Appendix

A. Proof for Section 4

A.1. Angles Between Two Equidimensional Subspaces

In this section, we introduce full definitions and lemmas on metrics between two subspaces, which will be useful in our following proof.

Principal Angles. Given two matrices $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathcal{O}_{d\times k}$ which are both full rank with $1 \le k \le d$, we define the i-th ($1 \le i \le k$) principal angle between $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ in a recursive manner:

$$\theta_i(\mathbf{U}, \widetilde{\mathbf{U}}) = \min\left\{ \arccos\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right) : \mathbf{x} \in \mathcal{R}(\mathbf{U}), \mathbf{y} \in \mathcal{R}(\widetilde{\mathbf{U}}), \mathbf{x} \perp \mathbf{x}_j, \mathbf{y} \perp \mathbf{y}_j, \forall j < i \right\} \tag{7}$$

where $\mathcal{R}(\mathbf{U})$ denotes by the space spanned by all columns of $\mathbf{U}$. In this definition, we require that $0 \le \theta_1 \le \cdots \le \theta_k \le \frac{\pi}{2}$ and that $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ and $\{\mathbf{y}_1, \cdots, \mathbf{y}_k\}$ are the associated principal vectors. Principal angles can be used to quantify the differences between two given subspaces.

We have following facts about the k-th principal angle between $\mathbf{U}$ and $\widetilde{\mathbf{U}}$:

Fact 1. *Let $\mathbf{U}^\perp$ denote by the complement subspace of $\mathbf{U}$ (so that $[\mathbf{U}, \mathbf{U}^\perp] \in \mathbb{R}^{d\times d}$ forms an orthonormal basis of $\mathbb{R}^d$) and so dose $\widetilde{\mathbf{U}}^\perp$,*

1. $\sin\theta_k(\mathbf{U}, \widetilde{\mathbf{U}}) = \|\mathbf{U}^\top \widetilde{\mathbf{U}}^\perp\| = \|\widetilde{\mathbf{U}}^\top \mathbf{U}^\perp\|$;
2. $\tan\theta_k(\mathbf{U}, \widetilde{\mathbf{U}}) = \| \left[(\mathbf{U}^\perp)^\top \widetilde{\mathbf{U}}\right] (\mathbf{U}^\top \widetilde{\mathbf{U}})^\dagger \|$ *where $\dagger$ denotes by the Moore–Penrose inverse.*
3. *For any reversible matrix $\mathbf{R} \in \mathbb{R}^{k\times k}$, $\tan\theta_k(\mathbf{U}, \widetilde{\mathbf{U}}) = \tan\theta_k(\mathbf{U}, \widetilde{\mathbf{U}}\mathbf{R})$.*

Projection Distance. Define the projection distance[9] between two subspaces by

$$\mathrm{dist}(\mathbf{U}, \widetilde{\mathbf{U}}) = \|\mathbf{U}\mathbf{U}^\top - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\|. \tag{8}$$

This metric has several equivalent expressions:

$$\mathrm{dist}(\mathbf{U}, \widetilde{\mathbf{U}}) = \|\mathbf{U}^\top \widetilde{\mathbf{U}}^\perp\| = \|\widetilde{\mathbf{U}}^\top \mathbf{U}^\perp\| = \sin\theta_k(\mathbf{U}, \widetilde{\mathbf{U}}).$$

More generally, for any two matrix $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times k}$, we define the projection distance between them as

$$\mathrm{dist}(\mathbf{A}, \mathbf{B}) = \|\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top - \mathbf{U}_\mathbf{B}\mathbf{U}_\mathbf{B}^\top\|$$

where $\mathbf{U}_\mathbf{A}, \mathbf{U}_\mathbf{B}$ are the orthogonal basis of $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{B})$ respectively.

Orthogonal Procrustes. Let $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathbb{R}^{d\times k}$ be two orthonormal matrices. $\mathcal{R}(\mathbf{U})$ is close to $\mathcal{R}(\widetilde{\mathbf{U}})$ does not necessarily imply $\mathbf{U}$ is close to $\widetilde{\mathbf{U}}$, since any orthonormal invariant of $\mathbf{U}$ forms a base of $\mathcal{R}(\mathbf{U})$. However, the converse is true. If we try to map $\widetilde{\mathbf{U}}$ to $\mathbf{U}$ using an orthogonal transformation, we arrive at the following optimization

$$\mathbf{O}^* = \operatorname*{argmin}_{\mathbf{O}\in\mathcal{O}_k} \|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}\|_F, \tag{9}$$

where $\mathcal{O}_k$ denotes the set of $k \times k$ orthogonal matrices. The following lemma shows there is an interesting relationship between the subspace distance and their corresponding basis matrices. It implies that as a metric on linear space, $\mathrm{dist}(\mathbf{U}, \widetilde{\mathbf{U}})$ is equivalent to $\|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}^*\|_2$ (or $\min_{\mathbf{O}\in\mathcal{O}_k} \|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}\|_2$) up to some universal constant. The optimization problem involved in is named as the orthogonal procrustes problem and has been well studied (Schönemann, 1966; Cape, 2020).

[9] Unlike the spectral norm or the Frobenius norm, the projection norm will not fall short of accounting for global orthonormal transformation. Check (Ye & Lim, 2014) to find more information about distance between two spaces.

Lemma 3. *Let $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathcal{O}_{d\times k}$ and $\mathbf{O}^*$ is the solution of eqn. (9). Then we have*

1. *$\mathbf{O}^*$ has a closed form given by $\mathbf{O}^* = \mathbf{W}_1\mathbf{W}_2^\top$ where $\widetilde{\mathbf{U}}^\top\mathbf{U} = \mathbf{W}_1\mathbf{\Sigma}\mathbf{W}_2^\top$ is the singular value decomposition of $\widetilde{\mathbf{U}}^\top\mathbf{U}$.*
2. *Define $d(\mathbf{U}, \widetilde{\mathbf{U}}) := \|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}^*\|_2$ where $\|\cdot\|_2$ is the spectral norm. Then we have*
$$d(\mathbf{U}, \widetilde{\mathbf{U}}) = \sqrt{2 - 2\sqrt{1 - \operatorname{dist}(\mathbf{U}, \widetilde{\mathbf{U}})^2}} = 2\sin\frac{\theta_k(\mathbf{U}, \widetilde{\mathbf{U}})}{2}.$$
3. *$d(\mathbf{U}_1, \mathbf{U}_2) = d(\mathbf{U}_2, \mathbf{U}_1)$ for any $\mathbf{U}_1, \mathbf{U}_2 \in \mathcal{O}_{d\times k}$.*
4. *$\operatorname{dist}(\mathbf{U}, \widetilde{\mathbf{U}}) \le d(\mathbf{U}, \widetilde{\mathbf{U}}) \le \sqrt{2}\operatorname{dist}(\mathbf{U}, \widetilde{\mathbf{U}})$.*
5. *Define*
$$\ell(\mathbf{U}, \widetilde{\mathbf{U}}) := \min_{\mathbf{O}\in\mathcal{O}_k} \|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}\|_2.$$
Then $\ell(\mathbf{U}, \widetilde{\mathbf{U}})$ is a metric satisfying
   - *$\ell(\mathbf{U}, \widetilde{\mathbf{U}}) \ge 0$ for all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathcal{O}_{d\times k}$. $\ell(\mathbf{U}, \widetilde{\mathbf{U}}) = 0$ if and only if $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\widetilde{\mathbf{U}})$.*
   - *$\ell(\mathbf{U}, \widetilde{\mathbf{U}}) = \ell(\widetilde{\mathbf{U}}, \mathbf{U})$ for all $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathcal{O}_{d\times k}$.*
   - *$\ell(\mathbf{U}_1, \mathbf{U}_2) \le \ell(\mathbf{U}_1, \mathbf{U}_3) + \ell(\mathbf{U}_3, \mathbf{U}_2)$ for any $\mathbf{U}_1, \mathbf{U}_2$ and $\mathbf{U}_3 \in \mathcal{O}_{d\times k}$.*
6. *$\frac{1}{\sqrt{k}}\operatorname{dist}(\mathbf{U}, \widetilde{\mathbf{U}}) \le \ell(\mathbf{U}, \widetilde{\mathbf{U}}) \le d(\mathbf{U}, \widetilde{\mathbf{U}}) \le \sqrt{2}\operatorname{dist}(\mathbf{U}, \widetilde{\mathbf{U}})$.*

Proof. The first item comes from Schönemann (1966). The second item comes from Cape (2020). The third and forth items follow from the second one. The fifth item follows directly from definition. For the rightest two $\le$ of the last item, we use $\ell(\mathbf{U}, \widetilde{\mathbf{U}}) \le d(\mathbf{U}, \widetilde{\mathbf{U}})$ and the forth item. For the leftest $\le$, we use $\min_{\mathbf{O}\in\mathcal{O}_k} \|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}\|_2 \ge \frac{1}{\sqrt{k}}\min_{\mathbf{O}\in\mathcal{O}_k} \|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}\|_F$ and $\min_{\mathbf{O}\in\mathcal{O}_k} \|\mathbf{U} - \widetilde{\mathbf{U}}\mathbf{O}\|_F \ge \operatorname{dist}(\mathbf{U}, \widetilde{\mathbf{U}})$ (which is referred from Proposition 2.2 of Vu et al. (2013)). □

A.2. Proof Technique and Useful Lemmas

Update Rule. Assume $1 = \operatorname{argmax}_{i\in[m]} p_i$. We overwrite $\mathbf{Y}_t^{(i)}$ when $t \in \mathcal{I}_T$ (line 4 in Algorithm 1). To distinguish the difference, we additionally use $\mathbf{V}_t^{(i)}$ to denote the updated but not communicated $\mathbf{Y}_t^{(i)}$. Then the update rule becomes for all $i \in [m]$,

$$\mathbf{V}_t^{(i)} = \mathbf{M}_i\mathbf{Z}_{t-1}^{(i)}; \tag{10}$$

$$\mathbf{Y}_t^{(i)} = \begin{cases} \mathbf{V}_t^{(i)} & \text{if } t \notin \mathcal{I}_T; \\ \sum_{i=1}^m p_i\mathbf{V}_t^{(i)}\mathbf{D}_t^{(i)} & \text{if } t \in \mathcal{I}_T. \end{cases} \tag{11}$$

$$\mathbf{Y}_t^{(i)} = \mathbf{Z}_t^{(i)}\mathbf{R}_t^{(i)}. \tag{12}$$

Here we abuse the notation a little bit and define $\mathbf{D}_t^{(i)}$ as

$$\mathbf{D}_t^{(i)} = \operatorname*{argmin}_{\mathbf{D}\in\mathcal{F}\cap\mathcal{O}_k} \|\mathbf{Z}_{t-1}^{(i)}\mathbf{D} - \mathbf{Z}_{t-1}^{(1)}\|_o \tag{13}$$

where $\|\cdot\|_o$ can be set as either the Frobenius norm $\|\cdot\|_F$ or the spectrum norm $\|\cdot\|_2$, though in the body text we use only $\|\cdot\|_F$. There are some observations about the update rule:

1. If $t \notin \mathcal{I}_T$, we have $\mathbf{M}_i\mathbf{Z}_{t-1}^{(i)} = \mathbf{V}_t^{(i)} = \mathbf{Y}_t^{(i)} = \mathbf{Z}_t^{(i)}\mathbf{R}_t^{(i)}$.
2. If $t \in \mathcal{I}_T$, we have $\mathbf{Y}_t^{(1)} = \cdots = \mathbf{Y}_t^{(m)} = \sum_{i=1}^m p_i\mathbf{V}_t^{(i)}\mathbf{D}_t^{(i)} = \sum_{i=1}^m p_i\mathbf{M}_i\mathbf{Z}_{t-1}^{(i)}\mathbf{D}_t^{(i)}$ and thus $\mathbf{R}_t^{(1)} = \cdots = \mathbf{R}_t^{(m)}$ and $\mathbf{Z}_t^{(1)} = \cdots = \mathbf{Z}_t^{(m)}$. It implies that $\mathbf{D}_{t+1}^{(i)} = \mathbf{I}_k$.
3. If $\mathcal{F} = \mathcal{O}_k$, then $\mathbf{D}_t^{(i)}$ is the OPT we introduced in Section 4. If $\mathcal{F} = \mathcal{D}_k$, then $\mathbf{D}_t^{(i)}$ is the sign-fixing. If $\mathcal{F} = \{\mathbf{I}_k\}$, then $\mathbf{D}_t^{(i)}$ is always equal to the identity matrix $\mathbf{I}_k$ and we arrive at the vanilla `LocalPower`. The unified view helps us give theoretical analysis in a unified way.

Virtual Sequence. To analyze convergence of `LocalPower`, we define a virtual sequences defined as the weighted aggregation of local eigenvector matrices, i.e.,

$$\overline{\mathbf{Y}}_t = \sum_{i=1}^{m} p_i \mathbf{Y}_t^{(i)} \mathbf{O}_t^{(i)}. \tag{14}$$

Here $\mathbf{O}_t^{(i)} \in \mathbb{R}^{k\times k}$ is defined as

$$\mathbf{O}_t^{(i)} = \begin{cases} \mathbf{I}_k & \text{if } t \in \mathcal{I}_T \\ \mathbf{D}_t^{(i)} & \text{if } t \notin \mathcal{I}_T. \end{cases}$$

If $t \in \mathcal{I}_T$, $\overline{\mathbf{Y}}_t = \mathbf{Y}_t^{(i)}$ for $i \in [m]$ and thus is obtainable. Otherwise, $\overline{\mathbf{Y}}_t$ is a shadow matrix facilitating analysis.

Recurrence Lemma. Lemma 4 shows that we can express $\overline{\mathbf{Y}}_{t+1}$ as a linear transformation of $\overline{\mathbf{Y}}_t$. The resulting expression is similar to the iterates of the noisy power method proposed in (Hardt & Price, 2014), which motivates us to apply their technique to prove the main convergence of `LocalPower`. Lemma 4 holds for any invertible $\mathbf{R}_t \in \mathbb{R}^{k\times k}$. But, to guarantee convergence, we should carefully determine $\mathbf{R}_t$. In Lemma 8, we will give a particular expression of $\mathbf{R}_t$, which plays a crucial role in helping us to bound the noise term $\mathbf{G}_t$.

Lemma 4 (Recurrence)**.** *For any invertible $\mathbf{R}_t \in \mathbb{R}^{k\times k}$, we have*

$$\overline{\mathbf{Y}}_{t+1} = \left(\mathbf{M}\overline{\mathbf{Y}}_t + \mathbf{G}_t\right)\mathbf{R}_t^{-1} \tag{15}$$

where $\mathbf{M} = \frac{1}{n}\mathbf{A}^\top\mathbf{A} \in \mathbb{R}^{d\times d}$ and

$$\mathbf{G}_t = \mathbf{H}_t + \mathbf{W}_t \tag{16}$$

with $\mathbf{H}_t = \sum_{i=1}^m p_i \mathbf{H}_t^{(i)}$ and $\mathbf{W}_t = \sum_{i=1}^m p_i \mathbf{W}_t^{(i)}$. Here for $i \in [m]$,

$$\mathbf{H}_t^{(i)} = (\mathbf{M}_i - \mathbf{M})\,\mathbf{Y}_t^{(i)}\mathbf{O}_t^{(i)} \quad \textit{and} \quad \mathbf{W}_t^{(i)} = \mathbf{V}_{t+1}^{(i)}\left[\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\right]. \tag{17}$$

Proof. First notice that we always have $\overline{\mathbf{Y}}_t = \sum_{i=1}^m p_i \mathbf{V}_t^{(i)}\mathbf{D}_t^{(i)}$. If $t \in \mathcal{I}_T$, $\mathbf{Y}_t^{(1)} = \cdots = \mathbf{Y}_t^{(m)}$ and $\mathbf{O}_t^{(i)} = \mathbf{I}_r$, implying the equation follows from eqn. (11) and eqn. (14). Otherwise, we have $\mathbf{Y}_t^{(i)} = \mathbf{V}_t^{(i)}$ and $\mathbf{O}_t^{(i)} = \mathbf{D}_t^{(i)}$, then $\overline{\mathbf{Y}}_t = \sum_{i=1}^m p_i \mathbf{Y}_t^{(i)}\mathbf{O}_t^{(i)} = \sum_{i=1}^m p_i \mathbf{V}_t^{(i)}\mathbf{D}_t^{(i)}$.

