Communication Efficient Primal-Dual Algorithm for Nonconvex Nonsmooth Distributed Optimization

Congliang Chen[1,2]
congliangchen@link.cuhk.edu.cn

Jiawei Zhang[1,2]
216019001@link.cuhk.edu.cn

Li Shen[3]
mathshenli@gmail.com

Peilin Zhao[3]
masonzhao@tencent.com

Zhi-Quan Luo[1,2]
luozq@cuhk.edu.cn

[1]The Chinese University of Hong Kong, Shenzhen [2]Shenzhen Research Institute of Big Data [3]Tencent AI Lab

1 PROOF OF THEOREM 1

Notation 1. *Define*

$$
\begin{aligned}
&g_i\left(x_i, z_i\right) = f_i\left(x_i\right) + h_i\left(x_i\right) + \frac{Np}{2}\|x_i - z_i\|^2, \\
&g\left(x, z\right) = \frac{1}{N}\sum_{i=1}^{N} g_i\left(x_i, z_i\right), \\
&d\left(y, z\right) = \min_{x\in\mathbb{R}^{nN}} K\left(x, y, z\right) + h\left(x\right), \\
&x\left(y, z\right) = \arg\min_{x\in\mathbb{R}^{nN}} K\left(x, y, z\right) + h\left(x\right), \\
&M\left(z\right) = \min_{x\in\mathbb{R}^{nN}, Ax=0} \left(f\left(x\right) + h\left(x\right) + \frac{p}{2}\|x - z\|^2\right), \\
&x^*\left(z\right) = \arg\min_{x\in\mathbb{R}^{nN},\ Ax=0} \left(f\left(x\right) + h\left(x\right) + \frac{p}{2}\|x - z\|^2\right), \\
&e^t = x^t - \hat{x}^t.
\end{aligned}
$$

First, we bound the compression error e^t in the following lemma.

Lemma 1. *The following equality always holds:*

$$
\sum_{t=1}^{T}\|e^t\|^2 \le \frac{\left(1-\delta\right)^2}{\delta^2}\sum_{t=1}^{T}\|x^t - x^{t-1}\|^2.
$$

Proof. Using the definition of e_t, we can obtain:

$$
\begin{aligned}
\|e^t\| = \|x^t - \hat{x}^t\| &= \|x^t - \hat{x}^{t-1} - C\left(x^t - \hat{x}^{t-1}\right)\| \\
&\le \left(1-\delta\right)\|x^t - \hat{x}^{t-1}\| = \left(1-\delta\right)\|x^t - x^{t-1} + \left(x^{t-1} - \hat{x}^{t-1}\right)\| \\
&\le \left(1-\delta\right)\|x^t - x^{t-1}\| + \left(1-\delta\right)\|e^{t-1}\|,
\end{aligned}
$$

where the first inequality is due to the assumption on the compression function.

By the induction, we can get

$$
\|e^t\| \le \sum_{i=1}^{t}\left(1-\delta\right)^{t-i+1}\|x^i - x^{i-1}\|.
$$

Then, using the convexity of square function and rearranging the summation terms, we can obtain

$$
\begin{aligned}
\sum_{t=1}^{T}\|e^t\|^2 &\le \sum_{t=1}^{T}\left(\sum_{i=1}^{t}\left(1-\delta\right)^{t-i+1}\|x^i - x^{i-1}\|\right)^2 \\
&\le \sum_{t=1}^{T}\left(\sum_{i=1}^{t}\left(1-\delta\right)^{t-i+1}\right)\sum_{i=1}^{t}\left(1-\delta\right)^{t-i+1}\|x^i - x^{i-1}\|^2 \\
&\le \frac{1-\delta}{\delta}\sum_{t=1}^{T}\sum_{i=1}^{t}\left(1-\delta\right)^{t-i+1}\|x^i - x^{i-1}\|^2 \\
&\le \frac{1-\delta}{\delta}\sum_{i=1}^{T}\left(\sum_{t=i}^{T}\left(1-\delta\right)^{t-i+1}\right)\|x^i - x^{i-1}\|^2 \\
&\le \frac{\left(1-\delta\right)^2}{\delta^2}\sum_{t=1}^{T}\|x^t - x^{t-1}\|^2.
\end{aligned}
$$

The proof is finished. □

Then, we give the lower bound on the change of primal function when updating iterates.

Lemma 2. *For any $t \geq 0$, if $c \leq \frac{1}{L_K}$, the following inequality always holds:*

$$\begin{aligned}
&h\left(x^t\right) - \left\langle \nabla_x K\left(x^t, y^t, z^t\right), x^{t+1} - x^t \right\rangle - h\left(x^{t+1}\right) - \frac{1}{2c}\left\|x^{t+1} - x^t\right\|^2 \\
&\geq \frac{1}{2c}\|x^{t+1} - x^t\|^2 - \alpha\left(A\hat{x}^{t+1}\right)^T A x^{t+1} + \frac{p}{2\beta}\|z^t - z^{t+1}\|^2.
\end{aligned}$$

Proof. Recall the update iteration of x^{t+1}:

$$x^{t+1} = \arg\min_x \left(\left\langle \nabla_x K\left(x^t, y^t, z^t\right), x - x^t \right\rangle + h(x) + \frac{1}{2c}\left\|x - x^t\right\|^2 \right).$$

By the optimality condition of strongly convex function, we obtain

$$h\left(x^t\right) - \left\langle \nabla_x K\left(x^t, y^t, z^t\right), x^{t+1} - x^t \right\rangle - h\left(x^{t+1}\right) - \frac{1}{2c}\left\|x^{t+1} - x^t\right\|^2 \geq \frac{1}{2c}\|x^{t+1} - x^t\|^2.$$

Besides, because $K(x, z, y)$ has Lipschitz gradient with respect to x, we can obtain

$$\begin{aligned}
&K\left(x^t, y^t, z^t\right) + h\left(x^t\right) - K\left(x^{t+1}, y^t, z^t\right) - h\left(x^{t+1}\right) \\
&\geq h\left(x^t\right) - h\left(x^{t+1}\right) - \left\langle \nabla_x K\left(x^t, y^t, z^t\right), x^{t+1} - x^t \right\rangle - \frac{L_K}{2}\left\|x^{t+1} - x^t\right\|^2 \\
&\geq h\left(x^t\right) - h\left(x^{t+1}\right) - \left\langle \nabla_x K\left(x^t, y^t, z^t\right), x^{t+1} - x^t \right\rangle - \frac{1}{2c}\left\|x^{t+1} - x^t\right\|^2 \\
&\geq \frac{1}{2c}\|x^{t+1} - x^t\|^2.
\end{aligned} \tag{1}$$

Next, according to the update of y^{t+1}, i.e. $y^{t+1} = y^t + \alpha A x^{t+1}$, we can obtain

$$K\left(x^{t+1}, y^t, z^t\right) - K\left(x^{t+1}, y^{t+1}, z^t\right) = -\alpha\left(A\hat{x}^{t+1}\right)^T A x^{t+1}. \tag{2}$$

In addition, by using the update of z^{t+1}, i.e., $z^{t+1} = z^t + \beta\left(\hat{x}^{t+1} - z^t\right)$, we can obtain

$$\begin{aligned}
x^{t+1} - z^t &= \frac{1}{\beta}\left(z^{t+1} - z^t\right), \\
x^{t+1} - z^{t+1} &= \frac{1-\beta}{\beta}\left(z^{t+1} - z^t\right).
\end{aligned}$$

By using above two equalities, we have

$$\begin{aligned}
&K\left(x^{t+1}, y^{t+1}, z^t\right) - K\left(x^{t+1}, y^{t+1}, z^{t+1}\right) \\
&= \frac{p}{2}\left(\|x^{t+1} - z^t\|^2 - \|x^{t+1} - z^{t+1}\|^2\right) \\
&= \frac{p}{2}\left(z^{t+1} - z^t\right)^T\left(\left(x^{t+1} - z^t\right) + \left(x^{t+1} - z^{t+1}\right)\right) \\
&= \frac{p}{2}\left(2/\beta - 1\right)\|z^t - z^{t+1}\|^2 \\
&\geq \frac{p}{2\beta}\|z^t - z^{t+1}\|^2.
\end{aligned} \tag{3}$$

Then, by combining inequalities (1), (2) and (3), we obtain the desired result. □

To bound the dual function and proximal function, we first give the bound on difference of dual function and proximal function during update in the following Lemma.

