that y_i is bounded, we have that in the limit, the following term is bounded

$$\sum_{i=1}^{N} \frac{1}{2}\|X_i^{r+1} y_i^{r+1} - z_i\|^2 + \gamma \|X_i^{r+1}\|_F^2.$$

This implies that the primal variable sequence $\{X_i^{r+1}\}$ are bounded for all i. To show the boundedness of the dual sequence, note that $\boldsymbol{\Omega}^{r+1} \in \operatorname{col}(\boldsymbol{A})$ (due to the initialization that $\boldsymbol{\Omega}^0 = \mathbf{0}$). Therefore using (71) we have

$$\sigma_{\min}(\boldsymbol{A}^T\boldsymbol{A})\|\boldsymbol{\Omega}^{r+1}\|_F^2 \le 2\|\boldsymbol{M}^{r+1}\|_F^2 + 2\beta\|\boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r)\|_F^2$$

Note that from the expression of $\boldsymbol{M}$ in (70), we see that $\{\boldsymbol{M}^{r+1}\}$ is bounded because both $\boldsymbol{X}^{r+1}$ and Y^{r+1} are bounded. Similarly, the second term on the rhs of the above inequality is bounded because $\boldsymbol{X}^{r+1} \to \boldsymbol{X}^r$. These two facts imply that $\{\boldsymbol{\Omega}^{r+1}\}$ is bounded as well.

Arguing the convergence to stationary point as well as the convergence rate follows exactly the same steps as in the proof of Theorem 1.

10.1. Prox-PDA-IP for Distributed Matrix Factorization

In this section we extend the Prox-PDA for distributed matrix factorization utilizing increasing penalty parameters, just as what we have done in Section 5. In particular, when replacing the penalty parameter β by an increasing sequence $\{\beta^r\}$ that satisfies (21), the resulting algorithm also generates bounded $\{\boldsymbol{X}^{r+1}\}$ and $\{\boldsymbol{\Omega}^{r+1}\}$, whose limit points are stationary points of problem (26). The detailed steps of this variant is given in Algorithm 3.

Algorithm 3 Prox-PDA-IP for Distributed Matrix Factorization

1: At iteration 0, initialize $\boldsymbol{\Omega}^0 = \mathbf{0}$, and $\boldsymbol{X}^0, y^0$
2: At each iteration $r+1$, update variables by:

$$\theta_i^r = \|X_i^r y_i^r - z_i\|^2, \quad \forall\, i; \tag{85a}$$

$$y_i^{r+1} = \arg\min_{\|y_i\|^2 \le \tau} \frac{1}{2}\|X_i^r y_i - z_i\|^2 + h_i(y_i) + \frac{\beta^{r+1}\theta_i^r}{2}\|y_i - y_i^r\|^2 + \frac{\beta^{r+1}}{2}\|X_i^r(y_i - y_i^r)\|^2, \quad \forall\, i; \tag{85b}$$

$$\mathbf{X}^{r+1} = \arg\min_{\mathbf{X}\in\mathbb{R}^{NM\times K}} f\left(\mathbf{X}, Y^{r+1}\right) + \langle \boldsymbol{\Omega}^r, \mathbf{A}\mathbf{X}\rangle + \frac{\beta^{r+1}}{2}\langle \mathbf{A}\mathbf{X}, \mathbf{A}\mathbf{X}\rangle + \frac{\beta^{r+1}}{2}\langle \mathbf{B}(\mathbf{X}-\mathbf{X}^r), \mathbf{B}(\mathbf{X}-\mathbf{X}^r)\rangle; \tag{85c}$$

$$\boldsymbol{\Omega}^{r+1} = \boldsymbol{\Omega}^r + \beta^{r+1}\mathbf{A}\mathbf{X}^{r+1}. \tag{85d}$$

Now, we provide the proof of convergence for Algorithm 3. Our convergence claim is given below.

Theorem 4 *Consider using Algorithm 3 to solve the distributed matrix factorization problem (27). Suppose that $h(Y)$ is lower bounded over $\operatorname{dom} h$, the penalty parameter $\{\beta^r\}$ satisfies (21), and that the matrix B satisfies*

$$B^T B \succ 0, \quad \text{and} \quad \|B^T B\| > 1. \tag{86}$$

Then in the limit, consensus will be achieved, i.e.,

$$X_i = X_j, \quad \forall\, (i,j) \in \mathcal{E}.$$

Further, the sequences $\{\mathbf{X}^{r+1}\}$ *and* $\{\mathbf{\Omega}^{r+1}\}$ *are both bounded, and every limit point generated by Algorithm 5 is a stationary point for problem* (27).

The proof essentially combines the analysis steps of Theorem 2 and 3. However the notation is significantly more complicated due to the increased number of terms involved in the analysis. We include the proof here for completeness.

Step 1. Bound the size of the successive difference of $\{\mathbf{\Omega}^r\}$. Similarly as in the proof of Theorem 3, the optimality condition for the $\mathbf{X}$-update can be written as

$$\mathbf{A}^T\mathbf{\Omega}^{r+1} = -\mathbf{M}^{r+1} - \beta^{r+1}\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r+1} - \mathbf{X}^r) \tag{87}$$

$$\mathbf{A}^T\mathbf{\Omega}^{r} = -\mathbf{M}^{r} - \beta^{r}\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r} - \mathbf{X}^{r-1}). \tag{88}$$

Subtracting the above equations, we obtain

$$\mathbf{A}^T\left(\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^r\right) = -\left(\mathbf{M}^{r+1}-\mathbf{M}^r\right) - \beta^r\mathbf{B}^T\mathbf{B}\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right) - \left(\beta^{r+1}-\beta^r\right)\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r+1}-\mathbf{X}^r)$$

Since $\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^r$ lies in the column space of $\mathbf{A}$, and the eigenvalues of A^TA equal to that of $\mathbf{A}^T\mathbf{A}$, we have

$$\sigma_{\min}(A^TA)\left\|\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^r\right\|_F^2 \le \left\|\mathbf{A}\left(\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^r\right)\right\|_F^2,$$

which results in

$$\begin{aligned}
&\tfrac{1}{\beta^{r+1}}\left\|\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^r\right\|_F^2 \\
&\le \tfrac{1}{\beta^{r+1}\sigma_{\min}(A^TA)}\left\|-\left(\mathbf{M}^{r+1}-\mathbf{M}^r\right) - \beta^r\mathbf{B}^T\mathbf{B}\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right) - \left(\beta^{r+1}-\beta^r\right)\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r+1}-\mathbf{X}^r)\right\|^2 \\
&\le \tfrac{3}{\beta^{r+1}\sigma_{\min}(A^TA)}\begin{pmatrix} \left\|\mathbf{M}^{r+1}-\mathbf{M}^r\right\|_F^2 + (\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right)\right\|_F^2 \\ +\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r+1}-\mathbf{X}^r)\right\|^2 \end{pmatrix}.
\end{aligned} \tag{89}$$

