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

Anonymous Authors

1 Related Works

[LLY: Hi Gan Luo, Please fill in related works you have read here. Don't forget to make some comments.]

1.1 Push-Pull Algorithm

$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} a_{ij}\mathbf{x}_j(k) - \eta \mathbf{y}_i(k)$

$\mathbf{y}_i(k+1) = \sum_{j=1}^{n} b_{ij}(\mathbf{y}_j(k) + \mathbf{r}_j(k))$, where $\mathbf{r}_j(k) = \nabla f_j(\mathbf{x}_j(k+1)) - \nabla f_j(\mathbf{x}_j(k))$

$a_{ij} > 0$ if $j \in \mathcal{N}_i^{\text{in}}$, $\sum_{j}^{n} a_{ij} = 1, \forall i$, $\underline{A} = (a_{ij})$ is row-stochastic, can be determined by the node i itself

$b_{ij} > 0$ if $i \in \mathcal{N}_j^{\text{out}}$, $\sum_{i}^{n} b_{ij} = 1, \forall j$, $\underline{B} = (b_{ij})$ is column-stochastic, we can just let$b_{ij} = \dfrac{1}{|\mathcal{N}_j^{\text{out}}|}$

Remark:$\mathbf{y}_i(k)$ will approach to $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k))$, the gradient of the aggregated cost function $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, so the first step of the algorithm will approach to a centralized gradient decent.

1.2 work

$\pi_l^T := (\pi_l^1, ..., \pi_l^n)$, then $\mathbf{w}^{(k)} = \mathbb{1}_n \pi_l^T \mathbf{x}^{(k)} = \mathbf{1}_n \sum_{i=1}^{n} \pi_l^i (x_i^{(k)})^T$, so every row in $\mathbf{w}^{(k)}$ are $\sum_{i=1}^{n} \pi_l^i (x_i^{(k)})^T$

So $\mathbf{x}^{(k+1)} - \mathbf{w}^{(k+1)}$ is just considering the difference: $x_i^{k+1} - \sum_{i=1}^{n} \pi_l^i x_i^{(k+1)}, \forall i \in [1, ..., n]$, and we need to use error in kth iteration to estimate the error in $k+1$th iteration.

$$
\begin{aligned}
\mathbf{x}^{(k+1)} - \mathbf{w}^{(k+1)} &= A\mathbf{x}^{(k)} - \alpha \mathbf{y}^{(k)} - \mathbf{w}^{(k)} + \alpha \mathbb{1} \pi_l^T \mathbf{y}^{(k)} \\
&= A\mathbf{x}^{(k)} - A\mathbf{w}^{(k)} + A\mathbf{w}^{(k)} - \alpha \mathbf{y}^{(k)} - \mathbf{w}^{(k)} + \alpha \mathbb{1} \pi_l^T \mathbf{y}^{(k)} \\
? &= (A - \mathbb{1}_n \pi_l^T)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) - \alpha (I_n - \mathbb{1}_n \pi_l^T)\mathbf{y}^{(k)}
\end{aligned}
$$

Then we should consider the decent in gradient.

2 How to understand Push-Pull Gradient Tracking

2.1 Notations

Suppose that we have primitive row-stochastic matrix A, primitive column-stochastic matrix B satisfying $\pi_l^\top A = \pi_l^\top, A\mathbb{1}_n = \mathbb{1}_n, \mathbb{1}_n^T B = \mathbb{1}_n^T, B\pi_r = \pi_r$. Here π_l, π_r is the Perron vector whose sum is 1.

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

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

Define $\mathbf{w}^{(k)} = \mathbb{1}_n \pi_l^T \mathbf{x}^{(k)}, \bar{y} = \frac{1}{n}\mathbb{1}_n^T \mathbf{y}, v_B^{(k)} = B^k \mathbb{1}_n, J^{(k)} = \frac{1}{n} v_B^{(k+1)} \mathbb{1}_n^T, c^{(k)} = \pi_l^T v_B^{(k)}$.

2.2 Consensus Error Part

Left multiply $\mathbb{1}_n \pi_l^T$ on both side of (1), we have:

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha \mathbb{1}_n \pi_l^T \mathbf{y}^{(k)} \tag{3}$$

Then we unfold the iteration:

$$\begin{aligned}
\mathbf{x}^{(k+1)} - \mathbf{w}^{(k+1)} &= (A - \mathbb{1}_n \pi_l^T)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) - \alpha (I_n - \mathbb{1}_n \pi_l^T)\mathbf{y}^{(k)} \\
&= \dots \\
&= -\alpha \sum_{i=0}^{k} (A - \mathbb{1}_n \pi_l^T)^i (I_n - \mathbb{1}_n \pi_l^T)\mathbf{y}^{(k-i)} \\
&= -\alpha \sum_{i=0}^{k} (A - \mathbb{1}_n \pi_l^T)^i (I_n - \mathbb{1}_n \pi_l^T)(\mathbf{y}^{(k-i)} - J^{(k-i)}\mathbf{y}^{(k-i)}) - \alpha \sum_{i=0}^{k} (A - \mathbb{1}_n \pi_l^T)^i (v_B^{(k-i)} - c^{(k-i)}\mathbb{1}_n)\bar{y}^{(k-i)} \\
&= -\alpha \sum_{i=0}^{k} (A - \mathbb{1}_n \pi_l^T)^i (I_n - \mathbb{1}_n \pi_l^T)(\mathbf{y}^{(k-i)} - J^{(k-i)}\mathbf{y}^{(k-i)}) - \alpha \sum_{i=0}^{k} (A - \mathbb{1}_n \pi_l^T)^i (v_B^{(k-i)} - c^{(k-i)}\mathbb{1}_n)\bar{g}^{(k-i)}
\end{aligned}$$

Where the last equation comes from (4). The first part is easy to control (see Sec 2.3). Could we find a bound for the second part?

