A. Riemannian Optimization Operators on Hypershpere Manifold

The unit norm constraint of a d-dimensional vector admits a geometry structure as a hypersphere. For the hypersphere manifold, the tangent space projector is $P_w(g) = (I - ww^\top)g$. The Riemannian metric in the hypersphere manifold is the same as the Euclidean metric, which means the inner product of two vectors in tangent space is calculated in the same way as in the Euclidean space. The vector transport is computed as $\mathcal{T}_w^u(\xi) = P_u(\xi) = (I - uu^\top)\xi$. The retraction (exponential map) is calculated as $R_w(\xi) = \cos(\|\xi\|)w + \frac{\sin(\|\xi\|)}{\|\xi\|}\xi$. The inverse retraction (logarithmic map) is computed as $R_w^{-1}(u) = \mathrm{D}(u,w)\frac{P_w(u-w)}{\|P_w(u-w)\|}$, where $\mathrm{D}(w,u)$ is the distance between two points in the manifold and it is computed as $\mathrm{D}(w,u) = \arccos(\langle w,u\rangle)$.

Note the distance between w and v in the hypersphere manifold is the angle between these two vectors. For any $w,u \in \mathcal{M}$ satisfying $w^\top u \geq 0$, we have

$$1 - (w^\top u)^2 \leq \mathrm{D}^2(w,u) \leq \frac{\pi^2}{4}\left(1 - (w^\top u)^2\right). \tag{7}$$

Therefore, bounding the $1-(w^\top v)^2$ is in fact bounding the distance between w and v on the hypersphere manifold.

B. Proof for Section 2.1 and Section 4

B.1. Proof for Lemma 1

Proof. To prove that $\hat{F}(w)$ is L-g-smooth with $L = \lambda$ is equivalent to prove that the largest eigenvalue of the Riemannian manifold Hessian of $\hat{F}(w)$, denoted by $\tilde{\nabla}^2\hat{F}(w)$, is λ. Based on the analytic form of Hessian matrix-vector product for hypersphere manifold in (Boumal et al., 2014; Absil et al., 2008), the largest eigenvalue is $\max_{u\in T_w\mathcal{M}}\langle u, \tilde{\nabla}^2\hat{F}(w)[u]\rangle$, which can be computed as

$$\begin{aligned}\langle u, \tilde{\nabla}^2\hat{F}(w)[u]\rangle &= \langle u, -(I-ww^\top)\hat{A}u + w^\top\hat{A}wu\rangle \\ &= -u^\top\hat{A}u + w^\top\hat{A}w \\ &\leq 0 + \lambda\end{aligned} \tag{8}$$

The second equality uses the fact that u is in the tangent space of w such that $u^\top w = 0$. And the inequality uses the fact that $\hat{A}$ is positive (semi-)definite and its largest eigenvalue is λ.

Therefore, we have $L = \lambda$.

□

B.2. Proof for Lemma 2

Lemma 2 is a simple corollary of Theorem 4 in (Zhang et al., 2016).

B.3. Proof for Lemma 3

Proof. Assume all data instances are i.i.d. sampled from some unknown distribution $\mathcal{D}$ and define $A = \mathbb{E}_{x\sim\mathcal{D}}[xx^\top]$. Then we have $\mathbb{E}[x_i^k(x_i^k)^\top] = A$. By matrix Hoeffding's inequality (Theorem 1.3 in (Tropp, 2012)), we have for each k, with probability at least $1-p$ over the i.i.d. samples in machine k,

$$||A_k - A||_2^2 \leq \frac{8\sigma^2\log(d/p)}{n}$$

where

$$\sigma^2 = \left\|\frac{1}{n}\left(\sum_{i=1}^n x_i^k(x_i^k)^\top - A\right)^2\right\|_2 \leq \max\{\max_i[(x_i^k(x_i^k)^\top)^2], A^2\} = b^2.$$

Therefore, $\sigma = b$ and we have

$$||A_k - A||_2^2 \leq \frac{8b^2\log(d/p)}{n}$$

Again using matrix Hoeffding's inequality on $\hat{A}$, we have, with probability at least $1-p$,

$$||\hat{A} - A||_2^2 \leq \frac{8b^2 \log(d/p)}{N}$$

Combining these two inequalities, we get

$$\|A_k - \hat{A}\|_2^2 \leq 2\|A_k - A\|_2^2 + 2\|A - \hat{A}\|_2^2 \leq \left(\frac{1}{n} + \frac{1}{N}\right) 16b^2 \log(d/p) \leq \frac{32b^2 \log(d/p)}{n}.$$

The last inequality comes from $n \leq N$.

□

Remark: When the data instances are not i.i.d., we can still have a similar bound on $\|A_k - \hat{A}\|_2^2$ if the data are arbitrarily distributed over all local machine based on the following lemma.

Lemma 9. *Assume the data instances are randomly partitioned over all local machines and their ℓ_2 norm is at most b. Then for each local machine k, with probability at least $1-p$ and n being sufficiently large such that $n \geq \log(2d/p)$, we have $\|A_k - \hat{A}\|_2^2 \leq \alpha^2 b^2$, where $\alpha^2 = \frac{4log(2d/p)}{n}$.*

Proof. If the data are randomly partitioned over all local machines, $\{x_i^k\}$ for $i = 1, ..., n$ are sampled without replacement from all data to construct A_k. Apply the without-replacement version of Bernstein's inequality for matrices (Theorem 1 in (Gross & Nesme, 2010)), we have

$$\Pr(\|A_k - \hat{A}\|_2 > \varepsilon) \leq \begin{cases} 2d\exp(-\frac{n\varepsilon^2}{4c_2}), & if\ \varepsilon \leq \frac{2c_2}{c_1} \\ 2d\exp(-\frac{n\varepsilon}{2c_1}), & if\ \varepsilon > \frac{2c_2}{c_1} \end{cases},$$

where $\max_i \|x_i x_i^\top\|_2 \leq c_1$ and

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \left(x_i x_i^\top - \hat{A} \right)^2 \right\|_2 \leq c_2.$$

