A GEOMETRIC PROPERTIES OF THE PARAMETRIZATION

We start by calculating the vector field induced by the parameterization $G(\cdot)$.

$$\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) = 2u_{g(i)}v_i\mathbf{e}_{g(i)} + u_{g(i)}^2\mathbf{e}_{L+i},$$

where $\mathbf{e}_i \in \mathbb{R}^{L+p}$ is only 1 on i^{th} entry and 0 elsewhere, and

$$\nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top]) = 2v_i\mathbf{E}_{g(i),g(i)} + 2u_{g(i)}\mathbf{E}_{g(i),L+i} + 2u_{g(i)}\mathbf{E}_{L+i,g(i)},$$

where $\mathbf{E}_{i,j} \in \mathbb{R}^{(L+p)\times(L+p)}$ is the one-hot matrix for i^{th} row and j^{th} column. For $i \neq j$ s.t. $g(i) = g(j)$,

$$\begin{aligned}\nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_j([\mathbf{u}^\top, \mathbf{v}^\top]) &= (2v_i\mathbf{E}_{g(i),g(i)} + 2u_{g(i)}\mathbf{E}_{g(i),L+i} + 2u_{g(i)}\mathbf{E}_{L+i,g(i)}) \\ &\quad \cdot (2u_{g(j)}v_j\mathbf{e}_{g(j)} + u_{g(j)}^2\mathbf{e}_{L+j}) \\ &= 4u_{g(j)}v_iv_j\mathbf{e}_{g(i)} + 4u_{g(i)}u_{g(j)}v_j\mathbf{e}_{L+i} \\ &= 4u_{g(i)}v_iv_j\mathbf{e}_{g(i)} + 4u_{g(i)}^2v_j\mathbf{e}_{L+i},\end{aligned}$$

similarly,

$$\nabla^2 G_j([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) = 4u_{g(i)}v_iv_j\mathbf{e}_{g(i)} + 4u_{g(i)}^2v_i\mathbf{e}_{L+j}.$$

Proof for Lemma 1. For two indices within the same group, i.e, $i \neq j$ and $g(i) = g(j)$, we obtain that

$$\begin{aligned}[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) &= \nabla^2 G_j([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) - \nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_j([\mathbf{u}^\top, \mathbf{v}^\top]) \\ &= 4u_{g(i)}^2v_j\mathbf{e}_{L+i} - 4u_{g(i)}^2v_i\mathbf{e}_{L+j},\end{aligned}$$

which is not always $\mathbf{0}$ when $v_i \neq v_j$. Therefore, $G(\cdot)$ is not commuting. □

Proof for Theorem 1. For $i \neq j$ and $g(i) \neq g(j)$, we have

$$[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) = \mathbf{0}.$$

For $i \neq j$ and $g(i) = g(j)$, we have that

$$[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) = v_j\nabla G_i - v_i\nabla G_j \in \operatorname{span}\{\nabla G_i\}_{i=1}^p.$$

By Corollary 4.13 in (Li et al., 2022) and Lemma 1, we show that there exists and initialization and a time-dependent loss that the gradient flow can not be analyzed by mirror flow. □

Alternatively, we can show directly that the necessary condition in Theorem 4.10 in Li et al. (2022) is violated, i.e.,

$$\langle \nabla G_j, [\nabla G_i, [\nabla G_i, \nabla G_j]]\rangle([\mathbf{u}^\top, \mathbf{v}^\top]) \neq 0$$

for some $[\mathbf{u}^\top, \mathbf{v}^\top]$ in every open set M.

We first obtain that

$$\begin{aligned}\nabla[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) &= 8u_{g(i)}v_j\mathbf{E}_{L+i,g(i)} + 4u_{g(i)}^2\mathbf{E}_{L+i,L+j} \\ &\quad - 8u_{g(i)}v_i\mathbf{E}_{L+j,g(i)} - 4u_{g(i)}^2\mathbf{E}_{L+j,L+i}.\end{aligned}$$

Therefore,

$$\begin{aligned}[\nabla G_i, [\nabla G_i, \nabla G_j]]([\mathbf{u}^\top, \mathbf{v}^\top]) &= \nabla[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) \\ &\quad - \nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top])[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) \\ &= (8u_{g(i)}v_j\mathbf{E}_{L+i,g(i)} + 4u_{g(i)}^2\mathbf{E}_{L+i,L+j} \\ &\quad - 8u_{g(i)}v_i\mathbf{E}_{L+j,g(i)} - 4u_{g(i)}^2\mathbf{E}_{L+j,L+i}) \\ &\quad \cdot (2u_{g(i)}v_i\mathbf{e}_{g(i)} + u_{g(i)}^2\mathbf{e}_{L+i}) \\ &\quad - (2v_i\mathbf{E}_{g(i),g(i)} + 2u_{g(i)}\mathbf{E}_{g(i),L+i} + 2u_{g(i)}\mathbf{E}_{L+i,g(i)}) \\ &\quad \cdot (4u_{g(i)}^2v_j\mathbf{e}_{L+i} - 4u_{g(i)}^2v_i\mathbf{e}_{L+j}) \\ &= 16u_{g(i)}^2v_iv_j\mathbf{e}_{L+i} - 16u_{g(i)}^2v_i^2\mathbf{e}_{L+j} - 4u_{g(i)}^4\mathbf{e}_{L+j} - 8u_{g(i)}^3v_j\mathbf{e}_{g(i)} \\ &= 16u_{g(i)}^2v_iv_j\mathbf{e}_{L+i} - (16u_{g(i)}^2v_i^2 + 4u_{g(i)}^4)\mathbf{e}_{L+j} - 8u_{g(i)}^3v_j\mathbf{e}_{g(i)}.\end{aligned}$$

Hence,

$$\begin{aligned}
&\langle \nabla G_j, [\nabla G_i, [\nabla G_i, \nabla G_j]]\rangle([\mathbf{u}^\top, \mathbf{v}^\top]) \\
=&\langle 2u_{g(i)}v_j\mathbf{e}_{g(i)} + u_{g(i)}^2\mathbf{e}_{L+j}, 16u_{g(i)}^2 v_i v_j \mathbf{e}_{L+i} - (16u_{g(i)}^2 v_i^2 + 4u_{g(i)}^4)\mathbf{e}_{L+j} - 8u_{g(i)}^3 v_j \mathbf{e}_{g(i)}\rangle \\
=&-16u_{g(i)}^4 v_j^2 - 16u_{g(i)}^4 v_i^2 - 4u_{g(i)}^6 < 0.
\end{aligned}$$

By Theorem 4.10 in Li et al. (2022), there exists an initialization such that no Legendre function R is able to make the gradient flow be written as a mirror flow with respect to R.

B Proof for Analysis of Gradient Flow

Proof for Lemma 2. Recall

$$\frac{\partial \mathcal{L}}{\partial u_l} = -\frac{2}{n} u_l \mathbf{v}_l^\top \mathbf{X}_l^\top \mathbf{r}(t), \quad \frac{\partial \mathcal{L}}{\partial \mathbf{v}_l} = -\frac{1}{n} u_l^2 \mathbf{X}_l^\top \mathbf{r}(t).$$

Therefore, we obtain that

$$\begin{aligned}
\frac{\partial \left\|\mathbf{v}_l(t)\right\|^2}{\partial t} &= 2\mathbf{v}_l^\top(t)\frac{\partial \mathbf{v}_l(t)}{\partial t} = 2\mathbf{v}_l^\top(t)\left(-\frac{\partial \mathcal{L}}{\partial \mathbf{v}_l}\right) \\
&= \frac{2}{n}u_l^2 \mathbf{v}_l^\top(t)\mathbf{X}_l^\top \mathbf{r}(t) \\
&= u_l\left(-\frac{\partial \mathcal{L}}{\partial u_l}\right) = \frac{\partial \frac{1}{2}u_l^2(t)}{\partial t}.
\end{aligned}$$

□

Proof for Lemma 3. We start with decomposing $\mathbf{v}_l(0)$

$$\begin{aligned}
\mathbf{v}_l(0) = \eta\frac{1}{n}\mathbf{X}_l^\top \mathbf{y} &= \eta \mathbf{w}_l^\star + \eta\left(\frac{1}{n}\mathbf{X}_l^\top \mathbf{X} - \mathbf{I}\right)\mathbf{w}_l^\star + \eta\sum_{l'\neq l}\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_{l'}\mathbf{w}_{l'}^\star + \eta\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi} \\
&= \eta\mathbf{w}_l^\star + \eta\mathbf{b}_l.
\end{aligned}$$

With this decomposition, we have that

$$\begin{aligned}
\langle \mathbf{v}_l(0), \mathbf{v}_l^\star\rangle^2 &= \eta^2((u_l^\star)^2 + \langle \mathbf{b}_l, \mathbf{v}_l^\star\rangle)^2 \\
\left\|\mathbf{v}_l(0)\right\|_2^2 &= \eta^2((u_l^\star)^4 + 2\langle \mathbf{b}_l, \mathbf{w}_l^\star\rangle + \left\|\mathbf{b}_l\right\|_2^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\langle \mathbf{v}_l(0), \mathbf{v}_l^\star\rangle^2}{\left\|\mathbf{v}_l(0)\right\|_2^2} &= \frac{\eta^2((u_l^\star)^2 + \langle \mathbf{b}_l, \mathbf{v}_l^\star\rangle)^2}{\eta^2((u_l^\star)^4 + 2\langle \mathbf{b}_l, \mathbf{w}_l^\star\rangle + \left\|\mathbf{b}_l\right\|_2^2)} \\
&= 1 - \frac{\left\|\mathbf{b}_l\right\|_2^2 - \langle \mathbf{b}_l, \mathbf{v}_l^\star\rangle^2}{(u_l^\star)^4 + 2\langle \mathbf{b}_l, \mathbf{w}_l^\star\rangle + \left\|\mathbf{b}_l\right\|_2^2} \\
&= 1 - \frac{\left\|\mathbf{b}_l/(u_l^\star)^2\right\|_2^2 - \langle \mathbf{b}_l/(u_l^\star)^2, \mathbf{v}_l^\star\rangle^2}{1 + 2\langle \mathbf{b}_l/(u_l^\star)^2, \mathbf{v}_l^\star\rangle + \left\|\mathbf{b}_l/(u_l^\star)^2\right\|_2^2} \\
&= 1 - \frac{1 - \langle \mathbf{b}_l/\left\|\mathbf{b}_l\right\|, \mathbf{v}_l^\star\rangle^2}{1 + 2\left\|\mathbf{b}_l\right\|/(u_l^\star)^2\langle \mathbf{b}_l/\left\|\mathbf{b}_l\right\|, \mathbf{v}_l^\star\rangle + \left\|\mathbf{b}_l\right\|^2/(u_l^\star)^4}\left\|\mathbf{b}_l/(u_l^\star)^2\right\|^2 \\
&\geq 1 - \left\|\mathbf{b}_l/(u_l^\star)^2\right\|_2^2,
\end{aligned}$$

where last inequality is from

$$\begin{aligned}
\frac{1-\alpha^2}{\beta^2 + 2\alpha\beta + 1} = \frac{1}{\frac{\beta^2+2\alpha\beta+1}{1-\alpha^2}} &= \frac{1}{1 + \frac{\beta^2+2\alpha\beta+\alpha^2}{1-\alpha^2}} \\
&= \frac{1}{1 + \frac{(\alpha+\beta)^2}{1-\alpha^2}} \leq 1,
\end{aligned}$$

for $0 \leq \alpha \leq 1$.

