A. Proofs of Section 3

Proof of Theorem 3.1. We first expand the loss function directly.

$$\begin{aligned}
&\ell(\mathbf{v}, \mathbf{a}) \\
=&\mathbb{E}\left[\frac{1}{2}\left(y - \mathbf{a}^\top \sigma(\mathbf{Z})\mathbf{w}\right)^2\right] \\
=&(\mathbf{a}^*)^\top \mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w}^*)\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]\mathbf{a}^* + \mathbf{a}^\top \mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right]\mathbf{a} - 2\mathbf{a}^\top \mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]\mathbf{a}^* \\
=&(\mathbf{a}^*)^\top \mathbf{A}(\mathbf{w}^*)\mathbf{a}^* + \mathbf{a}^\top \mathbf{A}(\mathbf{w})\mathbf{a} - 2\mathbf{a}^\top \mathbf{B}(\mathbf{w}, \mathbf{w}^*)\mathbf{w}^*.
\end{aligned}$$

where for simplicity, we denote

$$\mathbf{A}(\mathbf{w}) = \mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right] \tag{5}$$

$$\mathbf{B}(\mathbf{w}, \mathbf{w}^*) = \mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]. \tag{6}$$

For $i \neq j$, using the second identity of Lemma A.1, we can compute

$$\mathbf{A}(\mathbf{w})_{ij} = \mathbb{E}\left[\sigma\left(\mathbf{Z}_i^\top \mathbf{w}\right)\right]\mathbb{E}\left[\sigma\left(\mathbf{Z}_j^\top \mathbf{w}\right)\right] = \frac{1}{2\pi}\|\mathbf{w}\|_2^2$$

For $i = j$, using the second moment formula of half-Gaussian distribution we can compute

$$\mathbf{A}(\mathbf{w})_{ii} = \frac{1}{2}\|\mathbf{w}\|_2^2.$$

Therefore

$$\mathbf{A}(\mathbf{w}) = \frac{1}{2\pi}\|\mathbf{w}\|_2^2\left(\mathbf{1}\mathbf{1}^\top + (\pi - 1)\mathbf{I}\right).$$

Now let us compute $\mathbf{B}(\mathbf{w}, \mathbf{w}_*)$. For $i \neq j$, similar to $\mathbf{A}(\mathbf{w})_{ij}$, using the independence property of Gaussian, we have

$$\mathbf{B}(\mathbf{w}, \mathbf{w}_*)_{ij} = \frac{1}{2\pi}\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2.$$

Next, using the fourth identity of Lemma A.1, we have

$$\mathbf{B}(\mathbf{w}, \mathbf{w}^*)_{ii} = \frac{1}{2\pi}\left(\cos\phi(\pi - \phi) + \sin\phi\right)\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2.$$

Therefore, we can also write $\mathbf{B}(\mathbf{w}, \mathbf{w}^*)$ in a compact form

$$\mathbf{B}(\mathbf{w}, \mathbf{w}^*) = \frac{1}{2\pi}\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2\left(\mathbf{1}\mathbf{1}^\top + (\cos\phi(\pi - \phi) + \sin\phi - 1)\mathbf{I}\right).$$

Plugging in the formulas of $\mathbf{A}(\mathbf{w})$ and $\mathbf{B}(\mathbf{w}, \mathbf{w}^*)$ and $\mathbf{w} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$, we obtain the desired result. □

Proof of Theorem 3.2. We first compute the expect gradient for $\mathbf{v}$. From(Salimans & Kingma, 2016), we know

$$\frac{\partial \ell(\mathbf{v}, \mathbf{a})}{\partial \mathbf{v}} = \frac{1}{\|\mathbf{v}\|_2}\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\frac{\partial \ell(\mathbf{w}, \mathbf{a})}{\partial \mathbf{w}}.$$

Recall the gradient formula,

$$\begin{aligned}
&\frac{\partial \ell\left(\mathbf{Z}, \mathbf{w}, \mathbf{a}\right)}{\partial \mathbf{w}} \\
=&\left(\sum_{i=1}^{k} a_i^* \sigma\left(\mathbf{Z}_i \mathbf{w}\right)-\sum_{i=1}^{k} a_i^* \sigma\left(\mathbf{Z} \mathbf{w}^*\right)\right)\left(\sum_{i=1}^{k} a_i \mathbf{Z}_i \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w}\right\}\right) \\
=&\left(\sum_{i=1}^{k} a_i^2 \mathbf{Z}_i \mathbf{Z}_i^{\top} \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0\right\}+\sum_{i \neq j} a_i a_j \mathbf{Z}_i \mathbf{Z}_j^{\top} \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0, \mathbf{Z}_j^{\top} \mathbf{w} \geq 0\right\}\right) \mathbf{w} \qquad (7) \\
-&\left(\sum_{i=1}^{k} a_i a_i^* \mathbf{Z}_i \mathbf{Z}_i^{\top} \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0, \mathbf{Z}_i^{\top} \mathbf{w}^* \geq 0\right\}+\sum_{i \neq j} a_i a_j^* \mathbf{Z}_i \mathbf{Z}_j^* \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0, \mathbf{Z}_j^{\top} \mathbf{w}^* \geq 0\right\}\right) \mathbf{w}^*. \qquad (8)
\end{aligned}$$