It is easy to know that $c_1 = b$. As for c_2,

$$\begin{aligned} \left\| \frac{1}{N} \sum_{i=1}^{N} \left(x_i x_i^\top - \hat{A} \right)^2 \right\|_2 &= \left\| \mathbb{E} \left(x_i x_i^\top - \hat{A} \right)^2 \right\|_2 \\ &= \left\| \mathbb{E}\left[(x_i x_i^\top)^2 \right] - \hat{A}^2 \right\|_2 \\ &\leq \max\left\{ \left\| \mathbb{E}\left[(x_i (x_i)^\top)^2 \right] \right\|_2, \|\hat{A}^2\|_2 \right\}. \end{aligned}$$

Notice $\|\hat{A}^2\|_2 \leq \lambda^2$ and

$$\left\| \mathbb{E}\left[(x_i x_i^\top)^2 \right] \right\|_2 = \left\| \mathbb{E}\left[\|x_i\|^2 x_i x_i^\top \right] \right\|_2 \leq \max_i \{\|x_i\|^2\} \left\| \sum_{i=1}^{N} x_i x_i^\top \right\|_2 = b\lambda \leq b^2.$$

Since $b \geq \lambda$, we set $c_2 = b^2$.

Let the probability $p = \Pr(\|A_k - \hat{A}\|_2 > \varepsilon)$. When $\varepsilon \leq 2b$ is satisfied, with probability $1-p$, we have

$$\|A_k - \hat{A}\|_2 \leq \varepsilon = \sqrt{\frac{4b^2 \log(2d/p)}{n}}.$$

Since $\log(2d/p) \leq n$, it holds $\varepsilon \leq 2b$.

Therefore, with probability $1-p$, we have

$$\|A_k - \hat{A}\|_2^2 \leq \frac{4b^2 \log(2d/p)}{n}. \tag{9}$$

□

B.4. Proof for Lemma 4

Proof. Note that $\tilde{\nabla}\hat{F}(w) = P_w(\hat{A}w)$ and $\mathcal{T}^w_{\tilde{w}_s} = P_w$. Plugging these operators into $G(w)$ and $\tilde{\nabla}\hat{F}(w)$, we obtain

$$\begin{aligned}
\|G(w) - \tilde{\nabla}\hat{F}(w)\|^2 &= \|(I - ww^\top)(A_k - \hat{A})(w - \tilde{w}_s) + (I - ww^\top)\tilde{w}_s\tilde{w}_s^\top(A_k - \hat{A})\tilde{w}_s\|^2 \\
&= \|(I - ww^\top)(A_k - \hat{A})(w - \tilde{w}_s) + (I - ww^\top)\tilde{w}_s\tilde{w}_s^\top(A_k - \hat{A})\tilde{w}_s\|^2 \\
&\leq 2\|(I - ww^\top)(A_k - \hat{A})(w - \tilde{w}_s)\|^2 + 2\|(I - ww^\top)\tilde{w}_s\tilde{w}_s^\top(A_k - \hat{A})\tilde{w}_s\|^2 \\
&\leq 2\|I - ww^\top\|_2^2\|A_k - \hat{A}\|_2^2\|w - \tilde{w}_s\|^2 + 2\|(I - ww^\top)\tilde{w}_s\|^2\left(\tilde{w}_s^\top(A_k - \hat{A})\tilde{w}_s\right)^2 \\
&\leq 2\alpha^2b^2\|w - \tilde{w}_s\|^2 + 2\alpha^2b^2\|(I - ww^\top)\tilde{w}_s\|^2 \\
&= 2\alpha^2b^2\left(3 - 2w^\top\tilde{w}_s - (w^\top\tilde{w}_s)^2\right) \\
&\leq 6\alpha^2b^2\left(1 - (w^\top\tilde{w}_s)^2\right) \\
&\leq 6\alpha^2b^2\mathrm{D}^2(w, \tilde{w}_s).
\end{aligned}$$

The third inequality uses the result in Lemma 3 that $\|A_k - A\|_2^2 \leq \alpha^2b^2$. The fourth inequality uses the fact that $w^\top\tilde{w}_s \leq 1$ such that $w^\top\tilde{w}_s \geq (w^\top\tilde{w}_s)^2$. And the final inequality comes from Eq. (7).

□

B.5. Proof for Theorem 1

Proof. For simplification of notation, we denote $w_t^{s,k}$ by w_t. Note that in Algorithm 1, we have $g_t = G(w_t)$.

As presented in Lemma 1, $\hat{F}(w)$ is λ-smooth. By the gradient λ-Lipschitz, we have

$$\begin{aligned}
\hat{F}(w_{t+1}) &\leq \hat{F}(w_t) + \langle\tilde{\nabla}\hat{F}(w_t), R_{w_t}^{-1}(w_{t+1})\rangle + \frac{\lambda}{2}\left\|R_{w_t}^{-1}(w_{t+1})\right\|^2 \\
&= \hat{F}(w_t) - \eta\langle\tilde{\nabla}\hat{F}(w_t), G(w_t)\rangle + \frac{\eta^2\lambda}{2}\|G(w_t)\|^2 \\
&= \hat{F}(w_t) - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2 + \frac{\eta}{2}\|G(w_t) - \tilde{\nabla}\hat{F}(w_t)\|^2 + \eta(\frac{\eta\lambda}{2} - \frac{1}{2})\|G(w_t)\|^2 \\
&\leq \hat{F}(w_t) - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2 + 3\eta\alpha^2b^2\mathrm{D}^2(\tilde{w}_s, w_t) + \eta^2(\frac{\lambda}{2} - \frac{1}{2\eta})\|G(w_t)\|^2.
\end{aligned} \tag{10}$$

The second equality follows $\langle u_1, u_2\rangle = \frac{1}{2}\left(\|u_1\|^2 + \|u_2\|^2 - \|u_1 - u_2\|^2\right)$ for any two vectors u_1 and u_2. The second inequality follows Lemma 4. Notice that $G(w_t) \neq \tilde{\nabla}\hat{F}(w_t)$ is an obstacle from theoretically analyzing distributed algorithms, and thus the second equality distinguishes our proof from the proof of single-machine optimization of PCA problem in (Shamir, 2015; 2016; Xu et al., 2017; Zhang et al., 2016).

