A. Preliminaries

We first present the following proposition, which computes the explicit form of the loss function and the gradient of the loss function with respect to a and w.

Proposition 9 (Du et al. (2017))**.** *Let $\phi \in [0, \pi]$ be the angle between w and w^*. Then, the loss function $L\left(w, a\right)$ and the gradient w.r.t $\left(w, a\right)$, i.e., $\nabla_a L\left(w, a\right)$ and $\nabla_w L\left(w, a\right)$ have the following analytic forms.*

$$\begin{aligned} L\left(w,a\right) = & \frac{1}{2}[\frac{\left(\pi-1\right)\|w^*\|_2^2}{2\pi}\|a^*\|_2^2 + \frac{\left(\pi-1\right)}{2\pi}\|a\|_2^2 - \frac{\|w^*\|_2}{\pi}\left(g\left(\phi\right)-1\right)a^\top a^* \\ & + \frac{\|w^*\|_2^2}{2\pi}\left(\mathbf{1}^\top a^*\right)^2 + \frac{1}{2\pi}\left(\mathbf{1}^\top a\right)^2 - \frac{\|w^*\|_2}{\pi}\mathbf{1}^\top a^* a^\top \mathbf{1}], \\ \nabla_a L\left(w,a\right) = & \frac{1}{2\pi}\left(\mathbf{1}\mathbf{1}^\top + \left(\pi-1\right)I\right)a - \frac{1}{2\pi}\left(\mathbf{1}\mathbf{1}^\top + \left(g\left(\phi\right)-1\right)I\right)a^*\|w^*\|_2, \\ \nabla_w L\left(w,a\right) = & -\frac{a^\top a^*\left(\pi-\phi\right)}{2\pi}w^* + \left(\frac{\|a\|}{2} + \frac{\sum_{i\neq j}a_i a_j}{2\pi} - \frac{a^\top a^* \sin\phi}{2\pi}\frac{\|w^*\|_2}{\|w\|_2} - \frac{\sum_{i\neq j}a_i a_j^*}{2\pi}\frac{\|w^*\|_2}{\|w\|_2}\right)w, \end{aligned}$$

where $g\left(\phi\right) = \left(\pi-\phi\right)\cos\phi + \sin\phi$.

As can be seen, both $\nabla_w L\left(w, a\right)$ and $\nabla_a L\left(w, a\right)$ depend on ϕ, which is the angle between w and w^*. After injecting noise, we have

$$\mathbb{E}\phi_\xi = \mathbb{E}\left\{\arccos\frac{\left(w+\xi_t\right)^\top w^*}{\|w+\xi_t\|_2\|w^*\|_2}\right\} \neq \arccos\frac{w^\top w^*}{\|w\|_2\|w^*\|_2} = \phi.$$

As a direct result, we have

$$\begin{aligned} \mathbb{E}_{\xi_t,\epsilon_t}\nabla_a L\left(w_t+\xi_t, a_t+\epsilon_t\right) \neq \nabla_a L\left(w_t, a_t\right), \\ \mathbb{E}_{\xi_t,\epsilon_t}\nabla_w L\left(w_t+\xi_t, a_t+\epsilon_t\right) \neq \nabla_w L\left(w_t, a_t\right), \end{aligned}$$

which indicate that the perturbed gradient $\nabla_a\left(w_t+\xi_t, a_t+\epsilon_t\right)$, $\nabla_w\left(w_t+\xi_t, a_t+\epsilon_t\right)$ are biased estimates of the gradient (as we mentioned in Section 2).

For notational simplicity, we introduce an auxiliary iterate $\widetilde{w}_{t+1}$ and rewrite our perturbed GD algorithm as follows.

$$\begin{aligned} a_{t+1} &= a_t - \eta\nabla_a L\left(w_t+\xi_t, a_t+\epsilon_t\right), \\ \widetilde{w}_{t+1} &= w_t - \eta\left(I - w_t w_t^\top\right)\nabla_w L\left(w_t+\xi_t, a_t+\epsilon_t\right), \\ w_{t+1} &= \mathrm{Proj}_{\mathbb{S}_0(1)}\left(\widetilde{w}_{t+1}\right). \end{aligned}$$

In the later proof, we use $\mathcal{F}_t = \sigma\{\left(w_\tau, a_\tau\right)\big|\tau \leq t\}$. as the sigma algebra generated by previous t iterations and $V(\rho) = \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)}\rho^p$ as the volume of p-dimensional ball $\mathbb{B}_0(\rho)$.