Also, from (72) we have

$$\left\|\mathbf{M}^{r+1}-\mathbf{M}^r\right\|_F^2 \le \sum_{i=1}^N 4(\tau^2+4\gamma^2)\left\|X_i^{r+1}-X_i^r\right\|_F^2 + 4\theta_i^r\left\|y_i^{r+1}-y_i^r\right\|^2 + 4\tau\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2. \tag{90}$$

Thus, we have the following bound

$$\begin{aligned}
&\tfrac{1}{\beta^{r+1}}\left\|\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^r\right\|_F^2 \\
&\le \tfrac{12}{\beta^{r+1}\sigma_{\min}(A^TA)}\sum_{i=1}^N(\tau^2+4\gamma^2)\left\|X_i^{r+1}-X_i^r\right\|_F^2 + \theta_i^r\left\|y_i^{r+1}-y_i^r\right\|^2 + \tau\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2 \\
&+\tfrac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}(A^TA)}\left\|\mathbf{B}^T\mathbf{B}\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right)\right\|_F^2 + \tfrac{3\left(\beta^{r+1}-\beta^r\right)^2}{\beta^{r+1}\sigma_{\min}(A^TA)}\left\|\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r+1}-\mathbf{X}^r)\right\|_F^2.
\end{aligned} \tag{91}$$

Step 2. We analyze the descent of the augmented Lagrangian.

First, as in (76), we utilize the strong convexity of the objective function of the $\mathbf{X}$-update (cf. (85c)) and obtain the following descent estimate for the $\mathbf{X}$-update

$$L_{\beta^{r+1}}\left(\mathbf{X}^{r+1},Y^{r+1},\mathbf{\Omega}^r\right) - L_{\beta^{r+1}}\left(\mathbf{X}^r,Y^{r+1},\mathbf{\Omega}^r\right) \le -\tfrac{\beta^{r+1}+2\gamma}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r\right\|_F^2. \tag{92}$$

Note that compared with (76) we have replaced β with β^{r+1}.

Similarly, we have the following estimate for the descent of the Y-update

$$\begin{aligned}
& L_{\beta^{r+1}}\left(\mathbf{X}^r, Y^{r+1}, \mathbf{\Omega}^r\right) - L_{\beta^{r+1}}\left(\mathbf{X}^r, Y^r, \mathbf{\Omega}^r\right) \\
&= \sum_{i=1}^N \left(\tfrac{1}{2}\left\|X_i^r y_i^{r+1} - z_i\right\|^2 + h_i(y_i^{r+1}) - \tfrac{1}{2}\left\|X_i^r y_i^r - z_i\right\|^2 - h_i(y_i^r)\right) \\
&\le \sum_{i=1}^N \left(\begin{array}{l} \frac{1}{2}\left\|X_i^r y_i^{r+1} - z_i\right\|^2 + h_i(y_i^{r+1}) + \frac{\beta^{r+1}\theta_i^r}{2}\left\|y_i^{r+1} - y_i^r\right\|^2 + \frac{\beta^{r+1}}{2}\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 \\ -\frac{1}{2}\left\|X_i^r y_i^r - z_i\right\|^2 - h_i(y_i^r) \end{array}\right) \\
&\le \sum_{i=1}^N \left(\begin{array}{l} \left\langle (X_i^r)^T\left(X_i^r y_i^{r+1} - z_i\right) + \beta^{r+1}\theta_i^r\left(y_i^{r+1} - y_i^r\right) + \beta^{r+1}(X_i^r)^T\left(X_i^r y_i^{r+1} - X_i^r y_i^r\right), y_i^{r+1} - y_i^r\right\rangle \\ -\frac{1}{2}\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 - \frac{\beta^{r+1}\theta_i^r}{2}\left\|y_i^{r+1} - y_i^r\right\|^2 - \frac{\beta^{r+1}}{2}\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 + \left\langle \varsigma_i^{r+1}, y_i^{r+1} - y_i^r\right\rangle \end{array}\right) \\
&\le \sum_{i=1}^N \left(-\tfrac{1+\beta^{r+1}}{2}\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 - \tfrac{\beta^{r+1}\theta_i^r}{2}\left\|y_i^{r+1} - y_i^r\right\|^2\right)
\end{aligned} \tag{93}$$

where in the second to last inequality, $\varsigma_i^{r+1} \in \partial h_i(y_i^{r+1})$; the last inequality is true due to the optimality condition of the Y-update.

Next we analyze the descent of the augmented Lagrangian. Let us first decompose the successive difference of the augmented Lagrangian into the following

$$\begin{aligned}
& L_{\beta^{r+1}}\left(\mathbf{X}^{r+1}, Y^{r+1}, \mathbf{\Omega}^{r+1}\right) - L_{\beta^r}\left(\mathbf{X}^r, Y^r, \mathbf{\Omega}^r\right) \\
&= L_{\beta^{r+1}}\left(\mathbf{X}^{r+1}, Y^{r+1}, \mathbf{\Omega}^{r+1}\right) - L_{\beta^{r+1}}\left(\mathbf{X}^{r+1}, Y^{r+1}, \mathbf{\Omega}^r\right) \\
&+ L_{\beta^{r+1}}\left(\mathbf{X}^{r+1}, Y^{r+1}, \mathbf{\Omega}^r\right) - L_{\beta^{r+1}}\left(\mathbf{X}^r, Y^{r+1}, \mathbf{\Omega}^r\right) \\
&+ L_{\beta^{r+1}}\left(\mathbf{X}^r, Y^{r+1}, \mathbf{\Omega}^r\right) - L_{\beta^{r+1}}\left(\mathbf{X}^r, Y^r, \mathbf{\Omega}^r\right) \\
&+ L_{\beta^{r+1}}\left(\mathbf{X}^r, Y^r, \mathbf{\Omega}^r\right) - L_{\beta^r}\left(\mathbf{X}^r, Y^r, \mathbf{\Omega}^r\right) \\
&\le \tfrac{1}{\beta^{r+1}}\left\|\mathbf{\Omega}^{r+1} - \mathbf{\Omega}^r\right\|_F^2 - \tfrac{\beta^{r+1}+2\gamma}{2}\left\|\mathbf{X}^{r+1} - \mathbf{X}^r\right\|_F^2 - \sum_{i=1}^N \left(\tfrac{1+\beta^{r+1}}{2}\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 + \tfrac{\beta^{r+1}\theta_i^r}{2}\left\|y_i^{r+1} - y_i^r\right\|^2\right) \\
&+ \tfrac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\left\|\mathbf{\Omega}^r - \mathbf{\Omega}^{r-1}\right\|_F^2
\end{aligned}$$