Inspired by the proof of RSVRG (Zhang et al., 2016) for noncovex problem, we define a Lyapunov function as

$$R_t = \hat{F}(w_t) + r_t\|R_{\tilde{w}_s}^{-1}(w_t)\|^2, \tag{11}$$

with a series of auxiliary parameters r_t satisfying $r_m = 0$, $r_t = (1 + \beta)r_{t+1} + 3\eta\alpha^2b^2$ for $t = 1, 2, ..., m$, where $\beta = 1/m$. Note that the definition of auxiliary parameters are different from those defined in RSVRG (Zhang et al., 2016).

To bound R_t, we need to bound $\hat{F}(w_t)$ and $\|R_{\tilde{w}_s}^{-1}(w_t)\|^2$. The latter one is equivalent to the Riemannian distance between $\tilde{w}_s$ and w_t. And to bound it, we need to use the trigonometric geometry. Note that the trigonometric geometry in a Riemannian manifold is fundamentally different from the Euclidean space. However, with Lemma 5 proposed in (Zhang & Sra, 2016), the side lengths of a geodesic triangle can be upper bounded if the curvature on the manifold is lower bounded by some constant. For the hypersphere, the curvature is constant as 1, and therefore for any vectors $w, u, z \in \mathcal{M}$, the following inequality holds (Zhang et al., 2016):

$$\mathrm{D}^2(w, u) \leq \mathrm{D}^2(w, z) + \mathrm{D}^2(w, z) - 2\langle R_z^{-1}(u), R_z^{-1}(w)\rangle. \tag{12}$$

We then have

$$\begin{aligned}\|R_{\tilde{w}_s}^{-1}(w_{t+1})\|^2 &\leq \|R_{\tilde{w}_s}^{-1}(w_t)\|^2 + \|R_{w_t}^{-1}(w_{t+1})\|^2 - 2\langle R_{w_t}^{-1}(\tilde{w}_s), R_{w_t}^{-1}(w_{t+1})\rangle \\ &= \|R_{\tilde{w}_s}^{-1}(w_t)\|^2 + \eta^2\|G(w_t)\|^2 + 2\langle R_{w_t}^{-1}(\tilde{w}_s), \eta G(w_t)\rangle \\ &\leq \|R_{\tilde{w}_s}^{-1}(w_t)\|^2 + \eta^2\|G(w_t)\|^2 + \frac{1}{\beta}\eta^2\|G(w_t)\|^2 + \beta\|R_{\tilde{w}_s}^{-1}(w_t)\|^2 \\ &= (1+\frac{1}{\beta})\eta^2\|G(w_t)\|^2 + (1+\beta)\|R_{\tilde{w}_s}^{-1}(w_t)\|^2\end{aligned} \tag{13}$$

The first inequality follows the trigonometric geometry in hypersphere manifold. The second inequality comes from a simple application of Cauchy-Schwarz and Youngs inequality that $2\langle a, b\rangle \leq \frac{1}{\beta}\|a\|^2 + \beta\|b\|^2$.

Plugging (10) and (13) into R_t, we have

$$\begin{aligned}R_{t+1} &= \hat{F}(w_{t+1}) + r_{t+1}\|R_{\tilde{w}_s}^{-1}(w_{t+1})\|^2 \\ &\leq \hat{F}(w_t) - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2 + \left((1+\beta)r_{t+1} + 3\eta\alpha^2 b^2\right)\|R_{\tilde{w}_s}^{-1}(w_t)\|^2 \\ &\qquad + \eta^2\left(r_{t+1}(1+\frac{1}{\beta}) + \frac{\lambda}{2} - \frac{1}{2\eta}\right)\|G(w_t)\|^2 \\ &\leq R_t - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2.\end{aligned} \tag{14}$$

The second inequality is by the definition of R_t and the following inequality,

$$r_{t+1}(1+\frac{1}{\beta}) + \frac{\lambda}{2} \leq \frac{1}{2\eta}. \tag{15}$$

The proof of the inequality (15) is in Appendix B.5.1.

Sum up (14) from $t = 0$ to $t = m$, we have

$$R_m \leq R_0 - \sum_{t=0}^{m-1}\frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2.$$

Substituting $\tilde{w}_s = w_0$ and $r_m = 0$, the inequality above is equivalent to

$$\hat{F}(w_m) \leq \hat{F}(\tilde{w}_s) - \sum_{t=0}^{m-1}\frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2. \tag{16}$$

Note that inequality (16) holds for any $w_m = w_m^{s,k}$ where $k = 1, 2, ..., K$.

Since the global variable $\tilde{w}_{s+1}$ is randomly sampled from the local output of the s-th epoch of local computation $\{w_m^{s,k}\}_{k=1}^K$, we have,

$$\mathbb{E}[\hat{F}(\tilde{w}_{s+1})] \leq \frac{1}{K}\sum_{k=1}^{K}\hat{F}(w_m^{s,k}) \leq \hat{F}(\tilde{w}_s) - \frac{1}{K}\sum_{k=1}^{K}\sum_{t=0}^{m-1}\frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t^{s,k})\|^2. \tag{17}$$

By the definition of output in Algorithm 1, we obtain

$$\begin{aligned}\mathbb{E}[\|\tilde{\nabla}\hat{F}(w_a)\|^2] &= \frac{1}{KSm}\sum_{s=0}^{S-1}\sum_{k=1}^{K}\sum_{t=0}^{m-1}\|\tilde{\nabla}\hat{F}(w_t^{s,k})\|^2 \\ &\leq \frac{2}{\eta Sm}\sum_{s=0}^{S-1}\left[\hat{F}(\tilde{w}_s) - \hat{F}(\tilde{w}_{s+1})\right] \\ &\leq \frac{2\lambda}{\rho Sm}\sum_{s=0}^{S-1}\left(\hat{F}(\tilde{w}_s) - \hat{F}(w^*)\right).\end{aligned} \tag{18}$$

The first inequality uses (17) and the second inequality follows the definition of η and the fact that $\hat{F}(\tilde{w}_{s+1}) \geq \hat{F}(w^*)$.

□

Remark: As the proof of Theorem 1 begins with the L-g-smoonthness property of the objective function and the optimization accuracy is measured by the squred norm of the gradient, it seems that the proof could be simply generalized to other nonconvex problem. However, this is not the fact. In the second inequality of (10), we apply Lemma 4, which is a special property of our objective function. And (12) is a property on hypershpere, which may not be simply generalized to other Riemannian manifold. Therefore, to generalize the proof of Theorem 1 to other objective function, it is required to verify whether Lemma 4 and (12) are satisfied.