We always have $\mathbf{V}_{t+1}^{(i)} = \mathbf{M}_i\mathbf{Z}_t^{(i)} = \mathbf{M}_i\mathbf{Y}_t^{(i)}(\mathbf{R}_t^{(i)})^{-1}$. Then for any invertible $\mathbf{R}_t$, we have

$$\begin{aligned}
\overline{\mathbf{Y}}_{t+1} &= \sum_{i=1}^m p_i \mathbf{V}_{t+1}^{(i)}\mathbf{D}_{t+1}^{(i)} \\
&= \sum_{i=1}^m p_i \mathbf{M}_i\mathbf{Y}_t^{(i)}(\mathbf{R}_t^{(i)})^{-1}\mathbf{D}_{t+1}^{(i)} \\
&= \sum_{i=1}^m p_i \mathbf{M}_i\mathbf{Y}_t^{(i)}\mathbf{O}_t^{(i)}\mathbf{R}_t^{-1} + \sum_{i=1}^m p_i \mathbf{M}_i\mathbf{Y}_t^{(i)}(\mathbf{R}_t^{(i)})^{-1}\left[\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\right]\mathbf{R}_t^{-1} \\
&\overset{(a)}{=} \sum_{i=1}^m p_i \left(\mathbf{M}\mathbf{Y}_t^{(i)}\mathbf{O}_t^{(i)} + \mathbf{H}_t^{(i)}\right)\mathbf{R}_t^{-1} + \sum_{i=1}^m p_i \mathbf{M}_i\mathbf{Z}_t^{(i)}\left[\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\right]\mathbf{R}_t^{-1} \\
&= \sum_{i=1}^m p_i \left(\mathbf{M}\mathbf{Y}_t^{(i)}\mathbf{O}_t^{(i)} + \mathbf{H}_t^{(i)}\right)\mathbf{R}_t^{-1} + \sum_{i=1}^m p_i \mathbf{M}_i\mathbf{Z}_t^{(i)}\left[\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\right]\mathbf{R}_t^{-1} \\
&\overset{(b)}{=} \left(\mathbf{M}\overline{\mathbf{Y}}_t + \mathbf{H}_t + \mathbf{W}_t\right)\mathbf{R}_t^{-1}
\end{aligned}$$

where (a) results from the definition of $\mathbf{H}_t^{(i)}$; and (b) simplifies the equation via defining $\mathbf{H}_t = \sum_{i=1}^m p_i \mathbf{H}_t^{(i)}$ and $\mathbf{W}_t = \sum_{i=1}^m p_i \mathbf{W}_t^{(i)}$. Setting $\mathbf{G}_t = \mathbf{H}_t + \mathbf{W}_t$ completes the proof.

□

Convergence Lemma. The following lemma is an variant of Lemma 2.2 in Hardt & Price (2014). Given the relation $\overline{\mathbf{Y}}_{t+1} = \left(\mathbf{M}\overline{\mathbf{Y}}_t + \mathbf{G}_t\right)\mathbf{R}_t^{-1}$, Hardt & Price (2014) requires $\overline{\mathbf{Y}}_t$ to have orthonormal columns, i.e., $\overline{\mathbf{Y}}_t^\top \overline{\mathbf{Y}}_t = \mathbf{I}_r$. However, it is unlikely to hold in our case. As a remedy, we slightly change the lemma to allow arbitrary $\overline{\mathbf{Y}}_t$. This will also change the condition on $\mathbf{G}_t$.

Lemma 5. *Let $\mathbf{U}_k \in \mathbb{R}^{d\times k}$ be the top-k eigenvectors of a positive semi-definite matrix $\mathbf{M}$. For $t \geq 1$, assume $\overline{\mathbf{Y}}_t$ satisfies eqn. (15) and $\mathbf{G}_t \in \mathbb{R}^{d\times k}$ satisfy*

$$4\|\mathbf{U}_k^\top \mathbf{G}_t \overline{\mathbf{Y}}_t^\dagger\|_2 \leq (\sigma_k - \sigma_{k+1})\cos\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_t\right) \quad \textit{and} \quad 4\|\mathbf{G}_t\overline{\mathbf{Y}}_t^\dagger\|_2 \leq (\sigma_k - \sigma_{k+1})\epsilon \tag{18}$$

where $\overline{\mathbf{Y}}_t^\dagger$ is the Moore–Penrose inverse of $\overline{\mathbf{Y}}_t$ and $\epsilon < 1$. Then

$$\tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_{t+1}\right) \leq \max\left(\epsilon, \max\left(\epsilon, \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{1/4}\right)\tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_t\right)\right).$$

Proof. Let $\overline{\mathbf{Y}}_t = \overline{\mathbf{Z}}_t\overline{\mathbf{R}}_t$ be the QR factorization of $\overline{\mathbf{Y}}_t$ so that $\overline{\mathbf{Z}}_t$ has orthonormal columns. The recurrence relation becomes $\overline{\mathbf{Y}}_{t+1} = \left(\mathbf{M}\overline{\mathbf{Z}}_t\overline{\mathbf{R}}_t + \mathbf{G}_t\right)\mathbf{R}_t^{-1} = \left(\mathbf{M}\overline{\mathbf{Z}}_t + \mathbf{G}_t\overline{\mathbf{R}}_t^{-1}\right)\overline{\mathbf{R}}_t\mathbf{R}_t^{-1}$. By the fact 1, we have $\tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_{t+1}\right) = \tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_{t+1}\mathbf{R}_t\overline{\mathbf{R}}_t^{-1}\right) = \tan\theta_k\left(\mathbf{U}_k, \mathbf{M}\overline{\mathbf{Z}}_t + \mathbf{G}_t\overline{\mathbf{R}}_t^{-1}\right)$. By requiring

$$4\|\mathbf{U}_k^\top\mathbf{G}_t\overline{\mathbf{R}}_t^{-1}\|_2 \leq (\sigma_k - \sigma_{k+1})\cos\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Z}}_t\right) \quad \text{and} \quad 4\|\mathbf{G}_t\overline{\mathbf{R}}_t^{-1}\|_2 \leq (\sigma_k - \sigma_{k+1})\epsilon,$$

we have from Lemma 2.2 in Hardt & Price (2014) that

$$\tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_{t+1}\right) \leq \max\left(\epsilon, \max\left(\epsilon, \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{1/4}\right)\tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Z}}_t\right)\right).$$

Noting that $\mathcal{R}(\overline{\mathbf{Y}}_t) = \mathcal{R}(\overline{\mathbf{Z}}_t)$, we have $\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_t\right) = \theta_k\left(\mathbf{U}_k, \overline{\mathbf{Z}}_t\right)$ and thus

$$\cos\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_t\right) = \cos\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Z}}_t\right) \quad \text{and} \quad \tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Y}}_t\right) = \tan\theta_k\left(\mathbf{U}_k, \overline{\mathbf{Z}}_t\right).$$

Finally, using $\|\mathbf{U}_k^\top\mathbf{G}_t\overline{\mathbf{Y}}_t^\dagger\|_2 = \|\mathbf{U}_k^\top\mathbf{G}_t\overline{\mathbf{R}}_t^{-1}\|_2$ and $\|\mathbf{G}_t\overline{\mathbf{Y}}_t^\dagger\|_2 = \|\mathbf{G}_t\overline{\mathbf{R}}_t^{-1}\|_2$ completes the proof.

□

Other Useful Lemma. Lemma 6 handles $\tan\theta_k(\mathbf{U}, \mathbf{Z}_0)$ with randomly generate $\mathbf{Z}_0$, while Lemma 7 give a upper bound of $\|\overline{\mathbf{Y}}_t^\dagger\mathbf{M}\|_2$.

Lemma 6 (Lemma 2.4 in Hardt & Price (2014))**.** *For an arbitrary orthonormal $\mathbf{U}$ and random subspace $\mathbf{Z}_0 \in \mathbb{R}^{d\times r}$, with probability grater than $1 - \tau^{-\Omega(r+1-k)} - e^{-\Omega(d)}$, we have that*

$$\tan\theta_k(\mathbf{U}, \mathbf{Z}_0) \leq \frac{\tau\sqrt{d}}{\sqrt{r} - \sqrt{k-1}}.$$

Lemma 7. *Recall that $\kappa = \|\mathbf{M}\|_2\|\mathbf{M}^\dagger\|_2$ and $\eta = \max_{i\in[m]}\|\mathbf{M}_i - \mathbf{M}\|_2/\|\mathbf{M}\|_2$. Define*

$$\mu_t = 1 - \eta\kappa - \sum_{j=1}^m p_j\|\mathbf{Z}_{t-1}^{(j)}\mathbf{D}_t^{(j)} - \mathbf{Z}_{t-1}^{(1)}\|_2$$

and assume $\mu_t > 0$. Then it follows that $\|\overline{\mathbf{Y}}_t^\dagger\mathbf{M}\|_2 \leq \frac{1}{\mu_t}$.

Proof. For any matrix $\mathbf{X} \in \mathbb{R}^{d\times k}$, we have

$$\|\mathbf{X}^\dagger\|_2 = \max_{\mathbf{x}\in\mathbb{R}^k}\frac{\|\mathbf{w}\|_2}{\|\mathbf{X}\mathbf{w}\|_2} = \max_{\|\mathbf{X}\mathbf{w}\|_2=1}\|\mathbf{w}\|_2 = \max\{\|\mathbf{w}\|_2 : \|\mathbf{X}\mathbf{w}\|_2 \leq 1\}.$$

Notice that $\overline{\mathbf{Y}}_t^\dagger \mathbf{M} = (\mathbf{M}^\dagger \overline{\mathbf{Y}}_t)^\dagger$ and $\overline{\mathbf{Y}}_t = \sum_{j=1}^m p_j \mathbf{M}_j \mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)}$. We then have

$$\begin{aligned}
\|\overline{\mathbf{Y}}_t^\dagger \mathbf{M}\|_2 &= \|(\mathbf{M}^\dagger \overline{\mathbf{Y}}_t)^\dagger\|_2 \\
&= \max\{\|\mathbf{w}\|_2 : \|\mathbf{M}^\dagger \overline{\mathbf{Y}}_t \mathbf{w}\|_2 \le 1\} \\
&= \max\{\|\mathbf{w}\|_2 : \|(\mathbf{M}^\dagger \sum_{j=1}^m p_j \mathbf{M}_j \mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)})\mathbf{w}\|_2 \le 1\} \\
&\overset{(a)}{\le} \max\{\|\mathbf{w}\|_2 : \|\sum_{j=1}^m p_j \mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} \mathbf{w}\|_2 - \eta\kappa\|\mathbf{w}\|_2 \le 1\} \\
&\overset{(b)}{\le} \max\{\|\mathbf{w}\|_2 : \|\mathbf{w}\|_2 (1 - \eta\kappa - \sum_{j=1}^m p_j \|\mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} - \mathbf{Z}_{t-1}^{(1)}\|_2) \le 1\} \\
&\le \frac{1}{1 - \eta\kappa - \sum_{j=1}^m p_j \|\mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} - \mathbf{Z}_{t-1}^{(1)}\|_2} \le \frac{1}{\mu_t}
\end{aligned}$$

where (a) follows because of

$$\|(\mathbf{M}^\dagger \sum_{j=1}^m p_j \mathbf{M}_j \mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)})\mathbf{w}\|_2 \ge \|\sum_{j=1}^m p_j \mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} \mathbf{w}\|_2 - \sum_{i=1}^m p_i \|\mathbf{M}^\dagger(\mathbf{M}_j - \mathbf{M})\|_2 \|\mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} \mathbf{w}\|_2$$

and $\|\mathbf{M}^\dagger(\mathbf{M}_j - \mathbf{M})\|_2 \le \|\mathbf{M}^\dagger\|_2 \|(\mathbf{M}_j - \mathbf{M})\|_2 \le \eta\kappa$; and (b) holds since

$$\begin{aligned}
\|\sum_{j=1}^m p_j \mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} \mathbf{w}\|_2 &\ge \|\sum_{j=1}^m p_j \mathbf{Z}_{t-1}^{(1)} \mathbf{w}\|_2 - \|\sum_{j=1}^m p_j (\mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} - \mathbf{Z}_{t-1}^{(1)})\mathbf{w}\|_2 \\
&\ge \|\mathbf{w}\|_2 - \sum_{j=1}^m p_j \|\mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} - \mathbf{Z}_{t-1}^{(1)}\|_2 \|\mathbf{w}\|_2 \\
&= \|\mathbf{w}\|_2 (1 - \sum_{j=1}^m p_j \|\mathbf{Z}_{t-1}^{(j)} \mathbf{D}_t^{(j)} - \mathbf{Z}_{t-1}^{(1)}\|_2).
\end{aligned}$$

□

A.3. The Choice of $\mathbf{R}_t$

In this section, we specify the choice of $\mathbf{R}_t$ and analyze the residual error bound $\|\mathbf{D}_{t+1}^{(i)} \mathbf{R}_t - \mathbf{R}_t^{(i)} \mathbf{O}_t^{(i)}\|_2$. Lemma 8 specifies the way we set $\mathbf{R}_t$. Given a baseline data matrix $\mathbf{M}_o$, $\mathbf{R}_t$ is the shadow matrix that depicts what the upper triangle matrix ought to be, if we start from the nearest synchronized matrix and perform QR factorization using the matrix $\mathbf{M}_o$. We will set $\mathbf{M}_o = \mathbf{M}_t^{(1)}$ (by assuming $1 = \underset{i \in [m]}{\operatorname{argmax}}\, p_i$) and analyze $\|\mathbf{W}_t^{(i)} \overline{\mathbf{Y}}_t^\dagger\|_2$ and $\|\mathbf{H}_t^{(i)} \overline{\mathbf{Y}}_t^\dagger\|_2$ in terms of $\|\mathbf{Z}_t^{(i)} \mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t^{(1)}\|_2$. Latter we will bound $\|\mathbf{Z}_t^{(i)} \mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t^{(1)}\|_2$ when $\mathcal{F}$ is differently set.

Lemma 8 (Choice of $\mathbf{R}_t$)**.** *Fix any t and let $t_0 = \tau(t) \in \mathcal{I}_T$ be the latest synchronization step before t, then $t \ge \tau(t)$.*

- *If $t = t_0$, we define $\mathbf{R}_t = \mathbf{R}_t^{(i)}$ for any $i \in [m]$ since all $\mathbf{R}_t^{(i)}$'s are equal.*
- *If $t > t_0$, given a baseline data matrix $\mathbf{M}_o$, we define $\mathbf{R}_t \in \mathbb{R}^{r \times r}$ recursively as the following. Let $\mathbf{Y}_{t_0} = \overline{\mathbf{Y}}_{t_0} = \mathbf{Z}_{t_0} \mathbf{R}_{t_0}$, and for $l = t_0, t_0 + 1, \cdots, t$, we use the following QR factorization to define $\mathbf{R}_t$'s:*

$$\mathbf{V}_{l+1} = \mathbf{M}_o \mathbf{Z}_l = \mathbf{Z}_{l+1} \mathbf{R}_{l+1}.$$

Then for any $i \in [m]$, we have

$$\|\mathbf{D}_{t+1}^{(i)} \mathbf{R}_t - \mathbf{R}_t^{(i)} \mathbf{O}_t^{(i)}\|_2 \le \sigma_1(\mathbf{M}_o) \|\mathbf{Z}_t^{(i)} \mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t\|_2 + \left[\|\mathbf{M}_o - \mathbf{M}_i\|_2 + \sigma_1(\mathbf{M}_i) \|\mathbf{Z}_{t-1}^{(i)} \mathbf{D}_t^{(i)} - \mathbf{Z}_{t-1}\|_2\right] 1_{t \notin \mathcal{I}_T}. \quad (19)$$

Proof. We are going to bound $\|\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\|_2$ in two cases depending on whether $t \in \mathcal{I}_T$. If $t \in \mathcal{I}_T$, implying $t = t_0 := \tau(u)$, then $\mathbf{O}_t^{(i)} = \mathbf{D}_{t+1}^{(i)} = \mathbf{I}_r$ and $\mathbf{R}_t = \mathbf{R}_t^{(i)}$. Therefore, $\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)} = \mathbf{0}$.

Otherwise, $t \notin \mathcal{I}_T$ and thus $t > t_0$. Let's fix some $i \in [m]$ and denote $\Delta\mathbf{M} = \mathbf{M}_i - \mathbf{M}_o$. Based on `LocalPower`, we have $\mathbf{Y}_{t_0}^{(i)} = \overline{\mathbf{Y}}_{t_0} = \mathbf{Z}_{t_0}^{(i)}\mathbf{R}_{t_0}^{(i)}$, and for $l = t_0, t_0+1, \cdots, t$,

$$\mathbf{V}_{l+1}^{(i)} = \mathbf{M}_i\mathbf{Z}_l^{(i)} = \mathbf{Z}_{l+1}^{(i)}\mathbf{R}_{l+1}^{(i)}.$$

Then,

$$\begin{aligned}
\mathbf{Z}_l^{(i)}\mathbf{R}_l^{(i)}\mathbf{O}_l^{(i)} &= \mathbf{M}_i\mathbf{Z}_{l-1}^{(i)}\mathbf{O}_l^{(i)} \\
&= (\mathbf{M}_o + \Delta\mathbf{M})(\mathbf{Z}_{l-1} + \Delta\mathbf{Z}_{l-1}) \\
&= \mathbf{M}_o\mathbf{Z}_{l-1} + \Delta\mathbf{M}\cdot\mathbf{Z}_{l-1} + \mathbf{M}_i\cdot\Delta\mathbf{Z}_{l-1} \\
&:= \mathbf{M}_o\mathbf{Z}_{l-1} + \mathbf{E}_{l-1} = \mathbf{Z}_l\mathbf{R}_l + \mathbf{E}_{l-1}
\end{aligned}$$

where $\mathbf{E}_{l-1} = \Delta\mathbf{M}\cdot\mathbf{Z}_{l-1} + \mathbf{M}_i\cdot\Delta\mathbf{Z}_{l-1}$ and $\Delta\mathbf{Z}_{l-1} = \mathbf{Z}_{l-1}^{(i)}\mathbf{O}_l^{(i)} - \mathbf{Z}_{l-1}$.