where in the last inequality we have used the estimate given in (92) and (93). Plugging in the estimate (91), and (90), we have

$$\begin{aligned}
& L_{\beta^{r+1}}\left(\mathbf{X}^{r+1}, Y^{r+1}, \mathbf{\Omega}^{r+1}\right) - L_{\beta^r}\left(\mathbf{X}^r, Y^r, \mathbf{\Omega}^r\right) \\
&\le \tfrac{3}{\beta^{r+1}\sigma_{\min}(A^TA)}\left(\begin{array}{l} \left\|\mathbf{M}^{r+1} - \mathbf{M}^r\right\|_F^2 + (\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\left((\mathbf{X}^{r+1} - \mathbf{X}^r) - (\mathbf{X}^r - \mathbf{X}^{r-1})\right)\right\|_F^2 \\ +\left(\beta^{r+1} - \beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r+1} - \mathbf{X}^r)\right\|_F^2 \end{array}\right) \\
&- \tfrac{\beta^{r+1}+2\gamma}{2}\left\|\mathbf{X}^{r+1} - \mathbf{X}^r\right\|_F^2 - \sum_{i=1}^N \left(\tfrac{1+\beta^{r+1}}{2}\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 + \tfrac{\beta^{r+1}\theta_i^r}{2}\left\|y_i^{r+1} - y_i^r\right\|^2\right) \\
&+ \tfrac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\left\|\mathbf{\Omega}^r - \mathbf{\Omega}^{r-1}\right\|_F^2 \\
&\le \tfrac{3}{\beta^{r+1}\sigma_{\min}(A^TA)}\sum_{i=1}^N \left(4(\tau^2 + 4\gamma^2)\left\|X_i^{r+1} - X_i^r\right\|_F^2 + 4\theta_i^r\left\|y_i^{r+1} - y_i^r\right\|^2 + 4\tau\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2\right) \\
&+ \tfrac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}(A^TA)}\left\|\mathbf{B}^T\mathbf{B}\left((\mathbf{X}^{r+1} - \mathbf{X}^r) - (\mathbf{X}^r - \mathbf{X}^{r-1})\right)\right\|_F^2 + \tfrac{3\left(\beta^{r+1}-\beta^r\right)^2}{\beta^{r+1}\sigma_{\min}(A^TA)}\left\|\mathbf{B}^T\mathbf{B}(\mathbf{X}^{r+1} - \mathbf{X}^r)\right\|_F^2 \\
&- \tfrac{\beta^{r+1}+2\gamma}{2}\left\|\mathbf{X}^{r+1} - \mathbf{X}^r\right\|_F^2 - \sum_{i=1}^N \left(\tfrac{1+\beta^{r+1}}{2}\left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 + \tfrac{\beta^{r+1}\theta_i^r}{2}\left\|y_i^{r+1} - y_i^r\right\|^2\right) \\
&+ \tfrac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\left\|\mathbf{\Omega}^r - \mathbf{\Omega}^{r-1}\right\|_F^2 \\
&= \left(\tfrac{12\theta_i^r}{\beta^{r+1}\sigma_{\min}(A^TA)} - \tfrac{\beta^{r+1}\theta_i^r}{2}\right)\sum_{i=1}^N \left\|y_i^{r+1} - y_i^r\right\|^2 + \left(\tfrac{12\tau}{\beta^{r+1}\sigma_{\min}(A^TA)} - \tfrac{1+\beta^{r+1}}{2}\right)\sum_{i=1}^N \left\|X_i^r\left(y_i^{r+1} - y_i^r\right)\right\|^2 \\
&+ \tfrac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}(A^TA)}\left\|\mathbf{B}^T\mathbf{B}\left((\mathbf{X}^{r+1} - \mathbf{X}^r) - (\mathbf{X}^r - \mathbf{X}^{r-1})\right)\right\|_F^2 \\
&+ \left(\tfrac{12(\tau^2+4\gamma^2)}{\beta^{r+1}\sigma_{\min}(A^TA)} + \tfrac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\beta^{r+1}\sigma_{\min}(A^TA)} - \tfrac{\beta^{r+1}+2\gamma}{2}\right)\left\|(\mathbf{X}^{r+1} - \mathbf{X}^r)\right\|_F^2 \\
&+ \tfrac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\left\|\mathbf{\Omega}^r - \mathbf{\Omega}^{r-1}\right\|_F^2.
\end{aligned} \tag{94}$$

Step 3. We construct the remaining part of the potential function. From the optimality condition of the $\mathbf{X}$-update, we obtain

$$\left\langle \mathbf{A}^T\mathbf{\Omega}^{r+1} + \mathbf{M}^{r+1} + \beta^{r+1}\mathbf{B}^T\mathbf{B}\left(\mathbf{X}^{r+1} - \mathbf{X}^r\right), \mathbf{X}^{r+1} - \mathbf{X}\right\rangle \le 0, \tag{95}$$

$$\left\langle \mathbf{A}^T \boldsymbol{\Omega}^r + \mathbf{M}^r + \beta^r \mathbf{B}^T \mathbf{B} \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right), \mathbf{X}^r - \mathbf{X} \right\rangle \leq 0. \tag{96}$$

Plugging $\mathbf{X} = \mathbf{X}^r$ and $\mathbf{X} = \mathbf{X}^{r+1}$ into (95) and (96) and adding them together, we have

$$\begin{aligned} &\left\langle \mathbf{A}^T \left(\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r \right), \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle \\ &\leq -\left\langle \mathbf{M}^{r+1} - \mathbf{M}^r, \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle - \left\langle \mathbf{B}^T \mathbf{B} \left(\beta^{r+1} \left(\mathbf{X}^{r+1} - \mathbf{X}^r \right) - \beta^r \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right) \right), \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle. \end{aligned} \tag{97}$$