Now we calculate expectation of Equation (7) and (8) separately. For (7), by first two formulas of Lemma A.1, we have

$$\begin{aligned}
&\left(\sum_{i=1}^{k} a_i^2 \mathbf{Z}_i \mathbf{Z}_i^{\top} \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0\right\}+\sum_{i \neq j} a_i a_j \mathbf{Z}_i \mathbf{Z}_j^{\top} \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0, \mathbf{Z}_j^{\top} \mathbf{w} \geq 0\right\}\right) \mathbf{w} \\
=&\sum_{i=1}^{k} a_i^2 \cdot \frac{\mathbf{w}}{2}+\sum_{i \neq j} a_i a_j \frac{\mathbf{w}}{2 \pi}.
\end{aligned}$$

For (8), we use the second and third formula in Lemma A.1 to obtain

$$\begin{aligned}
&\left(\sum_{i=1}^{k} a_i a_i^* \mathbf{Z}_i \mathbf{Z}_i^{\top} \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0, \mathbf{Z}_i^{\top} \mathbf{w}^* \geq 0\right\}+\sum_{i \neq j} a_i a_j^* \mathbf{Z}_i \mathbf{Z}_j^* \mathbb{I}\left\{\mathbf{Z}_i^{\top} \mathbf{w} \geq 0, \mathbf{Z}_j^{\top} \mathbf{w}^* \geq 0\right\}\right) \mathbf{w}^* \\
=&\mathbf{a}^{\top} \mathbf{a}^*\left(\frac{1}{\pi}\left(\pi-\phi\right) \mathbf{w}^*+\frac{1}{\pi} \sin \phi \frac{\left\|\mathbf{w}^*\right\|_2}{\|\mathbf{w}\|_2} \mathbf{w}\right)+\sum_{i \neq j} a_i a_j^* \frac{1}{2 \pi} \frac{\left\|\mathbf{w}^*\right\|_2}{\|\mathbf{w}\|_2} \mathbf{w}.
\end{aligned}$$

In summary, aggregating them together we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell\left(\mathbf{Z}, \mathbf{w}, \mathbf{a}\right)}{\partial \mathbf{w}}\right] \\
=&\frac{1}{2 \pi} \mathbf{a}^{\top} \mathbf{a}^*\left(\pi-\phi\right) \mathbf{w}^*+\left(\frac{\|\mathbf{a}\|_2^2}{2}+\frac{\sum_{i \neq j} a_i a_j}{2 \pi}+\frac{\mathbf{a}^{\top} \mathbf{a}^* \sin \phi}{2 \pi} \frac{\left\|\mathbf{w}^*\right\|_2}{\|\mathbf{w}\|_2}+\frac{\sum_{i \neq j} a_j a_j^*}{2 \pi} \frac{\left\|\mathbf{w}_*\right\|_2}{\|\mathbf{w}\|_2}\right) \mathbf{w}.
\end{aligned}$$

As a sanity check, this formula matches Equation (16) of (Brutzkus & Globerson, 2017) when $\mathbf{a}=\mathbf{a}^*=\mathbf{1}$.

Next, we calculate the expected gradient of $\mathbf{a}$. Recall the gradient formula of $\mathbf{a}$

$$\begin{aligned}
\frac{\partial \ell(\mathbf{Z}, \mathbf{w}, \mathbf{a})}{\mathbf{a}} =&\left(\mathbf{a}^{\top} \sigma\left(\mathbf{Z} \mathbf{w}\right)-\left(\mathbf{a}^*\right)^{\top} \sigma\left(\mathbf{Z} \mathbf{w}^*\right)\right) \sigma\left(\mathbf{Z} \mathbf{w}\right) \\
=&\sigma\left(\mathbf{Z} \mathbf{w}\right) \sigma\left(\mathbf{Z} \mathbf{w}\right)^{\top} \mathbf{a}-\sigma\left(\mathbf{Z} \mathbf{w}\right) \sigma\left(\mathbf{Z} \mathbf{w}^*\right)^{\top} \mathbf{a}^*
\end{aligned}$$

Taking expectation we have

$$\frac{\partial \ell\left(\mathbf{w}, \mathbf{a}\right)}{\partial \mathbf{a}}=\mathbf{A}\left(\mathbf{w}\right) \mathbf{a}-\mathbf{B}\left(\mathbf{w}, \mathbf{w}^*\right) \mathbf{a}^*$$

where $\mathbf{A}\left(\mathbf{w}\right)$ and $\mathbf{B}\left(\mathbf{w}, \mathbf{w}^*\right)$ are defined in Equation (5) and (6). Plugging in the formulas for $\mathbf{A}\left(\mathbf{w}\right)$ and $\mathbf{B}\left(\mathbf{w}, \mathbf{w}^*\right)$ derived in the proof of Theorem 3.1 we obtained the desired result.

□

Lemma A.1 (Useful Identities). *Given* $\mathbf{w}$, $\mathbf{w}^*$ *with angle* ϕ *and* $\mathbf{Z}$ *is a Gaussian random vector, then*

$$\begin{aligned}
\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\mathbf{w} =& \frac{1}{2}\mathbf{w} \\
\mathbb{E}\left[\mathbf{z}\mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right] =& \frac{1}{\sqrt{2\pi}}\frac{\mathbf{w}}{\|\mathbf{w}\|_2} \\
\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\mathbf{w}_* =& \frac{1}{2\pi}(\pi-\phi)\mathbf{w}^* + \frac{1}{2\pi}\sin\phi\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2}\mathbf{w} \\
\mathbb{E}\left[\sigma\left(\mathbf{z}^\top \mathbf{w}\right)\sigma\left(\mathbf{z}^\top \mathbf{w}_*\right)\right] =& \frac{1}{2\pi}\left(\cos\phi\left(\pi-\phi\right)+\sin\phi\right)\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2
\end{aligned}$$