B. d-Dimensional Polar Coordinate and Some Important Lemmas

To calculate the expectation in our following analysis, we often need the d-dimension polar coordinate system. Specifically, if we write a vector ν under Cartesian coordinate as $\nu = (\nu_1, \nu_2, ..., \nu_d)$, then under the polar coordinate, ν can be written as $\nu = (r, \theta_1, \theta_2, ..., \theta_{d-1})$, where

$$\begin{aligned} \nu_1 &= r\cos(\theta_1), \\ \nu_i &= r\Pi_{j=1}^{i-1}\sin(\theta_j)\cos(\theta_i),\ i = 2, ..., d-1, \\ \nu_d &= r\Pi_{j=1}^{d-1}\sin(\theta_j), \end{aligned}$$

where $r \geq 0$, $0 \leq \theta_i \in [0, \pi],\ i = 1, 2, ..., d-2$, $\theta_{d-1} \in [0, 2\pi]$.

To use polar coordinate to calculate integral, we also need the following Jacobian Matrix.

$$\frac{\partial(\nu_1, \nu_2, \nu_3 ..., \nu_d)}{\partial(r, \theta_1, \theta_2, ..., \theta_{d-1})} = r^{d-1} \sin^{d-2} \theta_1 \sin^{d-3} \theta_2 \cdots \sin \theta_{d-2}.$$

The following important equation is required.

$$I_n \triangleq \int_0^{\pi} \sin^n (x)\, dx = \frac{\sqrt{\pi}\Gamma\left(\frac{1+n}{2}\right)}{\Gamma\left(1+\frac{n}{2}\right)}.$$

Then we have the following useful lemma here.

Lemma 10. *Let $f(\theta)$ be a positive bounded function defined on $[0,\pi]$, that is there exits a constant $C \geq 0$ such that $0 \leq f(\theta) \leq C$, $\forall \theta \in [0,\pi]$. For any $\epsilon > 0$ and positive integer d, define*

$$A_d(f) \triangleq \frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{d/2}} \int_0^{\pi} \cdots \int_0^{\pi} \int_0^{2\pi} f(\theta_1) \sin^{d-2} \theta_1 \sin^{d-3} \theta_2 \cdots \sin \theta_{d-2} d\theta_1 \cdots d\theta_{d-1},$$

$$M_d \triangleq \int_0^1 r^{d-1} dr, \quad L_d(\epsilon) \triangleq \int_0^{1-\epsilon} r^{d-1} dr, \quad H_d(\epsilon) \triangleq \int_0^{1+\epsilon} r^{d-1} dr.$$

Then we have

$$A_d(f) L_d(\epsilon) + O(\epsilon d) > A_d(f) M_d > A_d(f) H_d(\epsilon) - O(\epsilon d). \tag{10}$$

Proof. For simplicity, we only give the proof of the left side. The proof of the right side follows similar lines.

We compute $A_d(f)$.

$$\begin{aligned} 0 \leq A_d(f) &= \frac{2\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}-1}} I_{d-3} \cdots I_1 \int_0^{\pi} f(\theta_1) \sin^{d-2} \theta_1 d\theta_1 \\ &\leq \frac{2\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}-1}} C I_{d-2} \cdots I_1 \leq Cd. \end{aligned}$$

We give the lower bound on $L_d(\epsilon) - M_d$.

$$L_d(\epsilon) - M_d = \int_0^{1-\epsilon} r^{d-1} dr - \int_0^1 r^{d-1} dr = \frac{(1-\epsilon)^d - 1}{d} \geq -\epsilon.$$