The lhs of (97) can be expressed as

$$\begin{aligned} &\left\langle \mathbf{A}^T \left(\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r \right), \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle \\ &= \left\langle \beta^{r+1} \mathbf{A}\mathbf{X}^{r+1}, \mathbf{A}\mathbf{X}^{r+1} - \mathbf{A}\mathbf{X}^r \right\rangle \\ &= \beta^{r+1} \left\| \mathbf{A}\mathbf{X}^{r+1} \right\|_F^2 - \beta^{r+1} \left\langle \mathbf{A}\mathbf{X}^{r+1}, \mathbf{A}\mathbf{X}^r \right\rangle \\ &= \tfrac{\beta^{r+1}}{2} \left\| \mathbf{A}\mathbf{X}^{r+1} \right\|_F^2 - \tfrac{\beta^{r+1}}{2} \left\| \mathbf{A}\mathbf{X}^r \right\|_F^2 + \tfrac{\beta^{r+1}}{2} \left\| \mathbf{A} \left(\mathbf{X}^{r+1} - \mathbf{X}^r \right) \right\|_F^2 \\ &= \tfrac{\beta^{r+1}}{2} \left\| \mathbf{A}\mathbf{X}^{r+1} \right\|_F^2 - \tfrac{\beta^r}{2} \left\| \mathbf{A}\mathbf{X}^r \right\|_F^2 + \tfrac{\left(\beta^r - \beta^{r+1}\right)}{2} \left\| \mathbf{A}\mathbf{X}^r \right\|_F^2 + \tfrac{\beta^{r+1}}{2} \left\| \mathbf{A} \left(\mathbf{X}^{r+1} - \mathbf{X}^r \right) \right\|_F^2. \end{aligned} \tag{98}$$

Noting the following fact (cf. (78))

$$-\left\langle \mathbf{M}^{r+1} - \mathbf{M}^r, \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle \leq \frac{1}{d} \sum_{i=1}^N \left(\theta_i^r \left\| y_i^{r+1} - y_i^r \right\|^2 + \tau \left\| X_i^r \left(y_i^{r+1} - y_i^r \right) \right\|^2 \right) + \frac{d}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_F^2, \tag{99}$$

where $d > 0$ is a constant. The rhs of (97) can be rewritten as

$$\begin{aligned} &-\left\langle \mathbf{M}^{r+1} - \mathbf{M}^r, \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle - \left\langle \mathbf{B}^T \mathbf{B} \left(\beta^{r+1} \left(\mathbf{X}^{r+1} - \mathbf{X}^r \right) - \beta^r \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right) \right), \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle \\ &\leq \tfrac{1}{d} \sum_{i=1}^N \left(\theta_i^r \left\| y_i^{r+1} - y_i^r \right\|^2 + \tau \left\| X_i^r \left(y_i^{r+1} - y_i^r \right) \right\|^2 \right) + \tfrac{d}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_F^2 \\ &-\beta^{r+1} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_{\mathbf{B}^T\mathbf{B}}^2 + \beta^r \left\langle \mathbf{B}^T \mathbf{B} \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right), \mathbf{X}^{r+1} - \mathbf{X}^r \right\rangle \\ &\leq \tfrac{1}{d} \sum_{i=1}^N \left(\theta_i^r \left\| y_i^{r+1} - y_i^r \right\|^2 + \tau \left\| X_i^r \left(y_i^{r+1} - y_i^r \right) \right\|^2 \right) + \tfrac{d}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_F^2 - \left(\beta^{r+1} - \beta^r \right) \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_{\mathbf{B}^T\mathbf{B}}^2 \\ &+\tfrac{\beta^r}{2} \left(\left\| \mathbf{X}^r - \mathbf{X}^{r-1} \right\|_{\mathbf{B}^T\mathbf{B}}^2 - \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_{\mathbf{B}^T\mathbf{B}}^2 - \left\| \mathbf{X}^{r+1} - \mathbf{X}^r - \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right) \right\|_{\mathbf{B}^T\mathbf{B}}^2 \right) \\ &= \tfrac{1}{d} \sum_{i=1}^N \left(\theta_i^r \left\| y_i^{r+1} - y_i^r \right\|^2 + \tau \left\| X_i^r \left(y_i^{r+1} - y_i^r \right) \right\|^2 \right) + \tfrac{d}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_F^2 - \tfrac{\left(\beta^{r+1} - \beta^r\right)}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_{\mathbf{B}^T\mathbf{B}}^2 \\ &-\tfrac{\beta^{r+1}}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_{\mathbf{B}^T\mathbf{B}}^2 + \tfrac{\beta^r}{2} \left\| \mathbf{X}^r - \mathbf{X}^{r-1} \right\|_{\mathbf{B}^T\mathbf{B}}^2 - \tfrac{\beta^r}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r - \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right) \right\|_{\mathbf{B}^T\mathbf{B}}^2 \\ &\leq \tfrac{1}{d} \sum_{i=1}^N \left(\theta_i^r \left\| y_i^{r+1} - y_i^r \right\|^2 + \tau \left\| X_i^r \left(y_i^{r+1} - y_i^r \right) \right\|^2 \right) + \tfrac{d}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_F^2 - \tfrac{\beta^{r+1}}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_{\mathbf{B}^T\mathbf{B}}^2 \\ &+\tfrac{\beta^r}{2} \left\| \mathbf{X}^r - \mathbf{X}^{r-1} \right\|_{\mathbf{B}^T\mathbf{B}}^2 - \tfrac{\beta^r}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r - \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right) \right\|_{\mathbf{B}^T\mathbf{B}}^2, \end{aligned} \tag{100}$$

where the first inequality is obtained by plugging in (99); in the last inequality we have used the fact that $\beta^{r+1} \geq \beta^r$.