Proof. Consider an orthonormal basis of $\mathbb{R}^{d\times d}$: $\left\{\mathbf{e}_i\mathbf{e}_j^\top\right\}$ with $\mathbf{e}_1 \parallel \mathbf{w}$. Then for $i \neq j$, we know

$$\langle \mathbf{e}_i\mathbf{e}_j, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\rangle = 0$$

by the independence properties of Gaussian random vector. For $i = j = 1$,

$$\langle \mathbf{e}_i\mathbf{e}_j^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\rangle = \mathbb{E}\left[\left(\mathbf{z}^\top \mathbf{w}\right)^2 \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right] = \frac{1}{2}$$

where the last step is by the property of half-Gaussian. For $i = j \neq j$, $\langle \mathbf{e}_i\mathbf{e}_j^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\rangle = 1$ by standard Gaussian second moment formula. Therefore, $\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\mathbf{w} = \frac{1}{2}\mathbf{w}$. $\mathbb{E}\left[\mathbf{z}\mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right] = \frac{1}{\sqrt{2\pi}}\mathbf{w}$ can be proved by mean formula of half-normal distribution. To prove the third identity, consider an orthonormal basis of $\mathbb{R}^{d\times d}$: $\left\{\mathbf{e}_i\mathbf{e}_j^\top\right\}$ with $\mathbf{e}_1 \parallel \mathbf{w}_*$ and $\mathbf{w}$ lies in the plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$. Using the polar representation of 2D Gaussian random variables (r is the radius and θ is the angle with $\mathrm{d}P_r = r\exp(-r^2/2)$ and $\mathrm{d}P_\theta = \frac{1}{2\pi}$):

$$\begin{aligned}
\langle \mathbf{e}_1\mathbf{e}_1^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\rangle =& \frac{1}{2\pi}\int_0^\infty r^3\exp\left(-r^2/2\right)\mathrm{d}r \cdot \int_{-\pi/2+\phi}^{\pi/2}\cos^2\theta\mathrm{d}\theta \\
=& \frac{1}{2\pi}\left(\pi-\phi+\sin\phi\cos\phi\right), \\
\langle \mathbf{e}_1\mathbf{e}_2^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\rangle =& \frac{1}{2\pi}\int_0^\infty r^3\exp\left(-r^2/2\right)\mathrm{d}r \cdot \int_{-\pi/2+\phi}^{\pi/2}\sin\theta\cos\theta\mathrm{d}\theta \\
=& \frac{1}{2\pi}\left(\sin^2\phi\right), \\
\langle \mathbf{e}_2\mathbf{e}_2^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\rangle =& \frac{1}{2\pi}\int_0^\infty r^3\exp\left(-r^2/2\right)\mathrm{d}r \cdot \int_{-\pi/2+\phi}^{\pi/2}\sin^2\theta\mathrm{d}\theta \\
=& \frac{1}{2\pi}\left(\pi-\phi-\sin\phi\cos\phi\right).
\end{aligned}$$

Also note that $\mathbf{e}_2 = \frac{\bar{\mathbf{w}}-\cos\phi\mathbf{e}_1}{\sin\phi}$. Therefore

$$\begin{aligned}
\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\mathbf{w}_* =& \frac{1}{2\pi}\left(\pi-\phi+\sin\phi\cos\phi\right)\mathbf{w}^* + \frac{1}{2\pi}\sin^2\phi\cdot\frac{\bar{\mathbf{w}}-\cos\phi\mathbf{e}_1}{\sin\phi}\|\mathbf{w}^*\|_2 \\
=& \frac{1}{2\pi}\left(\pi-\phi\right)\mathbf{w}^* + \frac{1}{2\pi}\sin\phi\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2}\mathbf{w}.
\end{aligned}$$

For the fourth identity, focusing on the plane spanned by $\mathbf{w}$ and $\mathbf{w}_*$, using the polar decomposition, we have

$$\begin{aligned}
\mathbb{E}\left[\sigma\left(\mathbf{z}^\top \mathbf{w}\right)\sigma\left(\mathbf{z}^\top \mathbf{w}_*\right)\right] &= \frac{1}{2\pi}\int_0^\infty r^3\exp\left(-r^2/2\right)\mathrm{d}r \cdot \int_{-\pi/2+\phi}^{\pi/2}\left(\cos\theta\cos\phi+\sin\theta\sin\phi\right)\cos\theta\mathrm{d}\theta\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2 \\
&= \frac{1}{2\pi}\left(\cos\phi\left(\pi-\phi+\sin\phi\cos\phi\right)+\sin^3\phi\right)\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2.
\end{aligned}$$

□

B. Proofs of Qualitative Convergence Results

Proof of Lemma 5.1. When Algorithm 1 converges, since $\mathbf{a}^\top \mathbf{a}^* \neq 0$ and $\|\mathbf{v}\|_2 < \infty$, using the gradient formula in Theorem 3.2, we know that either $\pi - \phi = 0$ or $\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\mathbf{w}^* = \mathbf{0}$. For the second case, since $\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}$ is a projection matrix on the complement space of $\mathbf{v}$, $\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\mathbf{w}^* = \mathbf{0}$ is equivalent to $\theta(\mathbf{v}, \mathbf{w}^*) = 0$. Once the angle between $\mathbf{v}$ and $\mathbf{w}^*$ is fixed, using the gradient formula for $\mathbf{a}$ we have the desired formulas for saddle points. □

Proof of Lemma 5.2. By the gradient formula of $\mathbf{w}$, if $\mathbf{a}^\top \mathbf{a}^* > 0$, the gradient is of the form $c\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\mathbf{w}^*$ where $c > 0$. Thus because $\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}$ is the projection matrix onto the complement space of $\mathbf{v}$, the gradient update always makes the angle smaller. □

C. Proofs of Quantitative Convergence Results

C.1. Useful Technical Lemmas

We first prove the lemma about the convergence of ϕ^t.