Lemma 3. *Suppose $p > -\frac{L}{N}$, then for any $y, y' \in \mathbb{R}^{nM}$, the following inequality holds:*

$$\|y - y'\| \geq \sigma_4\|x(y, z) - x(y', z)\|, \tag{4}$$

where $\sigma_4 = \frac{(Np-L)}{N\sqrt{\lambda_1}}$.

Proof. First, we define $\hat{K}\left(x,y,z\right)=K\left(x,y,z\right)+h\left(x\right)$. Note that $\hat{K}\left(x,y,z\right)$ is a strongly convex function with respect to x. Then, it holds that

$$\begin{aligned}
&\hat{K}\left(x\left(y,z\right),y',z\right)-\hat{K}\left(x\left(y',z\right),y',z\right)\\
&=\hat{K}\left(x\left(y,z\right),y,z\right)-\hat{K}\left(x\left(y',z\right),y,z\right)-\left(\hat{K}\left(x\left(y',z\right),y',z'\right)-\hat{K}\left(x\left(y',z\right),y,z\right)\right)\\
&\quad+\left(\hat{K}\left(x\left(y,z\right),y',z\right)-\hat{K}\left(x\left(y,z\right),y,z\right)\right)\\
&=\hat{K}\left(x\left(y,z\right),y,z\right)-\hat{K}\left(x\left(y',z\right),y,z\right)-\left(y'-y\right)^{T}Ax\left(y',z\right)+\left(y'-y\right)^{T}Ax\left(y,z\right)\\
&\leq-\frac{(Np-L)}{2N}\|x\left(y',z\right)-x\left(y,z\right)\|^{2}+\left(y'-y\right)^{T}A\left(x\left(y,z\right)-x\left(y',z\right)\right)\\
&\leq-\frac{(Np-L)}{2N}\|x\left(y',z\right)-x\left(y,z\right)\|^{2}+\sqrt{\lambda_{1}}\|y'-y\|\|x\left(y,z\right)-x\left(y',z\right)\|.
\end{aligned}$$

Then, using the strongly convexity of $\hat{K}\left(x,y,z\right)$ on x with modular $\frac{Np-L}{N}$, we obtain

$$\hat{K}\left(x\left(y,z\right),y',z\right)-\hat{K}\left(x\left(y',z\right),y',z\right)\geq\frac{Np-L}{2N}\|x\left(y,z\right)-x\left(y',z\right)\|^{2}.$$

Then, combining the above two inequalities, we can obtain

$$\|y-y'\|\geq\frac{(Np-L)}{N\sqrt{\lambda_{1}}}\|x\left(y,z\right)-x\left(y',z\right)\|.$$

The proof is finished. ∎

Lemma 4. *Suppose $p>-\frac{L}{N}$, then for any $z,z'\in\mathbb{R}^{nN}$, the following inequalities hold:*

$$\|z-z'\|\geq\sigma_{5}\|x^{*}\left(z\right)-x^{*}\left(z'\right)\|, \tag{5}$$
$$\|z-z'\|\geq\sigma_{5}\|x\left(y,z\right)-x\left(y,z'\right)\|, \tag{6}$$

where $\sigma_{5}=\frac{Np-L}{Np}$.

Proof. According to the strongly convexity of function g, we can obtain:

$$\begin{aligned}
&g\left(x^{*}\left(z\right),z'\right)-g\left(x^{*}\left(z'\right),z'\right)\\
&=g\left(x^{*}\left(z\right),z\right)-g\left(x^{*}\left(z'\right),z\right)-\left(g\left(x^{*}\left(z'\right),z'\right)-g_{i}\left(x^{*}\left(z'\right),z\right)\right)\\
&\quad+\left(g\left(x^{*}\left(z\right),z'\right)-g\left(x^{*}\left(z\right),z\right)\right)\\
&=\left(g\left(x^{*}\left(z\right),z\right)-g\left(x^{*}\left(z'\right),z\right)\right)-\frac{p}{2}\left(-2\left(z'-z\right)^{T}x^{*}\left(z'\right)+\|z'\|^{2}-\|z\|^{2}\right)\\
&\quad+\frac{p}{2}\left(-2\left(z'-z\right)^{T}x^{*}\left(z\right)+\|z'\|^{2}-\|z\|^{2}\right)\\
&=\left(g\left(x^{*}\left(z\right),z\right)-g\left(x^{*}\left(z'\right),z\right)\right)+p\left(z'-z\right)^{T}\left(x^{*}\left(z'\right)-x^{*}\left(z\right)\right)\\
&\leq-\frac{Np-L}{2N}\|x^{*}\left(z\right)-x^{*}\left(z'\right)\|^{2}+p\left(z'-z\right)^{T}\left(x^{*}\left(z'\right)-x^{*}\left(z\right)\right).
\end{aligned}$$

On the other hand, using the strongly convexity of g, it holds that

$$g\left(x^{*}\left(z\right),z'\right)-g\left(x^{*}\left(z'\right),z'\right)\geq\frac{Np-L}{2N}\|x^{*}\left(z\right)-z^{*}\left(z'\right)\|^{2}.$$

Hence, we have

$$p\left(z'-z\right)^{T}\left(x^{*}\left(z'\right)-x^{*}\left(z\right)\right)\geq\frac{Np-L}{N}\|x^{*}\left(z\right)-x^{*}\left(z'\right)\|^{2}.$$

Further, according to Cauchy-Schwarz inequality, it implies that

$$\|x^{*}\left(z\right)-x^{*}\left(z'\right)\|\leq\frac{Np}{Np-L}\|z-z'\|.$$

With the same proof as it is used to prove (5), we directly get

$$\|x\left(y,z\right)-x\left(y,z'\right)\| \leq \frac{Np}{Np-L}\|z-z'\|.$$

□

With the above two lemmas, we give the bound on the difference of the dual function and the proximal function during the algorithm, as follows.