Note that

$$\mathbf{Z}_t^{(i)}\mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)} = \mathbf{Z}_t\mathbf{R}_t + \mathbf{E}_{t-1}.$$

Then we have

$$\begin{aligned}
\|\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\|_2 &= \|\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{Z}_t^{(i)}\mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\|_2 \\
&\overset{(a)}{=} \|\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{Z}_t\mathbf{R}_t - \mathbf{E}_{t-1}\|_2 \\
&\le \|(\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t)\mathbf{R}_t\|_2 + \|\mathbf{E}_{t-1}\|_2 \\
&\overset{(b)}{\le} \|\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t\|_2\|\mathbf{R}_t\|_2 + \|\Delta\mathbf{M}\|_2 + \|\mathbf{M}_i\|_2\|\mathbf{Z}_{t-1}^{(i)}\mathbf{O}_t^{(i)} - \mathbf{Z}_{t-1}\|_2 \\
&\overset{(c)}{\le} \sigma_1(\mathbf{M}_o)\|\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t\|_2 + \|\mathbf{M}_o - \mathbf{M}_i\|_2 + \sigma_1(\mathbf{M}_i)\|\mathbf{Z}_{t-1}^{(i)}\mathbf{D}_t^{(i)} - \mathbf{Z}_{t-1}\|_2
\end{aligned}$$

where (a) uses the equality of $\mathbf{Z}_t^{(i)}\mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}$; (b) uses the definition of $\mathbf{E}_{t-1}$ and $\mathbf{O}_t^{(i)} = \mathbf{D}_t^{(i)}$ (due to $t \notin \mathcal{I}_T$); and (c) uses $\|\mathbf{R}_t\|_2 \le \|\mathbf{M}_o\|_2 = \sigma_1(\mathbf{M}_o)$.

Combining the two cases, we have for all $t \in [T]$,

$$\|\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\|_2 \le \sigma_1(\mathbf{M}_o)\|\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t\|_2 + \left[\|\mathbf{M}_o - \mathbf{M}_i\|_2 + \sigma_1(\mathbf{M}_i)\|\mathbf{Z}_{t-1}^{(i)}\mathbf{D}_t^{(i)} - \mathbf{Z}_{t-1}\|_2\right] 1_{t\notin\mathcal{I}_T}.$$

□

Lemma 9. *Assume* $\eta = \max_{i\in[m]}\|\mathbf{M}_i - \mathbf{M}\|_2/\|\mathbf{M}\|_2$ *is sufficiently small and* $1 = \operatorname{argmax}_{i\in[m]} p_i$. *Define*

$$\rho_t = \|\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t^{(1)}\|_2,$$

we have

$$\|\mathbf{H}_t\overline{\mathbf{Y}}_t^{\dagger}\|_2 \le \frac{2\sigma_1\eta\kappa 1_{t\notin\mathcal{I}_T}}{1 - \eta\kappa - (1 - \max_{i\in[m]} p_i)\rho_{t-1}} \tag{20}$$

$$\|\mathbf{W}_t\overline{\mathbf{Y}}_t^{\dagger}\|_2 \le 4(1 - \max_{i\in[m]} p_i)\sigma_1\kappa\frac{\rho_t + (\rho_{t-1} + \eta)1_{t\notin\mathcal{I}_T}}{1 - \eta\kappa - (1 - \max_{i\in[m]} p_i)\rho_{t-1}}. \tag{21}$$

Proof. Without loss of generality, we assume $1 = \operatorname{argmax}_{i\in[m]} p_i$ and then set the baseline matrix in Lemma 8 as $\mathbf{M}_o = \mathbf{M}_1$ and use the $\mathbf{R}_t$ defined therein. Then Lemma 8 and Lemma 10 imply for all $i \in [m]$,

$$\|\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\|_2 \le \sigma_1(\mathbf{M}_o)\|\mathbf{Z}_t^{(i)}\mathbf{D}_{t+1}^{(i)} - \mathbf{Z}_t\|_2 + [\|\mathbf{M}_o - \mathbf{M}_i\|_2 + \sigma_1(\mathbf{M}_i)\|\mathbf{Z}_{t-1}^{(i)}\mathbf{D}_t^{(i)} - \mathbf{Z}_{t-1}\|_2]1_{t\notin\mathcal{I}_T}$$

$$\leq (1+\eta)\sigma_1\left[\rho_t + \rho_{t-1}1_{t\notin\mathcal{I}_T}\right] + \eta\sigma_1 1_{i\neq 1 \text{ and } t\notin\mathcal{I}_T}$$

where $\sigma_1 = \sigma_1(\mathbf{M})$ and $1_{i\neq 1 \text{ and } t\notin\mathcal{I}_T}$ is the indicator of event $\{i\neq 1\}\cap\{t\notin\mathcal{I}_T\}$.

Recall the definition of ρ_t. By Lemma 8 and Lemma 7, we have

$$\begin{aligned}
\|\mathbf{W}_t\overline{\mathbf{Y}}_t^\dagger\|_2 &= \|\sum_{i=1}^m p_i\mathbf{M}_i\mathbf{Z}_t^{(i)}\left[\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\right]\overline{\mathbf{Y}}_t^\dagger\mathbf{M}\mathbf{M}^{-1}\|_2 \\
&\leq \sum_{i=1}^m p_i\|\mathbf{M}^{-1}\|_2\|\mathbf{M}_i\|_2\|\overline{\mathbf{Y}}_t^\dagger\mathbf{M}\|_2\|\mathbf{D}_{t+1}^{(i)}\mathbf{R}_t - \mathbf{R}_t^{(i)}\mathbf{O}_t^{(i)}\|_2 \\
&\leq 2(1-p_1)\sigma_1\kappa\frac{\eta 1_{t\notin\mathcal{I}_T} + 2(\rho_t + \rho_{t-1}1_{t\notin\mathcal{I}_T})}{1-\eta\kappa-(1-p_1)\rho_{t-1}} \\
&\leq 4(1-p_1)\sigma_1\kappa\frac{\rho_t + (\rho_{t-1}+\eta)1_{t\notin\mathcal{I}_T}}{1-\eta\kappa-(1-p_1)\rho_{t-1}}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\mathbf{H}_t\overline{\mathbf{Y}}_t^\dagger\|_2 &= \|\sum_{i=1}^m p_i\left(\mathbf{M}_i-\mathbf{M}\right)\mathbf{Y}_t^{(i)}\mathbf{O}_t^{(i)}\overline{\mathbf{Y}}_t^\dagger\mathbf{M}\mathbf{M}^{-1}\|_2 \\
&\leq \sum_{i=1}^m p_i\|\mathbf{M}^{-1}\|_2\|(\mathbf{M}_i-\mathbf{M})\|_2\|\mathbf{Y}_t^{(i)}\mathbf{O}_t^{(i)}\|_2\|\overline{\mathbf{Y}}_t^\dagger\mathbf{M}\|_2 1_{t\notin\mathcal{I}_T} \\
&\leq \frac{(1+\eta)\sigma_1\kappa\eta 1_{t\notin\mathcal{I}_T}}{1-\eta\kappa-(1-p_1)\rho_{t-1}} \\
&\leq \frac{2\sigma_1\kappa\eta 1_{t\notin\mathcal{I}_T}}{1-\eta\kappa-(1-p_1)\rho_{t-1}}.
\end{aligned}$$

□

A.3.1. The Case When $\mathcal{F} = \mathcal{O}_k$

Lemma 10. *When setting $\mathcal{F}=\mathcal{O}_k$, no mater $\mathbf{D}_t^{(i)}$ is solved from eqn. (13) using $\|\cdot\|_F$ or $\|\cdot\|_2$, we have*

$$\|\mathbf{Z}_{t-1}^i\mathbf{D}_t^{(i)} - \mathbf{Z}_{t-1}^{(1)}\|_2 \leq \sqrt{2}\,\mathrm{dist}(\mathbf{Z}_{t-1}^{(i)}, \mathbf{Z}_{t-1}^{(1)}). \tag{22}$$

Proof. This follows directly from Lemma 3. □

Lemma 11 (Davis-Kahan $\sin(\theta)$ theorem)**.** *Let the top-k eigenspace of $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ be respectively $\mathbf{U}_k$ and $\widetilde{\mathbf{U}}_k$ (both of which are orthonormal). The k-largest eigenvalue of $\mathbf{M}$ is denoted by $\sigma_k(\mathbf{M})$ and similarly for $\sigma_k(\widetilde{\mathbf{M}})$. Define $\delta_k = \min\{|\sigma_k(\mathbf{M}) - \sigma_j(\widetilde{\mathbf{M}})| : j\geq k+1\}$, then*

$$\mathrm{dist}(\mathbf{U}_k, \widetilde{\mathbf{U}}_k) = \sin\theta_k(\mathbf{U}_k, \widetilde{\mathbf{U}}_k) \leq \frac{\|\mathbf{M}-\widetilde{\mathbf{M}}\|_2}{\delta_k}.$$

Lemma 12 (Perturbation theorem of projection distance)**.** *Let* $\mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{Y})$*, then*

$$\mathrm{dist}(\mathbf{X},\mathbf{Y}) \leq \min\{\|\mathbf{X}^\dagger\|_2, \|\mathbf{Y}^\dagger\|_2\}\|\mathbf{X}-\mathbf{Y}\|_2.$$

Proof. See Theorem 2.3 of Ji-guang (1987). □

Lemma 13. *Assume $\eta = \max_{i\in[m]}\|\mathbf{M}_i-\mathbf{M}\|_2/\|\mathbf{M}\|_2$ is sufficiently small. If $\mathbf{D}_t^{(i)}$ is solved from eqn. (13) with $\mathcal{F}=\mathcal{O}_k$, then eqn. (20) and eqn. (21) hold with*

$$\rho_t \leq \min\sqrt{2}\left\{\frac{2\kappa^p p\eta(1+\eta)^{p-1}}{(1-\eta)^p}, \frac{\eta\sigma_1}{\delta_k} + 2\gamma_k^{p/4}\max_{i\in[m]}\tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)})\right\}.$$

where

- $\delta_k = \min_{i\in[m]} \delta_k^{(i)}$ *with* $\delta_k^{(i)} = \min\{|\sigma_k(\mathbf{M}) - \sigma_j(\mathbf{M}_i)| : j \geq k+1\}$*;*
- $\gamma_k = \max\{\max_{i\in[m]} \frac{\sigma_{k+1}(\mathbf{M}_i)}{\sigma_k(\mathbf{M}_i)}, \frac{\sigma_{k+1}(\mathbf{M})}{\sigma_k(\mathbf{M})}\} \in (0,1)$*;*
- $\kappa = \|\mathbf{M}\|_2\|\mathbf{M}^\dagger\|_2$ *is the condition number of* $\mathbf{M}$*;*
- $p = t - \tau(u)$*,* $\tau(t) \in \mathcal{I}_T$ *is defined as the nearest synchronization time before* t*.*

Proof. By Lemma 5 and Lemma 10, we only need to bound $\max_{i\in[m]} \mathrm{dist}(\mathbf{Z}_t^{(i)}, \mathbf{Z}_t^{(1)})$. We will bound each $\mathrm{dist}(\mathbf{Z}_t^{(i)}, \mathbf{Z}_t^{(1)})$ uniformly in two ways. Then the minimum of the two upper bounds holds for their maximum that is exactly ρ_t.

Fix any $i \in [m]$ and $t \in [T]$. Let $\tau(t)$ be the latest synchronization step before t and $p = t - \tau(t)$ be the number of nearest local updates.

- For small p, by Lemma 12, it follows that

$$\begin{aligned}
\mathrm{dist}(\mathbf{Z}_t^i, \mathbf{Z}_t^{(1)}) &= \mathrm{dist}(\mathbf{M}_i^p \mathbf{Z}_{\tau(t)}, \mathbf{M}_1^p \mathbf{Z}_{\tau(t)}) \\
&\leq \mathrm{dist}(\mathbf{M}_i^p \mathbf{Z}_{\tau(t)}, \mathbf{M}^p \mathbf{Z}_{\tau(t)}) + \mathrm{dist}(\mathbf{M}^p \mathbf{Z}_{\tau(t)}, \mathbf{M}_1^p \mathbf{Z}_{\tau(t)}) \\
&\leq \min\{\|(\mathbf{M}_i^p \mathbf{Z}_{\tau(t)})^\dagger\|_2, \|(\mathbf{M}^p \mathbf{Z}_{\tau(t)})^\dagger\|_2\}\|(\mathbf{M}_i^p - \mathbf{M}^p)\mathbf{Z}_{\tau(t)}\|_2 \\
&\quad + \min\{\|(\mathbf{M}^p \mathbf{Z}_{\tau(t)})^\dagger\|_2, \|(\mathbf{M}_1^p \mathbf{Z}_{\tau(t)})^\dagger\|_2\}\|(\mathbf{M}^p - \mathbf{M}_1^p)\mathbf{Z}_{\tau(t)}\|_2 \\
&\leq 2\kappa^p \frac{(1+\eta)^p - 1}{(1-\eta)^p} \\
&\leq \frac{2\kappa^p p\eta(1+\eta)^{p-1}}{(1-\eta)^p}
\end{aligned}$$

  where $\kappa = \|\mathbf{M}\|_2\|\mathbf{M}^\dagger\|_2$ is the condition number of $\mathbf{M}$.

- For large p, let the top-k eigenspace of $\mathbf{M}_1$ and $\mathbf{M}_i$ be respectively $\mathbf{U}_k^{(1)}$ and $\mathbf{U}_k^{(i)}$ (both of which are orthonormal). The k-largest eigenvalue of $\mathbf{M}$ is denoted by $\sigma_k(\mathbf{M}_1)$ and similarly for $\sigma_k(\mathbf{M}_i)$. Then by Lemma 11, we have

$$\mathrm{dist}(\mathbf{U}_k, \mathbf{U}_k^{(i)}) \leq \frac{\|\mathbf{M}_i - \mathbf{M}\|}{\delta_k^{(i)}} \leq \frac{\eta\sigma_1}{\delta_k^{(i)}}.$$

  where $\sigma_1 = \sigma_1(\mathbf{M})$ and $\delta_k^{(i)} = \min\{|\sigma_j(\mathbf{M}_i) - \sigma_k(\mathbf{M})| : j \neq k\}$.

  Note that local updates are equivalent to noiseless power method. Then, using Lemma 5 and setting $\epsilon = 0$ and $\mathbf{G}_t = \mathbf{0}$ therein, we have

$$\tan\theta_k(\mathbf{Z}_t^i, \mathbf{U}_k^{(i)}) \leq \left(\frac{\sigma_{k+1}(\mathbf{M}_i)}{\sigma_k(\mathbf{M}_i)}\right)^{1/4} \tan\theta_k(\mathbf{Z}_{t-1}^i, \mathbf{U}_k^{(i)}).$$

  Hence,

$$\begin{aligned}
\mathrm{dist}(\mathbf{Z}_t^i, \mathbf{Z}_t^{(1)}) &\leq \mathrm{dist}(\mathbf{Z}_t^i, \mathbf{U}_k^{(i)}) + \mathrm{dist}(\mathbf{U}_k^{(i)}, \mathbf{U}_k^{(1)}) + \mathrm{dist}(\mathbf{U}_k^{(1)}, \mathbf{Z}_t^{(1)}) \\
&\leq \frac{\eta\sigma_1}{\delta_k^{(i)}} + \left(\frac{\sigma_{k+1}(\mathbf{M}_i)}{\sigma_k(\mathbf{M}_i)}\right)^{p/4} \tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)}) + \left(\frac{\sigma_{k+1}(\mathbf{M})}{\sigma_k(\mathbf{M})}\right)^{p/4} \tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(1)}) \\
&\leq \frac{\eta\sigma_1}{\min_{i\in[m]} \delta_k^{(i)}} + 2\gamma_k^{p/4} \max_{i\in[m]} \tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)}).
\end{aligned}$$

Combining the two cases, we have

$$\rho_t \leq \sqrt{2}\min\left\{\frac{2\kappa^p p\eta(1+\eta)^{p-1}}{(1-\eta)^p}, \frac{\eta\sigma_1}{\delta_k} + 2\gamma_k^{p/4} \max_{i\in[m]} \tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)})\right\}.$$

□

A.3.2. The Case When $\mathcal{F} = \{\mathbf{I}_k\}$

When $\mathcal{F}$ is only a singleton containing only $\mathbf{I}_k$, it is equivalent to set $\mathbf{D}_t^{(i)} = \mathbf{I}_r$ for all $t \in [T]$ and $i \in [m]$. In this case, the virtual sequence is actually a pure average: $\overline{\mathbf{Y}}_t = \sum_{i=1}^m p_i \mathbf{V}_t^{(i)}$.