Since

$$\|\mathbf{b}_l\|_2 \leq \delta_{in}(u_l^\star)^2 + L\delta_{out}(u_l^\star)^2 + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2,$$

we obtain that

$$\left\langle \frac{\mathbf{v}_l(0)}{\|\mathbf{v}_l(0)\|}, \mathbf{v}_l^\star \right\rangle \geq 1 - \left(\delta_{in} + L\delta_{out} + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2 /(u_l^\star)^2\right)^2.$$

□

Lemma 4. *Consider a simplified case where $\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_l = \mathbf{I}$, $\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_{l'} = \mathbf{O}, l \neq l'$, if $\mathbf{v}_l(0) = \eta\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$, then*

$$\mathbf{v}_l(t) = c\frac{1}{n}\mathbf{X}_l^\top \mathbf{y},$$

for some constant c.

Proof. From the gradient on the directions, we have that

$$\begin{aligned}\frac{\partial \mathbf{v}_l(t)}{\partial t} &= \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \mathbf{r}(t) = \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \mathbf{y} - \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \sum_{l'}\mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t) \\ &= \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \mathbf{y} - u_l^4(t)\mathbf{v}_l(t).\end{aligned}$$

Since $\mathbf{v}_l(0)$ is with the same direction as $\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$ at the initialization. Therefore, $\frac{\partial \mathbf{v}_l(t)}{\partial t}$ has the same direction as $\mathbf{v}_l(t)$. We obtain that $\mathbf{v}_l(t) = c\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$ for some constant c. □

Lemma 5. *If the gradient flow satisfies*

$$\frac{1}{2}\frac{\partial u^2(t)}{\partial t} \leq u^6(t) + \sqrt{2}u^4(t)B$$

for some constant $B > 0$, then for any $t \leq T = \frac{\log\frac{1}{\theta}}{2\theta^2 + \theta\sqrt{2}B}$ we have $u(t) \leq \sqrt{\theta}$ with initialization $u(0) = \theta$.

Proof. We wanted to find some time T such that when $t \leq T$, $u(t) \leq \sqrt{\theta}$. Since the gradient is bounded from above, we obtain that

$$\begin{aligned}\frac{1}{2}u^2(T) &\leq \frac{1}{2}\theta^2 \cdot \exp\left(\int_0^T 2u^4(t) + \sqrt{2}u^2(t)Bdt\right) \\ &\leq \frac{1}{2}\theta^2 \cdot \exp\left((2\theta^2 + \sqrt{2}\theta B)T\right) \leq \frac{1}{2}\theta.\end{aligned}$$

This gives us

$$T \leq \frac{\log\frac{1}{\theta}}{2\theta^2 + \theta\sqrt{2}B}.$$

□

Lemma 6. *Fix any $\tau < \frac{1}{2}$. Consider the gradient flow*

$$\frac{1}{2}\frac{\partial u^2(t)}{\partial t} \geq (1-2B)\sqrt{2}u^3(t)(u^\star)^2 - u^6(t) - \sqrt{2}u^3(t)B(u^\star)^2$$

for some constant $0 < B < \frac{1}{10}$ with initialization $u(0) = \theta < \frac{1}{2}u^\star$, we have that

$$\left|\frac{1}{\sqrt{2}}u^3(t) - (u^\star)^2\right| < (1 - 3B - \tau)(u^\star)^2,$$

after

$$t \geq T = \frac{2^{1/3}(u^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u^\star)^2\theta} + \frac{2\log_2\frac{1}{2\tau}}{3(u^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u^\star)^2\right)^{1/3}}.$$

Proof. For any $T \geq 0$, we have that

$$\frac{1}{2}u^2(T) \geq \frac{1}{2}\theta^2 \cdot \exp\left(\int_0^T (1-2B)2\sqrt{2}u(t)(u^\star)^2 - 2u^4(t) - 2\sqrt{2}u(t)B(u^\star)^2 dt\right).$$

When $u(t) < \frac{1}{2}u^\star$, we first aim to get T_1 such that $\frac{1}{\sqrt{2}}u^3(T_1) \geq \frac{1}{2}(u^\star)^2$. Therefore,

$$\begin{aligned}
&\frac{1}{2}\theta^2 \cdot \exp\left(\int_0^T (1-2B)2\sqrt{2}u(t)(u^\star)^2 - 2u^4(t) - 2\sqrt{2}u(t)B(u^\star)^2 dt\right) \\
&\geq \frac{1}{2}\theta^2 \cdot \exp\left(\left((1-2B)2\sqrt{2}(u^\star)^2 - \sqrt{2}(u^\star)^2 - 2\sqrt{2}B(u^\star)^2\right)\theta T_1\right) \\
&\geq \frac{1}{2}\left(\frac{\sqrt{2}}{2}(u^\star)^2\right)^{2/3}.
\end{aligned}$$

We obtain that

$$T \geq \frac{2^{1/3}(u^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u^\star)^2\theta}.$$

When $t \geq T_1$, we have that $\frac{1}{\sqrt{2}}u^3(t) \geq \frac{1}{2}(u^\star)^2$. Let us denote $\frac{1}{\sqrt{2}}u^3(0) = ((1-3B)-\eta)(u^\star)^2$, we wonder how many iterations T_d are needed to make $\frac{1}{\sqrt{2}}u^3(T_d) \geq \left((1-3B)-\frac{1}{2}\eta\right)(u^\star)^2$.

$$\begin{aligned}
&\frac{1}{2}\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{2/3} \cdot \exp\left(\int_0^T (1-2B)2\sqrt{2}u(t)(u^\star)^2 - 2u^4(t) - 2\sqrt{2}u(t)B(u^\star)^2 dt\right) \\
&\geq \frac{1}{2}\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{2/3} \cdot \exp\left(\left(\frac{1}{2}\eta(u^\star)^2\right)\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3} T_2\right) \\
&\geq \frac{1}{2}\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{2/3} \cdot \left(1 + \left(\frac{1}{2}\eta(u^\star)^2\right)\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3} T_2\right) \\
&\geq \frac{1}{2}\left(\sqrt{2}\left((1-3B)-\frac{1}{2}\eta\right)(u^\star)^2\right)^{2/3}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
T_2 &\geq \frac{\left((1-3B)-\frac{1}{2}\eta\right)^{2/3} - \left((1-3B)-\eta\right)^{2/3}}{\left((1-3B)-\eta\right)^{2/3}} \frac{1}{\frac{1}{2}\eta(u^\star)^2\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3}} \\
&\geq \frac{2}{3}\frac{\frac{1}{2}\eta}{\frac{1}{2}\eta(u^\star)^2\left((1-3B)-\eta\right)\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3}} \\
&\geq \frac{2}{3(u^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u^\star)^2\right)^{1/3}}.
\end{aligned}$$

Overall, we obtain that

$$\left|\frac{1}{\sqrt{2}}u^3(t) - (u^\star)^2\right| < (1-3B-\epsilon)(u^\star)^2,$$

after

$$t \geq T = T_1 + T_2 \log_2 \frac{1}{2\tau}.$$

□

Proof of Theorem 2. Denote $\zeta = 100\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty$. For $l \in S$, the gradient flow can be simplied as

$$\begin{aligned}
\frac{1}{2}\frac{\partial u_l^2(t)}{\partial t} &= \frac{2}{n}\mathbf{w}_l^\top(t)\mathbf{X}_l^\top\mathbf{r}(t) \\
&= 2\mathbf{w}_l^\top(t)(\mathbf{w}_l^\star - \mathbf{w}_l(t)) + \frac{2}{n}\mathbf{w}_l^\top\mathbf{X}_l^\top\boldsymbol{\xi} \\
&\geq 2u_l^2(t)(u_l^\star)^2\langle\mathbf{v}_l(t),\mathbf{v}_l^\star\rangle - 2u_l^4(t)\left\|\mathbf{v}_l(t)\right\|_2^2 - 2u_l^2(t)\left\|\mathbf{v}_l(t)\right\|_2\left\|\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right\|_2.
\end{aligned}$$

Since the initialization is balanced $\frac{1}{2}u_l^2(0) = \|\mathbf{v}_l(0)\|_2^2$, we know that from the balancing result Lemma 2,

$$\frac{1}{2}u_l^2(t) = \|\mathbf{v}_l(t)\|_2^2 .$$

Since the initialization of $\mathbf{v}_l(t)$ is aligned with direction $\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$, and with our assumption on orthogonal design, by Lemma 3 and Lemma 4, if $\left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2 \le B(u_l^\star)^2$, we can further simplify the gradient flow as

$$\begin{aligned}\frac{1}{2}\frac{\partial u_l^2(t)}{\partial t} &\ge \sqrt{2}(1-2B^2)u_l^3(t)(u_l^\star)^2 - u_l^6(t) - \sqrt{2}u_l^3(t)B \\ &\ge \sqrt{2}(1-2B)u_l^3(t)(u_l^\star)^2 - u_l^6(t) - \sqrt{2}u_l^3(t)B,\end{aligned}$$

where the last inequality holds when $B < 1$. We will verify that $B < 1$ holds in the following analysis.

If $\zeta \ge (u_{max}^\star)^2$, then our desired inequality is achieved at the initialization.