Proof of Lemma 5.5. We consider the dynamics of $\sin^2\phi^t$.

$$
\begin{aligned}
&\sin^2\phi^{t+1}\\
=&1 - \frac{\left(\left(\mathbf{v}^{t+1}\right)^\top \mathbf{w}^*\right)^2}{\|\mathbf{v}^{t+1}\|_2^2 \|\mathbf{w}^*\|_2^2}\\
=&1 - \frac{\left(\left(\mathbf{v}^{\mathbf{t}} - \eta\frac{\partial \ell}{\partial \mathbf{v}^{\mathbf{t}}}\right)^\top \mathbf{w}^*\right)^2}{\left(\|\mathbf{v}^t\|_2^2 + \eta^2\left(\frac{\partial \ell}{\partial \mathbf{v}^{\mathbf{t}}}\right)^2\right)\|\mathbf{w}^*\|_2^2}\\
=&1 - \frac{\left((\mathbf{v}^t)^\top \mathbf{v} + \eta\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi-\phi^t\right)}{2\pi\|\mathbf{v}\|_2}\cdot\sin^2\phi^t\|\mathbf{w}\|_2^2\right)^2}{\|\mathbf{v}^t\|_2^2\|\mathbf{w}^*\|_2^2 + \eta^2\left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi-\phi^t\right)}{2\pi}\right)^2\frac{\sin^2\phi^t\|\mathbf{w}^*\|_2^4}{\|\mathbf{v}^t\|_2^2}}\\
\leq&1 - \frac{\|\mathbf{v}^t\|_2^2\|\mathbf{w}^*\|_2^2\cos^2\phi^t + 2\eta\|\mathbf{w}^*\|_2^3\cdot\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}(\pi-\phi)}{2\pi}\cdot\sin^2\phi^t\cos\phi^t}{\|\mathbf{v}^t\|_2^2\|\mathbf{w}^*\|_2^2 + \eta^2\left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi-\phi^t\right)}{2\pi}\right)^2\frac{\sin^2\phi^t\|\mathbf{w}^*\|_2^4}{\|\mathbf{v}^t\|_2^2}}\\
=&\frac{\sin^2\phi^t - 2\eta\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2}\cdot\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}(\pi-\phi)}{2\pi}\cdot\sin^2\phi^t\cos\phi^t + \eta^2\left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*(\pi-\phi)}{2\pi}\right)^2\sin^2\phi^t\left(\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2}\right)^2}{1 + \eta^2\left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*(\pi-\phi)}{2\pi}\right)^2\sin^2\phi^t\left(\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2}\right)^2}\\
\leq&\sin^2\phi^t - 2\eta\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2}\cdot\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}\left(\pi-\phi\right)}{2\pi}\cdot\sin^2\phi^t\cos\phi^t + \eta^2\left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi-\phi\right)}{2\pi}\right)^2\sin^2\phi^t\left(\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2}\right)^2
\end{aligned}
$$

where in the first inequality we dropped term proportional to $O(\eta^4)$ because it is negative, in the last equality, we divided numerator and denominator by $\|\mathbf{v}^t\|_2^2\|\mathbf{w}^*\|_2^2$ and the last inequality we dropped the denominator because it is bigger than 1. Therefore, recall $\lambda^t = \frac{\|\mathbf{w}^*\|_2\left(\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\right)\left(\pi-\phi^t\right)}{2\pi\|\mathbf{v}^t\|_2^2}$ and we have

$$
\sin^2\phi^{t+1} \leq \left(1 - 2\eta\cos\phi^t\lambda^t + \eta^2\left(\lambda^t\right)^2\right)\sin^2\phi^t. \tag{9}
$$

To this end, we need to make sure $\eta \leq \frac{\cos\phi^t}{\lambda^t}$. Note that since $\|\mathbf{v}^t\|_2^2$ is monotonically increasing, it is lower bounded by 1. Next notice $\phi^t \leq \pi/2$. Finally, from Lemma C.2, we know $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \leq \left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)\|\mathbf{w}\|_2^2$. Combining these, we

have an upper bound

$$\lambda^t \leq \frac{\left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}{4}.$$

Plugging this back to Equation (9) and use our assumption on η, we have

$$\sin^2 \phi^{t+1} \leq \left(1 - \eta \cos\phi^t \lambda^t\right) \sin^2 \phi^t.$$

□

Lemma C.1. $\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* \geq \min\left\{\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \eta\left(\frac{g(\phi^t)-1}{\pi-1}\|\mathbf{a}^*\|_2^2 - \left(\mathbf{a}^t\right)^\top \mathbf{a}^*\right), \frac{g(\phi^t)-1}{\pi-1}\|\mathbf{a}^*\|_2^2\right\}$

Proof. Recall the dynamics of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$.

$$\begin{aligned}\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* &= \left(1 - \frac{\eta(\pi-1)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta(g(\phi^t)-1)}{2\pi}\|\mathbf{a}^*\|_2^2 + \frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top \mathbf{a}^*\right)^2 - \left(\mathbf{1}^\top \mathbf{a}^*\right)\left(\mathbf{1}^\top \mathbf{a}^t\right)\right) \\ &\geq \left(1 - \frac{\eta(\pi-1)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta(g(\phi^t)-1)}{2\pi}\|\mathbf{a}^*\|_2^2\end{aligned}$$

where the inequality is due to Lemma 5.4. If $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \geq \frac{g(\phi^t)-1}{\pi-1}\|\mathbf{a}^*\|_2^2$,

$$\begin{aligned}\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* &\geq \left(1 - \frac{\eta(\pi-1)}{2\pi}\right)\frac{g(\pi^t)-1}{\pi-1}\|\mathbf{a}^*\|_2^2 + \frac{\eta(g(\phi^t))}{\pi-1}\|\mathbf{a}^*\|_2^2 \\ &= \frac{g(\phi^t)-1}{\pi-1}\|\mathbf{a}^*\|_2^2.\end{aligned}$$