Therefore, combining (98) and (100), we obtain

$$\begin{aligned} &\tfrac{\beta^{r+1}}{2} \left\| \mathbf{A}\mathbf{X}^{r+1} \right\|_F^2 + \tfrac{\beta^{r+1}}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_{\mathbf{B}^T\mathbf{B}}^2 \\ &\leq \tfrac{\beta^r}{2} \left\| \mathbf{A}\mathbf{X}^r \right\|_F^2 + \tfrac{\beta^r}{2} \left\| \mathbf{X}^r - \mathbf{X}^{r-1} \right\|_{\mathbf{B}^T\mathbf{B}}^2 + \tfrac{1}{d} \sum_{i=1}^N \left(\theta_i^r \left\| y_i^{r+1} - y_i^r \right\|^2 + \tau \left\| X_i^r \left(y_i^{r+1} - y_i^r \right) \right\|^2 \right) + \tfrac{d}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r \right\|_F^2 \\ &-\tfrac{\beta^r}{2} \left\| \mathbf{X}^{r+1} - \mathbf{X}^r - \left(\mathbf{X}^r - \mathbf{X}^{r-1} \right) \right\|_{\mathbf{B}^T\mathbf{B}}^2 - \tfrac{\left(\beta^r - \beta^{r+1}\right)}{2(\beta^r)^2} \left\| \boldsymbol{\Omega}^r - \boldsymbol{\Omega}^{r-1} \right\|_F^2. \end{aligned}$$

Multiplying both sides by β^r, we have

$$
\begin{aligned}
&\tfrac{\beta^{r+1}\beta^r}{2}\left\|\mathbf{A}\mathbf{X}^{r+1}\right\|_F^2+\tfrac{\beta^r\beta^{r+1}}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r\right\|_{\mathbf{B}^T\mathbf{B}}^2\\
&\le\tfrac{\beta^r\beta^{r-1}}{2}\left\|\mathbf{A}\mathbf{X}^r\right\|_F^2+\tfrac{\beta^r\beta^{r-1}}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2+\tfrac{\beta^r}{d}\sum_{i=1}^N\left(\theta_i^r\left\|y_i^{r+1}-y_i^r\right\|^2+\tau\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2\right)\\
&+\tfrac{d\beta^r}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r\right\|_F^2-\tfrac{(\beta^r)^2}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r-\left(\mathbf{X}^r-\mathbf{X}^{r-1}\right)\right\|_{\mathbf{B}^T\mathbf{B}}^2-\tfrac{\left(\beta^r-\beta^{r+1}\right)}{2\beta^r}\left\|\mathbf{\Omega}^r-\mathbf{\Omega}^{r-1}\right\|_F^2\\
&+\tfrac{\beta^r\left(\beta^r-\beta^{r-1}\right)}{2}\left\|\mathbf{A}\mathbf{X}^r\right\|_F^2+\tfrac{\beta^r\left(\beta^r-\beta^{r-1}\right)}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2\\
&=\tfrac{\beta^r\beta^{r-1}}{2}\left\|\mathbf{A}\mathbf{X}^r\right\|_F^2+\tfrac{\beta^r\beta^{r-1}}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2+\tfrac{\beta^r}{d}\sum_{i=1}^N\left(\theta_i^r\left\|y_i^{r+1}-y_i^r\right\|^2+\tau\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2\right)\\
&+\tfrac{d\beta^r}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r\right\|_F^2-\tfrac{(\beta^r)^2}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r-\left(\mathbf{X}^r-\mathbf{X}^{r-1}\right)\right\|_{\mathbf{B}^T\mathbf{B}}^2+\tfrac{\left(\beta^{r+1}-\beta^{r-1}\right)}{2\beta^r}\left\|\mathbf{\Omega}^r-\mathbf{\Omega}^{r-1}\right\|_F^2\\
&+\tfrac{\beta^r\left(\beta^r-\beta^{r-1}\right)}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2
\end{aligned}\tag{101}
$$

Step 4. In this step, we construct and estimate the descent of the potential function. With $c>0$, let us define the potential function as

$$
\begin{aligned}
&P_{\beta^{r+1},c}\left(\mathbf{X}^{r+1},\mathbf{X}^r,Y^{r+1},\mathbf{\Omega}^{r+1}\right)\\
&=L_{\beta^{r+1}}\left(\mathbf{X}^{r+1},Y^{r+1},\mathbf{\Omega}^{r+1}\right)+\frac{c\beta^{r+1}\beta^r}{2}\left\|\mathbf{A}\mathbf{X}^{r+1}\right\|_F^2+\frac{c\beta^r\beta^{r+1}}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r\right\|_{\mathbf{B}^T\mathbf{B}}^2
\end{aligned}\tag{102}
$$