Hence, we have

$$(L_d(\epsilon) - M_d) A_d(f) \geq -Cd\epsilon. \tag{11}$$

□

C. Proof for Phase I

C.1. Proof of Theorem 5

Proof. We first derive the dissipativity w.r.t a in region

$$\mathcal{A}_{C_2,C_3} = \{(w,a)\ |{-4}\left(\mathbf{1}^\top a^*\right)^2 \leq \mathbf{1}^\top a^* \mathbf{1}^\top a - \left(\mathbf{1}^\top a^*\right)^2 \leq \frac{C_2}{p}\|a^*\|_2^2, \quad a^\top a^* \leq \frac{C_3}{p}\|a^*\|_2^2 \text{ or} \|a - a^*/2\|_2^2 \geq \|a^*\|_2^2, \quad w \in \mathbb{S}_0(1)\}.$$

Assume $(w,a) \in \mathcal{A}_{C_2,C_3}$, and $a^\top a^* \leq \frac{C_3}{p}\|a^*\|_2^2$, we have

$$\|a - a^*\|_2^2 \geq \left(1 - \frac{2C_3}{p}\right)\|a^*\|_2^2.$$

Combining the above inequality with $\mathbb{E}g\left(\phi_\xi\right) \geq 1 + \frac{C}{p}$ in Lemma 12, we get

$$
\begin{aligned}
\left\langle -\mathbb{E}_{\xi,\epsilon}\nabla_a L\left(w+\xi, a+\epsilon\right), a^*-a\right\rangle &= \frac{1}{2\pi}\left(\mathbf{1}^\top a - \mathbf{1}^\top a^*\right)^2 + \frac{1}{2\pi}\left(\left(\pi-1\right)a - \left(\mathbb{E}_\xi g\left(\phi_\xi\right)-1\right)a^*\right)^\top\left(a-a^*\right) \\
&= \frac{1}{2\pi}\left(\mathbf{1}^\top a - \mathbf{1}^\top a^*\right)^2 + \frac{1}{2\pi}\left(\pi - \mathbb{E}_\xi g\left(\phi_\xi\right)\right)a^\top\left(a-a^*\right) + \frac{\mathbb{E}_\xi g\left(\phi_\xi\right)-1}{2\pi}\|a-a^*\|_2 \\
&\geq -\frac{1}{2\pi}\left(\pi - \mathbb{E}_\xi g\left(\phi_\xi\right)\right)a^\top a^* + \frac{\mathbb{E}_\xi g\left(\phi_\xi\right)-1}{2\pi}\|a-a^*\|_2^2 \\
&= -\frac{1}{2\pi}\left(\pi - \mathbb{E}_\xi g\left(\phi_\xi\right)\right)a^\top a^* + \frac{\mathbb{E}_\xi g\left(\phi_\xi\right)-1}{4\pi}\|a-a^*\|_2^2 + \frac{\mathbb{E}_\xi g\left(\phi_\xi\right)-1}{4\pi}\|a-a^*\|_2^2 \\
&\geq -\frac{C_3}{2p}\|a^*\|_2^2 + \frac{C}{4\pi p}\left(1-\frac{2C_3}{p}\right)\|a^*\|_2^2 + \frac{C}{4\pi p}\|a-a^*\|_2^2 \geq \frac{C}{4\pi p}\|a-a^*\|_2^2
\end{aligned}
$$

for some constant $C_3 \leq \frac{C}{4\pi}$.

Moreover, if $(w,a) \in \mathcal{A}_{C_2,C_3}$ and $\|a - a^*/2\|_2^2 \geq \|a^*\|_2^2$, we have

$$a^\top\left(a-a^*\right) > 0.$$

Following the similar lines above, we have the same results. Thus, the dissipativity w.r.t a holds in region $\mathcal{A}_{C_2,C_3}$

Next, we derive the dissipativity w.r.t w in region

$$\mathcal{K}_{C_4,m,M} = \{(w,a) \mid a^\top a^* \in [m,M], \quad w^\top w^* \geq C_4, \quad w \in \mathbb{S}_0\left(1\right)\}.$$

Assume $(w,a) \in \mathcal{K}_{C_4,m,M}$ for some constant $C_4 \in (-1,1]$ and $0 < m < M$. We could write w as $w = \sum_{i=1}^p c_i v_i$, where $\{v_i\}_{i=1}^p$ is an orthonormal basis for $\mathbb{R}^p$, $w^* = v_1$, $\|w\|_2 = 1$ and $c_1 \geq C_4$. Without loss of generality, we assume $w^* = (1,0,...,0)^\top$. We have the following equation.