If $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \leq \frac{g(\phi^t)-1}{\pi-1}\|\mathbf{a}^*\|_2^2$, simple algebra shows $\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^*$ increases by at least

$$\eta\left(\frac{g(\phi^t)-1}{\pi-1}\|\mathbf{a}^*\|_2^2 - \left(\mathbf{a}^t\right)^\top \mathbf{a}^*\right).$$

□

A simple corollary is $\mathbf{a}^\top \mathbf{a}^*$ is uniformly lower bounded.

Corollary C.1. *For all* $t = 1, 2, \ldots$, $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \geq \min\left\{\left(\mathbf{a}^0\right)^\top \mathbf{a}^*, \frac{g(\phi^0)-1}{\pi-1}\|\mathbf{a}^*\|_2^2\right\}$.

This lemma also gives an upper bound of number of iterations to make $\mathbf{a}^\top \mathbf{a}^* = \Theta\left(\|\mathbf{a}^*\|_2^2\right)$.

Corollary C.2. *If* $g(\phi) - 1 = \Omega(1)$*, then after* $\frac{1}{\eta}$ *iterations,* $\mathbf{a}^\top \mathbf{a}^* = \Theta\left(\|\mathbf{a}^*\|_2^2\right)$.

Proof. Note if $g(\phi) - 1 = \Omega(1)$ and $\mathbf{a}^\top \mathbf{a}^* \leq \frac{1}{2} \cdot \frac{g(\phi)}{\pi-1}\|\mathbf{a}^*\|_2^2$, each iteration $\mathbf{a}^\top \mathbf{a}^*$ increases by $\eta\frac{g(\phi)}{\pi-1}\|\mathbf{a}^*\|_2^2$. □

We also need an upper bound of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$.

Lemma C.2. *For* $t = 0, 1, \ldots$, $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \leq \left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2$.

Proof. Without loss of generality, assume $\|\mathbf{w}^*\|_2 = 1$. Again, recall the dynamics of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$.

$$\begin{aligned}\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* &= \left(1 - \frac{\eta(\pi-1)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta(g(\phi^t)-1)}{2\pi}\|\mathbf{a}^*\|_2^2 + \frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top \mathbf{a}^*\right)^2 - \left(\mathbf{1}^\top \mathbf{a}^*\right)\left(\mathbf{1}^\top \mathbf{a}^t\right)\right) \\ &\leq \left(1 - \frac{\eta(\pi-1)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta(\pi-1)}{2\pi}\|\mathbf{a}^*\|_2^2 + \frac{\eta(\pi-1)}{2\pi}\left(\mathbf{1}^\top \mathbf{a}^*\right)^2.\end{aligned}$$

Now we prove by induction, suppose the conclusion holds at iteration t, $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \leq \|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2$. Plugging in we have the desired result. □

C.2. Convergence of Phase I

In this section we prove the convergence of Phase I.

Proof of Convergence of Phase I. Lemma C.3 implies after $O\left(\frac{1}{\cos\phi^0\beta^0}\right)$ iterations, $\cos\phi^t = \Omega(1)$, which implies $\frac{g(\phi^t)-1}{\pi-1} = \Omega(1)$. Using Corollary C.2, we know after $O\left(\frac{1}{\eta}\right)$ iterations we have $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \left\|\mathbf{w}^*\right\| = \Omega\left(\left\|\mathbf{w}^*\right\|_2^2 \left\|\mathbf{a}^*\right\|_2^2\right)$. □

The main ingredient of the proof of phase I is the follow lemma where we use a joint induction argument to show the convergence of ϕ^t and a uniform upper bound of $\left\|\mathbf{v}^t\right\|_2$.

Lemma C.3. *Let* $\beta^0 = \min\left\{\left(\mathbf{a}^0\right)^\top \mathbf{a}^*, \left(g(\phi^0)-1\right)\left\|\mathbf{a}^*\right\|_2^2\right\}\left\|\mathbf{w}^*\right\|_2^2$. *If the step size satisfies* $\eta \le \min\left\{\frac{\beta^*\cos\phi^0}{8\left(\|\mathbf{a}^*\|_2^2+(\mathbf{1}^\top\mathbf{a}^*)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{\cos\phi^0}{\left(\|\mathbf{a}^*\|_2^2+(\mathbf{1}^\top\mathbf{a}^*)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{2\pi}{k+\pi-1}\right\}$, *we have for* $t = 0, 1, \ldots$

$$\sin^2\phi^t \le \left(1-\eta\cdot\frac{\cos\phi^0\beta^0}{8}\right)^t \text{ and } \left\|\mathbf{v}^t\right\|_2 \le 2.$$

Proof. We prove by induction. The initialization ensure when $t=0$, the conclusion is correct. Now we consider the dynamics of $\left\|\mathbf{v}^t\right\|_2^2$. Note because the gradient of $\mathbf{v}$ is orthogonal to $\mathbf{v}$ (Salimans & Kingma, 2016), we have a simple dynamic of $\left\|\mathbf{v}^t\right\|_2^2$.