Lemma 14. *Let $\mathbf{A} \in \mathbb{R}^{d\times k}$ with $d \geq k$ be any matrix with full rank. Denote by its QR factorization as $\mathbf{A} = \mathbf{QR}$ where $\mathbf{Q}$ is an orthgonal metrix. Let $\mathbf{E}$ be some perturbation matrix and $\mathbf{A} + \mathbf{E} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}$ the resulting QR factorization of $\mathbf{A} + \mathbf{E}$. When $\|\mathbf{E}\|_2\|\mathbf{A}^\dagger\|_2 < 1$, $\mathbf{A} + \mathbf{E}$ is of full rank. What's more, it follows that*

$$\|\widetilde{\mathbf{Q}} - \mathbf{Q}\|_2 \leq \sqrt{2k}\frac{\|\mathbf{A}^\dagger\|_2\|\mathbf{E}\|_2}{1 - \|\mathbf{A}^\dagger\|_2\|\mathbf{E}\|_2}.$$

Proof. Actually, we have

$$\|\widetilde{\mathbf{Q}} - \mathbf{Q}\|_F \overset{(a)}{\leq} \frac{\sqrt{2}\|\mathbf{E}\|_F}{\|\mathbf{E}\|_2}\ln\frac{1}{1 - \|\mathbf{A}^\dagger\|_2\|\mathbf{E}\|_2} \overset{(b)}{\leq} \sqrt{2}\frac{\|\mathbf{A}^\dagger\|_2\|\mathbf{E}\|_F}{1 - \|\mathbf{A}^\dagger\|_2\|\mathbf{E}\|_2} \overset{(c)}{\leq} \sqrt{2k}\frac{\|\mathbf{A}^\dagger\|_2\|\mathbf{E}\|_2}{1 - \|\mathbf{A}^\dagger\|_2\|\mathbf{E}\|_2}$$

where (a) comes from Theorem 5.1 in Sun (1995); (b) uses $\ln(1+x) \leq x$ for all $x > -1$; and (c) uses $\|\mathbf{E}\|_F \leq \sqrt{k}\|\mathbf{E}\|_2$. □

Lemma 15. *Let $\eta = \max_{i\in[m]} \|\mathbf{M}_i - \mathbf{M}\|_2/\|\mathbf{M}\|_2$ be sufficiently small. If $\mathbf{D}_t^{(i)}$ is solved from eqn. (13) with $\mathcal{F} = \{\mathbf{I}_k\}$, then eqn. (20) and eqn. (21) hold with*

$$\rho_t \leq 4\sqrt{2k}p\kappa^p\eta(1+\eta)^{p-1}$$

where $\kappa = \|\mathbf{M}\|_2\|\mathbf{M}^\dagger\|_2$ is the condition number of $\mathbf{M}$, $p = t - \tau(u)$, $\tau(t) \in \mathcal{I}_T$ is defined as the nearest synchronization time before t.

Proof. By Lemma 5, we are going to bound $\rho_t = \max_{i\in[m]} \|\mathbf{Z}^{(i)} - \mathbf{Z}_t^{(1)}\|_2$. Fix any $i \in [m]$ and $t \in [T]$. We will bound $\|\mathbf{Z}^{(i)} - \mathbf{Z}_t^{(1)}\|_2$ uniformly so that the bound holds for their maximum.

Fix any $i \in [m]$ and $t \in [T]$. Let $\tau(t)$ be the latest synchronization step before t and $p = t - \tau(t)$ be the number of nearest local updates. Note that $\mathbf{Z}_t^{(i)}$ and $\mathbf{Z}_t^{(1)}$ are the Q-factor of the QR factorization of $\mathbf{M}_i^p\mathbf{Z}_{\tau(t)}$ and $\mathbf{M}_1^p\mathbf{Z}_{\tau(t)}$. Let $\mathbf{Z}_t$ be the Q-factor of the QR factorization of $\mathbf{M}^p\mathbf{Z}_{\tau(t)}$. Then Lemma 14 yields

$$\|\mathbf{Z}_t^{(i)} - \mathbf{Z}_t\|_2 \leq \sqrt{2k}\frac{\|(\mathbf{M}^p\mathbf{Z}_{\tau(t)})^\dagger\|_2\|(\mathbf{M}_i^p - \mathbf{M}^p)\mathbf{Z}_{\tau(t)}\|_2}{1 - \|(\mathbf{M}^p\mathbf{Z}_{\tau(t)})^\dagger\|_2\|(\mathbf{M}_i^p - \mathbf{M}^p)\mathbf{Z}_{\tau(t)}\|_2} := \sqrt{2k}\frac{\omega}{1-\omega}$$

where $\omega = \|(\mathbf{M}^p\mathbf{Z}_{\tau(t)})^\dagger\|_2\|(\mathbf{M}_i^p - \mathbf{M}^p)\mathbf{Z}_{\tau(t)}\|_2$ for short. If $\omega \leq 1/2$, then we have $\|\mathbf{Z}_t^{(i)} - \mathbf{Z}_t\|_2 \leq 2\sqrt{2k}\omega$. Otherwise, we have $\omega \geq 1/2$ and $\|\mathbf{Z}_t^{(i)} - \mathbf{Z}_t\|_2 \leq 2 \leq \sqrt{2k} \leq 2\sqrt{2k}\omega$. Then we have for all $i \in [m]$,

$$\|\mathbf{Z}_t^{(i)} - \mathbf{Z}_t\|_2 \leq 2\sqrt{2k}\|(\mathbf{M}^p\mathbf{Z}_{\tau(t)})^\dagger\|_2\|(\mathbf{M}_i^p - \mathbf{M}^p)\mathbf{Z}_{\tau(t)}\|_2.$$

Hence,

$$\begin{aligned}
\rho_t &= \|\mathbf{Z}_t^{(i)} - \mathbf{Z}_t^{(1)}\|_2 \\
&\leq \|\mathbf{Z}_t^{(i)} - \mathbf{Z}_t\|_2 + \|\mathbf{Z}_t - \mathbf{Z}_t^{(1)}\|_2 \\
&\leq 2\sqrt{2k}\left[\|(\mathbf{M}^p\mathbf{Z}_{\tau(t)})^\dagger\|_2\|(\mathbf{M}_i^p - \mathbf{M}^p)\mathbf{Z}_{\tau(t)}\|_2 + \|(\mathbf{M}^p\mathbf{Z}_{\tau(t)})^\dagger\|_2\|(\mathbf{M}_1^p - \mathbf{M}^p)\mathbf{Z}_{\tau(t)}\|_2\right] \\
&\leq 4\sqrt{2k}\kappa^p\left[(1+\eta)^p - 1\right] \\
&\leq 4\sqrt{2k}p\kappa^p\eta(1+\eta)^{p-1}
\end{aligned}$$

where $\kappa = \|\mathbf{M}\|_2\|\mathbf{M}^\dagger\|_2$ is the condition number of $\mathbf{M}$. □

A.4. Proof of Theorem 1 and Theorem 2

Proof. We provide a proof in four steps.

First step: Perturbed iterate analysis. Recall that we defined a virtual sequence by

$$\overline{\mathbf{Y}}_t = \sum_{i=1}^m p_i \mathbf{Y}_t^{(i)} \mathbf{O}_t^{(i)}.$$

Notice that this sequence never has to be computed explicitly, it is just a virtual sequence we use in the analysis. From Lemma 4, we construct the iteration of the virtual sequence $\{\overline{\mathbf{Y}}_t\}$ as

$$\overline{\mathbf{Y}}_{t+1} = \left(\mathbf{M}\overline{\mathbf{Y}}_t + \mathbf{G}_t\right)\mathbf{R}_t^{-1}$$

where $\mathbf{M} = \frac{1}{n}\mathbf{A}^\top\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathbf{G}_t$ is the noise term inccured by the variance among different nodes, and $\mathbf{R}_t$ is chosen according to Lemma 8. Recall that $\mathbf{G}_t = \mathbf{H}_t + \mathbf{W}_t$ is given in eqn. (16) with $\mathbf{H}_t = \sum_{i=1}^m p_i \mathbf{H}_t^{(i)}$ and $\mathbf{W}_t = \sum_{i=1}^m p_i \mathbf{W}_t^{(i)}$.

Second step: Bound the noise term $\mathbf{G}_t$. Let $p = \mathrm{gap}(\mathcal{I}_T)$ denotes by the longest interval between subsequent synchronization steps. In order to guarantee convergence, we should make sure the noise term $\mathbf{G}_t$ is small enough. In particular, we require

$$\|\mathbf{G}_t\overline{\mathbf{Y}}_t^\dagger\|_2 \le \frac{\sigma_k - \sigma_{k+1}}{5}\min\left(\frac{\sqrt{r}-\sqrt{k-1}}{\tau\sqrt{d}}, \epsilon\right) \tag{23}$$

By Lemma 13 or 15, we always have

$$\|\mathbf{H}_t\overline{\mathbf{Y}}_t^\dagger\|_2 \le \frac{2\sigma_1\kappa\eta 1_{t\notin\mathcal{I}_T}}{1-\eta\kappa-(1-\max_{i\in[m]}p_i)\rho_{t-1}}$$

$$\|\mathbf{W}_t\overline{\mathbf{Y}}_t^\dagger\|_2 \le 4(1-\max_{i\in[m]}p_i)\sigma_1\kappa\frac{\rho_t + (\rho_{t-1}+\eta)1_{t\notin\mathcal{I}_T}}{1-\eta\kappa-(1-\max_{i\in[m]}p_i)\rho_{t-1}}$$

We assume $\eta\kappa \le 1/3$ and additionally assume $(1-\max_{i\in[m]}p_i)\rho_{t-1} \le \frac{1}{3}$. Then the last two inequalities become

$$\|\mathbf{H}_t\overline{\mathbf{Y}}_t^\dagger\|_2 \le 6\sigma_1\kappa\eta 1_{t\notin\mathcal{I}_T} := 6\sigma_1\kappa\Psi_t$$

$$\|\mathbf{W}_t\overline{\mathbf{Y}}_t^\dagger\|_2 \le 12(1-\max_{i\in[m]}p_i)\sigma_1\kappa\left[\rho_t + (\rho_{t-1}+\eta)1_{t\notin\mathcal{I}_T}\right] := 12\sigma_1\kappa\Omega_t$$

Then in order to ensure eqn. (23), we only need to ensure

$$6\sigma_1\Psi_t + 12\sigma_1\Omega_t \le \frac{\sigma_k-\sigma_{k+1}}{5\kappa}\min\left(\frac{\sqrt{r}-\sqrt{k-1}}{\tau\sqrt{d}}, \epsilon\right).$$

A sufficient condition to that is

$$\Psi_t + \Omega_t \le \frac{1}{60}\frac{\sigma_k-\sigma_{k+1}}{\sigma_1\kappa}\min\left(\frac{\sqrt{r}-\sqrt{k-1}}{\tau\sqrt{d}}, \epsilon\right) = \mathcal{O}(\epsilon_0). \tag{24}$$

Finally, we argue that the condition $(1-\max_{i\in[m]}p_i)\rho_{t-1} \le \frac{1}{3}$ is indicated in the uniform boundedness of eqn. (24) (i.e., eqn. (24) holds for all $t \in [T]$). This is because

$$(1-\max_{i\in[m]}p_i)\rho_{t-1} \le \Omega_{t-1} \le \Psi_{t-1}+\Omega_{t-1} \le \frac{\epsilon_0}{60} < \frac{1}{3}.$$

Third step: Bound ρ_t. Let $\kappa = \|\mathbf{M}\|_2\|\mathbf{M}^\dagger\|_2$ be the condition number of $\mathbf{M}$ and $p = t-\tau(u)$ with $\tau(t) \in \mathcal{I}_T$ defined as the nearest synchronization time before t. Then, we can prove Theorem 2 now.

- If $\mathcal{F} = \mathcal{O}_k$, then

$$\rho_t \le \sqrt{2}\min\left\{\frac{2\kappa^p\eta(1+\eta)^{p-1}}{(1-\eta)^p}, \frac{\eta\sigma_1}{\delta_k} + 2\gamma_k^{p/4}\max_{i\in[m]}\tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)})\right\}.$$

with the parameters δ_k, γ_k given in Lemma 13. By requiring $\eta \leq 1/p$, we have $\frac{(1+\eta)^{p-1}}{(1-\eta)^p} \leq \frac{(1+1/p)^{p-1}}{(1-1/p)^p} \leq \mathrm{e}^2$. Define $C_t = \max_{i\in[m]} \tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)})$. Latter we will show that since `LocalPower` converges under Assumption 1, then $\lim_{t\to\infty} \sin\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k) \leq \epsilon$. Then, we have

$$\begin{aligned}
\limsup_{t\to\infty} C_t &= \limsup_{t\to\infty} \max_{i\in[m]} \tan\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)}) \\
&= \limsup_{t\to\infty} \max_{i\in[m]} \tan\arg\sin\sin\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k^{(i)}) \\
&\leq \limsup_{t\to\infty} \max_{i\in[m]} \tan\arg\sin(\sin\theta_k(\mathbf{Z}_{\tau(t)}, \mathbf{U}_k) + \sin\theta_k(\mathbf{U}_k, \mathbf{U}_k^{(i)})) \\
&\leq \max_{i\in[m]} \tan\arg\sin(\frac{\eta\sigma_1}{\delta_k} + \epsilon) = \mathcal{O}(\eta + \epsilon).
\end{aligned}$$

It can be seen that when p is sufficiently large, $\rho_t = \mathcal{O}(\eta)$ which is independent with p.

- If $\mathcal{F} = \{\mathbf{I}_k\}$, then
$$\rho_t \leq 4\sqrt{2k}p\kappa^p\eta(1+\eta)^{p-1} \leq 4\mathrm{e}\sqrt{2k}p\kappa^p\eta.$$

Simply put together, if $\Psi + \Omega \leq \epsilon_0$, we can firmly ensure eqn. (23) holds.

Forth step: Establish convergence. Let's first assume eqn. (18) holds. With eqn. (18), the following argument is quite similar to Hardt & Price (2014). Note that Specifically, we will see that at every step t of the algorithm,

$$\tan\theta_k(\mathbf{U}_k, \overline{\mathbf{Y}}_t) \leq \max\left(\epsilon, \tan\theta_k(\mathbf{U}_k, \mathbf{Z}_0)\right),$$

which implies for $\epsilon \leq \frac{1}{2}$ that

$$\cos\theta_k(\mathbf{U}_k, \overline{\mathbf{Z}}_t) \geq \min\left(1-\epsilon^2/2, \cos\theta_k(\mathbf{U}_k, \mathbf{Z}_0)\right) \geq \frac{7}{8}\cos\theta_k(\mathbf{U}_k, \mathbf{Z}_0)$$

so Lemma 5 applies at every step. This means that

$$\tan\theta_k(\mathbf{U}_k, \overline{\mathbf{Y}}_{t+1}) \leq \max\left(\epsilon, \delta\tan\theta_k(\mathbf{U}_k, \overline{\mathbf{Y}}_t)\right)$$

for $\delta = \max(\epsilon, (\sigma_{k+1}/\sigma_k)^{1/4})$. After $T \geq \log_{1/\delta}\frac{\tan\theta_k(\mathbf{U}_k, \mathbf{Z}_0)}{\epsilon}$ steps, the tangent will reach the accuracy ϵ and remain there. So we have

$$\|(\mathbf{I} - \mathbf{Z}_T\mathbf{Z}_T^\top)\mathbf{U}\| = \sin\theta_k(\mathbf{U}_k, \overline{\mathbf{Y}}_T) \leq \tan\theta_k(\mathbf{U}_k, \overline{\mathbf{Y}}_T) \leq \epsilon.$$

Plus the observation that

$$\log(1/\delta) \geq c\min(\log(1/\epsilon), \log(\sigma_k/\sigma_{k+1})) \geq c\min\left(1, \log\frac{1}{1-\gamma}\right) \geq c\min(1,\gamma) = c\gamma$$

where $\gamma = 1 - \sigma_{k+1}/\sigma_k$ and $c = \frac{1}{4}$, we can set $T \in \mathcal{I}_T$ and

$$T = \Omega\left(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}}\log(d\tau/\varepsilon)\right).$$

Finally we are going to show that once the noise term $\mathbf{G}_t$ is bounded as eqn. (23), eqn. (18) would naturally hold. From Lemma 6, we have

$$\tan\theta_k(\mathbf{U}, \mathbf{Z}_0) \leq \frac{\tau\sqrt{d}}{\sqrt{r} - \sqrt{k-1}}$$

with all but $\tau^{-\Omega(p+1-k)} + e^{-\Omega(d)}$ probability. Hence

$$\cos\theta_k(\mathbf{U}, \mathbf{Z}_0) \geq \frac{1}{1+\tan\theta_k(\mathbf{U}, \mathbf{Z}_0)} \geq \frac{\sqrt{r} - \sqrt{k-1}}{2\tau\sqrt{d}}.$$

□

A.5. Proof of Corollary 1

Proof. From Theorem 1, our algorithm has error no larger than ϵ. We then find the minimum ϵ that is a function of m, n, p by combining Theorem 2 and Lemma 2. For a fixed n/m, Lemma 2 bounds η in terms of s or equivalently n/m, implies $\eta = \sqrt{\frac{3\mu\rho}{s_i}\log\left(\frac{\rho m}{\delta}\right)} = \widetilde{\Theta}(\sqrt{\frac{\mu\rho}{s}}) = \widetilde{\Theta}(\sqrt{\frac{m\mu\rho}{n}})$. For sufficiently small ϵ, we have $\epsilon = \frac{\sigma_1\kappa}{\sigma_k-\sigma_{k+1}}\epsilon_0$. Let ϵ be sufficiently small such that eqn. (6) just holds. Then we have

$$\epsilon = \Theta(\epsilon_0) = \Theta(\eta + \sup_t(\rho_t + \rho_{t-1})) = \mathcal{O}(h_p(\eta) + \eta)$$

where the last equality follows from Theorem 2 which bounds ρ_t in terms of η. It is in the form of $\rho_t \leq h_p(\eta)$ where $h_1(\cdot) = 0$ and $h_p(\eta)$ typically increases in p and η. With OPT, $h_p(\eta) = O(\eta)$, while without OPT, $h_p(\eta) = \mathcal{O}(\sqrt{k}p\kappa^p\eta)$.