$$
\begin{aligned}
(I - ww^\top)(w^*-w) &= w^* - w - ww^\top w^* + ww^\top w = w^* - ww^\top w^* \\
&= (1,0,...,0)^\top - (c_1^2, c_1c_2, ..., c_1c_p)^\top = (1-c_1^2, -c_1c_2, ..., -c_1c_p)^\top,
\end{aligned}
$$

The norm of this vector is

$$
\begin{aligned}
\left\|(I-ww^\top)(w^*-w)\right\|_2 &= \sqrt{(1-c_1^2)^2 + c_1^2(c_2^2 + ... + c_p^2)} \\
&= \sqrt{(1-c_1^2)^2 + c_1^2(1-c_1^2)} = \sqrt{1-c_1^2}
\end{aligned}
$$

By $\mathbb{E}_\xi\left(\phi_\xi\right) \leq \frac{3\pi}{4}$ in Lemma 12, we have

$$
\begin{aligned}
&\left\langle -\mathbb{E}_{\xi,\epsilon}\left(I - ww^\top\right)\nabla_w L\left(w+\xi, a+\epsilon\right), w^*-w\right\rangle = \frac{a^\top a^*\left(\pi - \mathbb{E}_\xi\phi_\xi\right)}{2\pi}\left(1-c_1^2\right) \\
&+ \mathbb{E}_{\xi,\epsilon}\left(w^* - w^\top w^* w\right)^T\left(\frac{\|a+\epsilon\|_2^2}{2} + \frac{\sum_{i\neq j}\left(a_i+\epsilon_i\right)\left(a_j+\epsilon_j\right)}{2\pi} - \frac{\left(a+\epsilon\right)^\top a^*\sin\phi_\xi}{2\pi}\frac{1}{\|w+\xi\|_2} - \frac{\sum_{i\neq j}\left(a_i+\epsilon_i\right)a_j^*}{2\pi}\frac{1}{\|w+\xi\|_2}\right)\xi \\
&= \frac{a^\top a^*\left(\pi - \mathbb{E}_\xi\phi_\xi\right)}{2\pi}\left(1-c_1^2\right) + \mathbb{E}_\xi\left(w^* - w^\top w^* w\right)^T\left(-\frac{a^\top a^*\sin\phi_\xi}{2\pi}\frac{1}{\|w+\xi\|_2} - \frac{\sum_{i\neq j}a_i a_j^*}{2\pi}\frac{1}{\|w+\xi\|_2}\right)\xi
\end{aligned}
$$

We next show that

$$\mathbb{E}_\xi\frac{\left(w^* - w^\top w^* w\right)^T\xi}{\|w+\xi\|_2} = 0 \tag{12}$$

and

$$\mathbb{E}_\xi \frac{\sin\phi_\xi \left(w^* - w^\top w^* w\right)^T \xi}{\|w+\xi\|_2} \leq C\frac{\sqrt{1-c_1^2}}{\rho_w} \tag{13}$$

for some constant C.

For (12), recall that $V(\rho_w)$ is the volume of $\mathbb{B}_0(\rho_w)$. Then we have

$$\begin{aligned}
\mathbb{E}_\xi \frac{\left(w^* - w^\top w^* w\right)^\top \xi}{\|w+\xi\|_2} &= \int_{\mathbb{B}_0(\rho_w)} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|w+x\|_2} dx \\
&= \int_{\mathbb{B}_w(\rho_w)} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x - \left(w^* - w^\top w^* w\right)^\top w}{\|x\|_2} dx \\
&= \int_{\mathbb{B}_w(\rho_w)} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&= \int_{\mathbb{B}_w(\rho_w), (w^*-w^\top w^* w)^\top x > 0} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} \\
&\quad + \int_{\mathbb{B}_w(\rho_w), (w^*-w^\top w^* w)^\top x < 0} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx
\end{aligned}$$