If $(u_{min}^\star)^2 \le \zeta \le (u_{max}^\star)^2$, for these group that $\zeta \le (u_l^\star)^2$, applying Lemma 6 with

$$B = \frac{\left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2}{(u_l^\star)^2} \le \frac{\left\|\frac{1}{n}\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty}{(u_l^\star)^2} \le \frac{1}{100}, \quad \tau = \frac{\epsilon}{(u_l^\star)^2}$$

we obtain the convergence on magnitudes

$$\left|\, \|\mathbf{w}_l(t)\|_2 - \|\mathbf{w}_l^\star\|_2 \,\right| \le (3B+\epsilon)\|\mathbf{w}_l^\star\|_2 ,$$

after

$$\frac{2^{1/3}(u_l^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u_l^\star)^2\theta} + \frac{2\log_2 \frac{(u_l)^2}{2\epsilon}}{3(u_l^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u_l^\star)^2\right)^{1/3}}.$$

If $\zeta \le (u_{min}^\star)^2$, similarly applying Lemma 6, the number of iterations needed for entries on the support to converge is

$$T_l = \frac{2^{1/3}(u_{max}^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u_{min}^\star)^2\theta} + \frac{2\log_2 \frac{(u_{max})^2}{2\epsilon}}{3(u_{min}^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u_{min}^\star)^2\right)^{1/3}}.$$

We now have that for $l \in S$,

$$\left|\, \|\mathbf{w}_l(t)\|_2 - \|\mathbf{w}_l^\star\|_2 \,\right| \le (3B+\epsilon)\|\mathbf{w}_l^\star\|_2 ,$$

where $B = \frac{\left\|\frac{1}{n}\mathbf{X}^\top \mathbf{y}\right\|_\infty}{(u_{min}^\star)^2} \le \frac{1}{100}, \forall l \in S$.

Recall that the direction is lower bounded by Lemma 3 and Lemma 8,

$$\left\langle \frac{\mathbf{w}_l(t)}{\|\mathbf{w}_l(t)\|_2}, \frac{\mathbf{w}_l^\star}{\|\mathbf{w}_l^\star\|_2} \right\rangle \ge 1 - B^2.$$

Therefore, the error bound on the support is as follows,

$$\begin{aligned}\|\mathbf{w}_l(t) - \mathbf{w}_l^\star\|_\infty \le \|\mathbf{w}_l(t) - \mathbf{w}_l^\star\|_2 &= \left\| \left(\|\mathbf{w}_l(t)\|_2 - (u_l^\star)^2\right)\frac{\mathbf{v}_l(t)}{\|\mathbf{v}_l(t)\|} + (u_l^\star)^2\left\langle \frac{\mathbf{v}_l(t)}{\|\mathbf{v}_l(t)\|}, \mathbf{v}_l^\star \right\rangle \right\|_2 \\ &\le (3B+\tau)(u_l^\star)^2 + (u_l^\star)^2\sqrt{2 - 2\left\langle \frac{\mathbf{v}_l(t)}{\|\mathbf{v}_l(t)\|}, \mathbf{v}_l^\star \right\rangle} \\ &= (3B+\tau)(u_l^\star)^2 + (u_l^\star)^2\sqrt{2}B \le \left\|\frac{1}{n}\mathbf{X}^\top \mathbf{y}\right\|_\infty + \epsilon.\end{aligned}$$

For $l \notin S$, we derive a lower bound on the growth rate

$$\begin{aligned}\frac{1}{2}\frac{\partial u_l^2(t)}{\partial t} &= \frac{2}{n}\mathbf{w}_l^\top(t)\mathbf{X}_l^\top \mathbf{r}(t) \\ &= 2\|\mathbf{w}_l(t)\|_2^2 + \frac{2}{n}\mathbf{w}_l^\top \mathbf{X}_l^\top \boldsymbol{\xi} \\ &\le u_l^6(t) + \sqrt{2}u_l^4(t)B.\end{aligned}$$

By applying Lemma 5 with $B = \left\|\frac{1}{n}\mathbf{X}^\top \mathbf{y}\right\|_\infty$, we obtain that before

$$T_u = \frac{\log \frac{1}{\theta}}{2\theta^2 + \theta\sqrt{2B}}.$$

Since $\theta < \frac{\epsilon}{2(u_{max})^2}$, $T_l < T_u$ is ensured.

□

C Analysis of gradient descent

C.1 Monotonic updates

Lemma 7. *With an initialization $u(0) < u^\star$ and step size $\gamma \leq \frac{1}{4(u^\star)^2}$, the updating sequence*

$$u(t) = u(t-1) + 2\gamma u(t-1)[(u^\star)^2 - u^2(t-1)],$$

is always bounded above by $u^\star$.

Proof. We prove it by contradiction. Assume there is a time t s.t.

$$u(t) \leq u^\star, u(t+1) > u^\star.$$

Therefore,

$$u(t) + 2\gamma u(t)[(u^\star)^2 - u^2(t)] > u^\star.$$

Denote $\lambda = u(t)/u^\star$, we have that

$$1 + 2\gamma(u^\star)^2(1-\lambda^2) - 1/\lambda > 0$$

for some $\lambda \in (0, 1]$.

Let $f(\lambda) = 1 + 2\gamma(u^\star)^2(1-\lambda^2) - 1/\lambda$, we obtain the derivative

$$f'(\lambda) = -4\gamma(u^\star)^2\lambda + \frac{1}{\lambda^2} > 0.$$

However, $f_{max}(\lambda) = f(1) = 0$, and $f(\lambda) \leq 0$ for all $\lambda \in (0, 1]$, which gives our desired contradiction. □

C.2 Updates with bounded perturbations

To study the general non-orthogonal and noisy case, we first extend the lemmas above to gradient dynamics with bounded perturbations.

Consider the update on $\mathbf{v}(t)$ with bounded perturbations

$$\begin{aligned} \mathbf{z}(t+1) &= \mathbf{v}(t) + \eta_t u^2(t)((u^\star)^2\mathbf{v}^\star - u^2(t)\mathbf{v}(t)) + \eta_t u^2(t)\mathbf{b}_t \\ \mathbf{v}(t+1) &= \frac{\mathbf{z}(t+1)}{\|\mathbf{z}(t+1)\|}. \end{aligned} \tag{4}$$

and the updates on $u(t)$

$$u(t+1) = u(t) + 2\gamma u(t)\mathbf{v}^\top(t+1)\{(u^\star)^2\mathbf{v}^\star - u^2(t)\mathbf{v}(t+1)\} + 2\gamma u(t)e_t, \tag{5}$$

Note that if we choose $\eta_t = \frac{1}{u^4(t)}$, Eq. (4) is recast as

$$\begin{aligned} \mathbf{z}(t+1) &= \frac{(u^\star)^2}{u^2(t)}\mathbf{v}^\star + \frac{1}{u^2(t)}\mathbf{b}_t \\ \mathbf{v}(t+1) &= \frac{\mathbf{z}(t+1)}{\|\mathbf{z}(t+1)\|}. \end{aligned} \tag{6}$$

Lemma 8. *Consider the update in Eq.* (6)*, if* $\|\mathbf{b}_t\| \leq B(u^\star)^2$ *for some constant* $0 < B < 1$*, we have that*

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq 1 - B^2.$$

Proof. We have that

$$\begin{aligned}
\langle \mathbf{z}(t+1), \mathbf{v}^\star \rangle &= \frac{(u^\star)^2}{u^2(t)} + \frac{1}{u_l^2(t)} \langle \mathbf{b}_t, \mathbf{v}^\star \rangle \\
\|\mathbf{z}(t+1)\|^2 &= \frac{(u^\star)^4}{u^4(t)} + 2\frac{(u^\star)^2}{u^4(t)} \langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \frac{1}{u_l^4(t)} \|\mathbf{b}_t\|^2 ,
\end{aligned}$$

therefore,

$$\begin{aligned}
\frac{\langle \mathbf{z}(t+1), \mathbf{v}^\star \rangle^2}{\|\mathbf{z}(t+1)\|^2} &= \frac{\frac{(u^\star)^4}{u^4(t)} + 2\frac{(u^\star)^2}{u^4(t)} \langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \frac{1}{u_l^4(t)} \langle \mathbf{b}_t, \mathbf{v}^\star \rangle^2}{\frac{(u^\star)^4}{u^4(t)} + 2\frac{(u^\star)^2}{u^4(t)} \langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \frac{1}{u_l^4(t)} \|\mathbf{b}_t\|^2} \\
&= 1 - \frac{\|\mathbf{b}_t\|^2 - \langle \mathbf{b}_t, \mathbf{v}^\star \rangle^2}{(u^\star)^4 + 2(u^\star)^2 \langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \|\mathbf{b}_t\|^2} \\
&= 1 - \frac{\left\|\mathbf{b}_t/(u^\star)^2\right\|^2 - \langle \mathbf{b}_t/(u^\star)^2, \mathbf{v}^\star \rangle^2}{1 + 2\langle \mathbf{b}_t/(u^\star)^2, \mathbf{v}^\star \rangle + \|\mathbf{b}_t/(u^\star)^2\|^2} \\
&= 1 - \frac{1 - \langle \mathbf{b}_t/\|\mathbf{b}_t\|, \mathbf{v}^\star \rangle^2}{1 + 2\|\mathbf{b}_t\|/(u^\star)^2 \langle \mathbf{b}_t/\|\mathbf{b}_t\|, \mathbf{v}^\star \rangle + \|\mathbf{b}_t\|^2/(u^\star)^4} \left\|\mathbf{b}_t/(u^\star)^2\right\|^2 \\
&\geq 1 - \left\|\mathbf{b}_t/(u^\star)^2\right\|^2 \\
&\geq 1 - B^2.
\end{aligned}$$

Hence, we have that

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq \sqrt{1 - B^2} \geq 1 - B^2.$$

□

Lemma 9. *Consider the updates in Eq.* (5) *with* $|e_t| \leq B$*, if* $u^2(0) \leq (u^\star)^2$*, then* $u^2(t) \leq (u^\star)^2 + B$ *for all* t*. If* $u^2(0) \geq (u^\star)^2$ *and* $|\langle \mathbf{v}(t), \mathbf{b}_t \rangle| \leq B_2 \tau (u^\star)^2$*, then* $u^2(t) \geq (1 - B_2)(u^\star)^2 - B$ *for all* t*.*

Proof. Proof by contradiction similarly to Lemma 7. □

Lemma 10. *Fix the step size* γ *for the update on* $u(t)$*, and choose* $u(0) = \alpha \leq \frac{1}{5}u^\star$*. Consider the updates in Eq.* (5) *and Eq.* (4) *with* $|\langle \mathbf{v}(t), \mathbf{b}_t \rangle| \leq \frac{1}{20}(u^\star)^2$ *and* $|e_t| \leq \frac{1}{20}(u^\star)^2$*, then* $T \geq \frac{\log \frac{(u^\star)^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(u^\star)^2)}$*, we have that* $u^2(T) \geq \frac{1}{2}(u^\star)^2$*.*

Proof. Apply Lemma 8 with $B = \frac{1}{20}$,

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq 1 - B^2 = 1 - \frac{1}{400} \geq \frac{4}{5}$$