$$\begin{aligned}
\left\|\mathbf{v}^t\right\|_2^2 =& \left\|\mathbf{v}^{t-1}\right\|_2^2 + \eta^2\left\|\frac{\partial\ell(\mathbf{v},\mathbf{a})}{\partial\mathbf{v}}\right\|_2^2 \\
=& \left\|\mathbf{v}^{t-1}\right\|_2^2 + \eta^2\left(\frac{\left(\mathbf{a}^t\right)^\top\mathbf{a}^*\left(\pi-\phi^{t-1}\right)}{2\pi}\right)^2\frac{\sin^2\phi^t\left\|\mathbf{w}^*\right\|_2^2}{\left\|\mathbf{v}^t\right\|_2^2} \\
\le& \left\|\mathbf{v}^{t-1}\right\|_2^2 + \eta^2\left(\left\|\mathbf{a}^*\right\|_2^2 + \left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\left\|\mathbf{w}^*\right\|_2^2\sin^2\phi^{t-1} \\
=& 1+\eta^2\left(\left\|\mathbf{a}^*\right\|_2^2 + \left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\left\|\mathbf{w}^*\right\|_2^2\sum_{i=1}^{t-1}\sin^2\phi^i \\
\le& 1+\eta^2\left(\left\|\mathbf{a}^*\right\|_2^2 + \left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\left\|\mathbf{w}^*\right\|_2^2\frac{8}{\eta\cos\phi^0\beta^0} \\
\le& 2
\end{aligned}$$

where the first inequality is by Lemma C.2 and the second inequality we use our induction hypothesis. Recall $\lambda^t = \frac{\left\|\mathbf{w}^*\right\|_2\left(\left(\mathbf{a}^t\right)^\top\mathbf{a}^*\right)\left(\pi-\phi^t\right)}{2\pi\left\|\mathbf{v}^t\right\|_2^2}$. The uniform upper bound of $\left\|\mathbf{v}\right\|_2$ and the fact that $\phi^t \le \pi/2$ imply a lower bound $\lambda^t \ge \frac{\beta^0}{8}$. Plugging in Lemma 5.5, we have

$$\sin^2\phi^{t+1} \le \left(1-\eta\frac{\cos\phi^0\beta^0}{8}\right)\sin^2\phi^t \le \left(1-\eta\frac{\cos\phi^0\beta^0}{8}\right)^{t+1}.$$

We finish our joint induction proof. □

C.3. Analysis of Phase II

In this section we prove the convergence of phase II and necessary auxiliary lemmas.

Proof of Convergence of Phase II. At the beginning of Phase II, $\left(\mathbf{a}^{T_1}\right)^\top\mathbf{a}^*\left\|\mathbf{w}^*\right\| = \Omega\left(\left\|\mathbf{w}^*\right\|_2^2\left\|\mathbf{a}^*\right\|_2^2\right)$ and $g(\phi^{T_1})-1 = \Omega(1)$. Therefore, Lemma C.1 implies for all $t = T_1, T_1+1, \ldots$, $\left(\mathbf{a}^t\right)^\top\mathbf{a}^*\left\|\mathbf{w}^*\right\| = \Omega\left(\left\|\mathbf{w}^*\right\|_2^2\left\|\mathbf{a}^*\right\|_2^2\right)$. Combining with the fact that $\left\|\mathbf{v}\right\|_2 \le 2$ (c.f. Lemma C.3), we obtain a lower bound $\lambda_t \ge \Omega\left(\left\|\mathbf{w}^*\right\|_2^2\left\|\mathbf{a}^*\right\|_2^2\right)$ We also know that $\cos\phi^{T_1} = \Omega(1)$

and $\cos\phi^t$ is monotinically increasing (c.f. Lemma 5.2), so for all $t = T_1, T_1+1, \ldots$, $\cos\phi^t = \Omega(1)$. Plugging in these two lower bounds into Theorem 5.5, we have

$$\sin^2\phi^{t+1} \leq \left(1 - \eta C \left\|\mathbf{w}^*\right\|_2^2 \left\|\mathbf{a}^*\right\|_2^2\right) \sin^2\phi^t.$$

for some absolute constant C. Thus, after $O\left(\frac{1}{\eta\|\mathbf{w}^*\|_2^2\|\mathbf{a}^*\|_2^2}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations, we have $\sin^2\phi^t \leq \min\left\{\epsilon^{10}, \left(\epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}\right)^{10}\right\}$, which implies $\pi - g(\phi^t) \leq \min\left\{\epsilon, \epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}\right\}$. Now using Lemma C.4,Lemma C.5 and Lemma C.6, we have after $\widetilde{O}\left(\frac{1}{\eta k}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations $\ell(\mathbf{v},\mathbf{a}) \leq C_1\epsilon\left\|\mathbf{a}^*\right\|_2^2\left\|\mathbf{w}^*\right\|_2^2$ for some absolute constant C_1. Rescaling ϵ properly we obtain the desired result. □

C.3.1. Technical Lemmas for Analyzing Phase II

In this section we provide some technical lemmas for analyzing Phase II. Because of the positive homogeneity property, without loss of generality, we assume $\left\|\mathbf{w}^*\right\|_2 = 1$.