Lemma 5. *For any $t \geq 0$, it holds that*

$$\begin{aligned}
&d\left(y^{t+1},z^{t+1}\right)-d\left(y^{t},z^{t}\right)\\
&\geq \alpha\left(A\hat{x}^{t+1}\right)^{T}Ax\left(y^{t},z^{t+1}\right)-\frac{\alpha^{2}\sqrt{\lambda_{1}}}{2\sigma_{4}}\|A\hat{x}^{t+1}\|^{2}+\frac{p}{2}\left(z^{t+1}-z^{t}\right)^{T}\left(z^{t+1}+z^{t}-2x\left(y^{t},z^{t+1}\right)\right).
\end{aligned}$$

Proof. First, with the definition of $x(y^t, z^{t+1})$, we have

$$\begin{aligned}
&d\left(y^{t},z^{t+1}\right)-d\left(y^{t},z^{t}\right)\\
&=K\left(x\left(y^{t},z^{t+1}\right),y^{t},z^{t+1}\right)+h\left(x\left(y^{t},z^{t+1}\right)\right)-K\left(x\left(y^{t},z^{t}\right),y^{t},z^{t}\right)+h\left(x\left(y^{t},z^{t}\right)\right)\\
&\geq K\left(x\left(y^{t},z^{t+1}\right),y^{t},z^{t+1}\right)-K\left(x\left(y^{t},z^{t+1}\right),y^{t},z^{t}\right)\\
&=\frac{p}{2}\left(\|x\left(y^{t},z^{t+1}\right)-z^{t+1}\|^{2}-\|x\left(y^{t},z^{t+1}\right)-z^{t}\|^{2}\right)\\
&=\frac{p}{2}\left(z^{t+1}-z^{t}\right)^{T}\left(z^{t+1}+z^{t}-2x\left(y^{t},z^{t+1}\right)\right).
\end{aligned} \tag{7}$$

Besides, we can compute the gradient of $d\left(y,z\right)$ as

$$\nabla_{y}d\left(y,z\right)=Ax\left(y,z\right).$$

Then, for any y, y', we have

$$\|\nabla_{y}d\left(y,z\right)-\nabla_{y}d\left(y',z\right)\|=\|Ax\left(y,z\right)-Ax\left(y',z\right)\|\leq\left(\sqrt{\lambda_{1}}/\sigma_{4}\right)\|y-y'\|,$$

which is equivalent to say that $d(y, z)$ has Lipschitz gradient with respect to y with Lipschitz constant $\sqrt{\lambda_1}/\sigma_4$. According to the gradient Lipschitz continuity of $d(y, z)$, it holds

$$\begin{aligned}
&d\left(y^{t+1},z^{t+1}\right)-d\left(y^{t},z^{t+1}\right)\\
&\geq\langle y^{t+1}-y^{t},Ax\left(y^{t},z^{t+1}\right)\rangle-\frac{\sqrt{\lambda_{1}}}{2\sigma_{4}}\|y^{t+1}-y^{t}\|^{2}\\
&\geq\alpha\left(A\hat{x}^{t+1}\right)^{T}Ax\left(y^{t},z^{t+1}\right)-\frac{\alpha^{2}\sqrt{\lambda_{1}}}{2\sigma_{4}}\|A\hat{x}^{t+1}\|^{2}.
\end{aligned} \tag{8}$$

Combining (7) and (8), we get the desired result. □

Lemma 6. *For any $t \geq 0$, it holds that*

$$M\left(z^{t+1}\right)-M\left(z^{t}\right)\leq p\left(z^{t+1}-z^{t}\right)^{T}\left(z^{t}-x^{*}\left(z^{t}\right)\right)+\frac{p\tilde{L}}{2}\|z^{t}-z^{t+1}\|^{2},$$

where $\tilde{L} = \frac{Np}{Np-L} + 1$.

Proof. Recall the definition of $M\left(z\right)$, we can compute the gradient of $M\left(z\right)$ as follows:

$$\nabla M\left(z\right)=p\left(z-x^{*}\left(z\right)\right).$$

Then using Lemma 4, we can obtain

$$\begin{aligned}\|\nabla M(z)-\nabla M(z')\| &= \|p(z-x^*(z'))-p(z-x^*(z'))\|\\ &\le p(\|z-z'\|+\|x^*(z)-x^*(z')\|)\\ &\le p(1+1/\sigma_5)\|z-z'\|.\end{aligned}$$

Therefore, $M(z)$ is a gradient Lipschitz continuous function with Lipschitz constant $p(1+1/\sigma_5)$. Then, the result directly holds. □

The following lemmas give the dual error bound and primal approximation error bound.

Lemma 7. *For any $y \in \mathbb{R}^{nM}$, if $Ax(y,z)=r$, then $x(y,z)=\arg\min_{x:Ax=r} g(x,z)$.*

Proof. Recall the definition of $x(y,z)$:

$$x(y,z)=\arg\min_x\{g(x,z)+y^TAx\}.$$

Together with $Ax(y,z)=r$, $x(y,z)$ satisfies the optimality condition of the optimization problem. □

Therefore, we define $x^*(r,z)=x(y,z)$, if $Ax(y,z)=r$.

Lemma 8. *For any $r \in$ **Range** (A) there exists unique $v_i(r)$, $i=2,3,\cdots,N$, such that for any x that satisfies $Ax=r$. In addition, it holds that*

$$x_i = x_1+v_i(r).$$

Moreover, defining $v(r)=(v_1(r)^T,v_2(r)^T,\cdots,v_N(r)^T)^T$, it holds that

$$\|v(r)\| \le \frac{1}{\lambda_3}\|r\|,$$

for some $\lambda_3>0$.

Proof. Based on the construction of matrix A, it can be easily verify that when G is connected, $rank\left(\tilde{A}\right)=n(N-1)$.

Besides, for equation $Ax=r$, we can give the solution set as follows:

$$x_i = v_i(r)+b,$$

where b is an arbitrary vector in $\mathbb{R}^n$, and v is a solution with $v_1=0$.

Therefore, for any vector x that satisfies $Ax=r$, it can be written as $x_i=x_1+v_i(r)$, $i=1,2,3,\cdots,N$.

In addition, we can solve v by solving equation $\tilde{A}v(r)=\left(r^T,0_n^T\right)^T$, where

$$\tilde{A}=\begin{bmatrix} \multicolumn{2}{c}{A} \\ I_n & 0_{n\times n(N-1)}\end{bmatrix}.$$

When G is connected, it can be easily verified that $rank\left(\tilde{A}\right)=nN$, then $v(r)$ is unique.

Let $\tilde{W}$ be the matrix generated by removing the first column of W and $\tilde{v(r)}=(v_2^(r)^T,v_3(r)^T,\cdots,v_N(r)^T)^T$.

Then, because of the connectivity of the graph, $rank(\tilde{W})=N-1$, which is a full column rank matrix.

Because $v_1=0$, it holds that $(W\otimes I_n)v=(\tilde{W}\otimes I_n)\tilde{v}=r$ and $\|v\|=\|\tilde{v}\|$.

In addition, we define $\tilde{\mathcal{A}}=\tilde{W}^T\tilde{W}$, which is equivalent to remove the first column and first row of W^TW, and we define λ_3 as the smallest eigenvalue of $\tilde{\mathcal{A}}$.

We can obtain

$$\sqrt{\lambda_3}\|v(r)\| = \sqrt{\lambda_3}\|\tilde{v}\left(r\right)\| \leq \|(\tilde{W} \otimes I_n)\tilde{v}\left(r\right)\| = \|r\|.$$

Hence, we get

$$\|v(r)\| \leq \frac{1}{\sqrt{\lambda_3}}\|r\|,$$

for $\lambda_3 > 0$. □

Lemma 9. *Suppose $p > -L/N$, then for any $y \in \mathbb{R}^{nM}$ and $z \in \mathbb{R}^{nN}$, it holds that*

$$\|x\left(y,z\right) - x^*\left(z\right)\| \leq \sigma_1\|Ax(y,z)\|,$$

where $\sigma_1 = \frac{\left(3Np-L+\sqrt{N^2p^2-L^2}\right)}{2\sqrt{\lambda_3}(Np-L)}$.

Proof. Let $\Psi\left(u,z\right) = \frac{1}{N}\sum_{i=1}^{N} g_i\left(u,z_i\right),\ \Psi_r\left(u\right) = \frac{1}{N}\sum_{i=1}^{N} g_i\left(u + v_i\left(r\right), z_i\right)$.

