Towards better understanding of decentralized optimization using row and column stochastic matrices

Liyuan Liang, Gan Luo

1 Related Works

To be filled.

1.1 Push-Pull Algorithm

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} - \alpha \mathbf{y}^{(k)} \tag{1}$$

$$\mathbf{y}^{(k+1)} = B\mathbf{y}^{(k)} + \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)} \tag{2}$$

2 New Approach

Consider using $\bar{x}^{(k)} := \frac{1}{n}\mathbb{1}_n^T \mathbf{x}^{(k)}$ and $\bar{\mathbf{x}}^{(k)} := \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T \mathbf{x}^{(k)}$ as the true parameter.

2.1 Descent Lemma

Lemma 1.

$$f(\bar{x}^{(k+1)}) \leq f(\bar{x}^{(k)}) - \frac{\alpha}{4}\|\bar{y}^{(k)}\|^2 - \frac{\alpha}{4}\|\nabla f(\bar{x}^{(k)})\|^2 + (\frac{\alpha L^2}{2} + \frac{2}{\alpha n^2})\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2 \tag{3}$$

Proof. Notice that

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \frac{1}{n}\mathbb{1}_n^T A(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}) - \alpha \bar{y}^{(k)} \tag{4}$$

By L-smooth inequality, we have

$$f(y) \leq f(x) + \langle y - x, \nabla f(x)\rangle + \frac{L}{2}\|y - x\|^2 \tag{5}$$

Thus, using the smoothness of $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, we have

$$
\begin{aligned}
f(\bar{x}^{(k+1)}) &\le f(\bar{x}^{(k)}) - \alpha\langle \bar{y}^{(k)}, \nabla f(\bar{x}^{(k)})\rangle + \langle \frac{1}{n}\mathbb{1}_n^T A(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}), \nabla f(\bar{x}^{(k)})\rangle + \frac{\alpha^2 L}{2}\|\bar{y}^{(k)}\|^2 \\
&\overset{A-M}{\le} f(\bar{x}^{(k)}) - \frac{\alpha - \alpha^2 L}{2}\|\bar{y}^{(k)}\|^2 - \frac{\alpha}{2}\|\nabla f(\bar{x}^{(k)})\|^2 + \frac{\alpha}{2}\|\bar{y}^{(k)} - \nabla f(\bar{x}^{(k)})\|^2 \\
&\quad + \frac{\alpha}{4}\|\nabla f(\bar{x}^{(k)})\|^2 + \frac{2}{\alpha}\|\frac{1}{n}\mathbb{1}_n^T A(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)})\|^2 \\
&\overset{\alpha \le \frac{1}{2L}}{\le} f(\bar{x}^{(k)}) - \frac{\alpha}{4}\|\bar{y}^{(k)}\|^2 - \frac{\alpha}{4}\|\nabla f(\bar{x}^{(k)})\|^2 + \frac{\alpha}{2}\|\bar{y}^{(k)} - \nabla f(\bar{x}^{(k)})\|^2 \\
&\quad + \frac{2}{\alpha n^2}\|\mathbb{1}_n^T A(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)})\|^2 \\
&= f(\bar{x}^{(k)}) - \frac{\alpha}{4}\|\bar{y}^{(k)}\|^2 - \frac{\alpha}{4}\|\nabla f(\bar{x}^{(k)})\|^2 + \frac{\alpha}{2n}\|\sum_{i=1}^{n}(\nabla f_i(x_i^{(k)}) - \nabla f_i(\bar{x}^{(k)})\|^2 \\
&\quad + \frac{2}{\alpha n^2}\|(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)})\|^2 \\
&\le f(\bar{x}^{(k)}) - \frac{\alpha}{4}\|\bar{y}^{(k)}\|^2 - \frac{\alpha}{4}\|\nabla f(\bar{x}^{(k)})\|^2 + (\frac{\alpha L^2}{2} + \frac{2}{\alpha n^2})\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|^2
\end{aligned} \tag{6}
$$

This suffices to controlling the size of consensus error. $\square$

2.2 Consensus Error

Lemma 2.

$$
\sum_{k=0}^{T}\|(I-R)\mathbf{x}^{(k+1)}\|^2 \le \frac{\alpha^2\kappa_A^2}{(1-\beta)^2}\sum_{k=0}^{T}\|(I-R)\mathbf{y}^{(k)}\|^2 \tag{7}
$$

Proof. The consensus error can be expressed by $(I-R)\mathbf{x}^{(k)}$ where $R := \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$.

$$
(I-R)\mathbf{x}^{(k+1)} = (A - RA)(I-R)\mathbf{x}^{(k)} - \alpha(I-R)\mathbf{y}^{(k)} \tag{8}
$$

Notice that $(A-RA)^k(I-R) = (I-R)(A-A_\infty)^k(I-R)$, its size decays exponentially fast. The same for the consensus error.