Combining the estimate in (94) and (101), we obtain

$$
\begin{aligned}
&P_{\beta^{r+1},c}\left(\mathbf{X}^{r+1},\mathbf{X}^r,Y^{r+1},\mathbf{\Omega}^{r+1}\right)-P_{\beta^r,c}\left(\mathbf{X}^r,\mathbf{X}^{r-1},Y^r,\mathbf{\Omega}^r\right)\\
&=L_{\beta^{r+1}}\left(\mathbf{X}^{r+1},Y^{r+1},\mathbf{\Omega}^{r+1}\right)-L_{\beta^r}\left(\mathbf{X}^r,Y^r,\mathbf{\Omega}^r\right)+\tfrac{c\beta^{r+1}\beta^r}{2}\left\|\mathbf{A}\mathbf{X}^{r+1}\right\|_F^2\\
&+\tfrac{c\beta^r\beta^{r+1}}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r\right\|_{\mathbf{B}^T\mathbf{B}}^2-\tfrac{c\beta^r\beta^{r-1}}{2}\left\|\mathbf{A}\mathbf{X}^r\right\|_F^2-\tfrac{c\beta^r\beta^{r-1}}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2\\
&\le\left(\tfrac{12\theta_i^r}{\beta^{r+1}\sigma_{\min}(A^TA)}-\tfrac{\beta^{r+1}\theta_i^r}{2}\right)\sum_{i=1}^N\left\|y_i^{r+1}-y_i^r\right\|^2+\left(\tfrac{12\tau}{\beta^{r+1}\sigma_{\min}(A^TA)}-\tfrac{1+\beta^{r+1}}{2}\right)\sum_{i=1}^N\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2\\
&+\tfrac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}(A^TA)}\left\|\mathbf{B}^T\mathbf{B}\left(\left(\mathbf{X}^{r+1}-\mathbf{X}^r\right)-\left(\mathbf{X}^r-\mathbf{X}^{r-1}\right)\right)\right\|_F^2\\
&+\left(\tfrac{12(\tau^2+4\gamma^2)}{\beta^{r+1}\sigma_{\min}(A^TA)}+\tfrac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\beta^{r+1}\sigma_{\min}(A^TA)}-\tfrac{\beta^{r+1}+2\gamma}{2}\right)\left\|\left(\mathbf{X}^{r+1}-\mathbf{X}^r\right)\right\|_F^2+\tfrac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\left\|\mathbf{\Omega}^r-\mathbf{\Omega}^{r-1}\right\|_F^2\\
&+\tfrac{c\beta^r}{d}\sum_{i=1}^N\left(\theta_i^r\left\|y_i^{r+1}-y_i^r\right\|^2+\tau\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2\right)+\tfrac{dc\beta^r}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r\right\|_F^2\\
&-\tfrac{c(\beta^r)^2}{2}\left\|\mathbf{X}^{r+1}-\mathbf{X}^r-\left(\mathbf{X}^r-\mathbf{X}^{r-1}\right)\right\|_{\mathbf{B}^T\mathbf{B}}^2+\tfrac{c\left(\beta^{r+1}-\beta^{r-1}\right)}{2\beta^r}\left\|\mathbf{\Omega}^r-\mathbf{\Omega}^{r-1}\right\|_F^2\\
&+\tfrac{c\beta^r\left(\beta^r-\beta^{r-1}\right)}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2\\
&\le\left(\tfrac{12\theta_i^r}{\beta^{r+1}\sigma_{\min}(A^TA)}-\tfrac{\beta^{r+1}\theta_i^r}{2}+\tfrac{c\beta^r\theta_i^r}{d}\right)\sum_{i=1}^N\left\|y_i^{r+1}-y_i^r\right\|^2\\
&+\left(\tfrac{12\tau}{\beta^{r+1}\sigma_{\min}(A^TA)}-\tfrac{1+\beta^{r+1}}{2}+\tfrac{c\beta^r\tau}{d}\right)\sum_{i=1}^N\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2\\
&+\left(\tfrac{3(\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\right\|}{\beta^{r+1}\sigma_{\min}(A^TA)}-\tfrac{c(\beta^r)^2}{2}\right)\left\|\left(\left(\mathbf{X}^{r+1}-\mathbf{X}^r\right)-\left(\mathbf{X}^r-\mathbf{X}^{r-1}\right)\right)\right\|_{\mathbf{B}^T\mathbf{B}}^2\\
&+\left(\tfrac{12(\tau^2+4\gamma^2)}{\beta^{r+1}\sigma_{\min}(A^TA)}+\tfrac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\beta^{r+1}\sigma_{\min}(A^TA)}-\tfrac{\beta^{r+1}+2\gamma}{2}+\tfrac{dc\beta^r}{2}\right)\left\|\left(\mathbf{X}^{r+1}-\mathbf{X}^r\right)\right\|_F^2\\
&+\left(\tfrac{\beta^{r+1}-\beta^{r-1}}{2(\beta^r)^2}+\tfrac{c\left(\beta^{r+1}-\beta^{r-1}\right)}{2\beta^r}\right)\left\|\mathbf{\Omega}^r-\mathbf{\Omega}^{r-1}\right\|_F^2+\tfrac{c\beta^r\left(\beta^r-\beta^{r-1}\right)}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2
\end{aligned}\tag{103}
$$

where the first inequality is obtained by plugging in (94) and (101); in the last inequality we have used the fact that $\beta^{r+1}\ge\beta^r$. Taking the sum of r from t to $T+1$ (for some $T>t>1$) and utilize the estimate in (91), we have

$$
\begin{aligned}
&P_{\beta^{T+1},c}\left(\mathbf{X}^{T+1},\mathbf{X}^{T},Y^{T+1},\mathbf{\Omega}^{T+1}\right)-P_{\beta^{t},c}\left(\mathbf{X}^{t},\mathbf{X}^{t-1},Y^{t},\mathbf{\Omega}^{t}\right)\\
&\leq\sum_{r=t}^{T}\left(\begin{array}{l}
\left(\frac{12\theta_i^r}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{\beta^{r+1}\theta_i^r}{2}+\frac{c\beta^r\theta_i^r}{d}\right)\sum_{i=1}^{N}\left\|y_i^{r+1}-y_i^r\right\|^2\\
+\left(\frac{12\tau}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{1+\beta^{r+1}}{2}+\frac{c\beta^r\tau}{d}\right)\sum_{i=1}^{N}\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2\\
+\left(\frac{3(\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\right\|}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{c(\beta^r)^2}{2}\right)\left\|\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right)\right\|^2_{\mathbf{B}^T\mathbf{B}}\\
+\left(\frac{12(\tau^2+4\gamma^2)}{\beta^{r+1}\sigma_{\min}(A^TA)}+\frac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{\beta^{r+1}+2\gamma}{2}+\frac{dc\beta^r}{2}\right)\left\|(\mathbf{X}^{r+1}-\mathbf{X}^r)\right\|_F^2\\
+\left(\frac{\beta^{r+1}-\beta^{r-1}}{2(\beta^r)^2}+\frac{c\left(\beta^{r+1}-\beta^{r-1}\right)}{2\beta^r}\right)\left\|\mathbf{\Omega}^r-\mathbf{\Omega}^{r-1}\right\|_F^2+\frac{c\beta^r\left(\beta^r-\beta^{r-1}\right)}{2}\left\|\mathbf{X}^r-\mathbf{X}^{r-1}\right\|^2_{\mathbf{B}^T\mathbf{B}}
\end{array}\right)\\
&\leq\sum_{r=t}^{T}\left(\begin{array}{l}
\left(\frac{12\theta_i^r}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{\beta^{r+1}\theta_i^r}{2}+\frac{c\beta^r\theta_i^r}{d}+\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{12\theta_i^r}{\sigma_{\min}(A^TA)}\right)\sum_{i=1}^{N}\left\|y_i^{r+1}-y_i^r\right\|^2\\
+\left(\frac{12\tau}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{1+\beta^{r+1}}{2}+\frac{c\beta^r\tau}{d}+\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{12\tau}{\sigma_{\min}(A^TA)}\right)\sum_{i=1}^{N}\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2\\
+\left(\begin{array}{l}\frac{3(\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\right\|}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{c(\beta^r)^2}{2}\\+\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{3(\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\right\|}{\sigma_{\min}(A^TA)}\end{array}\right)\left\|\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right)\right\|^2_{\mathbf{B}^T\mathbf{B}}\\
+\left(\begin{array}{l}\frac{12(\tau^2+4\gamma^2)}{\beta^{r+1}\sigma_{\min}(A^TA)}+\frac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{\beta^{r+1}+2\gamma}{2}\\+\frac{dc\beta^r}{2}+\frac{c\beta^{r+1}\left(\beta^{r+1}-\beta^r\right)\left\|\mathbf{B}^T\mathbf{B}\right\|}{2}+\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{12(\tau^2+4\gamma^2)}{\sigma_{\min}(A^TA)}\\+\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\sigma_{\min}(A^TA)}\end{array}\right)\left\|(\mathbf{X}^{r+1}-\mathbf{X}^r)\right\|_F^2
\end{array}\right)\\
&+\frac{c\beta^t\left(\beta^t-\beta^{t-1}\right)}{2}\left\|\mathbf{X}^t-\mathbf{X}^{t-1}\right\|^2_{\mathbf{B}^T\mathbf{B}}+\left(\frac{\beta^{t+1}-\beta^{t-1}}{2(\beta^t)^2}+\frac{c\left(\beta^{t+1}-\beta^{t-1}\right)}{2\beta^t}\right)\left\|\mathbf{\Omega}^t-\mathbf{\Omega}^{t-1}\right\|_F^2.
\end{aligned}
\tag{104}
$$