If we use any decay strategy in which p converges to 1 finally, then `LocalPower` is reduced to `DPI` finally and thus of course achieves zero error asymptotically. □

B. Statistical Error Between the Empirical Matrix and the Population One

Recall that $\mathbf{M} = \frac{1}{n}\mathbf{A}\mathbf{A}^\top = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ is the empirical correlation matrix and $\mathbf{M}_* = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\mathbf{x}\mathbf{x}^\top$ is the population one. By Matrix Hoeffding theorem, we can bound $\|\mathbf{M} - \mathbf{M}_*\|$ in terms of samples.

Lemma 16 (Matrix Hoeffding inequality Tropp (2012))**.** *Let $\mathcal{D}$ be a distribution over vectors with squared ℓ_2 norm at most b. Let $\mathbf{M}_* = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\mathbf{x}\mathbf{x}^\top$ and $\mathbf{M} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ where $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are sampled i.i.d. from $\mathcal{D}$, then it holds that*

$$\mathbb{P}\left(\|\mathbf{M}_* - \mathbf{M}\| \geq t\right) \leq d \cdot \exp\left(-\frac{t^2 n}{16b^2}\right).$$

Let the top-k eigenspace of $\mathbf{M}$ and $\mathbf{M}_*$ be respectively $\mathbf{V}_k$ and $\mathbf{V}_{k,*}$ (both of which are orthonormal). Let $\widehat{\mathbf{V}}$ be any estimated top-k eigenvector matrix (for example, $\mathbf{Z}_T$ produced by `LocalPower`). If we care how accurately $\widehat{\mathbf{V}}$ approximate $\mathbf{V}_{k,*}$, by the triangle inequality,

$$\operatorname{dist}(\widehat{\mathbf{V}}, \mathbf{V}_{k,*}) \leq \underbrace{\operatorname{dist}(\widehat{\mathbf{V}}, \mathbf{V}_k)}_{\text{optimization error}} + \underbrace{\operatorname{dist}(\mathbf{V}_k, \mathbf{V}_{k,*})}_{\text{statistical error}}$$

Theorem 1 characterizes the diminishing speed of the optimization term, however, has nothing to do with the statistical error. The latter is controlled by the available samples through the combination of the Davis-Kahan $\sin(\theta)$ theorem (Lemma 11) and $\|\mathbf{M}_* - \mathbf{M}\|$. In particular, with probability greater than $1 - \delta$, the statistical error is no larger than

$$\frac{1}{\delta_k} 4b\sqrt{\frac{\ln\frac{d}{\delta}}{n}}.$$

If only a single machine attends the training, $n = s$, while if m machines cooperate, $n = ms$. From the last inequality, the statistical error is reduced by a factor of $\sqrt{m}$.

C. Dependence on $\sigma_k - \sigma_{k+1}$

Our result depends on $\sigma_k - \sigma_{k+1}$ even when $r > k$ where r is the number of columns used in subspace iteration. This is mainly because we borrow tools from Hardt & Price (2014) to prove the theory. In the analysis of Hardt & Price (2014), the required iteration depends on the consecutive eigengap $\sigma_k - \sigma_{k+1}$ even when $r > k$ where r is the number of columns used in subspace iteration. Note that $\sigma_k - \sigma_{k+1}$ can be unimaginably small in practical large-scale problems. Balcan et al. (2016a) improved the result to a slightly milder dependency on $\sigma_k - \sigma_{q+1}$ by proposing a novel characterization measuring the discrepancy between the running rank-r subspace $\mathbf{Z}_t$ and target top-k eigenspace $\mathbf{U}_k$, where q is any intermediate integer between k and r. If we borrow the idea from the improved analysis of Balcan et al. (2016a), we can refine the result. In that case, the needed computation rounds will depend on $\sigma_k - \sigma_{q+1}$ as a result. All the above discussion can be easily parallel.

Theorem 3. *Let Assumption 1 hold with sufficiently small $\eta\kappa \leq \frac{1}{3}$ where $\kappa = \|\mathbf{M}\|\|\mathbf{M}^\dagger\|$ is the condition number of $\mathbf{M}$. Let Assumption 1 holds with $\tau > 0$ and the following ϵ_0*

$$\epsilon_0 = \frac{\sqrt{r} - \sqrt{q-1}}{\tau\sqrt{d}} \min\left\{\frac{\sigma_k - \sigma_{q+1}}{\sigma_1}\epsilon, \frac{\sigma_q}{\sigma_1}\right\}.$$

Let $k \leq q \leq r$. If we borrow the refined analysis in Balcan et al. (2016a), then for sufficiently small ϵ satisfying

$$\epsilon = \mathcal{O}\left(\frac{\sigma_q}{\sigma_k} \cdot \min\left\{\frac{1}{\log(\sigma_k/\sigma_q)}, \frac{1}{\log(\tau d)}\right\}\right),$$

when

$$T = \Omega\left(\frac{\sigma_k}{\sigma_k - \sigma_{q+1}} \log\left(\frac{\tau d}{\epsilon}\right)\right)$$

after $|\mathcal{I}_T|$ rounds of communication, with probability at least $1 - \tau^{-\Omega(r+1-q)} - e^{-\Omega(d)}$, we have

$$\operatorname{dist}(\mathbf{Z}_T, \mathbf{U}_k) = \sin\theta_k(\mathbf{Z}_T, \mathbf{U}_k) = \left\| \left(\mathbf{I}_d - \mathbf{Z}_T\mathbf{Z}_T^\top\right)\mathbf{U}_k \right\| \leq \epsilon.$$

Proof. We use Corollary A.1 in Balcan et al. (2016a) instead of Lemma 5 in the third step of the proof of Theorem 1. □

D. Related Work

Truncated SVD or principal component analysis (PCA) is one of the most important and popular techniques in data analysis and machine learning. A multitude of researches focus on iterative algorithms such as power iterations or its variants (Golub & Van Loan, 2012; Saad, 2011). These deterministic algorithms inevitably depends on the spectral gap, which can be quite large in large scale problems. Another branch of algorithm seek alternatives in stochastic and incremental algorithms (Oja & Karhunen, 1985; Arora et al., 2013; Shamir, 2015; 2016; De Sa et al., 2018). Some work could achieve eigengap-free convergence rate and low-iteration-complexity (Musco & Musco, 2015; Shamir, 2016; Allen-Zhu & Li, 2016).

Large-scale problems necessitate cooperation among multiple worker nodes to overcome the obstacles of data storage and heavy computation. For a review of distributed algorithms for PCA, one could refer to Wu et al. (2018). One feasible approach is divide-and-conquer algorithm which performs a one-shot averaging of the individual top-k eigenvectors (or subspace) returned by worker nodes (Garber et al., 2017; Fan et al., 2019; Bhaskara & Wijewardena, 2019; Charisopoulos et al., 2020). In particular, the concurrent work (Charisopoulos et al., 2020) proposes to average local eigenvector matrices via OPT as ours, though they focus on one-shot scenario and obtain better error analysis. The divide-and-conquer algorithms have only one round of communication. To reach a certain accuracy, it often requires that the per-machine sample size s to grow with the number of machines m (Garber et al., 2017), which means it is only effective in large local dataset regime.

Another line of results for distributed eigenspace estimation uses iterative algorithms that perform multiple communication rounds. They require much smaller sample size and can often achieve arbitrary accuracy. For example, in our work, we only require the per-machine sample size s depends on m in a very mild way like $\mathcal{O}(\ln m)$, however, Garber et al. (2017) requires $s = \widetilde{\mathcal{O}}(m)$ to reach a comparable result. Some works make use of shift-and-invert framework (S&I) for PCA (Garber & Hazan, 2015; Garber et al., 2016; Allen-Zhu & Li, 2016). S&I methods turn the problem of computing the leading eigenvector to that of approximately solving a small system of linear equations. This, in turn, could be solved by arbitrary convex solvers (Xu, 2018), and, therefore, can be extended in distributed settings naturally. Garber et al. (2017) coupled S&I methods with a distributed first-order convex solver, giving guarantees in terms of communication costs. Gang et al. (2019) turns the problem of distributed PCA into a constraint optimization problem (by letting each device hold a independent parameter and adding a constraint that all local parameter should be same), and then uses gradient-based methods to solve it iteratively. Chen et al. (2021) combined S&I methods with a distributed approximate Newton method where the communication cost is saved by only using the Hessian information on the first machine. Very recently, Grammenos et al. (2019) proposed a federated, asynchronous, and differential privacy algorithm for distributed PCA. Methodologically, the algorithm is not power-iteration-based. Instead, their algorithm incrementally computes local model updates using streaming procedure and adaptively estimates its leading principle components. In particular, they assume the clients are arranged in a tree-like structure, while we did not make such assumption.

Recently, the technique of local updates emerges as a simple but powerful tool in distributed empirical risk minimization (McMahan et al., 2017; Zhou & Cong, 2017; Stich, 2018; Wang & Joshi, 2018b; Yu et al., 2019; Li et al., 2019a;b; Khaled et al., 2019). Distributed algorithms with local updates typically alternate between local computation and periodical communication. Therefore, local updates allow less frequent communication but incur more computation due to the inevitably accumulated residual errors. This paper uses local updates for the distributed power iteration. However, our analysis is totally different from the local SGD algorithms (Zhou & Cong, 2017; Stich, 2018; Wang & Joshi, 2018b; Yu et al., 2019; Li et al., 2019a;b; Khaled et al., 2019). A main challenge in analyzing `LocalPower` is that the local SGD

algorithms for empirical risk minimization often involve an explicit form of (stochastic) gradients. For SVD or PCA, a canonical example of non-convex problems, the gradient cannot be explicitly expressed, so the existing techniques cannot be applied (Simchowitz et al., 2017). Instead, we borrowed tools from the noisy power method (Hardt & Price, 2014; Balcan et al., 2016a) and carefully analyze the residual errors.

In our paper, we only consider the centralize PCA, where there is a server connecting all other nodes. However, the technique of local updates can also be used in other settings like decentralized or streaming PCA (Gang et al., 2019; Raja & Bajwa, 2020).

E. Experiments

E.1. Experimental Settings

We conduct experiments to demonstrate the communication efficiency of `LocalPower`. We use several datasets from the LIBSVM website and summarize them in Table 4. Our focus is the needed communication round required to minimize the optimization error and analyze `LocalPower` through different lens. For comparison, we consider the following baselines:

1. Weighted Distributed Averaging (Bhaskara & Wijewardena, 2019);
2. Unweighted Distributed Averaging (Fan et al., 2019);
3. Distributed Randomized SVD.
4. Distributed power Iteration (the case of `LocalPower` when $p = 1$)

For completeness, we include the former three algorithms in the next subsection. We also study the effect of different choice of m, p, and the decay strategy. Throughout, we use either $\mathcal{I}_T = \{0, p, 2p, \cdots, T\}$ or the decay strategy.

Preprocessing. The data are randomly shuffled and partitioned among m nodes. We scale each feature by dividing it by the maximum value of each coordinate, so that each feature locates between $[-1, 1]$. In particular, we will first find the maximum value for each feature coordinate among all workers in the system and share it with all participants. All the experiments use the same initialization $\mathbf{Z}_0 \in \mathbb{R}^{d\times r}$ (for any $r > k$) which contains a set of randomly generated orthonormal bases.

Experimental. All the experiments are conducted on a single machine. We fix the target rank to $k = 5$. We plot $\mathrm{dist}(\mathbf{Z}_t, \mathbf{U}_k) = \|(1 - \mathbf{Z}_t\mathbf{Z}_t^\top)\mathbf{U}_k\| = \sin\theta_k(\mathbf{Z}_t, \mathbf{U}_k)$ against the number of communications to evaluate communication efficiency. In Table 4, we list the information of (n, d) for the datasets we use, all satisfying $n \ll d$. Though we focus on large n regime, latter we also test large d regimes namely $n \approx d$ for completeness. In Table 5, we estimate η by $\max_{i\in[m]} \|\mathbf{M}_i - \mathbf{M}\|_2/\|\mathbf{M}\|_2$. Under uniform sampling, when we fix n, the larger m (equals to smaller s), the larger η.

Table 4. A summary of used data sets from the LIBSVM website.

Data set	n	d	Data set	n	d
A9a	32561	123	Abalone	2114	8
Acoustic	78823	50	Aloi	108000	128
Combined	78823	100	Connect-4	7990	125
Covtype	581,012	54	Housing	506	13
Ijcnn1	49990	22	MNIST	60,000	780
Poker	25010	10	Space-ga	3107	6
Splice	1000	24	W8a	49749	300
MSD	463,715	90			

Table 5. The value of η under uniform partitions on fifteen datasets. In the following experiments, we uniformly distribute n samples into $m = \max(\lfloor \frac{n}{1000} \rfloor, 3)$ so that each device has about 1000 samples. It implies m ranges from 20 to 100, which is the range we consider here. To fill the following table, we distributed n samples into m devices and estimate it by $\eta = \max_{i\in[m]} \|\mathbf{M} - \mathbf{M}_i\|_2 / \|\mathbf{M}\|_2$. It can be seen that for a fixed n, the larger m, the larger η.

Dataset	$m = 20$	$m = 40$	$m = 60$	$m = 80$	$m = 100$
A9a	0.034	0.0563	0.0701	0.0906	0.0998
Abalone	0.1089	0.23	0.2458	0.2629	0.3556
Acoustic	0.0063	0.0107	0.0134	0.0179	0.0199
Aloi	0.0479	0.0659	0.1023	0.1162	0.203
Combined	0.006	0.0089	0.0113	0.014	0.0158
Connect-4	0.0376	0.054	0.0771	0.0791	0.0899
Covtype	0.0078	0.011	0.0159	0.0164	0.0202
Housing	0.3117	0.3747	0.5062	0.6442	0.6741
Ijcnn1	0.016	0.0288	0.0348	0.0363	0.0489
MNIST	0.0396	0.0584	0.0689	0.0896	0.0904
Poker	0.0369	0.0519	0.0702	0.0803	0.0904
Space-ga	0.0855	0.1317	0.1495	0.2111	0.3446
Splice	0.1627	0.2484	0.3154	0.3957	0.4717
W8a	0.1046	0.1664	0.1937	0.2515	0.3167
MSD	0.0007	0.0009	0.0012	0.0014	0.0015

E.2. One-shot Baseline Algorithms

Algorithm 2 Unweighted Distributed Averaging (`UDA`) (Fan et al., 2019)

1: **Input:** distributed dataset $\{\mathbf{A}_i\}_{i=1}^m$ with $\mathbf{A}_i \in \mathbb{R}^{s_i \times d}$, target rank k.
2: **Local:** Each device computes the rank-k SVD of $\mathbf{M}_i = \frac{1}{s_i}\mathbf{A}_i^\top \mathbf{A}_i$ as $\widehat{\mathbf{V}}_i \mathbf{\Sigma}_i \widehat{\mathbf{V}}_i^\top$ with $\mathbf{\Sigma}_i \in \mathbb{R}^{k\times k}$ and $\widehat{\mathbf{V}}_i \in \mathbb{R}^{d\times k}$.
3: **Server:** The central server computes $\widetilde{\mathbf{M}} = \frac{1}{m}\sum_{i=1}^n \widehat{\mathbf{V}}_i \widehat{\mathbf{V}}_i^\top$, then output the top k eigenvalues and the corresponding eigenvectors of $\widetilde{\mathbf{M}}$.