Then, because of the uniqueness of $v_i\left(r\right)$, it is straightforward to obtain

$$\begin{aligned}
x_1^*\left(r,z\right) &= \arg\min_u \Psi_r\left(u\right)\\
x_1^*\left(z\right) &= \arg\min_u \Psi\left(u\right).
\end{aligned}$$

Because $p > -L/N$, $g_i(u, z_i)$ is a strongly convex function with modular $-L + Np$. Thus Ψ and Ψ_r are strongly convex. Using the strong convexity of Ψ and Ψ_r, we obtain

$$\begin{aligned}
\Psi\left(x_1^*\left(r,z\right)\right) - \Psi\left(x_1^*\left(z\right)\right) &\geq \left(-L + Np\right)\|x_1^*\left(r,z\right) - x_1^*\left(z\right)\|^2\\
\Psi_r\left(x_1^*\left(z\right)\right) - \Psi_r\left(x_1^*\left(r,z\right)\right) &\geq \left(-L + Np\right)\|x_1^*\left(r,z\right) - x_1^*\left(z\right)\|^2.
\end{aligned}$$

Combining the above two inequality we can obtain

$$\begin{aligned}
&2\left(-L + Np\right)\|x_1^*\left(r,z\right) - x_1^*\left(z\right)\|^2\\
&\leq \Psi\left(x_1^*\left(r,z\right)\right) - \Psi\left(x_1^*\left(z\right)\right) + \Psi_r\left(x_1^*\left(z\right)\right) - \Psi_r\left(x_1^*\left(r,z\right)\right)\\
&= \Psi\left(x_1^*\left(r,z\right)\right) - \Psi_r\left(x_1^*\left(r,z\right)\right) - \left(\Psi\left(x_1^*\left(z\right)\right) - \Psi_r\left(x_1^*\left(z\right)\right)\right).
\end{aligned}$$

Then, using the smoothness and strongly convexity of g_i for $i = 2, 3, ..., N$, we obtain

$$\begin{aligned}
&\Psi\left(x_1^*\left(r,z\right)\right) - \Psi_r\left(x_1^*\left(r,z\right)\right) - \left(\Psi\left(x_1^*\left(z\right)\right) - \Psi_r\left(x_1^*\left(z\right)\right)\right)\\
&= \frac{1}{N}\sum_{i=2}^{N}\left(g_i\left(x_1^*\left(r,z\right), z_i\right) - g_i\left(x_1^*\left(r,z\right) + v_i\left(r\right), z_i\right)\right) - \frac{1}{N}\sum_{i=2}^{N}\left(g_i\left(x_1^*\left(z\right), z_i\right) - g_i\left(x_1^*\left(z\right) + v_i\left(r\right), z_i\right)\right)\\
&\leq \frac{1}{N}\sum_{i=2}^{N}\left(\langle\nabla g_i\left(x_1^*\left(r,z\right) + v_i\left(r\right)\right), -v_i\left(r\right)\rangle + \frac{L+Np}{2}\|v_i\left(r\right)\|^2\right) - \frac{1}{N}\sum_{i=2}^{N}\langle\nabla g_i\left(x_1^*\left(z\right) + v_i\left(r\right)\right), -v_i\left(r\right)\rangle\\
&\leq \frac{1}{N}\sum_{i=2}^{N}(L + Np)\|x_1^*\left(r,z\right) - x_1^*\left(z\right)\|\|v_i\left(r\right)\| + \frac{(L+Np)}{2}\|v_i\left(r\right)\|^2.
\end{aligned}$$

Then, using the convexity of square function we obtain

$$\sum_{i=1}^{N}\|v_i(r)\| \leq \sqrt{N}\|v(r)\|.$$

Together with the above inequalities, we obtain

$$\begin{aligned}
&2\left(-L + Np\right)\|x_1^*\left(r,z\right) - x_1^*\left(z\right)\|^2\\
&\leq \frac{L+Np}{\sqrt{N}}\|x_1^*\left(r,z\right) - x_1^*\left(z\right)\|\|v\left(r\right)\| + \frac{(L+Np)}{2N}\|v\left(r\right)\|^2\\
&\leq \frac{L+Np}{\sqrt{N\lambda_3}}\|x_1^*\left(r,z\right) - x_1^*\left(z\right)\|\|r\| + \frac{(L+Np)}{2N\lambda_3}\|r\|^2.
\end{aligned}$$

By solving the above quadratic inequality, we can obtain

$$\|x_1^*(r,z) - x_1^*(z)\| \leq \frac{\left(L + Np + \sqrt{N^2p^2 - L^2}\right)}{2\sqrt{N\lambda_3}(Np - L)}\|r\|.$$

By the definition of v, we have $x(y,z) = x^*(r,z) = \left(x_1^*(r,z)^T, x_1^*(r,z)^T, \cdots, x_1^*(r,z)^T\right)^T + v$ and $x^*(z) = \left(x_1^*(z)^T, x_1^*(z)^T, \cdots, x_1^*(z)^T\right)^T$.

Therefore, with triangular inequality, we have

$$\|x(y,z) - x^*(z)\| \leq \sqrt{N}\|x_1^*(r,z) - x_1^*(z)\| + \|v\| \leq \sigma_1\|r\| = \sigma_1\|Ax(y,z)\|,$$

where $\sigma_1 = \frac{\left(3Np - L + \sqrt{N^2p^2 - L^2}\right)}{2\sqrt{\lambda_3}(Np-L)}$. Then, the proof is finished. □

Lemma 10 (Smooth version of Lemma 9)**.** *Suppose $p > -L/N$ and $h(\cdot) = 0$, then for any $y \in \mathbb{R}^{nM}$ and $z \in \mathbb{R}^{nN}$, it holds that*

$$\|x(y,z) - x^*(z)\| \leq \sigma_1\|Ax(y,z)\|,$$

where $\sigma_1 = \frac{\left(3Np - L + \sqrt{N^2p^2 - L^2}\right)}{2\sqrt{\lambda_2}(Np-L)}$.

Proof. First, we define $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$, similar to Lemma 8 for any $r \in \mathbf{Range}(A)$ there exists a unique $v_i(r)$, $i = 1, 2, 3, \cdots, N$, such that for any x that satisfies $Ax = r$, it holds that

$$x_i = \bar{x} + v_i(r).$$

For any vector u with $Au = 0$, we have $u^T(x - (\bar{x}^T, \bar{x}^T, \cdots, \bar{x}^T)^T) = 0$. Then, it holds $u^T v = 0$.

Therefore, it holds that

$$\|v\| \leq \frac{1}{\lambda_2}\|Av\| = \frac{1}{\lambda_2}\|r\|,$$

where λ_2 is the smallest nonzero eigenvalue of $A^T A$.

Similar to the proof of Lemma 9, we define $\Psi(u,z) = \frac{1}{N}\sum_{i=1}^N g_i(u, z_i)$, $\Psi_r(u) = \frac{1}{N}\sum_{i=1}^N g_i(u + v_i(r), z_i)$, $\bar{x}^*(r,z) = \arg\min_u \Psi_r(u)$ and $\bar{x}^*(z) = \arg\min_u \Psi(u)$.

With the same decomposition in Lemma 9, it holds that

$$\|\bar{x}^*(r,z) - \bar{x}^*(z)\| \leq \frac{\left(L + Np + \sqrt{N^2p^2 - L^2}\right)}{2\sqrt{N\lambda_2}(Np - L)}\|r\|.$$

Together with the definition of $x^*(r,z) = (\bar{x}^*(r,z)^T, \bar{x}^*(r,z)^T, \cdots, \bar{x}^*(r,z)^T)^T + v$ and $x^*(z) = (\bar{x}^*(z)^T, \bar{x}^*(z)^T, \cdots, \bar{x}^*(z)^T)^T$, we can get the result. □

Lemma 11. *For a differentiable convex function f (defined on the $\mathbb{R}^n$) with L-Lipschitz gradient, the following inequality always holds*

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

Proof. For fix x, we define function $\Gamma(z) = f(z) - \langle \nabla f(x), z \rangle$.