Starting from $t = 1$, we have that

$$\mathbf{v}^\top(t)\{(u^\star)^2 \mathbf{v}^\star - u^2(t)\mathbf{v}(t)\} \geq \frac{4}{5}(u^\star)^2 - u^2(t),$$

therefore, we obtain an lower bound of the growth rate on $u(t)$, which reads

$$\begin{aligned}
u(t+1) &\geq u(t) + 2\gamma u(t) \left(\frac{4}{5}(u^\star)^2 - u^2(t) - \frac{1}{20}(u^\star)^2 \right) \\
&= u(t) \left(1 + 2\gamma \left(\frac{3}{4}(u^\star)^2 - u^2(t) \right) \right) \\
&\geq u(t) \left(1 + \gamma \frac{1}{2}(u^\star)^2 \right).
\end{aligned}$$

Therefore, the requirement on the number of iterations is recast as

$$\begin{aligned}
&\alpha^2\left(1+\gamma\frac{1}{2}(u^\star)^2\right)^{2T} \geq \frac{1}{2}(u^\star)^2 \\
\Longleftrightarrow &2T \geq \frac{\log\frac{(u^\star)^2}{2\alpha^2}}{\log(1+\gamma\frac{1}{2}(u^\star)^2)} \\
\Longleftrightarrow &T \geq \frac{\log\frac{(u^\star)^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(u^\star)^2)}.
\end{aligned}$$

With these requirements, by Lemma 9, we also have that $u^2(t) \leq \frac{3}{2}(u^\star)^2, \forall t \geq 0$. □

Lemma 11. *Fix the step size γ for the update on $u(t)$, and choose the initialization $u(0)$ such that $|(u^\star)^2 - u^2(0)| \leq \tau(u^\star)^2$ where $0 < \tau \leq 1/2$. Consider the updates in Eq. (5) and Eq. (4) with $|\langle \mathbf{v}(t), \mathbf{b}_t\rangle| \leq \frac{1}{10}\tau(u^\star)^2$ and $|e_t| \leq \frac{1}{10}\tau(u^\star)^2$, then after $T \geq \frac{5}{2\gamma(u^\star)^2}$, we have that $\langle \mathbf{v}(t), \mathbf{v}^\star\rangle \geq 1-\frac{1}{5}\tau^2$ for all $t \leq T$ and $|u^2(T) - (u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$.*

Proof. When $u^2(0) \leq (u^\star)^2$, by applying to Lemma 8, we have that

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star\rangle \geq 1-\left(\frac{1}{10}\tau\right)^2 \geq 1-\frac{1}{5}\tau^2,$$

therefore,

$$\begin{aligned}
u(t+1) &\geq u(t) + 2\gamma u(t)\left(\left(1-\frac{1}{5}\tau\right)(u^\star)^2 - u^2(t) - \frac{1}{10}\tau(u^\star)^2\right) \\
&= u(t)\left(1+2\gamma\left(\left(1-\frac{3}{10}\tau\right)(u^\star)^2 - u^2(t)\right)\right).
\end{aligned}$$

Further, we want to find an lower bound requirement on T s.t.

$$\left((u^\star)^2-\tau(u^\star)^2\right)\left(1+2\gamma\left(\left(1-\frac{3}{10}\tau\right)(u^\star)^2 - \left((u^\star)^2-\frac{1}{2}\tau\right)(u^\star)^2\right)\right)^{2T} \geq (u^\star)^2 - \frac{1}{2}\tau(u^\star)^2,$$

which can be relaxed as

$$\begin{aligned}
&\left((u^\star)^2-\tau(u^\star)^2\right)\left(1+\frac{2}{5}\gamma T\tau(u^\star)^2\right) \geq (u^\star)^2 - \frac{1}{2}\tau(u^\star)^2 \\
\Longleftrightarrow &1+\frac{2}{5}\gamma T\tau(u^\star)^2 \geq \frac{(u^\star)^2-\frac{1}{2}\tau(u^\star)^2}{(u^\star)^2-\tau(u^\star)^2} \\
\Longleftrightarrow &\frac{2}{5}\gamma T\tau(u^\star)^2 \geq \frac{\frac{1}{2}\tau(u^\star)^2}{((u^\star)^2-\tau(u^\star)^2)} \\
\Longleftrightarrow &T \geq \frac{5}{4\gamma(u^\star)^2(1-\tau)} \\
\Longrightarrow &T \geq \frac{5}{2\gamma(u^\star)^2}.
\end{aligned}$$

When $u^2(0) > (u^\star)^2$, we have that

$$\begin{aligned} u(t+1) &\leq u(t) + 2\gamma u(t)\left((u^\star)^2 - u^2(t) + \frac{1}{10}\tau(u^\star)^2\right) \\ &= u(t)\left(1 + 2\gamma\left(\left(1+\frac{1}{10}\tau\right)(u^\star)^2 - u^2(t)\right)\right). \\ &\leq u(t)\left(1 - \frac{4}{5}\gamma\tau(u^\star)^2\right). \end{aligned}$$

Similarly, we want to get

$$\begin{aligned} &(u^\star)^2 + \frac{1}{2}\tau(u^\star)^2 \geq \left((u^\star)^2 + \tau(u^\star)^2\right)\left(1 - \frac{4}{5}\gamma T\tau(u^\star)^2\right) \\ \Longleftrightarrow &\frac{(u^\star)^2 + \frac{1}{2}\tau(u^\star)^2}{(u^\star)^2 + \tau(u^\star)^2} \geq 1 - \frac{4}{5}\gamma T\tau(u^\star)^2 \\ \Longleftrightarrow &\frac{4}{5}\gamma T\tau(u^\star)^2 \geq \frac{\frac{1}{2}\tau(u^\star)^2}{(u^\star)^2 + \tau(u^\star)^2} \\ \Longleftrightarrow &T \geq \frac{5}{8\gamma(u^\star)^2(1+\tau)} \\ \Longrightarrow &T \geq \frac{5}{8\gamma(u^\star)^2}. \end{aligned}$$

If $u(0) <= u^\star$ and $u(t) > u^\star, t < T$, or $u(0) > u^\star$ and $u(t) \leq u^\star, t < T$, we have already have $|u^2(t) - u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$. By Lemma 9, $|u^2(T) - u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$ remains to hold.

Hence, after $T \geq \frac{5}{2\gamma(u^\star)^2}$, we have $|u^2(T) - u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$. □

C.3 Analysis of perturbations

We decompose the updates into several terms for later investigation.

The gradient of $\mathcal{L}(\cdot)$ on each $\mathbf{v}_l$ is

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{v}_l} &= -\frac{1}{n}u_l^2\mathbf{X}_l^\top\left(\mathbf{y} - \sum_{l'\neq l} u_{l'}^2\mathbf{X}_{l'}\mathbf{v}_{l'}\right) + \frac{1}{n}u_l^4\mathbf{X}_l^\top\mathbf{X}_l\mathbf{v}_l \\ &= -\frac{1}{n}u_l^2\mathbf{X}_l^\top\left(\mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2\mathbf{X}_{l'}\mathbf{v}_{l'}\right) \end{aligned}$$

When $l \in S$, the gradient update on each $\mathbf{v}_l$ is

$$\begin{aligned} \mathbf{z}_l(t+1) &= \mathbf{v}_l(t) + \eta_{l,t}u_l^2(t)\frac{1}{n}\mathbf{X}_l^\top\left(\mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2(t)\mathbf{X}_{l'}\mathbf{v}_{l'}(t)\right) \\ &= \mathbf{v}_l(t) + \eta_{l,t}u_l^2(t)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)) \\ &+ \eta_{l,t}u_l^2(t)\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l - \mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)) \\ &+ \eta_{l,t}u_l^2(t)\sum_{l'\neq l, l'\in S}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t)) \\ &- \eta_{l,t}u_l^2(t)\sum_{l'\in S^c}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t) \\ &+ \eta_{l,t}u_l^2(t)\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}. \end{aligned}$$

The gradient of $\mathcal{L}(\cdot)$ on each u_l is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial u_l} &= -\frac{2}{n}u_l \left\langle \mathbf{X}_l \mathbf{v}_l, \mathbf{y} - \sum_{l' \neq l} u_{l'}^2 \mathbf{X}_{l'} \mathbf{v}_{l'} \right\rangle + \frac{2}{n} u_l^3 \left\| \mathbf{X}_l \mathbf{v}_l \right\|^2 \\ &= -\frac{2}{n}u_l \left\langle \mathbf{X}_l \mathbf{v}_l, \mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2 \mathbf{X}_{l'} \mathbf{v}_{l'} \right\rangle\end{aligned}$$

When $l \in S$, the gradient update on u_l reads

$$\begin{aligned}u_l(t+1) &= u_l(t) + \gamma \frac{2}{n} u_l(t) \left\langle \mathbf{X}_l \mathbf{v}_l(t+1), \mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2(t) \mathbf{X}_{l'} \mathbf{v}_{l'}(t+1) \right\rangle \\ &= u_l(t) + 2\gamma u_l(t) \mathbf{v}_l^\top(t+1)((u_l^\star)^2 \mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t+1)) \\ &+ 2\gamma u_l(t) \mathbf{v}_l^\top(t+1) \left(\frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_l - \mathbf{I} \right) ((u_l^\star)^2 \mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t+1)) \\ &+ 2\gamma u_l(t) \mathbf{v}_l^\top(t+1) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \neq l, l' \in S} \mathbf{X}_{l'} ((u_{l'}^\star)^2 \mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t+1)) \\ &- 2\gamma u_l(t) \mathbf{v}_l^\top(t+1) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \in S^c} \mathbf{X}_{l'} u_{l'}^2(t) \mathbf{v}_{l'}(t+1) \\ &+ 2\gamma u_l(t) \frac{1}{n} \mathbf{v}_l^\top(t+1) \mathbf{X}_l^\top \boldsymbol{\xi}.\end{aligned}$$

We now rewrite the definition of bounded perturbation in Eq. (4, 5), where the bounded perturbation $e_{l,t}$ on updates of $u_l(t)$ reads

$$\begin{aligned}e_{l,t} &= \mathbf{v}_l^\top(t+1) \left(\frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_l - \mathbf{I} \right) ((u_l^\star)^2 \mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t+1)) \\ &+ \mathbf{v}_l^\top(t+1) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \neq l, l' \in S} \mathbf{X}_{l'} ((u_{l'}^\star)^2 \mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t+1)) \\ &- \mathbf{v}_l^\top(t+1) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \in S^c} \mathbf{X}_{l'} u_{l'}^2(t) \mathbf{v}_{l'}(t+1) \\ &+ \frac{1}{n} \mathbf{v}_l^\top(t+1) \mathbf{X}_l^\top \boldsymbol{\xi},\end{aligned}$$

and the bounded perturbation $\mathbf{b}_{l,t}$ on updates of $\mathbf{v}_l(t)$ reads

$$\begin{aligned}\mathbf{b}_{l,t} &= \left(\frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_l - \mathbf{I} \right) ((u_l^\star)^2 \mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)) \\ &+ \sum_{l' \neq l, l' \in S} \frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_{l'} ((u_{l'}^\star)^2 \mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t)) \\ &- \sum_{l' \in S^c} \frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_{l'} u_{l'}^2(t) \mathbf{v}_{l'}(t) \\ &+ \frac{1}{n} \mathbf{X}_l^\top \boldsymbol{\xi}.\end{aligned}$$

We show in Lemma 11 that when the perturbations are bounded, the direction is roughly accurate ($\langle \mathbf{v}_l(t), \mathbf{v}^\star \rangle$ is large) and $u_l(t)$ converges exponentially. Now we show below that when the direction is roughly accurate and $u_l(t)$ is close to $u_l^\star$, the perturbations are bounded.