Algorithm 3 Weighted Distributed Averaging (`WDA`) (Bhaskara & Wijewardena, 2019)

1: **Input:** distributed dataset $\{\mathbf{A}_i\}_{i=1}^m$ with $\mathbf{A}_i \in \mathbb{R}^{s_i \times d}$, target rank k.
2: **Local:** Each device computes the rank-k SVD of $\mathbf{M}_i = \frac{1}{s_i}\mathbf{A}_i^\top \mathbf{A}_i$ as $\widehat{\mathbf{V}}_i \mathbf{\Sigma}_i \widehat{\mathbf{V}}_i^\top$ with $\mathbf{\Sigma}_i \in \mathbb{R}^{k\times k}$ and $\widehat{\mathbf{V}}_i \in \mathbb{R}^{d\times k}$.
3: **Server:** The central server computes $\widetilde{\mathbf{M}} = \frac{1}{m}\sum_{i=1}^n \widehat{\mathbf{V}}_i \mathbf{\Sigma}_i \widehat{\mathbf{V}}_i^\top$, then output the top k eigenvalues and the corresponding eigenvectors of $\widetilde{\mathbf{M}}$.

Algorithm 4 Distributed Randomized SVD (`DR-SVD`) (A distributed variant of Randomized SVD in Halko et al. (2011))

1: **Input:** distributed dataset $\{\mathbf{A}_i\}_{i=1}^m$, $\mathbf{A} = [\mathbf{A}_1^\top, \cdots, \mathbf{A}_m^\top]^\top \in \mathbb{R}^{n\times d}$ with target rank k, $\mathbf{A}_i \in \mathbb{R}^{s_i \times d}$ and $r = k + \lfloor \frac{d-k}{4} \rfloor$.
2: The server generates a $d \times r$ random Gaussian matrix $\mathbf{\Omega}$;
3: The server learns $\mathbf{Y} = \mathbf{A}\mathbf{A}^\top \mathbf{A}\mathbf{\Omega}$ and obtains an orthonormal $\mathbf{Q} \in \mathbb{R}^{n\times r}$ by QR decomposition on $\mathbf{Y}$;
4: Let $\mathbf{Q} = [\mathbf{Q}_1^\top, \cdots, \mathbf{Q}_m^\top]^\top$ with $\mathbf{Q}_i \in \mathbb{R}^{s_i \times r}$ and each worker receives $\mathbf{Q}_i$;
5: The i-th worker computes $\mathbf{B}_i = \mathbf{Q}_i^\top \mathbf{A}_i \in \mathbb{R}^{r\times d}$ for all $i \in [m]$;
6: The server aggregate $\mathbf{B} = \sum_{i=1}^m \mathbf{B}_i = \mathbf{Q}^\top \mathbf{A}$ and perform SVD: $\mathbf{B} = \widetilde{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T$;
7: Set $\widehat{\mathbf{U}} = \mathbf{Q}\widetilde{\mathbf{U}}$;
8: **Output:** the first k columns of $(\widehat{\mathbf{U}}, \widehat{\mathbf{\Sigma}}, \widehat{\mathbf{V}})$.

E.3. Additional Experiments Results

Table 6. Error comparison among three one-shot baseline algorithms and our `LocalPower`. We uniformly distribute n samples into $m = \max(\lfloor \frac{n}{1000} \rfloor, 3)$ devices so that each device has about 1000 samples. We show the mean errors of ten repeated experiments with its standard deviation enclosed in parentheses. Here we use $p = 4$ for all variants of `LocalPower` and sufficiently large T's which guarantee `LocalPower` converges. For better visualization, we show the box plot of final errors of ten repeated experiments in Figure 4.

Datasets	`LocalPower` with $p = 4$			DR-SVD	UDA	WDA
	OPT	Sign-fixing	Vanilla			
A9a	**4.09e-03** (**4.20e-04**)	5.82e-03 (1.41e-03)	8.13e-02 (3.44e-02)	4.63e-02 (9.24e-03)	2.64e-02 (1.58e-02)	2.40e-02 (1.50e-02)
Abalone	**3.16e-03** (2.89e-03)	3.85e-03 (**2.54e-03**)	3.03e-02 (5.70e-02)	3.20e-01 (2.30e-01)	1.03e-01 (9.38e-02)	1.03e-01 (9.18e-02)
Acoustic	**1.83e-03** (4.40e-04)	2.03e-03 (3.90e-04)	2.38e-03 (8.50e-04)	1.54e-02 (6.59e-03)	7.76e-03 (2.64e-03)	6.67e-03 (2.41e-03)
Aloi	3.07e-02 (1.10e-02)	6.57e-02 (1.06e-02)	5.24e-02 (1.10e-02)	**1.92e-03** (**4.30e-04**)	4.80e-02 (1.10e-02)	4.37e-02 (4.73e-03)
Combined	6.01e-03 (1.59e-03)	**5.57e-03** (**1.05e-03**)	2.47e-02 (3.40e-02)	5.19e-02 (6.23e-03)	4.63e-02 (2.97e-02)	4.16e-02 (2.76e-02)
Connect-4	**1.27e-02** (4.52e-03)	1.81e-02 (3.79e-03)	1.70e-02 (4.35e-03)	1.61e-02 (2.96e-03)	1.65e-01 (3.48e-02)	1.56e-01 (3.26e-02)
Covtype	7.38e-03 (8.50e-04)	**6.23e-03** (**3.30e-04**)	1.28e-02 (1.88e-03)	1.82e-01 (8.73e-02)	6.09e-02 (9.70e-03)	5.60e-02 (9.41e-03)
Housing	**1.18e-02** (**5.45e-03**)	2.76e-02 (1.14e-02)	3.84e-02 (5.11e-02)	5.66e-01 (2.62e-01)	9.16e-02 (5.09e-02)	5.89e-02 (3.25e-02)
Ijcnn1	**1.53e-01** (1.87e-01)	1.95e-01 (2.45e-01)	3.23e-01 (2.24e-01)	1.21e+00 (1.70e-01)	3.85e-01 (7.62e-02)	3.67e-01 (**7.59e-02**)
MNIST	2.62e-03 (3.40e-04)	4.85e-03 (8.00e-04)	5.08e-03 (7.90e-04)	**5.00e-05** (**0.00e+00**)	1.08e-02 (3.00e-03)	8.91e-03 (2.53e-03)
Poker	**6.45e-03** (**1.90e-03**)	1.08e-02 (3.34e-03)	5.33e-02 (3.63e-02)	1.25e+00 (1.61e-01)	2.39e-02 (3.00e-03)	2.00e-02 (2.19e-03)
Space-ga	**2.80e-04** (**1.40e-04**)	5.10e-04 (2.90e-04)	6.50e-04 (3.60e-04)	7.40e-01 (2.14e-01)	2.83e-02 (2.46e-02)	3.82e-02 (2.72e-02)
Splice	**1.61e-02** (**5.46e-03**)	2.87e-02 (8.93e-03)	7.45e-02 (9.26e-02)	4.52e-01 (1.37e-01)	1.56e-01 (7.08e-02)	1.34e-01 (6.26e-02)
W8a	1.90e-02 (2.46e-03)	1.75e-02 (1.76e-03)	**1.68e-02** (**1.29e-03**)	7.13e-02 (2.06e-02)	1.52e-01 (4.37e-02)	1.51e-01 (4.11e-02)
MSD	9.90e-03 (1.21e-03)	**9.62e-03** (**5.20e-04**)	1.44e-02 (1.58e-03)	3.01e-02 (9.64e-03)	1.55e-02 (1.39e-03)	1.92e-02 (1.14e-03)

Table 7. Error comparison among `LocalPower` with the decay strategy and three different $\mathcal{F}$. We uniformly distribute n samples into $m = \max(\lfloor \frac{n}{1000} \rfloor, 3)$ devices so that each device has about 1000 samples. We show the mean errors of ten repeated experiments with its standard deviation enclosed in parentheses. Here we use $p = 4$ for all variants of `LocalPower` and sufficiently large T's which guarantee `LocalPower` converges.

Datasets	`LocalPower` with the decay strategy		
	OPT	Sign-fixing	Vanilla
A9a	4.84e-03 (1.40e-02)	1.52e-03 (4.08e-03)	**3.11e-04** (**4.84e-04**)
Abalone	**3.50e-10** (4.10e-10)	4.14e-10 (**4.00e-10**)	6.12e-10 (6.77e-10)
Acoustic	**1.40e-05** (**2.16e-05**)	1.92e-05 (3.72e-05)	2.28e-05 (4.91e-05)
Aloi	**5.82e-10** (**5.17e-10**)	1.71e-09 (2.20e-09)	2.36e-09 (2.14e-09)
Combined	3.68e-03 (5.63e-03)	7.74e-03 (1.70e-02)	**2.99e-03** (**3.88e-03**)
Connect-4	4.90e-03 (8.47e-03)	3.58e-03 (4.35e-03)	**3.09e-03** (**3.16e-03**)
Covtype	5.57e-04 (1.55e-03)	**4.95e-05** (**5.40e-05**)	8.01e-05 (8.62e-05)
Housing	**1.38e-05** (**2.88e-05**)	2.20e-05 (5.66e-05)	2.08e-05 (5.68e-05)
Ijcnn1	3.56e-01 (1.97e-01)	3.33e-01 (**1.67e-01**)	**3.32e-01** (1.72e-01)
MNIST	2.06e-05 (2.38e-05)	**1.72e-05** (1.62e-05)	1.72e-05 (**1.62e-05**)
Poker	**3.08e-03** (**1.49e-03**)	3.22e-03 (1.82e-03)	3.22e-03 (1.93e-03)
Space-ga	**3.47e-14** (2.13e-14)	3.56e-14 (**2.11e-14**)	3.87e-14 (2.27e-14)
Splice	**4.11e-07** (**5.29e-07**)	8.88e-07 (1.24e-06)	1.01e-06 (1.34e-06)
W8a	**1.70e-03** (**2.46e-03**)	1.85e-02 (4.94e-02)	6.09e-03 (9.60e-03)
MSD	2.75e-05 (3.34e-05)	**2.47e-05** (3.27e-05)	3.02e-05 (**2.10e-05**)

[11] Actually, it means setting $\mathcal{F}$ for `LocalPower` as $\mathcal{O}_k, \mathcal{D}_k$ and $\{\mathbf{I}_k\}$ respectively (see eqn. (13) for the reason).

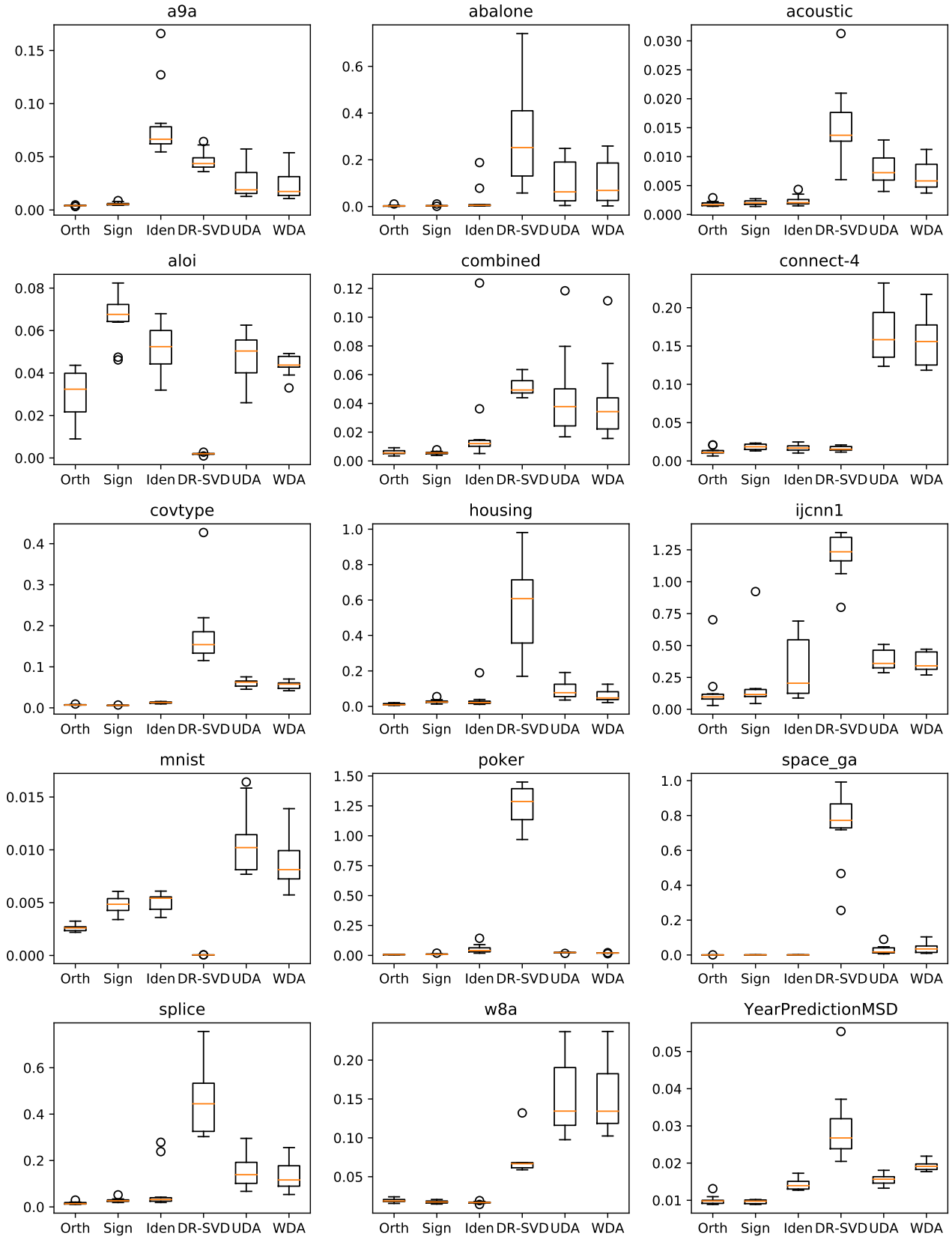

Figure 4. Box plot of Table 6 for better visualization. Here Orth, Sign and Iden represents OPT, sign-fixing and the vanilla `LocalPower` respectively.[11] We can see that for most datasets, `LocalPower` with $p = 4$ obtains smallest error and more stability. We can obtain zero error if we use the decay strategy.