Then $\Gamma(z)$ is a convex function with L-Lipschitz gradient. Beside, the minimum value of function $\Gamma(\cdot)$ will be $\Gamma(x) = f(x) - \langle \nabla f(x), x \rangle$.

Meanwhile, using the Lipschitz smoothness of $\Gamma(\cdot)$, for fix y, we have

$$\Gamma(z) \leq \Gamma(y) + \langle \nabla\Gamma(y), z - y \rangle + \frac{L}{2}\|z - y\|^2.$$

Therefore, we can obtain

$$
\begin{aligned}
f(x) - \langle \nabla f(x), x \rangle &= \min_z \Gamma(z) \le \min_z \{ \Gamma(y) + \langle \nabla \Gamma(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \} \\
&\le \Gamma(y) - \frac{1}{2L} \|\nabla \Gamma(y)\|^2 = f(y) - \langle \nabla f(x), y \rangle + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2.
\end{aligned}
$$

By rearranging the terms, we can obtain

$$
\frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \le f(y) - f(x) + \langle \nabla f(x), x - y \rangle. \tag{9}
$$

By swapping x and y, we have

$$
\frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \le f(x) - f(y) + \langle \nabla f(y), y - x \rangle. \tag{10}
$$

Thus, adding up (9) and (10), we can attain the result. □

Lemma 12. *Suppose $p > -\frac{L}{N}$ and $c \le \frac{1}{L_K} = \frac{N}{L+Np}$, then the following inequalities hold:*

$$
\|x^{t+1} - x^t\| \ge \sigma_2 \|x^t - x\left(y^t, z^t\right)\|, \tag{11}
$$

$$
\|x^{t+1} - x^t\| \ge \sigma_3 \|x^{t+1} - x\left(y^t, z^t\right)\|, \tag{12}
$$

where

$$
\begin{aligned}
\sigma_2 &= \frac{c\left(Np - L\right)}{2N}, \\
\sigma_3 &= \sigma_2 / \left(1 - \sigma_2\right).
\end{aligned}
$$

Proof. For (11), we define $\hat{g}(x; v) = \|x\|^2 - 2x^T v + h(x)$, $v_1 = x^t - c\nabla_x K(x^t, y^t, z^t)$ and $v_2 = x(y^t, z^t) - c\nabla_x K(x(y^t, z^t), y^t, z^t)$.

Then, with the update iteration of x^t we have

$$
x^{t+1} = \arg\min_x \hat{g}(x, v_1).
$$

Beside, with the definition of $x(y^t, z^t)$ we have

$$
x(y^t, z^t) = \arg\min_x \hat{g}(x, v_2).
$$

It is obvious that $\hat{g}(\cdot, v)$ is a strongly convex function with modular 1. Then, using the strong convexity of $\hat{g}(\cdot, v)$, we have

$$
\hat{g}(x^{t+1}; v_2) - \hat{g}(x(y^t, z^t); v_2) \ge \|x^{t+1} - x(y^t, z^t)\|^2, \tag{13}
$$

$$
\hat{g}(x(y^t, z^t); v_1) - \hat{g}(x^{t+1}; v_1) \ge \|x^{t+1} - x(y^t, z^t)\|^2. \tag{14}
$$

On the other hand, by the definition of $\hat{g}$, we have

$$
\hat{g}(x^{t+1}; v_1) - \hat{g}(x^{t+1}; v_2) = -\langle x^{t+1}, v_1 - v_2 \rangle, \tag{15}
$$

$$
\hat{g}(x(y^t, z^t); v_1) - \hat{g}(x(y^t, z^t); v_2) = -\langle x(y^t, z^t), v_1 - v_2 \rangle. \tag{16}
$$

Combining (13), (14), (15) and (16), we have

$$
\|x^{t+1} - x(y^t, z^t)\|^2 \le \langle x^{t+1} - x(y^t, z^t), v_1 - v_2 \rangle. \tag{17}
$$

Then, using Cauchy–Schwartz inequality, we have

$$
\|x^{t+1} - x(y^t, z^t)\| \le \|v_1 - v_2\|. \tag{18}
$$

On the other hand, using the definition of v_1 and v_2, it holds that

$$\begin{aligned}\|v_1 - v_2\|^2 = &\|x^t - x(y^t, z^t)\|^2 - 2c\langle x^t - x(y^t, z^t), \nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\rangle \\ &+ c^2 \|\nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\|^2.\end{aligned} \tag{19}$$

According to the Lipschitz continuity of $\nabla_x K(\cdot, y, z)$ and convexity of $K(\cdot, y, z)$, and Lemma 11, we have

$$\frac{N}{L + Np}\|\nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\|^2 \leq \langle x^t - x(y^t, z^t), \nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\rangle.$$

Substituting the above inequality into (19), we have

$$\begin{aligned}&\|v_1 - v_2\|^2 \\ &\leq \|x^t - x(y^t, z^t)\|^2 - (2c - c^2\frac{L + Np}{N})\langle x^t - x(y^t, z^t), \nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\rangle \\ &\leq \|x^t - x(y^t, z^t)\|^2 - c\langle x^t - x(y^t, z^t), \nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\rangle,\end{aligned}$$

where the last ineqaulity holds because $c \leq \frac{N}{L+Np}$.

Then, because $K(\cdot, y, z)$ is a convex function with $(p - \frac{L}{N})$-Lipschitz gradient, according to Lemma 11, we have

$$\langle x^t - x(y^t, z^t), \nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\rangle \geq (p - \frac{L}{N})\|x^t - x(y^t, z^t)\|^2.$$

Therefore, we have

$$\begin{aligned}&\|v_1 - v_2\|^2 \\ &\leq \|x^t - x(y^t, z^t)\|^2 - c\langle x^t - x(y^t, z^t), \nabla_x K(x^t, y^t, z^t) - \nabla_x K(x(y^t, z^t), y^t, z^t)\rangle \\ &\leq (1 - c(p - \frac{L}{N}))\|x^t - x(y^t, z^t)\|^2.\end{aligned}$$

Note that $1 - c(p - \frac{L}{N}) < (1 - c(p - \frac{L}{N})/2)^2$. Then we have

$$\|v_1 - v_2\|^2 \leq (1 - c(p - \frac{L}{N}))^2\|x^t - x(y^t, z^t)\|^2.$$

Hence,

$$\|x^{t+1} - x(y^t, z^t)\| \leq \|v_1 - v_2\| \leq (1 - c(p - \frac{L}{N})/2)\|x^t - x(y^t, z^t)\|. \tag{20}$$

Besides, according to triangular inequality, we have

$$\begin{aligned}&\|x^t - x^{t+1}\| \\ \geq\ &\|x^t - x(y^t, z^t)\| - \|x^{t+1} - x(y^t, z^t)\| \\ \geq\ &\frac{c(p - \frac{L}{N})}{2}\|x^t - x(y^t, z^t)\|,\end{aligned}$$

which yields Eq. (11).