B.5.1. PROOF FOR INEQUALITY (15)

Proof. Note that $\eta = \rho/\lambda$ where ρ satisfies $12\alpha^2b^2\rho^2m^2/\lambda^2 + \rho \leq 1$. By definition of r_t, we have $r_m = 0$ and $r_t = (1+\beta)r_{t+1} + 3\eta\alpha^2b^2$. Recursively calculating from $t = m$ to $t = 0$, we have

$$\begin{aligned} r_t = 3\eta\alpha^2b^2\frac{(1+\beta)^{m-t}-1}{\beta} &= \frac{3\alpha^2b^2\rho m}{\lambda}\left((1+\frac{1}{m})^{m-t}-1\right) \\ &\leq \frac{e-1}{2}\frac{3\alpha^2b^2\rho m}{\lambda} \\ &\leq \frac{3\alpha^2b^2\rho m}{\lambda} \end{aligned} \tag{19}$$

The second equality follows the definition of η and β. The first inequality uses the limitation of an increasing function $(1+\frac{1}{x})^x$ is the Euler's number, namely $\lim_{x\to+\infty}(1+\frac{1}{x})^x = e$. And the second inequality comes from the fact that $e \leq 3$.

With the upper bound of r_t for any t, we have

$$\begin{aligned} r_{t+1}(1+\frac{1}{\beta}) + \frac{\lambda}{2} &\leq \frac{3\alpha^2b^2\rho}{\lambda}m(1+m) + \frac{\lambda}{2} \\ &\leq \frac{6\alpha^2b^2\rho m^2}{\lambda} + \frac{\lambda}{2} \\ &\leq \frac{1-\rho}{2\rho}\lambda + \frac{\lambda}{2} = \frac{\lambda}{2\rho} = \frac{1}{2\eta} \end{aligned} \tag{20}$$

The second inequality follows thta $m \geq 1$, and the third inequality uses the setting $12\alpha^2b^2\rho^2m^2/\lambda^2 + \rho \leq 1$. □

B.6. Proof for Theorem 2

Before proceed to the proof, we first propose a lemma to bound $\|\tilde{\nabla}f_{k,i}(w) - \mathcal{T}_u^w\tilde{\nabla}f_{k,i}(u)\|^2$ for any $w, u \in \mathcal{M}$.

Lemma 10. *Given any $w, u \in \mathcal{M}$, it holds that*

$$\|\tilde{\nabla}f_{k,i}(w) - \mathcal{T}_u^w\tilde{\nabla}f_{k,i}(u)\|^2 \leq 6b^2D^2(w,u)$$

Proof. Define $A_{k,i} = x_i^k(x_i^k)^\top$. Then similar to the proof of Lemma 4, we have

$$\begin{aligned} &\|\tilde{\nabla}f_{k,i}(w) - \mathcal{T}_u^w\tilde{\nabla}f_{k,i}(u)\|^2 \\ =&\|-(I-ww^\top)A_{k,i}w + (I-ww^\top)(I-uu^\top)A_{k,i}u\|^2 \\ =&\|(I-ww^\top)A_{k,i}(w-u) + (I-ww^\top)uu^\top A_{k,i}u\|^2 \\ \leq&2\|(I-ww^\top)A_{k,i}(w-u)\|^2 + 2\|(I-ww^\top)uu^\top A_{k,i}u\|^2 \\ \leq&2\|I-ww^\top\|_2^2\|A_{k,i}\|_2^2\|w-u\|^2 + 2\|(I-ww^\top)u\|^2(u^\top A_{k,i}u)^2 \\ =&2b^2\left(\|w-u\|^2 + \|(I-ww^\top)u\|^2\right) \\ =&2b^2\left(3-2w^\top u-(w^\top u)^2\right) \\ \leq&6b^2\left(1-(w^\top u)^2\right) \\ \leq&6b^2\mathrm{D}^2(u,w). \end{aligned} \tag{21}$$

□

Now we are ready to provide proof for Theorem 2.

Proof. Again, for simplification of notation, we denote $w_t^{s,k}$ by w_t. With (3) replacing Step 10 in Algorithm 1, the local variable is updated with

$$g_t = \frac{1}{B}\sum_{i\in\mathcal{I}_t}\tilde{\nabla} f_{k,i}(w_t) - \mathcal{T}_{\tilde{w}_s}^{w_t}\left(\frac{1}{B}\sum_{i\in\mathcal{I}_t}\tilde{\nabla} f_{k,i}(\tilde{w}_s) - \tilde{\nabla}\hat{F}(\tilde{w}_s)\right). \tag{22}$$

Notice that $\mathbb{E}[g_t] = G(w_t)$, where $G(w)$ is defined in (6). Then the variance of g_t is upper bounded as

$$\begin{aligned}
\mathbb{E}\left[\|g_t - G(w_t)\|^2\right] &= \mathbb{E}[\Big\|\frac{1}{B}\sum_{i\in\mathcal{I}_t}\Big(\tilde{\nabla} f_{k,i}(w_t) - \mathcal{T}_{\tilde{w}_s}^{w_t}\tilde{\nabla} f_{k,i}(\tilde{w}_s)\Big) + \mathcal{T}_{\tilde{w}_s}^{w_t}\tilde{\nabla}\hat{F}(\tilde{w}_s) \\
&\qquad - \Big(\tilde{\nabla} f_k(w_t) - \mathcal{T}_{\tilde{w}_s}^{w_t}(\tilde{\nabla} f_k(\tilde{w}_s) - \tilde{\nabla}\hat{F}(\tilde{w}_s))\Big)\Big\|^2] \\
&= \frac{1}{B^2}\mathbb{E}\left[\Big\|\sum_{i\in\mathcal{I}_t}\Big(\tilde{\nabla} f_{k,i}(w_t) - \mathcal{T}_{\tilde{w}_s}^{w_t}\tilde{\nabla} f_{k,i}(\tilde{w}_s) - \Big(\tilde{\nabla} f_k(w_t) - \mathcal{T}_{\tilde{w}_s}^{w_t}\tilde{\nabla} f_k(\tilde{w}_s)\Big)\Big)\Big\|^2\right] \\
&\le \frac{1}{B^2}\mathbb{E}\left[\Big\|\sum_{i\in\mathcal{I}_t}\Big(\tilde{\nabla} f_{k,i}(w_t) - \mathcal{T}_{\tilde{w}_s}^{w_t}\tilde{\nabla} f_{k,i}(\tilde{w}_s)\Big)\Big\|^2\right] \\
&\le \frac{6b^2}{B}\mathbb{E}\left[\mathrm{D}^2(w_t,\tilde{w}_s)\right]
\end{aligned} \tag{23}$$

The first inequality uses $\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 = \mathbb{E}\|\xi\|^2 - \|\mathbb{E}\xi\|^2 \le \mathbb{E}\|\xi\|^2$ and the second inequality uses Lemma 10.