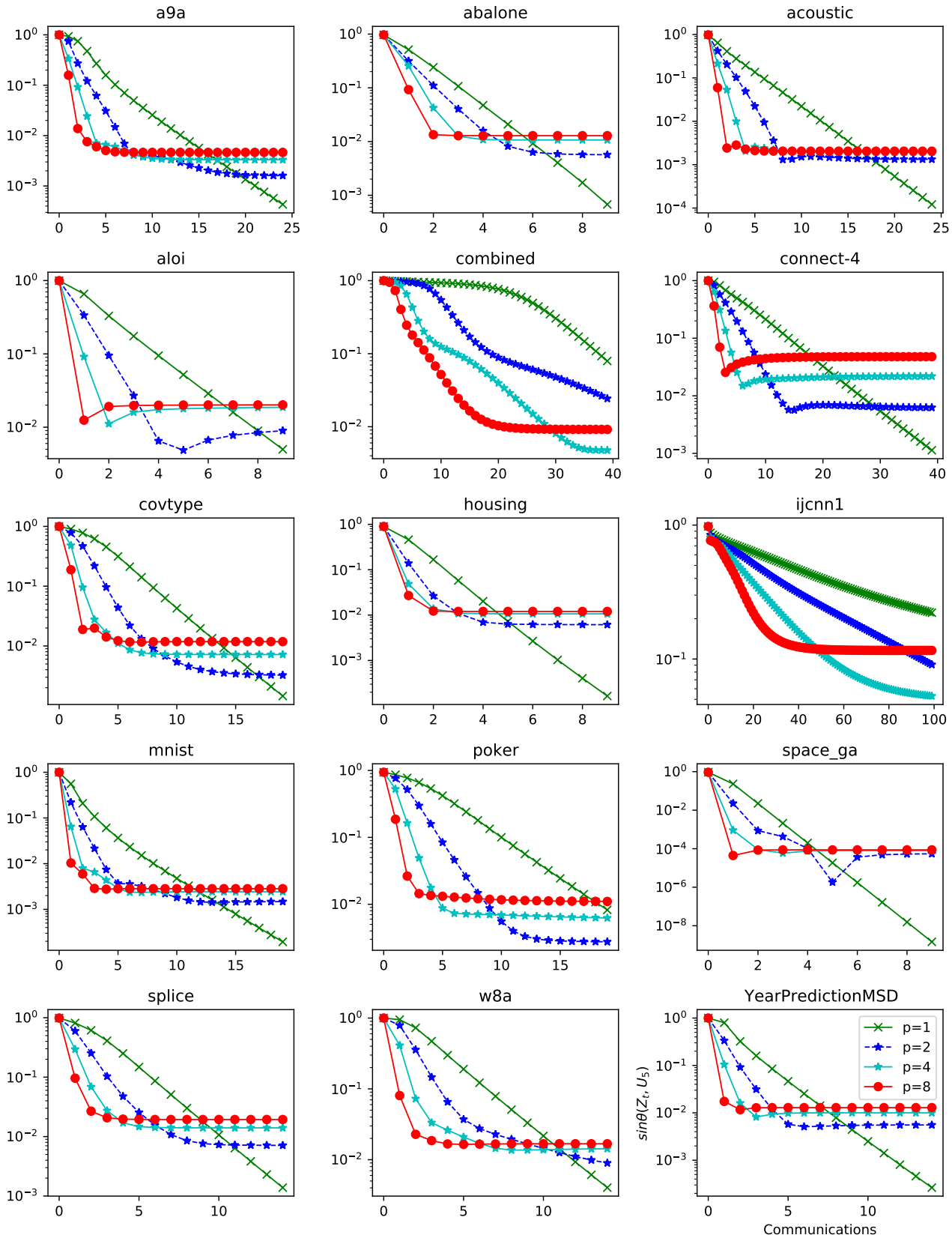

Figure 5. Vary p for `LocalPower` with OPT. Typically, the larger p, the larger error, which is consistent with our theory. Typically, `LocalPower` with OPT achieves the smallest error among our three proposed methods.

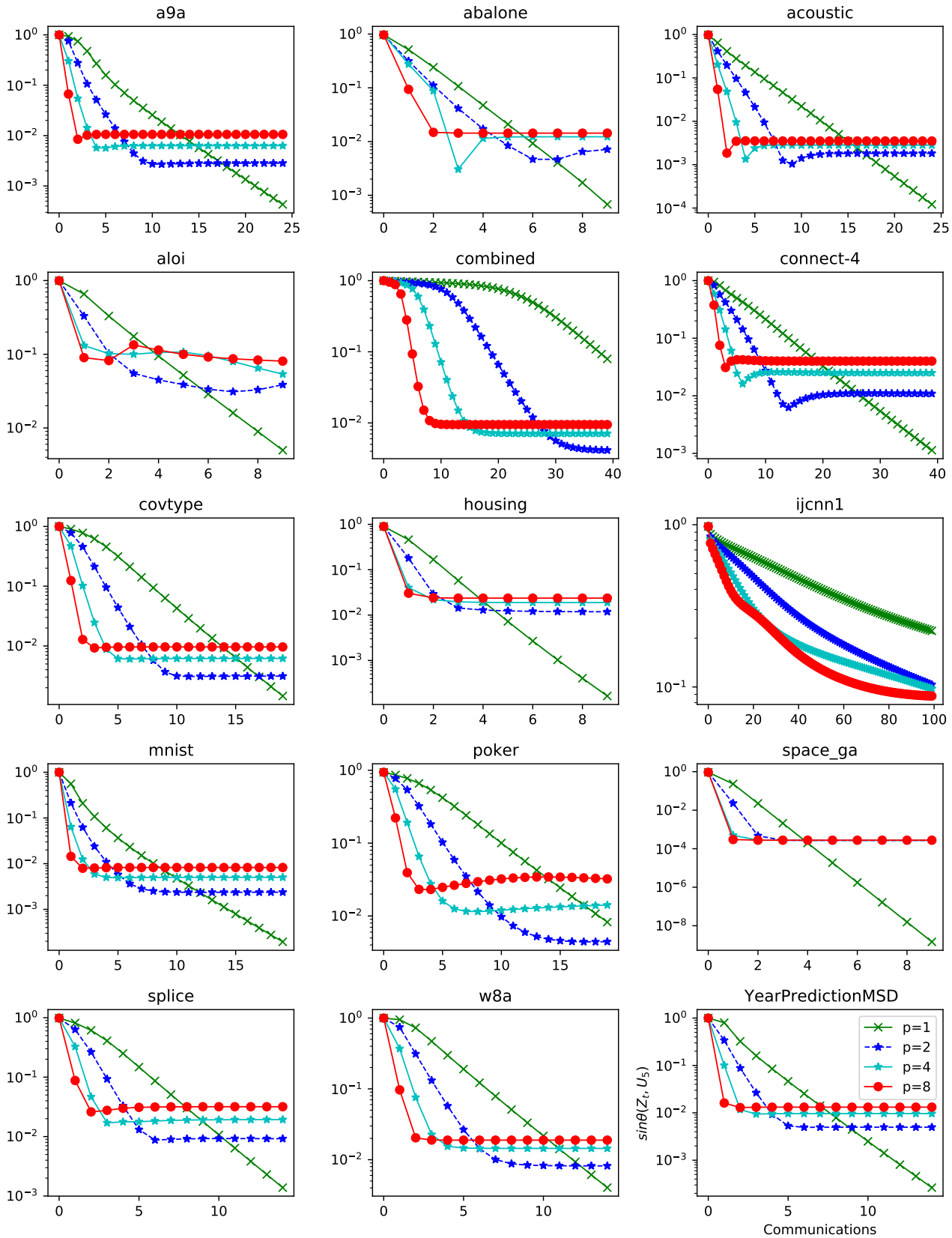

Figure 6. Vary p for `LocalPower` with sign-fixing Similar to Figure 5, the larger p, the larger error, which is consistent with our theory. `LocalPower` with sign-fixing is much computation efficient than that with OPT. Sign-fixing can be viewed as a good practical of surrogate of OPT.

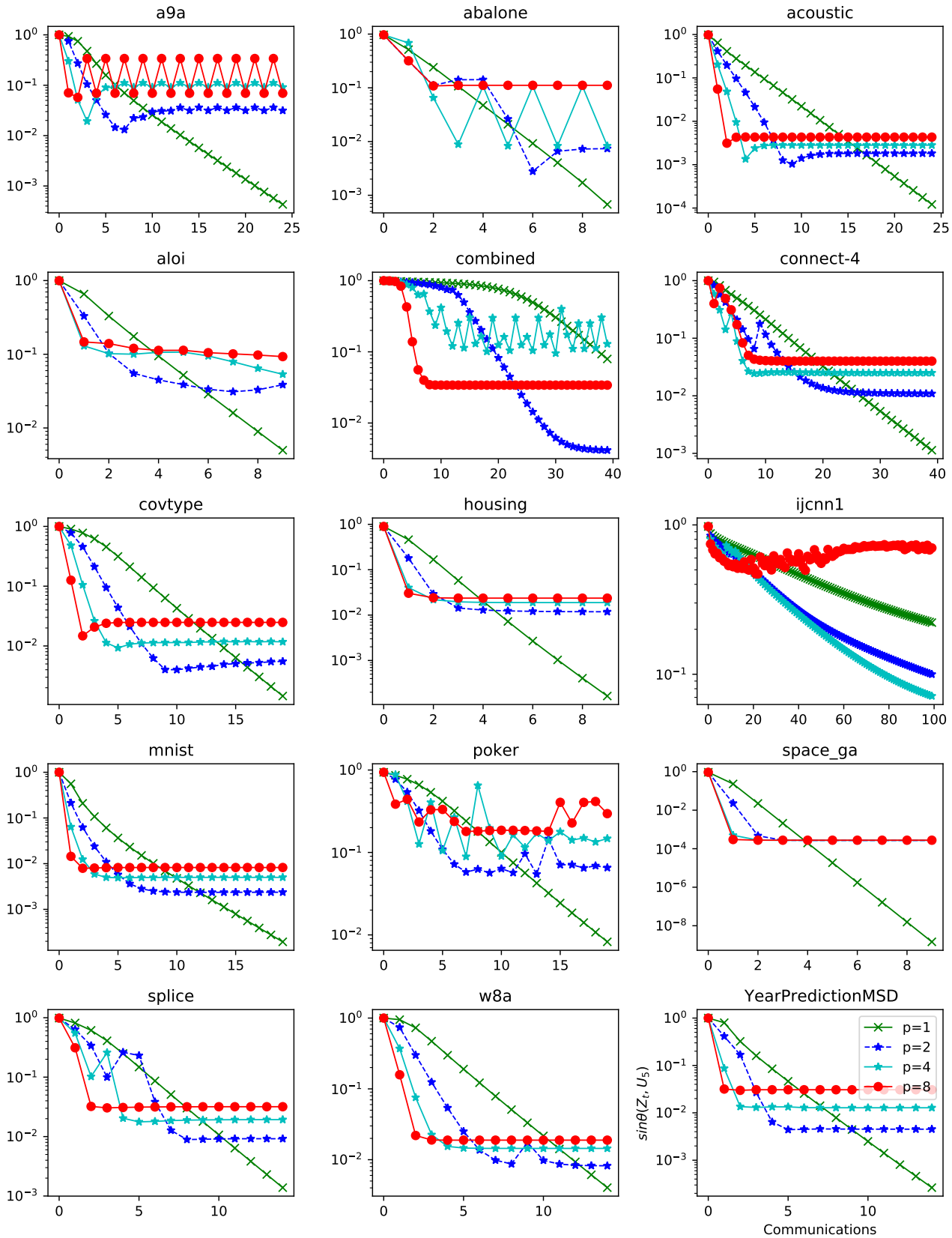

Figure 7. Vary p for vanilla `LocalPower` . For most datasets, vanilla `LocalPower` converges and the similar pattern that the larger p, the larger error occurs. However, for large p, it fluctuates and even diverges on some datasets (including A9a, Abalone, Combined, Ijcnn1 and Poker). This is because η can't meet required smallness. As argued, `LocalPower` with OPT or sign-fixing typically is more stable than the vanilla one, since it requires less strict smallness of η. Besides, we can use the decay strategy or decreases the number of devices.

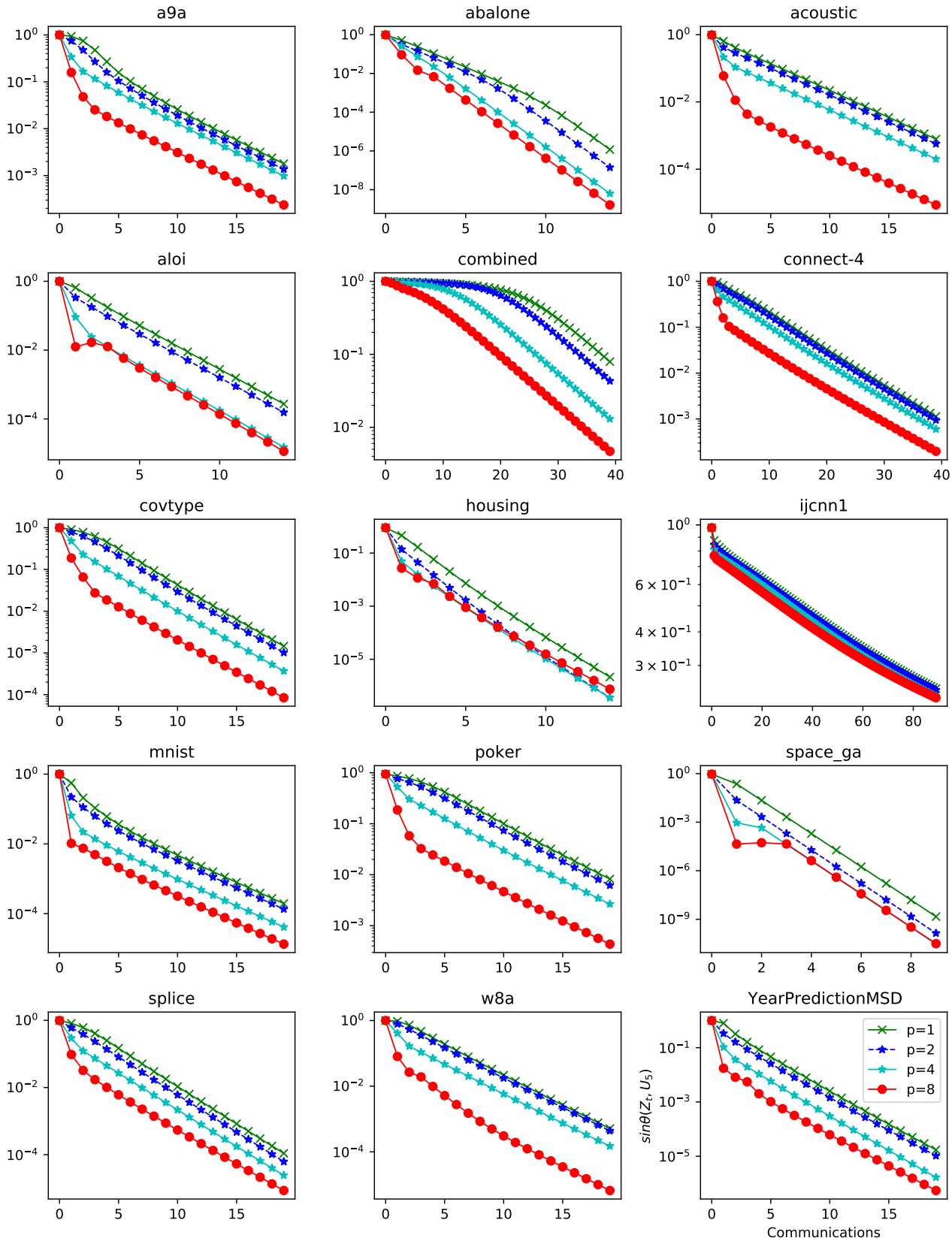

Figure 8. Decay strategy for `LocalPower` with OPT. For most datasets, `LocalPower` with OPT converges faster and achieves much less error than non-decay counterparts (see Figure 5). Theoretically, `LocalPower` with decay strategy can achieve zero error.

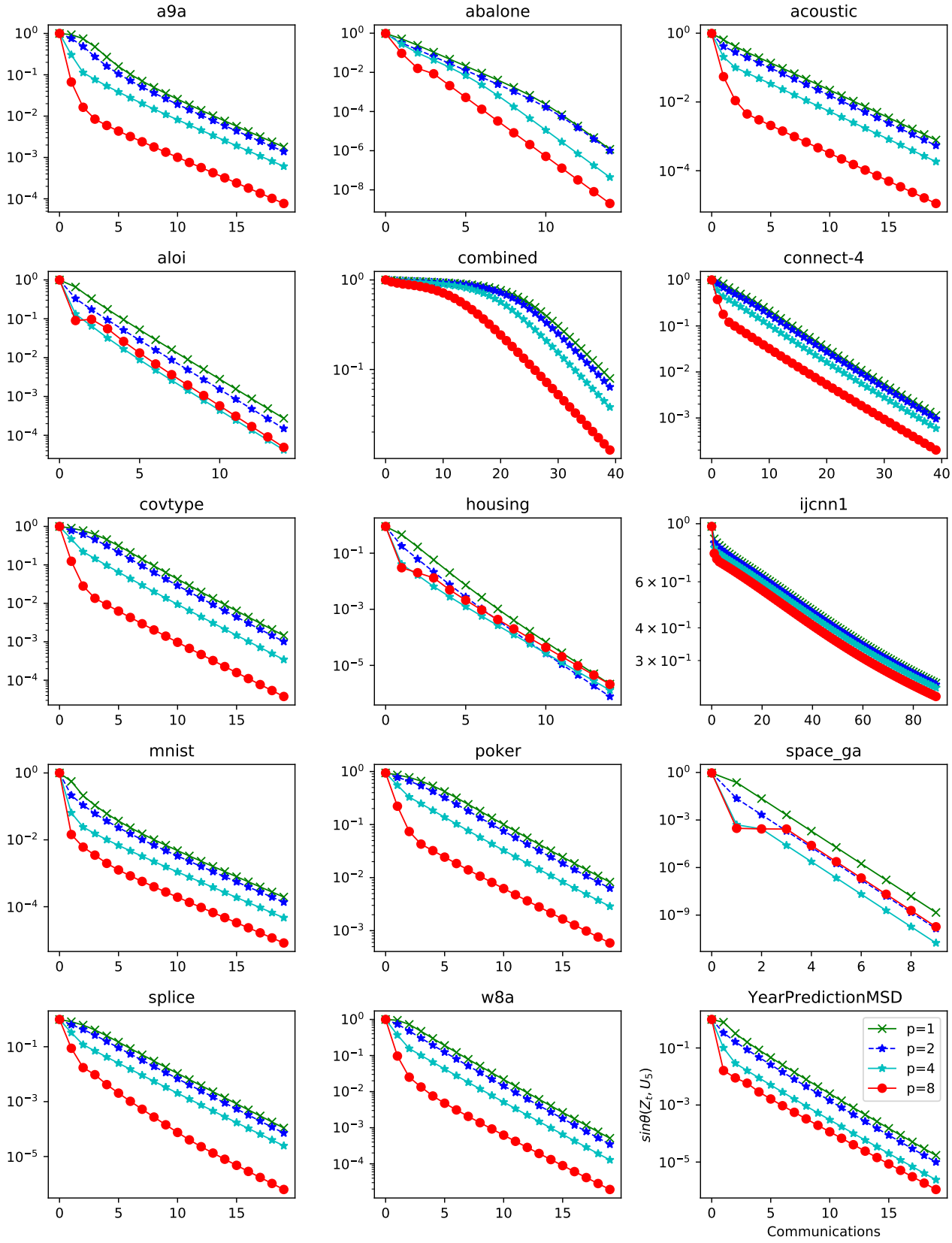

Figure 9. Decay strategy for `LocalPower` with sign-fixing. For most datasets, `LocalPower` with sign-fixing converges faster and achieves much less error than non-decay counterparts (see Figure 6). Theoretically, `LocalPower` with decay strategy can achieve zero error.