2.3 Gradient Tracking Part

Firstly, by multiply $n^{-1}\mathbb{1}_n^T$ on each side of (2), we have the following relation:

$$\bar{y}^{(k+1)} = \bar{y}^{(k)} + \bar{g}^{(k+1)} - \bar{g}^{(k)} \tag{4}$$

Thus we have $\bar{y}^{(k)} = \bar{g}^{(k)}$. This means the average of $y^{(k)}$ is exactly tracking the global sum of gradients. It will be fantastic if $\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} - \alpha \bar{\mathbf{y}}^{(k)}$ because this nearly reduces to gradient descent. However, the model parameters are updated with $\mathbf{y}_k$ instead of $\bar{\mathbf{y}}^{(k)}$. This is a gap and the error of $\mathbf{y} - \alpha \bar{\mathbf{y}}$ should be considered. According to Lemma B.2 in our previous work, we have

Lemma 1.

$$\|\mathbf{y}^{(k+1)} - J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2 = \sum_{i=0}^{k} \|(B^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\|_\pi^2 + 2\sum_{i=0}^{k}\langle (B^{k+1-i} - J^{(k+1)})(\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}), (B^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\rangle_\pi \tag{5}$$

To bound the first part, we have the following lemma:

Lemma 2. *The first term in Lemma 1 can be bounded as follow:*

$$\begin{aligned}&\mathbb{E}\|(B^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\|_\pi^2 \\ \leq&\beta_\pi^{2(k+1-i)}(1+\delta)^2\left(3nc_\pi^2\sigma^2 + 12L^2\kappa_\pi^2\beta_\pi^2\gamma^2\mathbb{E}\|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\right) \\ &+ \beta_\pi^{2(k+1-i)}(1+\delta)^2\left(6L^2\gamma^2\mathbb{E}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 + C_2\mathbb{E}\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2\right)\end{aligned} \tag{6}$$

Proof.

$$\mathbb{E}\|(B^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\|_\pi^2 \leq \beta_\pi^{2(k+1-i)}(1+\delta)^2\mathbb{E}\|\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)}\|_\pi^2, \forall i \geq 0. \tag{7}$$

Since both $\nabla f^{(i+1)}$ and $\mathbf{g}^{(i)}$ are $\mathcal{F}^{(i+1)}$-measurable and $\mathbb{E}[\mathbf{g}^{(i+1)}|\mathcal{F}^{(i+1)}] = \nabla f^{(i+1)}$, we have:

$$\begin{aligned}&\mathbb{E}\|\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)}\|_\pi^2 \leq \mathbb{E}\|\mathbf{g}^{(i+1)} - \nabla f^{(i+1)}\|_\pi^2 + \mathbb{E}\|\nabla f^{(i+1)} - \mathbf{g}^{(i)}\|_\pi^2 \\ \leq& nc_\pi^2\sigma^2 + \mathbb{E}\|\nabla f^{(i+1)} - \mathbf{g}^{(i)}\|_\pi^2 \\ \leq& nc_\pi^2\sigma^2 + 2\mathbb{E}\|\nabla f^{(i+1)} - \nabla f^{(i)}\|_\pi^2 + 2\mathbb{E}\|\nabla f^{(i)} - \mathbf{g}^{(i)}\|_\pi^2 \\ \leq& 3nc_\pi^2\sigma^2 + 2L^2\mathbb{E}\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\|_\pi^2,\end{aligned}$$

where the first inequality uses bounded noise assumption and the last inequality uses L-smoothness assumption.

$$\begin{aligned}&\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\|_\pi^2 \leq 3\|\mathbf{x}^{(i+1)} - \mathbf{w}^{(i+1)}\|_\pi^2 + 3\|\mathbf{w}^{(i+1)} - \mathbf{w}^{(i)}\|_\pi^2 + 3\|\mathbf{w}^{(i)} - \mathbf{x}^{(i)}\|_\pi^2 \\ =&3\|\mathbf{w}^{(i+1)} - \bar{\mathbf{x}}^{(i+1)}\|_\pi^2 + 3\gamma^2\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 + 3\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}(i)\|_\pi^2 \\ \leq& 6\kappa_\pi^2\beta_\pi^2\gamma^2\|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2 + 3\gamma^2\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 + (3 + 6\kappa_\pi^2\beta_\pi^2(1+\delta)^2)\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2,\end{aligned} \tag{8}$$

where the second inequality uses Jensen inequality. The proof follows by using (8) in (7). □

References