By the λ-smoothness of $\hat{F}(w)$, we can derive

$$\begin{aligned}
\mathbb{E}[\hat{F}(w_{t+1})] &\le \mathbb{E}\left[\hat{F}(w_t) + \langle\tilde{\nabla}\hat{F}(w_t), R_{w_t}^{-1}(w_{t+1})\rangle + \frac{\lambda}{2}\left\|R_{w_t}^{-1}(w_{t+1})\right\|^2\right] \\
&= \mathbb{E}\left[\hat{F}(w_t) - \eta\langle\tilde{\nabla}\hat{F}(w_t), G(w_t)\rangle + \frac{\eta^2\lambda}{2}\|g_t\|^2\right] \\
&= \mathbb{E}\left[\hat{F}(w_t) - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2 + \frac{\eta}{2}\|G(w_t) - \tilde{\nabla}\hat{F}(w_t)\|^2 + \eta(\frac{\eta\lambda}{2} - \frac{1}{2})\|g_t\|^2 + \frac{\eta}{2}(\|g_t\|^2 - \|G(w_t)\|^2)\right] \\
&\le \mathbb{E}\left[\hat{F}(w_t) - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2 + \frac{\eta}{2}\|G(w_t) - \tilde{\nabla}\hat{F}(w_t)\|^2 + \eta(\frac{\eta\lambda}{2} - \frac{1}{2})\|g_t\|^2 + \frac{\eta}{2}(\|g_t - G(w_t)\|^2)\right] \\
&\le \mathbb{E}\left[\hat{F}(w_t) - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2 + 3\eta\alpha^2 b^2\mathrm{D}^2(w_t,\tilde{w}_s) + \eta(\frac{\eta\lambda}{2} - \frac{1}{2})\|g_t\|^2 + \frac{3\eta b^2}{B}\mathrm{D}^2(w_t,\tilde{w}_s)\right] \\
&\le \mathbb{E}\left[\hat{F}(w_t) - \frac{\eta}{2}\|\tilde{\nabla}\hat{F}(w_t)\|^2 + 3\eta(\alpha^2 + \frac{1}{B})b^2\mathrm{D}^2(w_t,\tilde{w}_s) + \eta(\frac{\eta\lambda}{2} - \frac{1}{2})\|g_t\|^2\right]
\end{aligned} \tag{24}$$

The second equality is by $\langle u_1, u_2\rangle = \frac{1}{2}\left(\|u_1\|^2 + \|u_2\|^2 - \|u_1 - u_2\|^2\right)$ and by subtracting and adding $\frac{\eta}{2}\|g_t\|^2$. The second inequality uses $\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 = \mathbb{E}\|\xi\|^2 - \|\mathbb{E}\xi\|^2 \le \mathbb{E}\|\xi\|^2$. And the third inequality uses Lemma 4 and (23). Note that (24) is same as (10) except that α^2 is replaced by $\alpha^2 + \frac{1}{B}$. Therefore, the subsequent proof is similar to the proof of Theorem 1 in Appendix B.5.

□

B.7. Proof for Theorem 3

Proof. Since Algorithm 2 calls Algorithm 1, by Theorem 1 we have

$$\mathbb{E}[\|\tilde{\nabla}\hat{F}(\hat{w}_{r+1})\|^2] \le \frac{2\lambda}{\rho m S}\left(\hat{F}(\hat{w}_r) - \hat{F}(w^*)\right). \tag{25}$$

By Lemma 2 and definition of gradient-dominated function, we have

$$\hat{F}(\hat{w}_{r+1}) - \hat{F}(w^*) \leq \frac{2}{\delta}\|\tilde{\nabla}\hat{F}(\hat{w}_{r+1})\|^2. \tag{26}$$

Combining (25) and (26) and telescoping products from $r = 0$ to $r = R - 1$, we obtain

$$\mathbb{E}\left[\hat{F}(\hat{w}_R) - \hat{F}(w^*)\right] \leq \left(\frac{4\lambda}{\rho\delta mS}\right)^R \left(\hat{F}(w_0) - \hat{F}(w^*)\right)$$

□

C. Proof for Sections 4.1, 4.2 and 4.3

C.1. Proof of Lemma 5

Proof. Given any unit norm vector w, rewrite it as $w = \sum_{i=1}^d c_i v_i$, where c_i are scalers satisfying $\sum_{i=1}^d c_i^2 = 1$ and v_i is the eigenvector corresponding to i-th largest eigenvalue λ_i. By definition, $v = v_1$ and $\lambda = \lambda_1$. Then we have $(w^\top v)^2 = c_1^2$. For $\hat{F}(w) - \hat{F}(v)$, it is lower bounded as

$$\begin{aligned}
\hat{F}(w) - \hat{F}(v) &= \frac{1}{2}(\lambda_1 - w^\top \hat{A} w) = \frac{1}{2}(\lambda_1 - \sum_{i=1}^d c_i^2 \lambda_i) \\
&= \frac{1}{2}(\lambda_1 - c_1^2\lambda_1 - \sum_{i=2}^d c_i^2 \lambda_i) \\
&\geq \frac{1}{2}(\lambda_1 - c_1^2\lambda_1 - \sum_{i=2}^d c_i^2 \lambda_2) \\
&= \frac{1}{2}\left((1 - c_1^2)\lambda_1 - (1 - c_1^2)\lambda_2)\right) = \frac{\delta}{2}(1 - (w^\top v)^2).
\end{aligned}$$