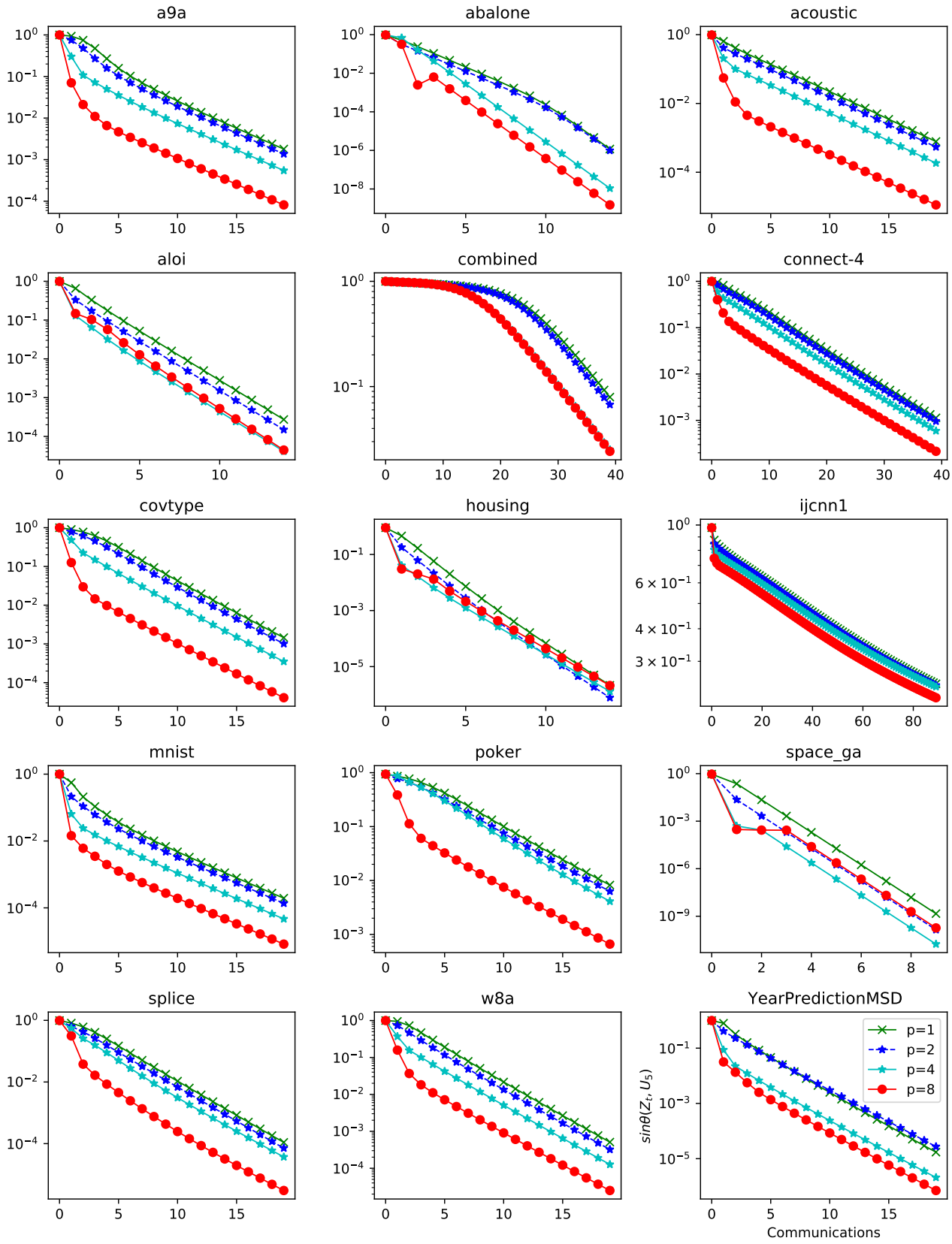

Figure 10. Decay strategy for vanilla `LocalPower` . For most datasets, vanilla `LocalPower` converges faster and more stable than non-decay counterparts (see Figure 6). It typically achieves much less error than non-decay counterparts. Theoretically, `LocalPower` with decay strategy can achieve zero error.

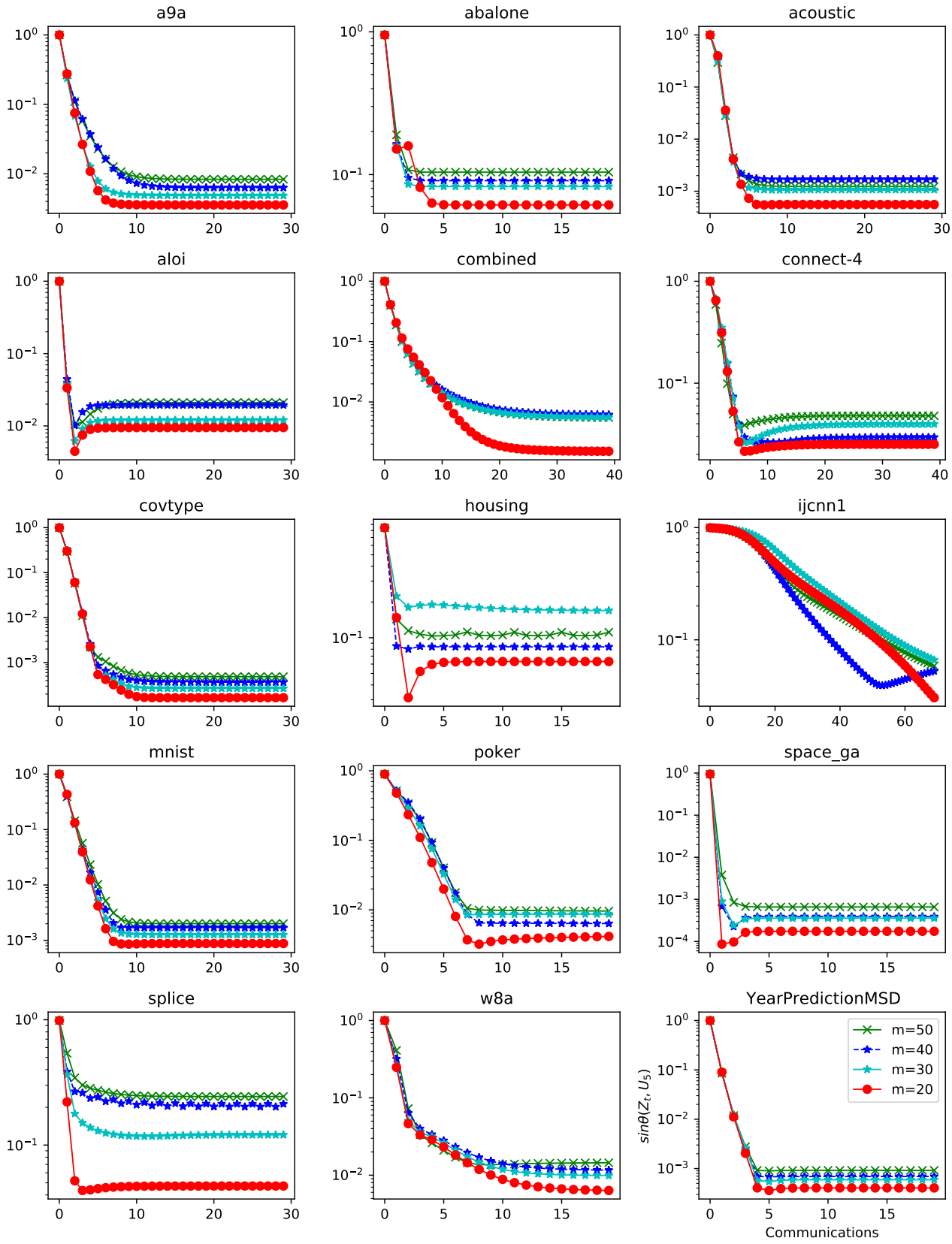

Figure 11. Various m for `LocalPower` with OPT. Typically, the smaller m has smaller errors.

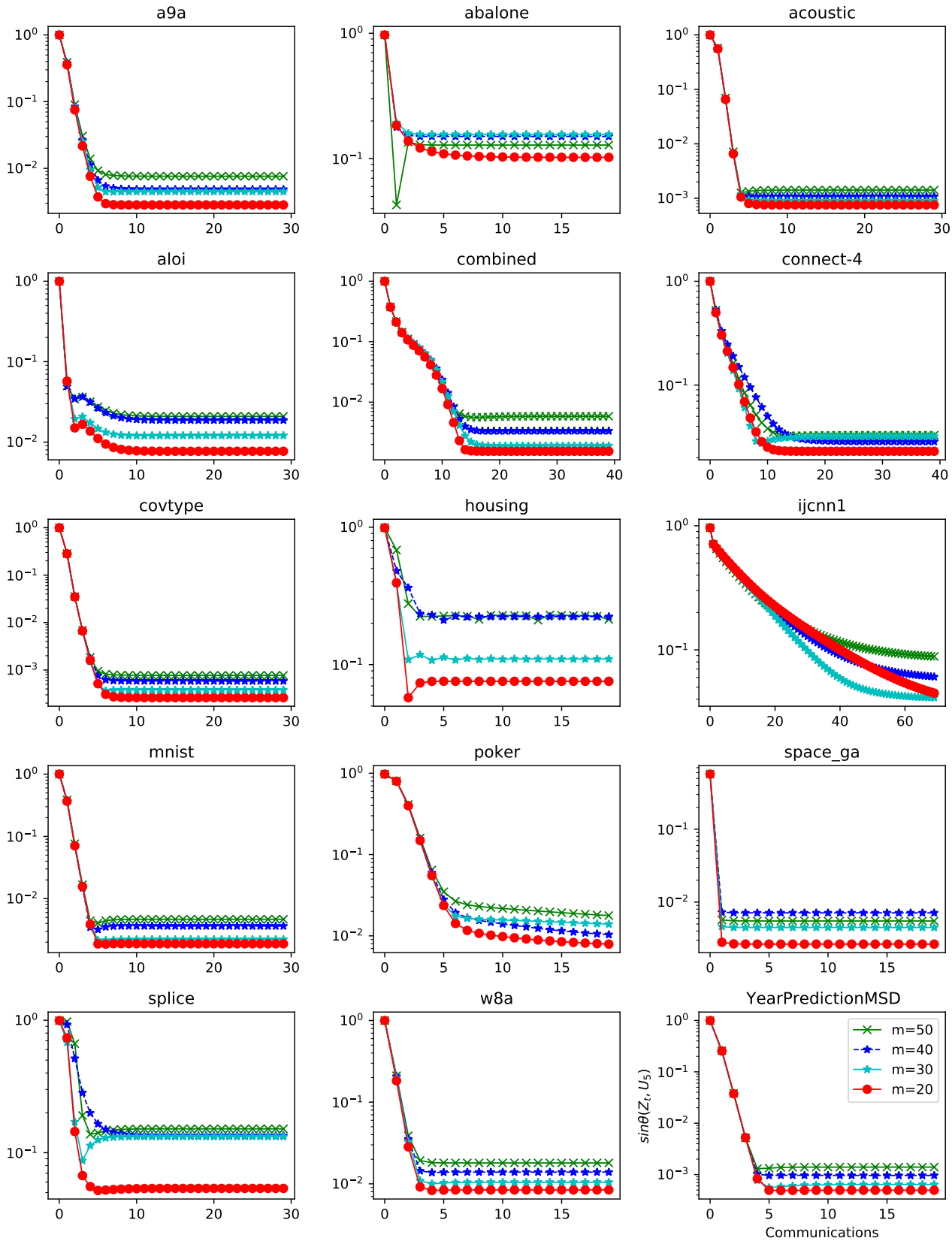

Figure 12. Various m for `LocalPower` with sign-fixing. Typically, the smaller m has smaller errors.

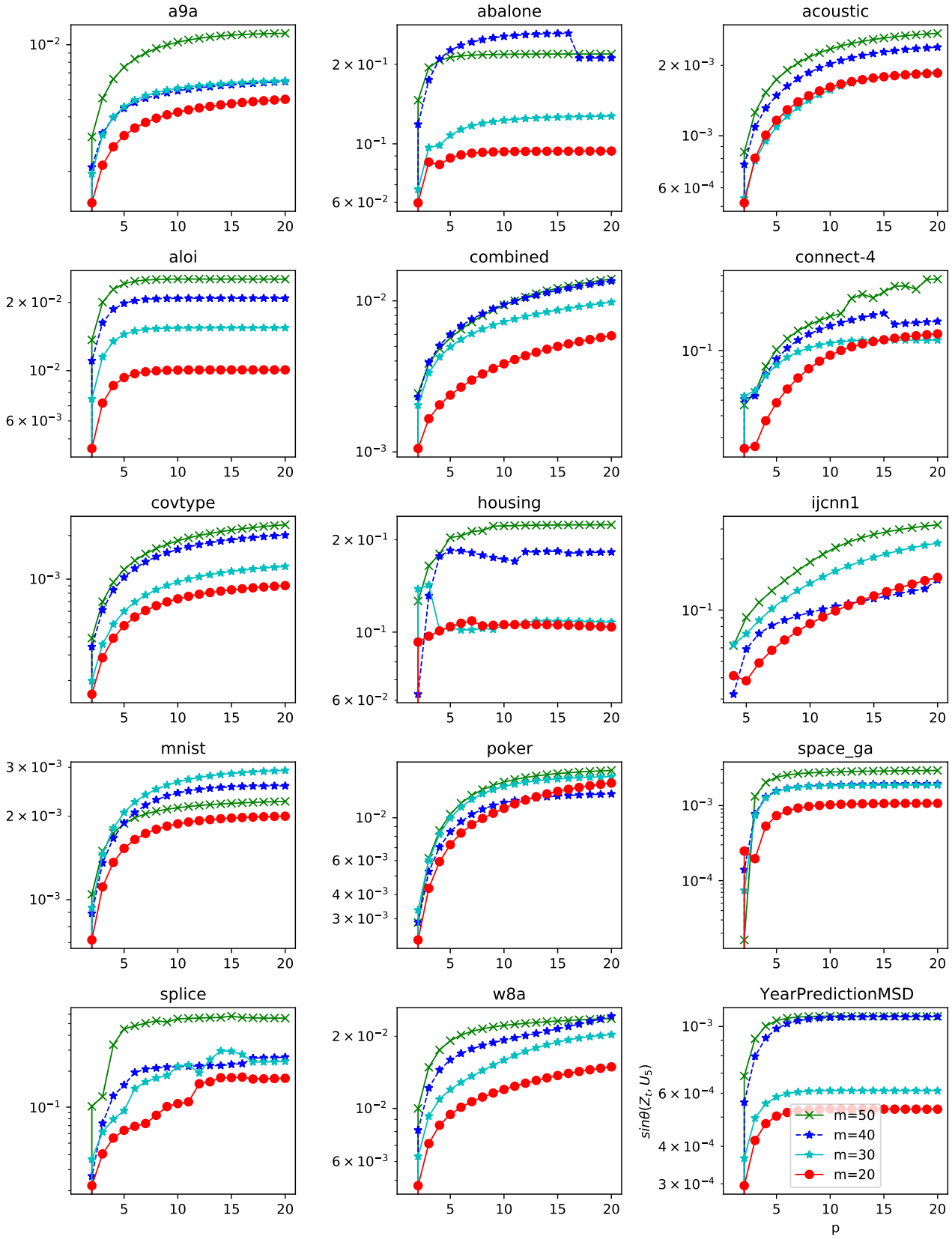

Figure 13. Error dependence of `LocalPower` with OPT.