Lemma 12. *Assume* $\delta_{in} \leq \frac{(u^\star_{min})^2}{120(u^\star_{max})^2}$ *and* $\delta_{out} \leq \frac{(u^\star_{min})^2}{120s(u^\star_{max})^2}$, $\alpha < \frac{1}{2}\sqrt{\frac{\tau_0}{L}}u^\star_l$, $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \leq \frac{1}{80}\tau_0(u^\star_l)^2$ *and* $|(u^\star_l)^2 - u^2_l(0)| \leq \tau(u^\star_l)^2$ *for each* $l \in [L]$ *where* $0 < \tau_0 \leq \tau \leq 1/2$. *If* $\langle \mathbf{v}_l(t), \mathbf{v}^\star_l\rangle \geq 1 - \frac{1}{5}\tau^2$, *then* $|\langle \mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \leq \frac{1}{10}\tau(u^\star_l)^2$ *and* $|e_{l,t}| \leq \frac{1}{10}\tau(u^\star_l)^2$.

Proof. We first verify

$$\begin{aligned}
\left\|(u^\star_l)^2\mathbf{v}^\star_l - u^2_l(t)\mathbf{v}_l(t)\right\| &= \left\|\{(u^\star_l)^2 - u^2_l(t)\}\mathbf{v}^\star_l - u^2_l(t)\{\mathbf{v}_l(t) - \mathbf{v}^\star_l\}\right\| \\
&\leq |(u^\star_l)^2 - u^2_l(t)| + u^2_l(t)\left\|\mathbf{v}_l(t) - \mathbf{v}^\star_l\right\| \\
&\leq \tau(u^\star_l)^2 + u^2_l(t)\sqrt{2 - 2\langle \mathbf{v}_l(t), \mathbf{v}^\star_l\rangle} \\
&\leq \tau(u^\star_l)^2 + \frac{3}{2}(u^\star_l)^2\frac{\sqrt{2}}{\sqrt{5}}\tau \qquad (7) \\
&\leq 3\tau(u^\star_l)^2.
\end{aligned}$$

By Assumption 1, we have that

$$\left|\mathbf{v}^\top_l(t)\left(\frac{1}{n}\mathbf{X}^\top_l\mathbf{X}_l - \mathbf{I}\right)((u^\star_l)^2\mathbf{v}^\star_l - u^2_l(t)\mathbf{v}_l(t)) + \mathbf{v}^\top_l(t)\sum_{l'\neq l, l'\in S}\frac{1}{n}\mathbf{X}^\top_l\mathbf{X}_{l'}((u^\star_{l'})^2\mathbf{v}^\star_{l'} - u^2_{l'}(t)\mathbf{v}_{l'}(t))\right|$$

$$\leq 3\delta_{in}\tau(u^\star_{max})^2 + 3s\delta_{out}\tau(u^\star_{max})^2 \leq \frac{1}{40}\tau(u^\star_l)^2 + \frac{1}{40}\tau(u^\star_l)^2 = \frac{1}{20}\tau(u^\star_l)^2.$$

For the other two terms, we have that

$$\left|\mathbf{v}^\top_l(t)\sum_{l'\in S^c}\frac{1}{n}\mathbf{X}^\top_l\mathbf{X}_{l'}u^2_{l'}(t)\mathbf{v}_{l'}(t)\right| \leq \delta(L-s)\alpha^2 \leq \frac{1}{80}\tau(u^\star_l)^2,$$

and

$$\begin{aligned}
\left|\mathbf{v}^\top_l(t)\frac{1}{n}\mathbf{X}^\top_l\boldsymbol{\xi}\right| &\leq \left\|\mathbf{v}^\top_l(t)\right\|_1\left\|\frac{1}{n}\mathbf{X}^\top_l\boldsymbol{\xi}\right\|_\infty \\
&\leq \left\|\mathbf{v}^\top_l(t)\right\|_2\left\|\frac{1}{n}\mathbf{X}^\top_l\boldsymbol{\xi}\right\|_\infty \\
&\leq \frac{1}{80}\tau(u^\star_l)^2.
\end{aligned}$$

Therefore,

$$|e_{l,t}| = |\langle \mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \leq \frac{1}{20}\tau(u^\star_l)^2 + \frac{1}{80}\tau(u^\star_l)^2 + \frac{1}{80}\tau(u^\star_l)^2 \leq \frac{1}{10}\tau(u^\star_l)^2.$$

□

Lemma 11 shows that when the upper bound of perturbation is fixed, $u_l(t)$ grows. Now we show that after $u_l(t)$ grows, the upper bound of perturbations will be decreased.

Lemma 13. *Assume* $\delta_{in} \leq \frac{(u^\star_{min})^2}{120(u^\star_{max})^2}$ *and* $\delta_{out} \leq \frac{(u^\star_{min})^2}{120s(u^\star_{max})^2}$, $\alpha < \frac{\sqrt{\tau_0}}{2\sqrt{L}}u^\star_l$, $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \leq \frac{1}{80}\tau_0(u^\star_l)^2$ *and* $\langle \mathbf{v}_l(t), \mathbf{v}^\star_l\rangle \geq 1 - \frac{1}{5}\tau^2$. *If we achieve that* $|(u^\star_l)^2 - u^2_l(0)| \leq \frac{1}{2}\tau(u^\star_l)^2$ *for each* $l \in [L]$ *where* $0 < \tau_0 \leq \tau \leq 1/2$, *then* $|\langle \mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \leq \frac{1}{20}\tau(u^\star_l)^2$ *and* $|e_{l,t}| \leq \frac{1}{20}\tau(u^\star_l)^2$.

Proof. Similarly to the proof of Lemma 11,

$$\begin{aligned}
\left\|(u^\star_l)^2\mathbf{v}^\star_l - u^2_l(t)\mathbf{v}_l(t)\right\| &\leq \frac{1}{2}\tau(u^\star_l)^2 + u^2_l(t)\sqrt{2 - 2\langle \mathbf{v}_l(t), \mathbf{v}^\star_l\rangle} \\
&\leq \frac{1}{2}\tau(u^\star_l)^2 + \frac{3}{2}(u^\star_l)^2\frac{1}{\sqrt{5}}\tau \\
&\leq \frac{3}{2}\tau(u^\star_l)^2.
\end{aligned}$$

By Assumption 1, we have that

$$\left|\mathbf{v}_l^\top(t)\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l-\mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star-u_l^2(t)\mathbf{v}_l(t))+\mathbf{v}_l^\top(t)\sum_{l'\neq l,l'\in S}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star-u_{l'}^2(t)\mathbf{v}_{l'}(t))\right|$$
$$\leq\frac{3}{2}\delta_{in}\tau(u_{max}^\star)^2+\frac{3}{2}s\delta_{out}\tau(u_{max}^\star)^2\leq\frac{1}{40}\tau(u_l^\star)^2,$$

where $\delta\leq\frac{1}{60s}$. Similarly, we obtain that

$$|e_{l,t}|=|\langle\mathbf{v}_l(t),\mathbf{b}_{l,t}\rangle|\leq\frac{1}{40}\tau(u_l^\star)^2+\frac{1}{80}\tau(u_l^\star)^2+\frac{1}{80}\tau(u_l^\star)^2\leq\frac{1}{20}\tau(u_l^\star)^2.$$

□

By Lemma 10, we know that after certain iterations, we have that $|u^2(t)-(u^\star)^2|\leq\frac{1}{2}(u^\star)^2$. Starting from there, we will apply Lemma 11 and Lemma 12 iteratively until we have our desired accuracy.

We just need to verify when $\tau=\frac{1}{2}$, the condition of either Lemma 11 and Lemma 12 is satisfied. Note that the condition of Lemma 10 already satisfies the condition of Lemma 11 at $\tau=\frac{1}{2}$. Note the condition of Lemma 10 is satisfied when $\delta_{in}\leq\frac{(u_{min}^\star)^2}{120(u_{max}^\star)^2}$ and $\delta_{out}\leq\frac{(u_{min}^\star)^2}{120s(u_{max}^\star)^2}$, $\alpha\leq\frac{1}{4}(u_{min}^\star)^2$, $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty\leq\frac{1}{80}\tau_0(u_{min}^\star)^2$.