It can be observed that the coefficient in front of $\left\|\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right)\right\|^2_{\mathbf{B}^T\mathbf{B}}$ becomes negative for sufficiently large (but finite) t.

Suppose that r is large enough such that $\frac{12(\tau^2+4\gamma^2)}{\beta^{r+1}\sigma_{\min}(A^TA)}+\frac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{\beta^{r+1}+2\gamma}{2}<-\frac{\beta^{r+1}}{3}$ and choose $c=\min\{1/(2d),1/(12\kappa\|B^TB\|)\}$, then we have

$$
\begin{aligned}
&\frac{12(\tau^2+4\gamma^2)}{\beta^{r+1}\sigma_{\min}(A^TA)}+\frac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{\beta^{r+1}+2\gamma}{2}+\frac{dc\beta^r}{2}+\frac{c\beta^{r+1}\left(\beta^{r+1}-\beta^r\right)\left\|\mathbf{B}^T\mathbf{B}\right\|}{2}\\
&<-\frac{\beta^{r+1}}{3}+\frac{\beta^r}{4}+\frac{\beta^r}{24}=-\frac{\beta^{r+1}}{24}
\end{aligned}
$$

We can also show that for sufficiently large r, the following is true

$$
\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{12(\tau^2+4\gamma^2)}{\sigma_{\min}(A^TA)}+\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{3\left(\beta^{r+1}-\beta^r\right)^2\left\|\mathbf{B}^T\mathbf{B}\right\|_F^2}{\sigma_{\min}(A^TA)}\leq\frac{\beta^{r+1}}{48},
$$

$$
\frac{3(\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\right\|}{\beta^{r+1}\sigma_{\min}(A^TA)}-\frac{c(\beta^r)^2}{2}+\left(\frac{\beta^{r+2}-\beta^r}{2\beta^{r+1}}\right)\left(\frac{1}{\beta^{r+1}}+c\right)\frac{3(\beta^r)^2\left\|\mathbf{B}^T\mathbf{B}\right\|}{\sigma_{\min}(A^TA)}\leq-\frac{c(\beta^r)^2}{48}.
$$

Furthermore, if we choose $d=\max\{1,\sqrt{\tau}\}$, the coefficients in front of the following terms would be negative for sufficiently large r

$$
\sum_{i=1}^{N}\left\|y_i^{r+1}-y_i^r\right\|^2,\ \sum_{i=1}^{N}\left\|X_i^r\left(y_i^{r+1}-y_i^r\right)\right\|^2,\ \left\|\left((\mathbf{X}^{r+1}-\mathbf{X}^r)-(\mathbf{X}^r-\mathbf{X}^{r-1})\right)\right\|^2_{\mathbf{B}^T\mathbf{B}},\ \left\|(\mathbf{X}^{r+1}-\mathbf{X}^r)\right\|_F^2.
$$

In conclusion, we have that for sufficiently large but finite t_0, we have

$$
\begin{aligned}
&P_{\beta^{T+1},c}\left(\mathbf{X}^{T+1},\mathbf{X}^{T},Y^{T+1},\mathbf{\Omega}^{T+1}\right)-P_{\beta^{t_0},c}\left(\mathbf{X}^{t},\mathbf{X}^{t-1},Y^{t},\mathbf{\Omega}^{t}\right)\\
&\leq \sum_{t=t_0}^{T} -\frac{\beta^{r+1}}{48}\left\|(\mathbf{X}^{r+1}-\mathbf{X}^{r})\right\|_F^2-\frac{c(\beta^{r})^2}{48}\left\|\left((\mathbf{X}^{r+1}-\mathbf{X}^{r})-(\mathbf{X}^{r}-\mathbf{X}^{r-1})\right)\right\|_{\mathbf{B}^T\mathbf{B}}^2\\
&+\frac{c\beta^{t_0}\left(\beta^{t_0}-\beta^{t_0-1}\right)}{2}\left\|\mathbf{X}^{t_0}-\mathbf{X}^{t_0-1}\right\|_{\mathbf{B}^T\mathbf{B}}^2+\left(\frac{\beta^{t_0+1}-\beta^{t_0-1}}{2(\beta^{t_0})^2}+\frac{c\left(\beta^{t_0+1}-\beta^{t_0-1}\right)}{2\beta^{t_0}}\right)\left\|\mathbf{\Omega}^{t_0}-\mathbf{\Omega}^{t_0-1}\right\|_F^2.
\end{aligned}
$$

Step 5. In this step, we show that the potential function must be lower bounded. Observe that the augmented Lagrangian is given by