In addition, by (20), we have

$$\begin{aligned}&\|x^t - x^{t+1}\| \\ \geq\ &\frac{c(p - \frac{L}{N})}{2}\|x^t - x(y^t, z^t)\| \\ \geq\ &\frac{c(p - \frac{L}{N})/2}{1 - c(p - \frac{L}{N})/2}\|x^{t+1} - x(y^t, z^t)\| \\ =\ &\frac{\sigma_2}{1 - \sigma_2}\|x^{t+1} - x(y^t, z^t)\|,\end{aligned}$$

which gives the result (12). □

Proof of Theorem 1. Let $\Phi(x,y,z) = K(x,y,z) + h(x) - 2d(y,z) + 2M(z)$.
Recall the definition of $d(y,z)$ and $M(z)$

$$
\begin{aligned}
d(y,z) &= \min_{x\in\mathbb{R}^{nN}} K(x,y,z) + h(x),\\
M(z) &= \min_{x\in\mathbb{R}^{nN}, Ax=0} \left(f(x) + h(x) + \frac{p}{2}\|x-z\|^2 \right),
\end{aligned}
$$

it holds that

$$
\begin{aligned}
&M(z) \geq d(y,z),\\
&K(x,y,z) + h(x) \geq d(y,z).
\end{aligned}
$$

Then,

$$
\begin{aligned}
\Phi(x,y,z) &= K(x,y,z) + h(x) - 2d(y,z) + 2M(z)\\
&\geq K(x,y,z) + h(x) - d(y,z) + M(z) - d(y,z) + M(z) \geq M(z)\\
&\geq \min_x \{f(x) + h(x)\} \geq \underline{f}.
\end{aligned}
$$

Using Lemmas 2, 5, and 6, it holds that

$$
\begin{aligned}
&\Phi\left(x^t,y^t,z^t\right) - \Phi\left(x^{t+1},y^{t+1},z^{t+1}\right)\\
&\quad\geq \frac{1}{2c}\|x^{t+1}-x^t\|^2 - \alpha\left(A\hat{x}^{t+1}\right)^T Ax^{t+1} + \frac{p}{2\beta}\|z^t - z^{t+1}\|^2 + 2\alpha\left(A\hat{x}^{t+1}\right)^T Ax\left(y^t,z^{t+1}\right)\\
&\qquad - \frac{\alpha^2\sqrt{\lambda_1}}{\sigma_4}\|A\hat{x}^{t+1}\|^2 + p\left(z^{t+1}-z^t\right)^T\left(z^{t+1}+z^t-2x\left(y^t,z^{t+1}\right)\right)\\
&\qquad - 2p\left(z^{t+1}-z^t\right)^T\left(z^t - x^*\left(z^t\right)\right) - p\tilde{L}\|z^t-z^{t+1}\|^2\\
&\quad\geq \frac{1}{2c}\|x^{t+1}-x^t\|^2 + \frac{p}{2\beta}\|z^t-z^{t+1}\|^2 + \alpha\left(Ae^{t+1}\right)^T A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right) - \alpha\|Ax^{t+1}\|^2\\
&\qquad - \alpha\left(Ae^{t+1}\right)^T Ax\left(y^t,z^{t+1}\right) - \frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\left(\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\|^2 + \|Ae^{t+1}\|^2 + \|Ax\left(y^t,z^{t+1}\right)\|^2\right)\\
&\qquad + 2\alpha\left(Ax^{t+1}\right)^T Ax\left(y^t,z^{t+1}\right) + p\left(z^{t+1}-z^t\right)^T\left(z^{t+1}-z^t-2\left(x\left(y^{t+1},z^{t+1}\right)-x^*\left(z^t\right)\right)\right) - p\tilde{L}\|z^t-z^{t+1}\|^2\\
&\quad\geq \frac{1}{2c}\|x^{t+1}-x^t\|^2 + \frac{p}{2\beta}\|z^t-z^{t+1}\|^2 - \alpha\sqrt{\lambda_1}\|e^{t+1}\|\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\| - \alpha\|Ax^{t+1}\|^2\\
&\qquad - \alpha\sqrt{\lambda_1}\|e^{t+1}\|\|Ax\left(y^t,z^{t+1}\right)\| - \frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\left(\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\|^2 + \lambda_1\|e^{t+1}\|^2 + \|Ax\left(y^t,z^{t+1}\right)\|^2\right)\\
&\qquad + 2\alpha\left(Ax^{t+1}\right)^T Ax\left(y^t,z^{t+1}\right) + p\left(z^{t+1}-z^t\right)^T\left(z^{t+1}-z^t-2\left(x\left(y^t,z^{t+1}\right)-x^*\left(z^t\right)\right)\right) - p\tilde{L}\|z^t-z^{t+1}\|^2.
\end{aligned}
\tag{21}
$$

First, we can bound the terms related to the e^{t+1}:

$$
\begin{aligned}
&-\alpha\sqrt{\lambda_1}\|e^{t+1}\|\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\| - \alpha\sqrt{\lambda_1}\|e^{t+1}\|\|Ax\left(y^t,z^{t+1}\right)\| - \frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\lambda_1\|e^{t+1}\|^2\\
&\geq -\frac{\alpha}{2}\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\|^2 - \frac{\alpha}{2}\|Ax\left(y^t,z^{t+1}\right)\|^2 - \left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4} + \alpha\lambda_1\right)\|e^{t+1}\|^2.
\end{aligned}
$$

Then, for the terms related to A, we can obtain

$$
\begin{aligned}
&-\alpha\sqrt{\lambda_1}\|e^{t+1}\|\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\|-\alpha\|Ax^{t+1}\|^2+2\alpha\left(Ax^{t+1}\right)^T Ax\left(y^t,z^{t+1}\right)\\
&\quad-\alpha\sqrt{\lambda_1}\|e^{t+1}\|\|Ax\left(y^t,z^{t+1}\right)\|-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\left(\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\|^2+\lambda_1\|e^{t+1}\|^2+\|Ax\left(y^t,z^{t+1}\right)\|^2\right)\\
&\geq-\frac{\alpha}{2}\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\|^2-\frac{\alpha}{2}\|Ax\left(y^t,z^{t+1}\right)\|^2-\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\|e^{t+1}\|^2\\
&\quad-\left(\alpha\|Ax^{t+1}\|^2-2\alpha\left(Ax^{t+1}\right)^T Ax\left(y^t,z^{t+1}\right)+\alpha\|Ax\left(y^t,z^{t+1}\right)\|^2\right)+\alpha\|Ax\left(y^t,z^{t+1}\right)\|^2\\
&\quad-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\left(\|A\left(x^{t+1}-x\left(y^t,z^{t+1}\right)\right)\|^2+\|Ax\left(y^t,z^{t+1}\right)\|^2\right)\\
&=\left(\frac{\alpha}{2}-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\right)\|Ax\left(y^t,z^{t+1}\right)\|^2-\left(\frac{\alpha}{2}+\frac{3\alpha^2\lambda_1^{1/2}}{\sigma_4}\right)\|A(x^{t+1}-x\left(y^t,z^{t+1}\right))\|^2-\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\|e^{t+1}\|^2\\
&\geq\left(\frac{\alpha}{2}-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\right)\|Ax\left(y^t,z^{t+1}\right)\|^2-\left(\frac{\alpha\lambda_1}{2}+\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}\right)\|x^{t+1}-x\left(y^t,z^{t+1}\right)\|^2-\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\|e^{t+1}\|^2\\
&=\left(\frac{\alpha}{2}-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\right)\|Ax\left(y^t,z^{t+1}\right)\|^2-\left(\frac{\alpha\lambda_1}{2}+\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}\right)\|x^{t+1}-x(y^t,z^t)+x(y^t,z^t)-x\left(y^t,z^{t+1}\right)\|^2\\
&\quad-\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\|e^{t+1}\|^2\\
&\geq\left(\frac{\alpha}{2}-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\right)\|Ax\left(y^t,z^{t+1}\right)\|^2-\left(\alpha\lambda_1+\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4}\right)\|x^{t+1}-x(y^t,z^t)\|^2\\
&\quad-\left(\alpha\lambda_1+\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4}\right)\|x(y^t,z^t)-x\left(y^t,z^{t+1}\right)\|^2-\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\|e^{t+1}\|^2\\
&\geq\left(\frac{\alpha}{2}-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\right)\|Ax\left(y^t,z^{t+1}\right)\|^2-\left(\frac{\alpha\lambda_1}{\sigma_3}+\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4\sigma_3}\right)\|x^{t+1}-x^t\|^2-\left(\frac{\alpha\lambda_1}{\sigma_5}+\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4\sigma_5}\right)\|z^{t+1}-z^t\|^2\\
&\quad-\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\|e^{t+1}\|^2,
\end{aligned}
\tag{22}
$$

where the last inequality is due to Lemma 12 and Lemma 4.