And the upper bound is deduced as

$$\begin{aligned}
\hat{F}(w) - \hat{F}(v) &= \frac{1}{2}(\lambda_1 - w^\top \hat{A} w) = \frac{1}{2}(\lambda_1 - \sum_{i=1}^d c_i^2 \lambda_i) \\
&= \frac{1}{2}(\lambda_1 - c_1^2\lambda_1 - \sum_{i=2}^d c_i^2 \lambda_i) \\
&\leq \frac{1}{2}(\lambda_1 - c_1^2\lambda_1) = \frac{\lambda}{2}(1 - (w^\top v)^2).
\end{aligned}$$

□

C.2. Proof of Lemma 6

Proof. Let $\hat{w}_k = \text{sign}(w_k^\top w_1) w_k = \sum_{i=1}^d c_i^k v_i$, where $\sum_{i=1}^d (c_i^k)^2 = 1$ for all k and v_i is the eigenvector corresponding to i-th largest eigenvalue λ_i. By definition, $v = v_1$ and $\lambda = \lambda_1$. Since $(c_1^k)^2 = (v^\top w_k)^2 > \frac{1}{2}$ and $\hat{w}_k^\top \hat{w}_j \geq 0$ for any j and k, we obtain that c_1^k has same sign for all k. Assume $c_1^{(k)} > 0$, then for any fixed $\{c_1^k\}_{k=1}^K$, we have

$$\begin{aligned}
(v^\top \bar{w})^2 &= \frac{(\sum_{k=1}^K v^\top \hat{w}_k)^2}{\|\sum_{k=1}^K \hat{w}_k\|^2} = \frac{(\sum_{k=1}^K c_1^k)^2}{(\sum_{k=1}^K c_1^k)^2 + \sum_{i=2}^d (\sum_{k=1}^K c_i^k)^2} \\
&\geq \frac{(\sum_{k=1}^K c_1^k)^2}{(\sum_{k=1}^K c_1^k)^2 + (\sum_{k=1}^K \sqrt{1 - (c_1^k)^2})^2}
\end{aligned}$$

The inequality comes from the fact that $\| \sum_{k=1}^K u_k\|^2 \leq (\sum_{k=1}^K \|u_k\|)^2$, where we set the vector $u_k = [c_2^k, c_3^k, ..., c_d^k]^\top$ and the fact $\|u_k\| = \sqrt{\sum_{i=2}^d (c_i^k)^2} = \sqrt{1-(c_1^k)^2}$. Notice that

$$\frac{1}{K}\sum_{k=1}^K (v^\top w_k)^2 = \frac{1}{K}\sum_{k=1}^K (c_1^k)^2.$$

By define $c_1^k = \cos(\theta_k)$, where $\theta_k \in [0, \pi/4)$, we have $\sqrt{1-(c_1^k)^2}$. Then to prove the lemma is equivalent to proving

$$\begin{aligned}
& (v^\top \bar{w})^2 && \geq && \frac{1}{K}\sum_{k=1}^K (v^\top w_k)^2 \\
\Leftrightarrow \quad & \frac{(\sum_{k=1}^K c_1^k)^2}{(\sum_{k=1}^K c_1^k)^2 + (\sum_{k=1}^K \sqrt{1-(c_1^k)^2})^2} && \geq && \frac{1}{K}\sum_{k=1}^K (c_1^k)^2 \\
\Leftrightarrow \quad & \frac{(\sum_{k=1}^K \cos(\theta_k))^2}{(\sum_{k=1}^K \cos(\theta_k))^2 + (\sum_{k=1}^K \sin(\theta_k))^2} && \geq && \frac{\sum_{k=1}^K \cos^2(\theta_k)}{\sum_{k=1}^K \cos^2(\theta_k) + \sum_{k=1}^K \sin^2(\theta_k)} \\
\Leftrightarrow \quad & \frac{(\sum_{k=1}^K \cos(\theta_k))^2}{(\sum_{k=1}^K \sin(\theta_k))^2} && \geq && \frac{\sum_{k=1}^K \cos^2(\theta_k)}{\sum_{k=1}^K \sin^2(\theta_k)} \\
\Leftrightarrow \quad & \frac{\sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i)\cos(\theta_j)}{\sum_{i=1}^K \sum_{j=1}^K \sin(\theta_i)\sin(\theta_j)} && \geq && \frac{\sum_{k=1}^K \cos^2(\theta_k)}{\sum_{k=1}^K \sin^2(\theta_k)} \\
\Leftrightarrow \quad & \frac{\frac{1}{2}\sum_{i=1}^K \sum_{j=1}^K \big(\cos(\theta_i-\theta_j) + \cos(\theta_i+\theta_j)\big)}{\frac{1}{2}\sum_{i=1}^K \sum_{j=1}^K \big(\cos(\theta_i-\theta_j) - \cos(\theta_i+\theta_j)\big)} && \geq && \frac{\sum_{k=1}^K \cos^2(\theta_k)}{\sum_{k=1}^K \sin^2(\theta_k)}
\end{aligned}$$

$$\begin{aligned}
& \Leftrightarrow \sum_{i=1}^K \sum_{j=1}^K \big(\cos(\theta_i-\theta_j) + \cos(\theta_i+\theta_j)\big) \sum_{k=1}^K \sin^2(\theta_k) \geq \sum_{k=1}^K \cos^2(\theta_k) \sum_{i=1}^K \sum_{j=1}^K \big(\cos(\theta_i-\theta_j) - \cos(\theta_i+\theta_j)\big) \\
& \Leftrightarrow K\sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i+\theta_j) \geq \sum_{k=1}^K \big(\cos^2(\theta_k) - \sin^2(\theta_k)\big) \sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i-\theta_j) \\
& \Leftrightarrow K\sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i+\theta_j) \geq \sum_{k=1}^K \cos(2\theta_k) \sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i-\theta_j) \\
& \Leftrightarrow \sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i+\theta_j) \geq \sum_{i=1}^K \sum_{j=1}^K \Big(\frac{\cos(2\theta_i)+\cos(2\theta_j)}{2}\Big) \frac{\sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i-\theta_j)}{K^2}
\end{aligned}$$