Lemma C.4. *If $\pi - g(\phi^0) \leq \epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}$, after $T = O\left(\frac{1}{\eta k}\log\left(\frac{|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^0|}{\epsilon\|\mathbf{a}^*\|_2}\right)\right)$ iterations, $\left|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^T\right| \leq 2\epsilon\left\|\mathbf{a}^*\right\|_2$.*

Proof. Recall the dynamics of $\mathbf{1}^\top\mathbf{a}^t$.

$$\begin{aligned}\mathbf{1}^\top\mathbf{a}^{t+1} &= \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\mathbf{1}^\top\mathbf{a}^t + \frac{\eta(k+g(\phi^t)-1)}{2\pi}\mathbf{1}^\top\mathbf{a}^* \\ &= \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\mathbf{1}^\top\mathbf{a}^t + \frac{\eta(k+g(\phi^t)-1)}{2\pi}\mathbf{1}^\top\mathbf{a}^*.\end{aligned}$$

Assume $\mathbf{1}^\top\mathbf{a}^* > 0$ (the other case is similar). By Lemma 5.4 we know $\mathbf{1}^\top\mathbf{a}^t < \mathbf{1}^\top\mathbf{a}^*$ for all t. Consider

$$\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^{t+1} = \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\left(\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^*\right) + \frac{\eta(\pi - g(\phi^t))}{2\pi}\mathbf{1}^\top\mathbf{a}^*.$$

Therefore we have

$$\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^{t+1} - \frac{(\pi - g(\phi^t))\mathbf{1}^\top\mathbf{a}^*}{k+\pi-1} = \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\left(\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^* - \frac{(\pi - g(\phi^t))\mathbf{1}^\top\mathbf{a}^*}{k+\pi-1}\right).$$

After $T = O\left(\frac{1}{\eta k}\log\left(\frac{|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^0|}{\epsilon\|\mathbf{a}^*\|_2}\right)\right)$ iterations, we have $\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^t - \frac{(\pi - g(\phi^t))\mathbf{1}^\top\mathbf{a}^*}{k+\pi-1} \leq \epsilon\left\|\mathbf{a}^*\right\|_2$, which implies$\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^t \leq 2\epsilon\left\|\mathbf{a}^*\right\|_2$. □

Lemma C.5. *If $\pi - g(\phi^0) \leq \epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}$ and $\left|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^0\right| \leq \frac{\epsilon}{k}\left\|\mathbf{a}^*\right\|_2$, then after $T = O\left(\frac{1}{\eta}\log\left(\frac{\left\|\mathbf{a}^* - \mathbf{a}^0\right\|_2}{\epsilon\|\mathbf{a}^*\|_2}\right)\right)$ iterations, $\left\|\mathbf{a}^* - \mathbf{a}^0\right\|_2 \leq C\epsilon\left\|\mathbf{a}^*\right\|_2$ for some absolute constant C.*

Proof. We first consider the inner product

$$\begin{aligned}&\left\langle\frac{\partial\ell(\mathbf{v}^t,\mathbf{a}^t)}{\mathbf{a}^t}, \mathbf{a}^t - \mathbf{a}^*\right\rangle \\ &= \frac{\pi-1}{2\pi}\left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 - \frac{g(\phi^t)-\pi}{2\pi}(\mathbf{a}^*)^\top\left(\mathbf{a}^t - \mathbf{a}^*\right) + \left(\mathbf{a}^t - \mathbf{a}^*\right)\mathbf{1}\mathbf{1}^\top\left(\mathbf{a}^\top - \mathbf{a}^*\right) \\ &\geq \frac{\pi-1}{2\pi}\left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 - \frac{g(\phi^t)-\pi}{2\pi}\left\|\mathbf{a}^*\right\|_2\left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2.\end{aligned}$$

Next we consider the squared norm of gradient

$$\begin{aligned}\left\|\frac{\partial \ell\left(\mathbf{v}, \mathbf{a}\right)}{\partial \mathbf{a}}\right\|_2^2 =& \frac{1}{4\pi^2}\left\|\left(\pi-1\right)\left(\mathbf{a}^t-\mathbf{a}^*\right)+\left(\pi-g(\phi^t)\right)\mathbf{a}^*+\mathbf{1}\mathbf{1}^\top\left(\mathbf{a}^t-\mathbf{a}^*\right)\right\|_2^2 \\ \leq& \frac{3}{4\pi^2}\left(\left(\pi-1\right)^2\left\|\mathbf{a}^t-\mathbf{a}^*\right\|_2^2+\left(\pi-g(\phi^t)\right)^2\left\|\mathbf{a}^*\right\|_2^2+k^2\left(\mathbf{1}^\top\mathbf{a}^t-\mathbf{1}^\top\mathbf{a}^*\right)^2\right).\end{aligned}$$

Suppose $\left\|\mathbf{a}^t-\mathbf{a}^*\right\|_2 \leq \epsilon\left\|\mathbf{a}^*\right\|_2$, then

$$\begin{aligned}\langle\frac{\partial \ell\left(\mathbf{v}^t, \mathbf{a}^t\right)}{\mathbf{a}^t}, \mathbf{a}^t-\mathbf{a}^*\rangle \geq& \frac{\pi-1}{2\pi}\left\|\mathbf{a}^t-\mathbf{a}^*\right\|_2^2-\frac{\epsilon^2}{2\pi}\left\|\mathbf{a}^*\right\|_2^2 \\ \left\|\frac{\partial \ell\left(\mathbf{v}, \mathbf{a}\right)}{\partial \mathbf{a}}\right\|_2^2 \leq& 3\epsilon^2\left\|\mathbf{a}^*\right\|_2^2.\end{aligned}$$

Therefore we have

$$\begin{aligned}&\left\|\mathbf{a}^{t+1}-\mathbf{a}^*\right\|_2^2 \leq\left(1-\frac{\eta\left(\pi-1\right)}{2\pi}\right)\left\|\mathbf{a}^t-\mathbf{a}^*\right\|_2^2+4\eta\epsilon^2\left\|\mathbf{a}\right\|^2 \\ \Rightarrow&\left\|\mathbf{a}^{t+1}-\mathbf{a}^*\right\|_2^2-\frac{8\left(\pi-1\right)\epsilon^2\left\|\mathbf{a}^*\right\|_2^2}{\pi-1} \leq\left(1-\frac{\eta\left(\pi-1\right)}{2\pi}\right)\left(\left\|\mathbf{a}^t-\mathbf{a}^*\right\|_2^2-\frac{8\left(\pi-1\right)\epsilon^2\left\|\mathbf{a}^*\right\|_2^2}{\pi-1}\right).\end{aligned}$$