Further, for the terms related to z, we have

$$
\begin{aligned}
&p\left(z^{t+1}-z^t\right)^T\left(z^{t+1}-z^t-2\left(x\left(y^t,z^{t+1}\right)x^*\left(z^t\right)\right)\right)-p\tilde{L}\|z^t-z^{t+1}\|^2+\frac{p}{2\beta}\|z^t-z^{t+1}\|^2\\
&=p\left(z^{t+1}-z^t\right)^T\left(z^{t+1}-z^t-2\left(x\left(y^t,z^{t+1}\right)-x^*(z^{t+1})\right)-2\left(x^*(z^{t+1})-x^*\left(z^t\right)\right)\right)\\
&\qquad-p\tilde{L}\|z^t-z^{t+1}\|^2+\frac{p}{2\beta}\|z^t-z^{t+1}\|^2\\
&\geq\left(\frac{p}{2\beta}+p-p\tilde{L}\right)\|z^t-z^{t+1}\|^2-2p\left(z^{t+1}-z^t\right)^T\left(x\left(y^t,z^{t+1}\right)-x^*\left(z^{t+1}\right)\right)\\
&\qquad-2p\left(z^{t+1}-z^t\right)^T\left(x^*\left(z^{t+1}\right)-x^*\left(z^t\right)\right)\\
&\geq\left(\frac{p}{2\beta}+p-p\tilde{L}\right)\|z^t-z^{t+1}\|^2-2p\|z^{t+1}-z^t\|\|x\left(y^t,z^{t+1}\right)-x^*\left(z^{t+1}\right)\|\\
&\qquad-2p\|z^{t+1}-z^t\|\|x^*(z^{t+1})-x^*(z^t)\|.
\end{aligned}
$$

By using Lemma 4 and Lemma 3, it further holds that

$$
\begin{aligned}
& p\left(z^{t+1}-z^t\right)^T\left(z^{t+1}-z^t-2\left(x\left(y^t, z^{t+1}\right) x^*\left(z^t\right)\right)\right)-p\tilde{L}\|z^t-z^{t+1}\|^2+\frac{p}{2\beta}\|z^t-z^{t+1}\|^2 \\
\geq & \left(\frac{p}{2\beta}+p-p\tilde{L}\right)\|z^t-z^{t+1}\|^2-\frac{4p^2\sigma_1^2}{\alpha}\|z^{t+1}-z^t\|^2+\frac{\alpha}{4\sigma_1^2}\|x\left(y^t, z^{t+1}\right)-x^*\left(z^{t+1}\right)\|^2 \\
& -\frac{2p}{\sigma_5}\|z^{t+1}-z^t\|^2 \\
\geq & \left(\frac{p}{2\beta}+p-p\tilde{L}-\frac{4p^2\sigma_1^2}{\alpha}-\frac{2p}{\sigma_5}\right)\|z^t-z^{t+1}\|^2-\frac{\alpha}{4\sigma_1^2}\|x\left(y^t, z^{t+1}\right)-x^*\left(z^{t+1}\right)\|^2 \\
\geq & \left(\frac{p}{2\beta}+p-p\tilde{L}-\frac{4p^2\sigma_1^2}{\alpha}-\frac{2p}{\sigma_5}\right)\|z^t-z^{t+1}\|^2-\frac{\alpha}{4}\|Ax\left(y^t, z^{t+1}\right)\|^2.
\end{aligned} \tag{23}
$$

Besides, by taking α and β sufficient small so that we can obtain

$$
\begin{aligned}
& \frac{\alpha}{8} \geq \frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}, \\
& \frac{1}{2c}-\frac{\alpha\lambda_1}{\sigma_3}-\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4\sigma_3}-\frac{(1-\delta)^2}{\delta^2}\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right) \geq \frac{1}{4c}, \\
& \frac{p}{2\beta}+p-p\tilde{L}-\frac{4p^2\sigma_1^2}{\alpha}-\frac{2p}{\sigma_5}-\left(\frac{\alpha\lambda_1}{\sigma_5}+\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4\sigma_5}\right) \geq \frac{p}{4\beta}, \\
& \beta \leq 1.
\end{aligned} \tag{24}
$$

Then combining (21), (22), (23), (24), and Lemma 1, we can obtain

$$
\begin{aligned}
& \Phi\left(x^0, y^0, z^0\right)-\underline{f} \\
\geq & \sum_{t=0}^{T-1} \Phi\left(x^t, y^t, z^t\right)-\Phi\left(x^{t+1}, y^{t+1}, z^{t+1}\right) \\
\geq & \sum_{t=0}^{T-1}\left(\frac{1}{2c}-\frac{\alpha\lambda_1}{\sigma_3}-\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4\sigma_3}\right)\|x^{t+1}-x^t\|^2+\left(\frac{\alpha}{4}-\frac{3\alpha^2\sqrt{\lambda_1}}{\sigma_4}\right)\|Ax\left(y^t, z^{t+1}\right)\|^2 \\
& -\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\|e_{t+1}\|^2+\frac{p}{4\beta}\|z^t-z^{t+1}\|^2 \\
\geq & \sum_{t=0}^{T-1}\left(\frac{1}{2c}-\frac{\alpha\lambda_1}{\sigma_3}-\frac{6\alpha^2\lambda_1^{3/2}}{\sigma_4\sigma_3}-\frac{(1-\delta)^2}{\delta^2}\left(\frac{3\alpha^2\lambda_1^{3/2}}{\sigma_4}+\alpha\lambda_1\right)\right)\|x^{t+1}-x^t\|^2 \\
& +\frac{\alpha}{8}\|Ax\left(y^t, z^{t+1}\right)\|^2+\frac{p}{4\beta}\|z^t-z^{t+1}\|^2 \\
\geq & \sum_{t=0}^{T-1} \frac{1}{4c}\|x^{t+1}-x^t\|^2+\frac{\alpha}{8}\|Ax\left(y^t, z^{t+1}\right)\|^2+\frac{p}{4\beta}\|z^t-z^{t+1}\|^2.
\end{aligned}
$$

According to the above inequality, we define $C=\Phi\left(x^0, y^0, z^0\right)-\underline{f}$, then it holds that for any $T>0$, there exists an $s \in \{0,1,\cdots,T-1\}$ such that

$$
\begin{aligned}
& \|x^s-x^{s+1}\|^2 \leq 4cC/T, \\
& \|Ax\left(y^s, z^{s+1}\right)\|^2 \leq \frac{8}{\alpha}C/T, \\
& \|x^{s+1}-z^s\|^2=\frac{1}{\beta^2}\|z^{s+1}-z^s\|^2 \leq \frac{4}{\beta p}C/T, \\
& \|z^{s+1}-z^s\|^2 \leq \frac{4\beta}{p}C/T.
\end{aligned}
$$