$$
\begin{aligned}
&L_{\beta^{r+1}}\left(\mathbf{X}^{r+1},Y^{r+1},\mathbf{\Omega}^{r+1}\right)\\
&=f\left(\mathbf{X}^{r+1},Y^{r+1}\right)+\left\langle\mathbf{\Omega}^{r+1},\mathbf{A}\mathbf{X}^{r+1}\right\rangle+\frac{\beta^{r+1}}{2}\left\langle\mathbf{A}\mathbf{X}^{r+1},\mathbf{A}\mathbf{X}^{r+1}\right\rangle\\
&=f\left(\mathbf{X}^{r+1},Y^{r+1}\right)+\left\langle\mathbf{\Omega}^{r+1},\frac{\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^{r}}{\beta^{r+1}}\right\rangle+\frac{\beta^{r+1}}{2}\left\langle\mathbf{A}\mathbf{X}^{r+1},\mathbf{A}\mathbf{X}^{r+1}\right\rangle\\
&=f\left(\mathbf{X}^{r+1},Y^{r+1}\right)+\frac{1}{2\beta^{r+1}}\left(\left\|\mathbf{\Omega}^{r+1}\right\|_F^2-\left\|\mathbf{\Omega}^{r}\right\|_F^2+\left\|\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^{r}\right\|_F^2\right)+\frac{\beta^{r+1}}{2}\left\langle\mathbf{A}\mathbf{X}^{r+1},\mathbf{A}\mathbf{X}^{r+1}\right\rangle\\
&=f\left(\mathbf{X}^{r+1},Y^{r+1}\right)+\frac{1}{2\beta^{r+1}}\left\|\mathbf{\Omega}^{r+1}\right\|_F^2-\frac{1}{2\beta^{r}}\left\|\mathbf{\Omega}^{r}\right\|_F^2+\left(\frac{1}{2\beta^{r}}-\frac{1}{2\beta^{r+1}}\right)\left\|\mathbf{\Omega}^{r}\right\|_F^2+\frac{1}{2\beta^{r+1}}\left\|\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^{r}\right\|_F^2\\
&+\frac{\beta^{r+1}}{2}\left\langle\mathbf{A}\mathbf{X}^{r+1},\mathbf{A}\mathbf{X}^{r+1}\right\rangle\\
&\geq f\left(\mathbf{X}^{r+1},Y^{r+1}\right)+\frac{1}{2\beta^{r+1}}\left\|\mathbf{\Omega}^{r+1}\right\|_F^2-\frac{1}{2\beta^{r}}\left\|\mathbf{\Omega}^{r}\right\|_F^2+\frac{1}{2\beta^{r+1}}\left\|\mathbf{\Omega}^{r+1}-\mathbf{\Omega}^{r}\right\|_F^2+\frac{\beta^{r+1}}{2}\left\langle\mathbf{A}\mathbf{X}^{r+1},\mathbf{A}\mathbf{X}^{r+1}\right\rangle
\end{aligned}
$$

Thus, following the similar argument in the Step 5 of the proof of Theorem 2, we conclude that the potential function $L_{\beta^{r+1}}\left(\mathbf{X}^{r+1},Y^{r+1},\mathbf{\Omega}^{r+1}\right)$ is lower bounded for all r.

Now that we have shown the descent and the lower boundedness of the potential function, the rest of the proof follows the same arguments as Step 6 - Step 7 of Theorem 2, therefore is omitted.

References

Allen-Zhu, Z. and Hazan, E. Variance Reduction for Faster Non-Convex Optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML, 2016.

Antoniadis, A., Gijbels, I., and Nikolova, M. Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 63(3):585–615, 2009.

Aybat, N-S. and Hamedani, E-Y. A primal-dual method for conic constrained distributed optimization problems. *Advances in Neural Information Processing Systems*, 2016.

Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Method*. Academic Press, 1982.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

Bianchi, P. and Jakubowicz, J. Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE Transactions on Automatic Control*, 58(2):391–405, 2013.

Bjornson, E. and Jorswieck, E. Optimal resource allocation in coordinated multi-cell systems. *Foundations and Trends in Communications and Information Theory*, 9, 2013.

Cevher, V., Becker, S., and Schmidt, M. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43, 2014.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *The Proceeding of NIPS*, 2014.

Forero, P. A., Cano, A., and Giannakis, G. B. Consensus-based distributed support vector machines. *Journal of Machine Learning Research*, 11(May):1663–1707, 2010.

Giannakis, G. B., Ling, Q., Mateos, G., Schizas, I. D., and Zhu, H. Decentralized learning for wireless communications and networking. *arXiv preprint arXiv:1503.08855*, 2015.

H.-T. Wai, T.-H. Chang and Scaglione, A. A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning. In *the Proceedings of the IEEE ICASSP*, 2015.

Hajinezhad, D., Hong, M., Zhao, T., and Wang, Z. Nestt: A nonconvex primal-dual splitting method for distributed and stochastic optimization. In *Advances in Neural Information Processing Systems 29*, pp. 3215–3223. 2016.

Hong, M., Luo, Z.-Q., and Razaviyayn, M. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. 2014. technical report, University of Minnesota.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *the Proceedings of the Neural Information Processing (NIPS)*. 2013.

Ling, Q., Xu, Y., Yin, W., and Wen, Z. Decentralized low-rank matrix completion. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2925–2928, March 2012.

Ling, Q., Shi, W., Wu, G., and Ribeiro, A. DLM: Decentralized linearized alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(15):4051–4064, Aug 2015.

Lobel, I. and Ozdaglar, A. Distributed subgradient methods for convex optimization over random networks. *Automatic Control, IEEE Transactions on*, 56(6):1291–1306, 2011.

Lobel, I., Ozdaglar, A., and Feijer, D. Distributed multi-agent optimization with state-dependent communication. *Mathematical Programming*, 129(2):255–284, 2011.

Lorenzo, P. Di and Scutari, G. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Nedic, A. and Olshevsky, A. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.

Nedic, A. and Ozdaglar, A. Cooperative distributed multi-agent optimization. In *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2009a.

Nedic, A. and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009b.

Powell, M. M. D. An efficient method for nonlinear constraints in minimization problems. In *Optimization*. Academic Press, 1969.

Rahmani, M. and Atia, G. A decentralized approach to robust subspace recovery. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 802–807. IEEE, 2015.

Reddi, S., Sra, S., Poczos, B., and Smola, A. Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*, 2016.

Scardapane, S. and Lorenzo, P. Di. A framework for parallel and distributed training of neural networks. *arXiv preprint arXiv:1610.07448*, 2016.

Scardapane, S., Fierimonte, R., Lorenzo, P. Di, Panella, M., and Uncini, A. Distributed semi-supervised support vector machines. *Neural Networks*, 80:43–52, 2016.

Schmidt, M., Roux, N. Le, and Bach., F. Minimizing finite sums with the stochastic average gradient. 2013. Technical report, INRIA.

Shi, W., Ling, Q., Wu, G., and Yin, W. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2014.

Tatarenko, T. and Touri, B. Non-convex distributed optimization. 2015. arXiv Preprint: arXiv:1512.00895.

Yan, F., Sundaram, S., Vishwanathan, S. V. N., and Qi, Y. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2013.

Zhu, M. and Martinez, S. An approximate dual subgradient algorithm for multi-agent non-convex optimization. In *49th IEEE Conference on Decision and Control (CDC)*, pp. 7487–7492, 2010.

Zlobec, S. On the liu–floudas convexification of smooth programs. *Journal of Global Optimization*, 32(3):401–407, 2005.