C.4 Error analysis outside the support

We only care about the growth rate of $u_l(t)$ when $l\notin S$. When $l\in S^c$, the gradient updates on u_l reads

$$\begin{aligned}u_l(t+1)&=u_l(t)+\gamma\frac{2}{n}u_l(t)\left\langle\mathbf{X}_l\mathbf{v}_l(t),\mathbf{y}-\sum_{l'=1}^{L}u_{l'}^2(t)\mathbf{X}_{l'}\mathbf{v}_{l'}(t)\right\rangle\\&=u_l(t)-2\gamma u_l^3(t)\\&\quad-2\gamma u_l^3(t)\mathbf{v}_l^\top(t)\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l-\mathbf{I}\right)\mathbf{v}_l(t)\\&\quad+2\gamma u_l(t)\mathbf{v}_l^\top(t)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\in S}\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star-u_{l'}^2(t)\mathbf{v}_{l'}(t))\\&\quad-2\gamma u_l(t)\mathbf{v}_l^\top(t)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\neq l,l'\in S^c}\mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t)\\&\quad+2\gamma u_l(t)\frac{1}{n}\mathbf{v}_l(t)\mathbf{X}_l^\top\boldsymbol{\xi}.\end{aligned}$$

Consider the initialization is $u_l(0)=\alpha$, we wonder the smallest number t of iterations that we can ensure $u_l(t)\leq\sqrt{\alpha}$. Denote

$$\begin{aligned}e_{l,t}&=-u_l^2(t)-u_l^2(t)\mathbf{v}_l^\top(t)\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l-\mathbf{I}\right)\mathbf{v}_l(t)\\&\quad+\mathbf{v}_l^\top(t)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\in S}\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star-u_{l'}^2(t)\mathbf{v}_{l'}(t))\\&\quad-\mathbf{v}_l^\top(t)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\neq l,l'\in S^c}\mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t)\\&\quad+\frac{1}{n}\mathbf{v}_l^\top(t)\mathbf{X}_l^\top\boldsymbol{\xi}.\end{aligned}$$

We have that

$$|e_{l,t}|\leq\alpha+\alpha\delta_{in}+\alpha\delta_{out}(L-s)+\frac{3}{2}(u_{max}^\star)^2\delta_{out}s+\left\|\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right\|_\infty.$$

If $\alpha \leq \frac{1}{80L}(u^\star_{min})^2$, $\delta_{in} \leq \frac{(u^\star_{min})^2}{120(u^\star_{max})^2}$ and $\delta_{out} \leq \frac{(u^\star_{min})^2}{120s(u^\star_{max})^2}$, we have that

$$|e_{l,t}| \leq \frac{1}{20}(u^\star_{min})^2 + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_\infty. \tag{8}$$

Lemma 14. *Consider*

$$u(t+1) = u(t)(1+2\gamma e_t)$$

where $|e_t| \leq B$ and $u(0) = \alpha$. Let the step size $\gamma \leq \frac{1}{4B}$, then for any $t \leq T = \frac{1}{32\gamma B}\log\frac{1}{\alpha^4}$, we have $u(t) \leq \sqrt{u(0)}$.

Proof. We start by observing,

$$\begin{aligned}\sqrt{\alpha} \geq u(t) \geq \alpha(1+2\gamma B)^t \\ \iff t \leq \frac{\log\frac{1}{\sqrt{\alpha}}}{\log(1+2\gamma B)}.\end{aligned}$$

By using $\log x \leq x - 1$,

$$\frac{\log\frac{1}{\sqrt{\alpha}}}{\log(1+2\gamma B)} \geq \frac{1}{2\gamma B}\log\frac{1}{\sqrt{\alpha}} \geq \frac{1}{32\gamma B}\log\frac{1}{\alpha^4}.$$

□

D Proof for Theorems in Section 4

D.1 Proof of Theorem 3

Proof. If $\zeta \geq (u^\star_{max})^2$, at the initialization, we already have for $\forall l \in [L]$

$$\begin{aligned}\left\|u_l^2(0)\mathbf{v}_l(0) - (u_l^\star)^2\mathbf{v}_l^\star\right\|_\infty &\leq u_l^2(0) + (u_l^\star)^2 \leq \alpha^2 + (u^\star_{\max})^2 \\ &\leq 2(u^\star_{\max})^2 \leq 2\zeta \\ &\leq 160\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \vee 160\epsilon.\end{aligned}$$

If $\zeta \leq (u^\star_{max})^2$, for those $l \in S$ such that $\zeta \leq (u_l^\star)^2$, we can apply Lemma 10. After

$$T_1 = \frac{\log\frac{(u_l^\star)^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(u_l^\star)^2)},$$

we obtain that $\frac{1}{2}(u_l^\star)^2 \leq u_l^2(T_1) \leq \frac{3}{2}(u_l^\star)^2$, where we also have that $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \leq \frac{1}{80}(u_l^\star)^2$ for every l.

Let m_0 be the number s.t.

$$2^{-m_0-1}(u^\star_{max})^2 \leq \zeta \leq 2^{-m_0}(u^\star_{max})^2,$$

which can be written as $m_0 = \lfloor\log_2\frac{(u^\star_{max})^2}{\zeta}\rfloor$. We can apply Lemma 11 and Lemma 12 together m_0 times. Then further after

$$T_2 = \lfloor\log_2\frac{(u^\star_{max})^2}{\zeta}\rfloor\frac{5}{2\gamma(u_l^\star)^2},$$

we have that

$$\begin{aligned}|u_l^2(T_2) - (u_l^\star)^2| \leq 2^{-m_0}(u^\star_{max})^2 \leq 2\zeta \\ \langle\mathbf{v}_l(T_2), \mathbf{v}_l^\star\rangle \geq 1 - \frac{1}{5}2^{-2m_0}.\end{aligned}$$

Therefore,

$$\begin{aligned}\left\|u_l^2(T_2)\mathbf{v}_l(T_2) - (u_l^\star)^2\mathbf{v}_l^\star\right\|_\infty &\le \left\|u_l^2(T_2)\mathbf{v}_l(T_2) - (u_l^\star)^2\mathbf{v}_l^\star\right\|_2 \\ &\le \left\|(u_l^2(T_2) - (u_l^\star)^2)\mathbf{v}_l(T_2) - (u_l^\star)^2(\mathbf{v}_l^\star - \mathbf{v}_l(T_2))\right\|_2 \\ &\le 2^{-m_0}(u_{max}^\star)^2 + (u_l^\star)^2\sqrt{2 - 2\langle \mathbf{v}_l(T_2), \mathbf{v}_l^\star\rangle} \\ &\le 2^{-m_0}(u_{max}^\star)^2 + (u_l^\star)^2\frac{2}{5}2^{-m_0} \\ &\le 2\zeta.\end{aligned} \tag{9}$$

Note that the above inequality holds for every $l \in S$ such that $(u_l^\star)^2 \ge \zeta$. For those l such that $\zeta \ge (u_l^\star)^2$, we are not able to recover the true signal $(u_l^\star)^2$. the gradient dynamics on this group behaves as errors outside group, and bounded by Lemma 14.

For entries outside the support, we know that from Eq. (8),

$$B = \frac{1}{20}(u_{min}^\star)^2 + \left\|\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right\|_\infty \le \frac{1}{10}(\zeta \vee (u_{min}^\star)^2).$$

By Lemma 14, we have that before $T_3 \le \frac{1}{32\gamma B}\log\frac{1}{\alpha^4}$, $u_l(T_3) \le \sqrt{\alpha}$.

When $\zeta \le (u_{min}^\star)^2$, Eq. (9) holds for every $l \in S$. Therefore, a uniform number of iterations T_1 and T_2 for all groups is written as

$$T_1 = \frac{\log\frac{(u_{max}^\star)^2}{2\alpha^2}}{2\log(1 + \gamma\frac{1}{2}(\zeta \vee (u_{min}^\star)^2))},$$

and

$$T_2 = \lfloor\log_2\frac{(u_{max}^\star)^2}{\zeta}\rfloor\frac{5}{2\gamma(\zeta \vee (u_{min}^\star)^2)}.$$

All we left is to show that $T_3 \ge T_1 + T_2$. We observe that

$$\begin{aligned}T_1 = \frac{\log\frac{(u_{max}^\star)^2}{2\alpha^2}}{2\log(1 + \gamma\frac{1}{2}(\zeta \vee (u_{min}^\star)^2))} &\le \frac{1 + \gamma\frac{1}{2}(\zeta \vee (u_{min}^\star)^2))}{\gamma(\zeta \vee (u_{min}^\star)^2))}\log\frac{(u_{max}^\star)^2}{2\alpha^2} \\ &\le \frac{2}{\gamma(\zeta \vee (u_{min}^\star)^2)}\log\frac{(u_{max}^\star)^2}{2\alpha^2}\end{aligned}$$

where the first inequality is by $\log x \ge \frac{x-1}{x}$.

With our choice of small initialization on α, we have $T_1 \le \frac{1}{2}T_3$, due to $\alpha < \frac{1}{(u_{max}^\star)^8}$. We have $T_2 \le \frac{1}{2}T_3$, because of $\alpha < \frac{\zeta^4}{(u_{max})^8}$.

Hence, we obtain that after $T_l = T_1 + T_2 \ge \frac{\log\frac{(u_{max}^\star)^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(\zeta\vee(u_{min}^\star)^2))} + \lfloor\log_2\frac{(u_{max}^\star)^2}{\zeta}\rfloor\frac{5}{2\gamma(\zeta\vee(u_{min}^\star)^2)}$, and before $T_u = T_3 \le \frac{5}{16\gamma(\zeta\vee(u_{min}^\star)^2)}\log\frac{1}{\alpha^4}$,

$$\left\|u_l^2(t)\mathbf{v}_l(t) - (u_l^\star)^2\mathbf{v}_l^\star\right\|_\infty \lesssim \begin{cases}\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \vee \epsilon, & \text{if } l \in S. \\ \alpha, & \text{if } l \notin S.\end{cases}$$

□

D.2 Proof for Corollary 1

Here is a standard result for sub-Gaussian noise.

Lemma 15. *Let $\frac{1}{\sqrt{n}}\mathbf{X}$ be a $n \times p$ matrix with ℓ_2-normalized columns. Let $\boldsymbol{\xi} \in \mathbb{R}^n$ be a vector of independent σ^2-sub-Gaussian random variables. Then, with probability at least $1 - \frac{1}{8p^3}$*

$$\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \lesssim \sqrt{\frac{\sigma^2\log p}{n}}.$$

Proof of Lemma 15. Since the vector $\boldsymbol{\xi}$ are made of independent σ^2-sub-Gaussian random variables and any column of $\mathbf{X}$ is ℓ_2-normalized, the random variable $\frac{1}{\sqrt{n}}(\mathbf{X}^\top \boldsymbol{\xi})_i$ is still σ^2-sub-Gaussian.