Besides, recall the update of x^{s+1}, i.e. $x^{s+1} = \arg\min_x \left(\langle \nabla_{x_i} K\left(x^t, y^t, z^t\right), x_i - x_i^t \rangle + h_i\left(x_i\right) + \frac{1}{2c}\|x_i - x_i^t\|^2\right)$, with the optimally condition we can obtain

$$0 \in \nabla_x K\left(x^s, y^s, z^s\right) + \frac{1}{c}\left(x^{s+1} - x^s\right) + \partial h\left(x^{s+1}\right).$$

Therefore, let

$$\nu = \nabla_x K\left(x^{s+1}, y^s, z^s\right) - \nabla_x K\left(x^s, y^s, z^s\right) - \frac{1}{c}\left(x^{s+1} - x^s\right) - p\left(x^{s+1} - z^s\right).$$

we can obtain that

$$\nu \in \nabla f\left(x\right) + A^T y^s + \partial h\left(x^{s+1}\right).$$

Moreover, we have

$$\begin{aligned} \|v\| &\leq \left(\frac{L}{N} + p\right)\|x^{s+1} - x^s\| + \frac{1}{c}\|x^{s+1} - x^s\| + p\|x^{s+1} - z^s\| \\ &\leq \left(\left(\frac{L}{N} + p + \frac{1}{c}\right)\sqrt{4c} + \frac{2\sqrt{p}}{\sqrt{\beta}}\right)\sqrt{\frac{C}{T}}. \end{aligned}$$

On the other hand, it holds that

$$\begin{aligned} \|Ax^{s+1}\| &\leq \|Ax\left(y^s, z^{s+1}\right)\| + \|A\left(x^{s+1} - x\left(y^s, z^s\right)\right)\| + \|A(x(y^s, z^s) - x(y^s, z^{s+1}))\| \\ &\leq \sqrt{C}/\sqrt{T}\left(\frac{8}{\sqrt{\alpha}} + \frac{\sqrt{\lambda_1 4c}}{\sigma_3} + \frac{\sqrt{4\lambda_1\beta}}{\sqrt{p}\sigma_5}\right). \end{aligned}$$

Hence, letting

$$B = \left(\left(\left(\frac{L}{N} + p + \frac{1}{c}\right)\sqrt{4c} + \frac{2\sqrt{p}}{\sqrt{\beta}}\right)\sqrt{N} + \frac{L}{\sqrt{N\lambda_2}}\left(\frac{8}{\sqrt{\alpha}} + \frac{\sqrt{\lambda_1 4c}}{\sigma_3} + \frac{\sqrt{4\beta\lambda_1}}{\sqrt{p}\sigma_5}\right)\right)\sqrt{C}. \tag{25}$$

Then, $\left(x^{s+1}, y^s\right)$ is a $B/\sqrt{T}$ -solution. □

Proof of Corollary 1. In the smooth case, we define $\gamma = \frac{\lambda_1}{\lambda_2}$ and in non-smooth case we define $\gamma = \frac{\lambda_1}{\lambda_3} \geq \frac{\lambda_1}{\lambda_2}$.

It is easy to check α and β in corollary 1 that satisfy inequalities (24).

By plugging c, p, α and β in to equation (25), we can get the result in smooth case.

With $\frac{\lambda_1}{\lambda_3} \geq \frac{\lambda_1}{\lambda_2}$ we can get the result in the nonsmooth case.

□

2 Proof of Remark 1

Proof of Remark 1. The proof will only consider the smooth case, because the definition of $2\epsilon^2$-stationary points in Hong et al. (2017) and $4\epsilon^2$-stationary points in Tang et al. (2019) are in the setting of smooth objective function.

First we show that our definition is the sufficient condition for $2\epsilon^2$-stationary points in Hong et al. (2017), i.e. $\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i)\|^2 + \frac{L}{N\lambda_2}\sum_{(i,j)\in E}\|x_i - x_j\|^2 \leq 2\epsilon^2$.

From Definition 1, we have

$$N\|\nabla f(x) + A^T y\|^2 \leq \epsilon^2.$$

Besides, recall that $\mu = (\mu_1, \mu_2, \cdots, \mu_N) = A^T y$ in the algorithm and $A\mathbf{1} = \mathbf{0}$.

Thus, $\sum_{i=1}^N \mu_i = 0$.

Then, it holds that

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i)\right\|^2 &= \left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i)+\mu_i\right\|^2\\ &\le N\left(\frac{1}{N^2}\sum_{i=1}^{N}\|\nabla f_i(x_i)+\mu_i\|^2\right)\\ &= N\|\nabla f(x)+A^Ty\|^2\le \epsilon^2.\end{aligned}$$

On the other hand, we have

$$\frac{L^2}{N\lambda_2}\|Ax\|^2\le\epsilon^2.$$

Hence, it holds that

$$\frac{L}{N\lambda_2}\sum_{(i,j)\in E}\|x_i-x_j\|^2=\frac{L}{N\lambda_2}\|Ax\|^2\le\frac{\epsilon^2}{L}.$$

Because L can take arbitrary large value, we can assume $L>1$ and the following statement holds:

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i)\right\|^2+\frac{L}{N\lambda_2}\sum_{(i,j)\in E}\|x_i-x_j\|^2\le 2\epsilon^2.$$

Then, for the $4\epsilon^2$-stationary point in Tang et al. (2019), it is defined as

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{x})\right\|^2\le 4\epsilon^2,$$

where $\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$.

Let $\tilde{x}=(\bar{x}^T,\bar{x}^T,\cdots,\bar{x}^T)^T\in\mathbb{R}^{nN}$.

Recall the definition of $x=(x_1^T,x_2^T,\cdots,x_N^T)^T$ and A, then we obtain if $Av=0$ then $v^T(\tilde{x}-x)=0$.

Besides, we have $A\tilde{x}=0$.

Therefore, with definition of λ_2 (the smallest nonzero eigenvalue of A^TA), we have

$$\sum_{i=1}^{N}\|\bar{x}-x_i\|^2=\|\tilde{x}-x\|^2\le\frac{1}{\lambda_2}\|A(\tilde{x}-x)\|^2=\frac{1}{\lambda_2}\|Ax\|^2.$$

Combining the above inequality and the Lipschitz gradient of f_i, it holds that

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{x})\right\|^2 &= \left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{x})-\nabla f_i(x_i)+\nabla f_i(x_i)\right\|^2\\ &\le 2N\left(\frac{1}{N^2}\sum_{i=1}^{N}\|\nabla f_i(x_i)\|^2\right)+\frac{2}{N}\sum_{i=1}^{N}\nabla f_i(\bar{x})-\nabla f_i(x_i)\|^2\\ &\le 2N\|\nabla f(x)+A^Ty\|^2+\frac{2L^2}{N}\sum_{i=1}^{N}\|\bar{x}-x_i\|^2\\ &\le 2N\|\nabla f(x)+A^Ty\|^2+\frac{2L^2}{N\lambda_2}\|Ax\|^2\\ &\le 4\epsilon^2.\end{aligned}$$

Hence, our definition is the sufficient condition for the $2\epsilon^2$-solution in Hong et al. (2017) and the $4\epsilon^2$-solution in Tang et al. (2019). $\square$

References

Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning*, pp. 1529–1538, 2017.

Hanlin Tang, Xiangru Lian, Shuang Qiu, Lei Yuan, Ce Zhang, Tong Zhang, and Ji Liu. Deepsqueeze : Parallel stochastic gradient descent with double-pass error-compensated compression. *arXiv preprint arXiv:1907.07346*, 2019.