The last inequality holds due to the fact that $\cos(\theta) \leq 1$ such that

$$\frac{\sum_{i=1}^K \sum_{j=1}^K \cos(\theta_i-\theta_j)}{K^2} \leq 1,$$

and the fact that the function $\cos(\theta)$ is concave for $\theta \in [0, \pi/2)$ such that

$$\cos(\theta_i+\theta_j) = \cos\big(\frac{1}{2}\,2\theta_i + \frac{1}{2}\,2\theta_j\big) \geq \frac{\cos(2\theta_i)+\cos(2\theta_i)}{2}$$

□

C.3. Proof of Theorem 4

Proof. Assume that $(v^\top w_m^{s,k})^2 > 1/2$ holds for all s and k. Combining Lemma 6 and Lemma 5, we have that

$$\begin{aligned}\hat{F}(\tilde{w}_{s+1}) - \hat{F}(v) &\leq \frac{\lambda}{2}\big(1-(v^\top \tilde{w}_{s+1})^2\big)\\ &\leq \frac{\lambda}{2K}\sum_{k=1}^{K}\big(1-(v^\top w_m^{s,k})^2\big)\\ &\leq \frac{\lambda}{\delta K}\sum_{k=1}^{K}\hat{F}(w_m^{s,k}) - \hat{F}(v).\end{aligned}$$

That is

$$\hat{F}(\tilde{w}_{s+1}) \leq \frac{\lambda}{\delta K}\sum_{k=1}^{K}\hat{F}(w_m^{s,k}). \tag{27}$$

Replacing (18) with (27) in proof of Theorem 1, we can achieve the conclusion in Theorem 4.

□

C.4. Proof of Lemma 7

Proof. We study the convexity of $\hat{F}(w)$ on hypersphere manifold by the smallest eigenvalue of its Riemannian Hessian. By (8), we have

$$\langle u, \tilde{\nabla}^2\hat{F}(w)[u]\rangle = -u^\top \hat{A}u + w^\top \hat{A}w.$$

Define $\varepsilon = 1-(w^\top v)^2$. Rewrite w as $w = \sum_{i=1}^d c_i v_i$, where c_i are scalers satisfying $\sum_{i=1}^d c_i^2 = 1$ and $c_1 > 0$. Then $w^\top v_1 = c_1$ and $c_1^2 = 1-\varepsilon^2$.

Rewrite u as $u = \sum_{i=1}^d a_i v_i$. By $u^\top w = 0$, we have

$$\begin{aligned}u^\top w = 0 &\Leftrightarrow \sum_{i=1}^{d} a_i c_i = 0\\ &\Leftrightarrow (a_1c_1)^2 = (\sum_{i=2}^{d} a_i c_i)^2 \leq \sum_{i=2}^{d} a_i^2 \sum_{i=2}^{d} c_i^2 = (1-a_1^2)(1-c_1^2)\\ &\Leftrightarrow a_1^2 \leq 1-c_1^2\end{aligned} \tag{28}$$

Therefore, $(u^\top v_1)^2 \leq 1-c_1^2 = \varepsilon^2$. This indicates $u^\top Au \leq \varepsilon^2\lambda_1 + (1-\varepsilon^2)\lambda_2$.

Then

$$\begin{aligned}-u^\top Au + w^\top Aw &\geq -\varepsilon^2\lambda_1 - (1-\varepsilon^2)\lambda_2 + \sum_{i=1}^{d} c_i^2\lambda_i\\ &\geq -\varepsilon^2\lambda_1 - (1-\varepsilon^2)\lambda_2 + c_1^2\lambda_1\\ &= -\varepsilon^2\lambda_1 - (1-\varepsilon^2)\lambda_2 + (1-\varepsilon^2)\lambda_1\\ &= \delta - \varepsilon^2(\lambda_1+\delta)\end{aligned} \tag{29}$$

Let $\delta - \varepsilon^2(\lambda_1+\delta) \geq 0$, we obtain $\varepsilon^2 \leq \frac{\delta}{\lambda_1+\delta}$. That is, when $1-(w^\top v)^2 < \frac{\delta}{\lambda_1+\delta}$, the Riemannian Hessian of $\hat{F}(w)$ has non-negative smallest eigenvalue, indicating that $F(w)$ is g-convex.

Let $\delta - \varepsilon^2(\lambda_1+\delta) \geq \frac{\delta}{2}$, we obtain $\varepsilon^2 \leq \frac{1}{2}\frac{\delta}{\lambda_1+\delta}$. Since the smallest value of the Hessian is not smaller than $\delta/2$, $\hat{F}(w)$ is $\delta/2$-strongly geodesics-convex in the Riemannian ball $\{w : (w^\top v_1)^2 \geq 1-\frac{1}{2}\frac{\delta}{\lambda_1+\delta}\}$.

A byproduct during studying the convexity of $\hat{F}(w)$ is that among all stationary points v_i for $i = 1, 2, ..., d$ of the objective function $\hat{F}(w)$ on Riemannian manifold, only v_1 is not a saddle point, and v_i for $i = 2, 3, ..., d$ are saddle points.

□

C.5. Proof of Theorem 5

Proof. Assume that $w_m^{s,k} \in \mathcal{A}$ holds for all s and k, where $\mathcal{A} = \{w \in \mathcal{M} : (w^\top v)^2 \geq 1 - \frac{\delta}{\lambda+\delta}\}$. Based on Lemma 7, we have that $\hat{F}(w)$ is geodesic convexity for any $w \in \mathcal{A}$.

For $\bar{w}_k$ ($k = 1, 2, ..., K$) defined in (5), by applying the Jensen's inequality along geodesic on $\mathcal{M}$, we have

$$\hat{F}(\bar{w}_k) \leq \frac{k-1}{k}\hat{F}(\bar{w}_{k-1}) + \frac{1}{k}\hat{F}(w_m^{s,k}),$$

where $k = 2, 3, ..., K$. Sum up from $k = 2$ to $k = K$, we have

$$\hat{F}(\tilde{w}_{s+1}) = \frac{1}{K}\sum_{k=1}^{K}\hat{F}(w_m^{s,k}).$$

Based on the above inequaltiy, we know that (18) still holds after switching the averaging stategy from option I to option III. Thus, we can have same result as in Theorem 3.