For any x such that $\left(w^* - w^\top w^* w\right)^\top x > 0$, its symmetric point with respect to vector w is $\widetilde{x} = 2w^\top x w - x$. We further have

$$\left(w^* - w^\top w^* w\right)^\top \widetilde{x} = \left(w^* - w^\top w^* w\right)^\top \left(2w^\top x w - x\right) = -\left(w^* - w^\top w^* w\right)^\top x < 0.$$

By this symmetric property with respect to vector w, we know

$$\begin{aligned}
\mathbb{E}_\xi \frac{\left(w^* - w^\top w^* w\right)^\top \xi}{\|w+\xi\|_2} &= \int_{\mathbb{B}_0(\rho_w)} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|w+x\|_2} dx \\
&= \int_{\mathbb{B}_w(\rho_w), (w^*-w^\top w^* w)^\top x > 0} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} \\
&\quad + \int_{\mathbb{B}_w(\rho_w), (w^*-w^\top w^* w)^\top x < 0} \frac{1}{V(\rho_w)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&= 0.
\end{aligned}$$

Now we prove (13). Denote that $\phi_x = \angle(x, w^*)$. When $\rho_w > 1$, we have

$$
\begin{aligned}
&\mathbb{E}_\xi \frac{\sin\phi_\xi \left(w^* - w^\top w^* w\right)^\top \xi}{\|w+\xi\|_2} \\
&= \int_{\mathbb{B}_w(\rho_w)} \frac{\sin\phi_x}{V\left(\rho_w\right)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&= \int_{\mathbb{B}_w(\rho_w),(w^*-w^\top w^* w)^\top x>0} \frac{\sin\phi_x}{V\left(\rho_w\right)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} + \int_{\mathbb{B}_w(\rho_w),(w^*-w^\top w^* w)^\top x<0} \frac{\sin\phi_x}{V\left(\rho_w\right)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&\leq \int_{\mathbb{B}_0(\rho_w+1),(w^*-w^\top w^* w)^\top x>0} \frac{\sin\phi_x}{V\left(\rho_w\right)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx + \int_{\mathbb{B}_0(\rho_w-1),(w^*-w^\top w^* w)^\top x<0} \frac{\sin\phi_x}{V\left(\rho_w\right)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&\leq \int_{\mathbb{B}_0(\rho_w+1)\backslash\mathbb{B}_0(\rho_w-1),(w^*-w^\top w^* w)^\top x>0} \frac{1}{V\left(\rho_w\right)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&= \frac{\Gamma\left(\frac{p}{2}+1\right)}{\rho_w^p \pi^{p/2}} \left(\int_{\rho_w-1}^{\rho_w+1} r^{p-1} dr \int_0^{\pi/2} \sqrt{1-c_1^2}\cos(\theta_1)\sin^{p-2}\left(\theta_1\right) d\theta_1\right) \\
&\cdot \left(\int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \sin^{p-3}\theta_2 \cdots \sin\theta_{p-2} d\theta_1 \cdots d\theta_{p-1}\right) \qquad \text{(Convert to polar coordinate)} \\
&= \frac{\Gamma\left(\frac{p}{2}+1\right)}{\rho_w^p \pi^{p/2}} \frac{\left(\rho_w+1\right)^p - \left(\rho_w-1\right)^p}{p} \frac{\sqrt{1-c_1^2}}{p-1} I_{p-3}\cdots I_1 \\
&= \frac{\Gamma\left(\frac{p}{2}+1\right)}{\rho_w^p \pi^{p/2}} \frac{\left(\rho_w+1\right)^p - \left(\rho_w-1\right)^p}{p} \frac{\sqrt{1-c_1^2}}{p-1} \frac{\pi^{\frac{p-3}{2}}}{\Gamma\left(\frac{p-1}{2}\right)} = C\frac{\sqrt{1-c_1^2}}{\rho_w}
\end{aligned}
$$

Thus when $\rho_w \geq C\frac{M_a}{\gamma}$, combining (12) and (13), we have for any $\gamma > 0$

$$
-\frac{a^\top a^*}{2\pi}\mathbb{E}_\xi \frac{\sin\phi_\xi \left(w^* - w^\top w^* w\right)^T \xi}{\|w+\xi\|_2} - \frac{\sum_{i\neq j} a_i a_j^*}{2\pi}\mathbb{E}_\xi \frac{\left(w^* - w^\top w^* w\right)^T \xi}{\|w+\xi\|_2} \geq -\gamma.
$$