Thus after $O\left(\frac{1}{\eta}\left(\frac{1}{\epsilon}\right)\right)$ iterations, we must have $\left\|\mathbf{a}^{t+1}-\mathbf{a}^*\right\|_2^2 \leq C\epsilon\left\|\mathbf{a}^*\right\|_2$ for some large absolute constant C. Rescaling ϵ, we obtain the desired result. □

Lemma C.6. *If $\pi-g(\phi) \leq \epsilon$ and $\left\|\mathbf{a}-\mathbf{a}^*\left\|\mathbf{w}^*\right\|_2\right\| \leq \epsilon\left\|\mathbf{a}^*\right\|_2\left\|\mathbf{w}^*\right\|_2$, then the population loss satisfies $\ell\left(\mathbf{v}, \mathbf{a}\right) \leq C\epsilon\left\|\mathbf{a}^*\right\|_2^2\left\|\mathbf{w}^*\right\|_2^2$ for some constant $C>0$.*

Proof. The result follows by plugging in the assumptions in Theorem 3.1. □

D. Proofs of Initialization Scheme

Proof of Theorem 4.2. The proof of the first part of Theorem 4.2 just uses the symmetry of unit sphere and ball and the second part is a direct application of Lemma 2.5 of (Hardt & Price, 2014). Lastly, since $\mathbf{a}^0 \sim \mathcal{B}\left(\mathbf{0}\frac{\left|\mathbf{1}^\top\mathbf{a}^*\right|}{\sqrt{k}}\right)$, we have $\mathbf{1}^\top\mathbf{a}^0 \leq\left\|\mathbf{a}^0\right\|_1 \leq \sqrt{k}\left\|\mathbf{a}^0\right\|_2 \leq\left|\mathbf{1}^\top\mathbf{a}^*\right|\left\|\mathbf{w}^*\right\|_2$ where the second inequality is due to Hölder's inequality. □

E. Proofs of Converging to Spurious Local Minimum

Proof of Theorem 4.3. The main idea is similar to Theorem 4.1 but here we show $\mathbf{w} \to -\mathbf{w}^*$ (without loss of generality, we assume $\left\|\mathbf{w}^*\right\|_2=1$). Different from Theorem 4.1, here we need to prove the invariance $\mathbf{a}^\top\mathbf{a}^*<0$, which implies our desired result. We prove by induction, suppose $\left(\mathbf{a}^t\right)^\top\mathbf{a}^*>0$, $\left|\mathbf{1}^\top\mathbf{a}^t\right| \leq\left|\mathbf{1}^\top\mathbf{a}^*\right|$, $g\left(\phi^0\right) \leq \frac{-2\left(\mathbf{1}^\top\mathbf{a}\right)^2}{\left\|\mathbf{a}^*\right\|_2^2}+1$ and $\eta<\frac{k+\pi-1}{2\pi}$. Note $\left|\mathbf{1}^\top\mathbf{a}^t\right| \leq\left|\mathbf{1}^\top\mathbf{a}^*\right|$ are satisfied by Lemma 5.4 and $g\left(\phi^0\right) \leq \frac{-2\left(\mathbf{1}^\top\mathbf{a}\right)^2}{\left\|\mathbf{a}^*\right\|_2^2}+1$ by our initialization condition and induction hypothesis that implies ϕ^t is increasing. Recall the dynamics of $\left(\mathbf{a}^t\right)^\top\mathbf{a}^*$.

$$\begin{aligned}\left(\mathbf{a}^{t+1}\right)^\top\mathbf{a}^* =& \left(1-\frac{\eta\left(\pi-1\right)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top\mathbf{a}^*+\frac{\eta\left(g\left(\phi^t\right)-1\right)}{2\pi}\left\|\mathbf{a}^*\right\|_2^2+\frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top\mathbf{a}^*\right)^2-\left(\mathbf{1}^\top\mathbf{a}^t\right)\left(\mathbf{1}^\top\mathbf{a}^*\right)\right) \\ \leq& \frac{\eta\left(\left(g(\phi^t)-1\right)\left\|\mathbf{a}^*\right\|_2+2\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)}{2\pi}<0\end{aligned}$$

where the first inequality we used our induction hypothesis on inner product between $\mathbf{a}^t$ and $\mathbf{a}^*$ and $\left|\mathbf{1}^\top\mathbf{a}^t\right| \leq\left|\mathbf{1}^\top\mathbf{a}^*\right|$ and the second inequality is by induction hypothesis on ϕ^t. Thus when gradient descent algorithm converges, according Lemma 5.1, $\theta\left(\mathbf{v}, \mathbf{w}^*\right)=\pi, \mathbf{a}=\left(\mathbf{1}\mathbf{1}^\top+\left(\pi-1\right)\mathbf{I}\right)^{-1}\left(\mathbf{1}\mathbf{1}^\top-\mathbf{I}\right)\left\|\mathbf{w}^*\right\|_2\mathbf{a}^*$. Plugging these into Theorem 3.1, with some routine algebra, we show $\ell\left(\mathbf{v}, \mathbf{a}\right)=\Omega\left(\left\|\mathbf{w}^*\right\|_2^2\left\|\mathbf{a}^*\right\|_2^2\right)$. □