□

Remark: By applying Lemma 6 to the function $h(w) = 1 - (v^\top w)^2$ where $w \in \mathcal{M}$, we have that $h(w)$ is g-convex if $(v^\top w) \geq 1/2$. Therefore, running Algorithm 1 with option III and with other assumptions and parameter settings described in Theorem 4, we can achieve the same conclusion as presented in Theorem 4.

C.6. Proof of Lemma 8

Proof. Define the leading eigenvalue of A_1 by λ_1. By the result in Lemma 3, we have $\|A_1 - \hat{A}\|_2 \leq \alpha b$. That is for any unit norm vector w, we have

$$|w^\top A_1 w - w^\top \hat{A} w| \leq \alpha b. \tag{30}$$

If $\lambda_1 \geq \lambda$, we have

$$\lambda_1 - \lambda = v_1^\top A_1 v_1 - v^\top \hat{A} v \leq v_1^\top A_1 v_1 - v_1^\top \hat{A} v_1 \leq \alpha b$$

Similar, when $\lambda \geq \lambda_1$, we have

$$\lambda - \lambda_1 = v^\top \hat{A} v - v_1^\top A_1 v_1 \leq v^\top \hat{A} v - v^\top A_1 v \leq \alpha b \tag{31}$$

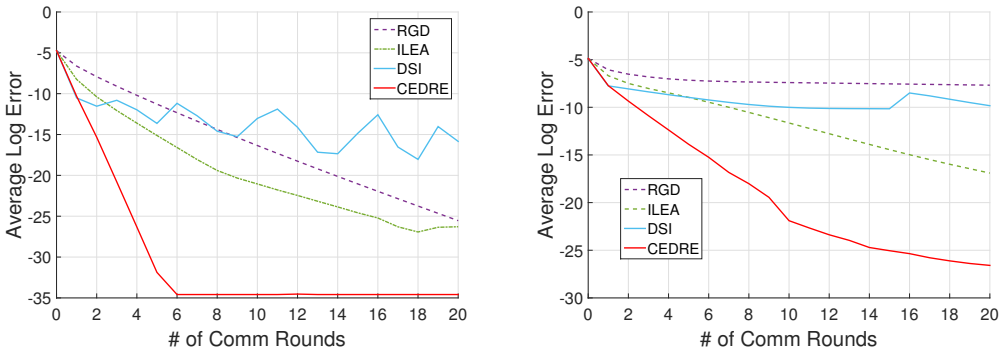
Therefore we have $|\lambda - \lambda_1| \leq \alpha b$. With this result, we have

$$\begin{aligned} \hat{F}(v_1) - \hat{F}(v) &= \frac{1}{2}(\lambda - v_1^\top \hat{A} v_1) \\ &= \frac{1}{2}(\lambda - \lambda_1 + v_1^\top A_1 v_1 - v_1^\top \hat{A} v_1) \\ &\leq \frac{1}{2}\left(|\lambda - \lambda_1| + |v_1^\top A_1 v_1 - v_1^\top \hat{A} v_1|\right) \\ &= \alpha b \end{aligned} \tag{32}$$

The second equality uses $\lambda_1 = v_1^\top A_1 v_1$ and the inequality uses (30) and (31).

□

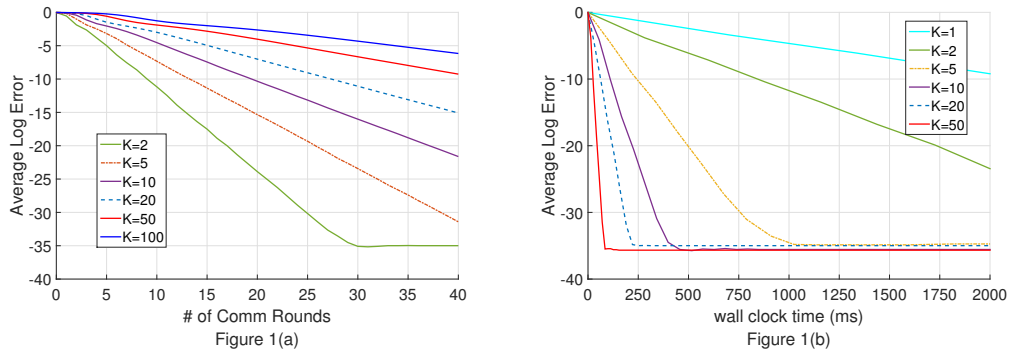
D. Extra Experiment Results on Synthetic datasets



(a) $\delta = 0.2$ (b) $\delta = 0.01$

Figure 5. Comparison results of different distributed optimization algorithms on synthetic datasets with different eigengap.

We compare CEDRE with baseline algorithms on the synthetic datasets described in Section 6.3 with $n = 2000$. The results are presented in Figure 5. When the eigengap is large (i.e. $\delta = 0.2$), the convergence regarding the communication cost of CEDRE outperforms its competitors by a large margin. To be specific, CEDRE converges to about -35 log error with only 6 communication rounds while other algorithms cannot converge to the same accuracy after communicating 20 communication rounds. When the eigengap is small (i.e. $\delta = 0.01$), CEDRE again outperforms other competitors. Therefore, CEDRE is more communication-efficient to compute the leading eigenvector in distributed settings.



(a) (b)

Figure 6. Results on synthetic datasets. (a) displays convergence of CEDRE v.s. number of communication rounds for different number of local mahines. (b) displays convergence of CEDRE v.s. computation time for different number of local mahines.

We also test the effect of the number of local machines on the convergence of CEDRE on the synthetic datasets. In this experiment, the number of total data instances is fixed at $200,000$. And the number of local machines K varies from 1 to 100. Specially, when testing the convergence vs. number of communication rounds (Figure 6(a)), the local computation iteration length m is set as $\lfloor 10\sqrt{n} \rfloor$. And when testing the convergence vs. local computation time (Figure 6(b)), the communication time is not calculated. But in practice, the communication time is much higher than the computation time and dominates the total running time of a distributed algorithm. The results in Figure 6 show that to achieve the same accuracy, with the increase of the number of local machines K, the number of communication rounds increases, but the wall-clock time, i.e. the computation cost in each local machine, decreases.