Then we have

$$
\left\langle -\mathbb{E}_{\xi,\epsilon}\left(I - ww^\top\right)\nabla_w L\left(w+\xi, a+\epsilon\right), w^* - w\right\rangle \geq \frac{m\left(1+C_4\right)}{16}\|w-w^*\|_2^2 - \gamma.
$$

□

C.2. Proof Sketch for Theorem 6

Proof Sketch. The next lemma shows that that our initialization (w_0, a_0) is guaranteed to fall in a superset of $\mathcal{A}_{C_2,C_3}$.

Lemma 11. *Given $a_0 \in \mathbb{B}_0\left(\frac{|\mathbf{1}^\top a^*|}{\sqrt{k}}\right)$ and $w_0 \in \mathbb{S}_0(1)$, we have for any constant $C_3 > 0$, $(w_0, a_0) \in \mathcal{A}_{C_3}$, where*

$$
\mathcal{A}_{C_3} = \left\{(w,a) \mid -4(\mathbf{1}^\top a^*)^2 \leq \mathbf{1}^\top a^* \mathbf{1}^\top a - (\mathbf{1}^\top a^*)^2 \leq \frac{C_3}{p}\|a^*\|_2^2,\ w \in \mathbb{S}_0(1)\right\}.
$$

Our subsequent analysis considers two cases: Case (1) $(w_0, a_0) \in \mathcal{A}_{C_2,C_3}$ and Case (2) $(w_0, a_0) \in \mathcal{A}_{C_3}\backslash\mathcal{A}_{C_2,C_3}$. Specifically, we first start with Case (1), and then show the algorithm will be able to escape from $\mathcal{A}_{C_2,C_3}$ in polynomial time and enter $\mathcal{A}_{C_3}\backslash\mathcal{A}_{C_2,C_3}$. Then we only need to proceed with Case (2).

Note that for $\mathcal{A}_{C_2,C_3}$, the dissipativity holds only for the perturbed gradient with respect to a. Though the dissipativity does not necessarily hold for w, we can show that the noise injection procedure guarantees a sufficiently accurate w for making progress in a, as shown in the next lemma.

Lemma 12. *Suppose $w, w^* \in \mathbb{S}_0(1)$ and $\xi \sim \mathrm{unif}(\mathbb{B}_0(\rho_w)) \in \mathbb{R}^p$. Define $\phi_\xi \triangleq \angle(w+\xi, w^*) \in [0,\pi]$, $\phi \triangleq \angle(w, w^*)$ and $g(\phi) = (\pi - \phi)cos(\phi) + sin(\phi)$. When $\rho_w \geq C_6 p^2$ for some constant C_6, there exists some constant C_8 such that*

$$1 + \frac{C_8}{p} \leq \mathbb{E}_\xi g(\phi_\xi) \leq \pi \quad \text{and} \quad \mathbb{E}_\xi \phi_\xi \leq \frac{3\pi}{4}$$

for all $\phi \in [0, \pi]$.

We remark that Lemma 12 is actually the key to the convergence analysis for Phase I. It helps prove both Theorems 5 and 6. The proof is highly non-trivial and very involved. See more details in Appendix C.4.2. Lemma 12 essentially shows that the noise injection prevents w from being attracted to v^*, and further prevents (w, a) from being attracted to the spurious local optimum.

We then analyze Case (1), where $(w_0, a_0) \in \mathcal{A}_{C_2, C_3}$.

Lemma 13. *Suppose $\rho_w^0 = C_w k p^2 \geq 1$, $\rho_a^0 = C_a$ and $(w_0, a_0) \in \mathcal{A}_{C_2,C_3}$. For any $\delta \in (0,1)$, we choose step size*

$$\eta = C_6 \Big(k^4 p^6 \cdot \max\Big\{1, p \log\frac{1}{\delta}\Big\}\Big)^{-1}$$

for some constant C_6. Then with at least probability $1 - \delta/3$, we have

$$m_a \leq a_t^\top a^* \leq M_a \quad \text{and} \quad (w_{\tau_{11}}, a_{\tau_{11}}) \in \mathcal{A}_{C_3} \backslash \mathcal{A}_{C_2,C_3} \tag{14}$$