It is a standard result that for any $\epsilon > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{n}}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty > \epsilon\right) \le 2p\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Setting $\epsilon = 2\sqrt{2\sigma^2\log(2p)}$, with probability at least $1-\frac{1}{8p^3}$ we have

$$\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \le \frac{1}{\sqrt{n}}2\sqrt{\sigma^2\log(2p)} \lesssim \sqrt{\frac{\sigma^2\log p}{n}}.$$

□

Proof of Corollary 1. Since $\boldsymbol{\xi}$ is made of independent σ^2-sub-Gaussian entries, by Lemma 15 with probability $1-1/(8p^3)$ we have

$$\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \le 2\sqrt{\frac{2\sigma^2\log(2p)}{n}}.$$

Hence, letting $\epsilon = 2\sqrt{\frac{2\sigma^2\log(2p)}{n}}$, we obtain that

$$\left\|(\mathbf{D}\mathbf{u}(t))^2 \odot \mathbf{v}(t) - \mathbf{w}^\star\right\|_2^2 \lesssim \sum_{l\in S}\epsilon^2 + \sum_{l\notin S}\alpha \le s\epsilon^2 + (L-s)\frac{\epsilon^2}{L^2} \lesssim \frac{s\sigma^2\log p}{n}.$$

□

D.3 Convergence for algorithm 2

Lemma 16. *Consider the update in Eq.* (4)*, choose the step size $\eta_t = \eta \le \frac{4}{9(u^\star)^4}$, if $\langle \mathbf{v}(t), \mathbf{v}^\star\rangle \ge 1-\frac{1}{5}\tau$, $|u^2(t)-(u^\star)^2| \le \tau(u^\star)^2$ and $\|\mathbf{b}_t\| \le \frac{1}{10}\tau(u^\star)^2$ for some constant $0<\tau<\frac{1}{2}$, we have that*

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star\rangle \ge 1-\frac{1}{5}\tau.$$

Proof. We first rewrite $\mathbf{z}(t+1)$ as

$$\mathbf{z}(t+1) = \eta u^2(t)(u^\star)^2\mathbf{v}^\star + (1-\eta u^4(t))\mathbf{v}(t) + \eta u^2(t)\mathbf{b}_t.$$

Therefore,

$$\begin{aligned}
\langle \mathbf{z}(t+1), \mathbf{v}^\star\rangle &\ge \eta u^2(t)(u^\star)^2 + (1-\eta u^4(t))\langle \mathbf{v}(t), \mathbf{v}^\star\rangle + \eta u^2(t)\langle \mathbf{b}_t, \mathbf{v}^\star\rangle \\
&\ge \eta u^2(t)(u^\star)^2 + (1-\eta u^4(t))\left(1-\frac{1}{5}\tau\right) - \eta u^2(t)\frac{1}{10}\tau(u^\star)^2 \\
\|\mathbf{z}(t+1)\| &\le \eta u^2(t)(u^\star)^2 + (1-\eta u^4(t)) + \eta u^2(t)\frac{1}{10}\tau(u^\star)^2.
\end{aligned}$$

We obtain that

$$\begin{aligned}
\langle \mathbf{v}(t+1), \mathbf{v}^\star\rangle = \frac{\langle \mathbf{z}(t+1), \mathbf{v}^\star\rangle}{\|\mathbf{z}(t+1)\|} &\ge 1 - \frac{\frac{1}{5}\tau(1-\eta u^4(t)) + 2\eta u^2(t)\frac{1}{10}\tau(u^\star)^2}{\eta u^2(t)(u^\star)^2 + (1-\eta u^4(t)) + \eta u^2(t)\frac{1}{10}\tau(u^\star)^2} \\
&\ge 1 - \frac{1-\eta u^4(t)+\eta u^2(t)(u^\star)^2}{\eta u^2(t)(u^\star)^2 + (1-\eta u^4(t)) + \eta u^2(t)\frac{1}{10}\tau(u^\star)^2}\frac{1}{5}\tau \\
&\ge 1-\frac{1}{5}\tau.
\end{aligned}$$

□

Note that compared with Lemma 8, under the condition $\|\mathbf{b}_t\| \le B(u^\star)^2$, we get $\langle \mathbf{v}(t+1), \mathbf{v}^\star\rangle \ge 1 - B$ instead of $\langle \mathbf{v}(t+1), \mathbf{v}^\star\rangle \ge 1 - B^2$. Accordingly, we need to a new version for Lemma 12 with a smaller bound on δ to make up the loss in Lemma 16.

Lemma 17. *Assume* $\delta_{in} \le \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120(u^\star_{max})^2}$ *and* $\delta_{out} \le \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120s(u^\star_{max})^2}$, $\alpha < \frac{1}{2}\sqrt{\frac{\tau_0}{L}}u_l^\star$, $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \le \frac{1}{80}\tau_0(u_l^\star)^2$ *and* $|(u_l^\star)^2 - u_l^2(0)| \le \tau(u_l^\star)^2$ *for each* $l \in [L]$ *where* $0 < \tau_0 \le \tau \le 1/2$. *If* $\langle \mathbf{v}_l(t), \mathbf{v}_l^\star\rangle \ge 1 - \frac{1}{5}\tau$, *then* $|\langle \mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \le \frac{1}{10}\tau(u_l^\star)^2$ *and* $|e_{l,t}| \le \frac{1}{10}\tau(u_l^\star)^2$.

Proof. Similarly to Lemma 12, we have that

$$\begin{aligned}\left\|(u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)\right\| &\le \tau(u_l^\star)^2 + u_l^2(t)\sqrt{2 - 2\langle \mathbf{v}_l(t), \mathbf{v}_l^\star\rangle} \\ &\le \tau(u_l^\star)^2 + \frac{3}{2}(u_l^\star)^2\frac{\sqrt{2}}{\sqrt{5}}\sqrt{\tau} \\ &\le \left(1 + 2\frac{1}{\sqrt{\tau_0}}\right)\tau(u_l^\star)^2.\end{aligned} \tag{10}$$

By Assumption 1, we have that

$$\left|\mathbf{v}_l^\top(t)\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l - \mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)) + \mathbf{v}_l^\top(t)\sum_{l'\neq l, l'\in S}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t))\right|$$
$$\le \left(1 + 2\frac{1}{\sqrt{\tau_0}}\right)\delta_{in}\tau(u^\star_{max})^2 + \left(1 + 2\frac{1}{\sqrt{\tau_0}}\right)s\delta_{out}\tau(u^\star_{max})^2 \le \frac{1}{20}\tau(u_l^\star)^2,$$

where $\delta \le \frac{\sqrt{\tau_0}(u^\star_{min})^2}{60s(u^\star_{max})^2}$. The other two terms follows exactly what we did in Lemma 12. Therefore,

$$|e_{l,t}| = |\langle \mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \le \frac{1}{20}\tau(u_l^\star)^2 + \frac{1}{80}\tau(u_l^\star)^2 + \frac{1}{80}\tau(u_l^\star)^2 \le \frac{1}{10}\tau(u_l^\star)^2.$$

□

Proof to Theorem 4. The proof is similar to that of Theorem 3. For the first stage, we apply Lemma 10, as nothing is changed from Theorem 3. For the second stage, instead of applying Lemma 11 and Lemma 12, we apply Lemma 16 and Lemma 17 iteratively. To apply these lemmas, we first observe that

$$\zeta \le \tau_0(u^\star_{max})^2 \iff \frac{\zeta}{(u^\star_{max})^2} \le \tau_0.$$

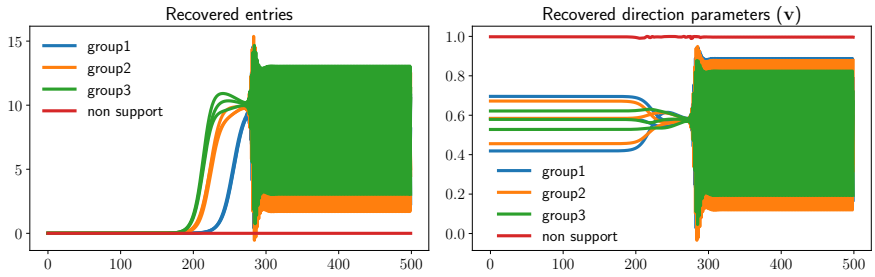
Therefore the requirement on δ's becomes $\delta_{in} \le \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120(u^\star_{max})^3}$ and $\delta_{out} \le \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120s(u^\star_{max})^3}$. The number of iterations and convergence results follow from the proof of Theorem 3.

□

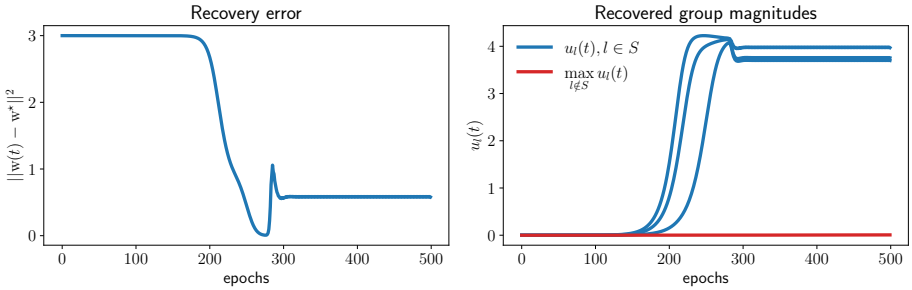
The criterion for switching time. We provide some motivation for the practical criterion. We first note that, the criterion in Theorem 4 actually indicates a lower bound of switching time. With more derivations, our results still hold if one choose to switch after the time when the criterion is first satisfied (instead of switching right at that time.) Let us focus on the entries on the support. In the proof of Theorem 3, one can also obtain the convergence on $u_l(t)$ as the positiveness of $u_l(t)$ can be ensured with a small step size γ (since the power-parametrization will recast the gradient updates into a multiplicative sequence). Therefore, with an appropriate choice of τ, the practical criterion $\max_{l\in S}\{|u_l(t+1) - u_l(t)|/|u_l(t) + \varepsilon|\} < \tau$ would imply the theoretical criterion $u_l(t)^2 \ge \frac{1}{2}u_l^\star(t)^2$ on the support, and therefore would indicate a possibly later switching time than what the theoretical criterion determines. For gradient updates outside the support, we observe slow growth rate and hence the practical rule is likely satisfied on the non-support entry, which we observe in the numerical experiments. Note that the switching only happens when both the support and non-support entries fulfill the criterion.

E MORE NUMERICAL RESULTS

E.1 STABILITY ISSUE OF ALGORITHM 1 AND STANDARD GD



(a) Numerical instability in direction estimations.



(b) Parameter estimation error remains small.

Figure 6: Numerical instability of algorithm 1

Stability issue of Algorithm 1. Figure 6 presents the recovered entries and direction parameters $\mathbf{v}(t)$ under the same setting as Figure 2. Because of the large learning rate on $\mathbf{v}$, the algorithm may not show a convergent result in the latter stage due to the irreducible error (perturbations). Although the parameter estimation is still reasonable with normalization on each $\mathbf{v}_l, l \in [L]$, we still aim to get a stable algorithm, which motivates our algorithm 2.

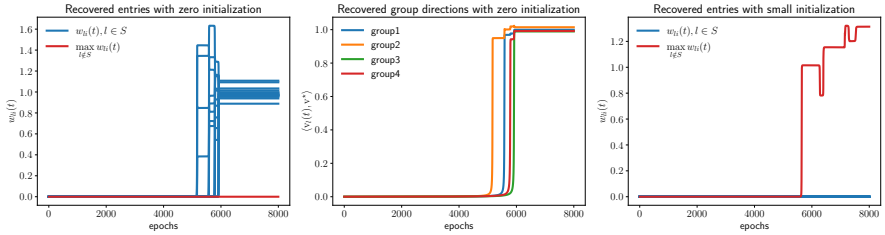


Figure 7: Gradient descent without weight normalization.