$$
\begin{aligned}
&\|(I-R)\mathbf{x}^{(k+1)}\|^2 \\
=&\alpha^2\|\sum_{i=0}^{k}(I-R)(A-A_\infty)^{k-i}(I-R)\mathbf{y}^{(i)}\|^2 \\
\overset{\text{Jensen}}{\le}& \frac{\alpha^2}{1-\beta_A}\sum_{i=0}^{k}\kappa_A^2\beta_A^{k-i}\|(I-R)\mathbf{y}^{(i)}\|^2 \\
=&\frac{\alpha^2\kappa_A^2}{1-\beta_A}\sum_{i=0}^{k}\beta_A^{k-i}\|(I-R)\mathbf{y}^{(i)}\|^2
\end{aligned} \tag{9}
$$

Here κ_A, β_A are some positive constants satisfying $\|(A - A_\infty)^k\|_2 \leq \kappa_A \beta_A^k$. For example, κ_A can be selected as the condition number of A and β is the second largest eigenvalue. All matrix norm here is Frobenius norm. □

2.3 Gradient Consensus Error

Lemma 3.

$$\sum_{k=0}^{T} \|(I-R)\mathbf{y}^{(k+1)}\|^2 \leq \frac{3\kappa_B^2 L^2(2\|A-RA\|_2^2+1+2n)}{(1-\beta)^2}\sum_{k=0}^{T}\|(I-R)\mathbf{x}^{(k)}\|^2 + \frac{6n^2\alpha^2\kappa_B^2L^2}{1-\beta}\sum_{k=0}^{T}\|\bar{y}^{(k)}\|^2 \tag{10}$$

[LLY: hi, Gan Luo, my proof here is correct but its technique may be not the best when we consider stochastic gradient. In stochastic case, using this proof, the order of $1-\beta$ may be higher than 2. Can you follow the proof in my paper (Lemma B.2-Lemma B.5) to provide a new proof in stochastic case? (and compare which one provides a better constant)]

Proof. By Cauchy inequality, (8) indicates that

$$\|(I-R)\mathbf{x}^{(k+1)}\|^2 \leq 2\|A-RA\|_2^2\|(I-R)\mathbf{x}^{(k)}\|^2 + 2\alpha^2\|(I-R)\mathbf{y}^{(k)}\|^2 \tag{11}$$

By Cauchy inequality, (4) indicates that

$$\|R(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})\|^2 \leq \frac{2}{n}\|\mathbb{1}_n^T A(I-R)\mathbf{x}^{(k)}\|^2 + 2n\alpha^2\|\bar{y}^{(k)}\|^2 \leq 2n\|(I-R)\mathbf{x}^{(k)}\|^2 + 2n\alpha^2\|\bar{y}^{(k)}\|^2 \tag{12}$$

$$\begin{aligned}
&\|(I-R)\mathbf{y}^{(k+1)}\|^2\\
=&\|(B-R)\mathbf{y}^{(k)} + (I-R)(\mathbf{g}^{(k+1)}-\mathbf{g}^{(k)})\|^2\\
=&\|\sum_{i=0}^{k}(B-B_\infty)^{k-i}(I-R)(\mathbf{g}^{(i+1)}-\mathbf{g}^{(i)})\|^2\\
\overset{\text{Jensen}}{\leq}&\frac{\kappa_B^2}{1-\beta}\sum_{i=0}^{k}\beta^{k-i}\|\mathbf{g}^{(i+1)}-\mathbf{g}^{(i)}\|^2\\
\leq&\frac{\kappa_B^2L^2}{1-\beta}\sum_{i=0}^{k}\beta^{k-i}\|\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)}\|^2\\
\leq&\frac{3\kappa_B^2L^2}{1-\beta}\sum_{i=0}^{k}\beta^{k-i}(\|(I-R)\mathbf{x}^{(i+1)}\|^2 + \|R(\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)})\|^2 + \|(I-R)\mathbf{x}^{(i)}\|^2)\\
\overset{(11),(12)}{\leq}&\frac{3\kappa_B^2L^2(2\|A-RA\|_2^2+1+2n)}{1-\beta}\sum_{i=0}^{k}\beta^{k-i}\|(I-R)\mathbf{x}^{(i)}\|^2 + \frac{6n^2\alpha^2\kappa_B^2L^2}{1-\beta}\sum_{i=0}^{k}\beta^{k-i}\|\bar{y}^{(i)}\|^2
\end{aligned} \tag{13}$$

□

2.4 Main Theorem

For non-stochastic case When $\alpha = \mathcal{O}(\frac{1}{L})$ is small enough, we have

Theorem 1.

$$\frac{1}{T+1}\sum_{k=0}^{T}\|\nabla f(\bar{x}^{(k)})\|^2 \leq \frac{4(f_0 - f^*)}{\alpha(T+1)} \tag{14}$$

[LLY: Please notice that we have not added stochastic noise. When there is some noise, we can show linear speedup easily.]

Proof. □

For stochastic gradient case, we when α satisfies some condition, wec have

Theorem 2.

$$\frac{1}{T+1}\sum_{k=0}^{T}\|\nabla f(\bar{x}^{(k)})\|^2 \leq (\frac{4\sigma^2(f_0 - f^*)}{n(T+1)})^{\frac{1}{2}} + \textit{network influence} \tag{15}$$

References