Standard gradient descent. To further understand how weight normalization affects the gradient dynamics, we conduct experiments using standard gradient descent without weight normalization. For that, we use the same setting as in Figure 4 and show the result in Figure 7. The left and middle figures are based on zero initialization on $\mathbf{v}$. We see a numerically convergent result, and the inner product between learned and true directions starts to grow from 0. As the directions guide the magnitude to grow, there is an extra stage for the directions to become roughly accurate. The choice of this initialization is necessary and subtle. The figure on the right is for small initialization 10^{-3}, where the entries outside support get significant magnitudes, and the algorithm fails.

E.2 AUTOENCODER WITH GROUPING LAYER

The grouping layers have been used in grouped CNN and grouped attention mechanisms (Wu et al., 2021; Xie et al., 2017; Lee et al., 2018), which usually leads to parameter efficiency and better accuracy. To demonstrate the practical value of such grouping layers, we conduct the following experiment about learning good representations on MNIST.

(Jing et al., 2020) proposed implicit rank-minimizing autoencoder (IRMAE), which is a deterministic autoencoder with implicit regularization. The idea is to apply more linear layers between encoder and decoder to penalize the rank of latent representation. A graphical illustration of the architecture is shown in Figure 8, where we explicitly show the last convolution layer and the linear layers in the latent space, which are absorbed into the last layer of the encoder in practice. This design is related to the power parametrization (Schwarz et al., 2021) trick to promote sparsity/low-rankness. One major advantage is that IRMAE produces a more interpretable latent representation, and the linear interpolation in the latent space gives a natural transition between two images.

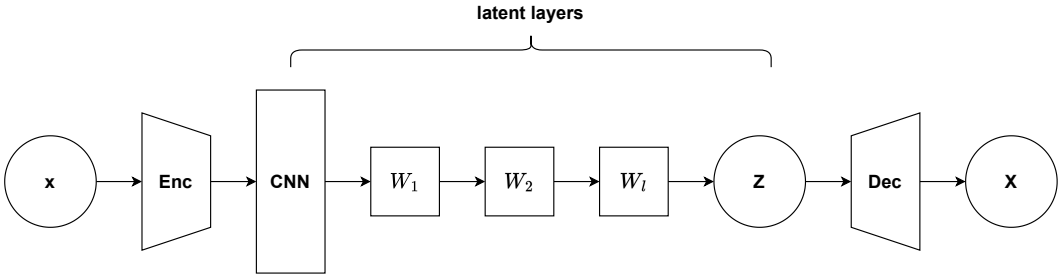


Figure 8: Implicit rank-minimizing autoencoder.

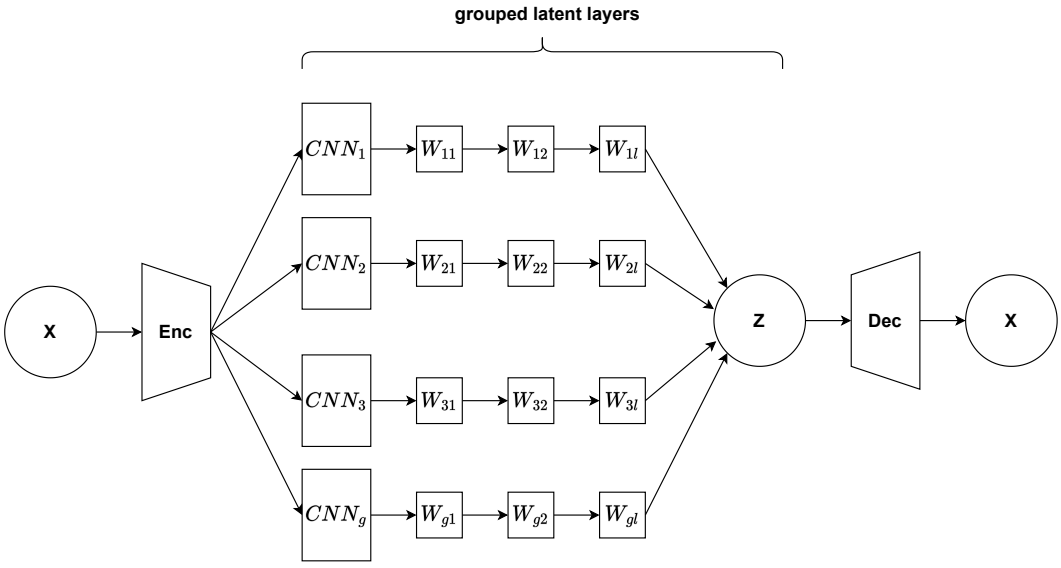


Figure 9: Implicit rank-minimizing autoencoder with grouping layers.

Inspired by our DGLNN, we design a CNN analog of it, which we call grouped autoencoder (GAE). The architecture is shown in Figure 9. The channels feed into the last convolutional layer of encoder is separable into g groups. The linear layers (power-parametrization) are applied within each group. Grouping channels of convolutional layers is a common practice to improve the parameter efficiency. With these grouping and power layers in the latent space, we expect it learns a better latent representation as IRMAE does.

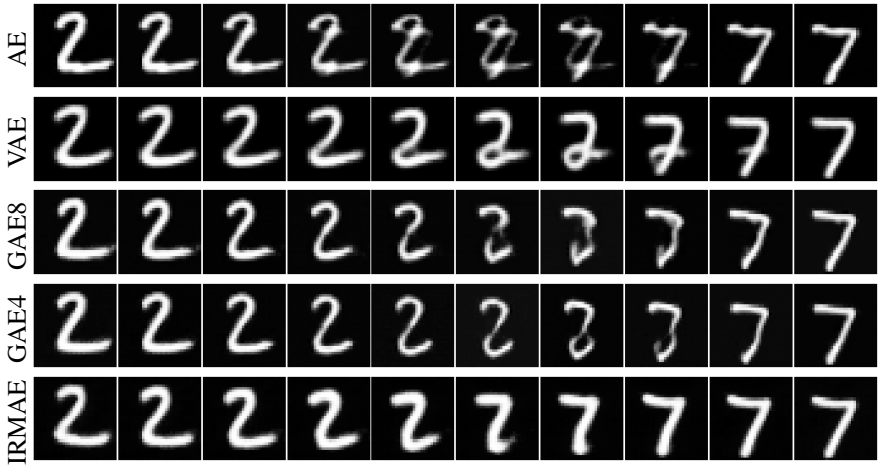


Figure 10: Linear interpolations between data points on the MNIST dataset. GAE4/8 stands for grouped autoencoder with 4/8 groups.

The linear interpolations between data points in the latent space are shown in Figure 10. We compare the grouped autoencoder (GAE) with autoencoder (AE), variantional autoencoder (VAE) and implicit rank-minimizing autoencoder (IRMAE). We see that GAE outperforms AE and VAE, and gives comparable results with IRMAE. However, GAE achieves a better parameter efficiency as shown in Table 2.

	# of params
IRMAE	786K
GAE4	196K
GAE8	98K

Table 2: Number of parameters of hidden layers in latent space.

E.3 Experiments with Gaussian measurements

Besides the numerical results shown in Section 5, we conduct the following experiments with sampling each entry of $\mathbf{X}$ from a standard normal distribution.

The effectiveness using Gaussian design. We follow the same setting with that Figure 3 except changing Rademacher random variables to Gaussian random variables. The convergence of Algorithm 2 is shown in Figure 11. We see that the recovered entries, group magnitudes and directions successfully converge to the true ones.

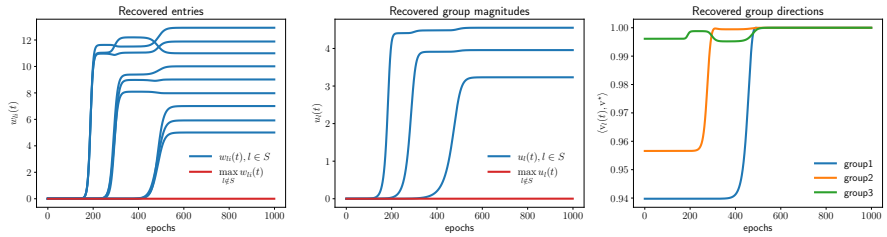


Figure 11: Convergence of algorithm 2 with Gaussian measurements

Comparisons with explicit regularization methods. We compare Algorithm 2 with proximal gradient descent implemented in (Carmichael et al., 2021) and primal-dual procedure (Molinari et al., 2021). Each entry of $\mathbf{X}$ is sampled from a standard Gaussian distribution. We set $n = 150$ and

$p = 300$, and the number of non-zero entries is 10, divided into 3 groups with size 4. We vary the variance in the noise to achieve different signal-to-noise ratios (SNR). The experiment is repeated 30 times at each noise level. The average and standard deviation of the estimation error are depicted in Figure 12. Our algorithm is consistently better than explicit regularization methods, whereas the primal-dual procedure has a comparable performance when SNR is large.

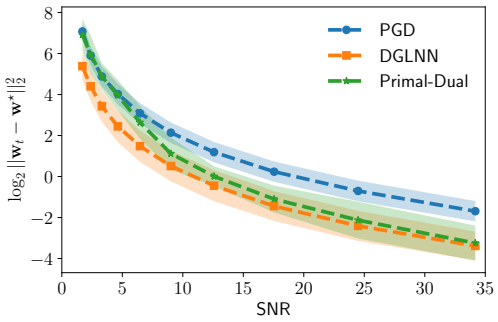


Figure 12: Comparisons with proximal gradient descent and iterative regularization.

To further discover the potential applications of our findings, we use a gene expression dataset from the Microarray experiments of mammalian eye tissue samples (Scheetz et al., 2006). The dataset consists of 120 samples with 100 predictors that are expanded from 20 genes using 5 basis B-splines, as described in (Yang & Zou, 2015). The goal is to predict the gene expression level of TRIM32, which causes Bardet-Biedl syndrome. We randomly split the data equally, and use the validation dataset for hyperparameter tuning and early stopping. We compare our approach with the commonly used proximal gradient descent and a primal-dual approach. The result is shown in Table 3. Our approach achieves the best performance among these three methods.

Test error	PGD	Primal-Dual	Our approach
MSE	0.03096	0.02868	0.02477

Table 3: Comparisons of MSE (mean squared error) on test set.