for all t's such that $\tau_{11} \leq t \leq T = \widetilde{O}(\eta^{-2})$, where

$$\tau_{11} = \widetilde{O}\Big(\frac{p}{\eta} \log\frac{1}{\delta}\Big).$$

As can be seen, after τ_{11} iterations, the algorithm enters $\mathcal{A}_{C_3} \backslash \mathcal{A}_{C_2,C_3}$. Then our following analysis will only consider Case (2), where $(w_0, a_0) \in \mathcal{A}_{C_3} \backslash \mathcal{A}_{C_2,C_3}$. We remark that although Theorem 4 no longer guarantees the dissipativity of the perturbed gradient with respect to a, Lemma 13 can ensure the optimization error of a_t within Phase I to be nonincreasing as long as $t \geq \tau_{11}$ with high probability.

We then continue to characterize the optimization error of w_t. Recall that the noise injection prevents $-w^*$ from being attracted to $-w^*$. Thus, we can guarantee that w_t is sufficiently distant from $-w^*$ after sufficiently many iterations, as shown in the next lemma.

Lemma 14. *Suppose $\rho_w^0 = C_w k p^2 \geq 1$, $\rho_a^0 = C_a$, $(w_0, a_0) \in \mathcal{A}_{C_3} \backslash \mathcal{A}_{C_2,C_3}$ and $m_a \leq a_t^\top a^* \leq M_a$ holds for all t's. For any $\delta \in (0,1)$, we choose step size*

$$\eta = C_6 \Big(k^4 p^6 \cdot \max\Big\{1, p \log\frac{1}{\delta}\Big\}\Big)^{-1}$$

for some constant C_6. Then with at least probability $1 - \delta/3$, there exists

$$\tau_{12} = \widetilde{O}\Big(\frac{p}{\eta} \log\frac{1}{\eta} \log\frac{1}{\delta}\Big)$$

such that $w_{\tau_{12}}^T w^* \geq C_4$ for some constant $C_4 \in (-1, 0)$.

Lemma 14 implies that the algorithm eventually attains $\mathcal{K}_{C_4, m_a, M_a}$, where the dissipativity of the perturbed gradient with respect to w. Then we can bound the optimization error of w_t by the next lemma.

Lemma 15. *Suppose $\rho_w^0 = C_w k p^2 \geq 1$, $\rho_a^0 = C_a$, $(w_0, a_0) \in \mathcal{K}_{C_4, m_a, M_a}$ and $m_a \leq a_t^\top a^* \leq M_a$ holds for all t's. For any $\delta \in (0,1)$, we choose step size*

$$\eta = C_6 \Big(k^4 p^6 \cdot \max\Big\{1, p \log\frac{1}{\delta}\Big\}\Big)^{-1}.$$

Then with at least probability $1 - \delta/3$, we have

$$\phi_t \leq 5\pi/12 \tag{15}$$

for all t's such that $\tau_{13} \leq t \leq T = \widetilde{O}(\eta^{-2})$, where

$$\tau_{13} = \widetilde{O}\Big(\frac{p}{\eta}\log\frac{1}{\delta}\Big).$$

Lemma 15 implies that after (w_t, a_t) enters $\mathcal{K}_{C_4,m_a,M_a}$, it starts to make progress towards w^*. Due to the large injected noise, however, the optimization error of w_t can only attain a large optimization error. Although the optimization error of a_t is also large, (w_t, a_t) can be guaranteed to escape from the spurious local optimum.

The proof of Lemmas 13–15 requires supermartingale-based analysis, which is very involved and technical. See more details in the appendix C.4. Combining all above lemmas, we take $T_1 = \tau_{11} + \tau_{12} + \tau_{13}$, and complete the proof of Theorem 6. □

C.3. Some Important Lemmas

These lemmas give proper bounds that we will use in our later proof.

Lemma 16 (Bound on $\left(\mathbf{1}^\top a^*\right)^2 - \mathbf{1}^\top a^* \mathbf{1}^\top a_t$)**.** *Suppose $-A \leq \mathbf{1}^\top a^* \mathbf{1}^\top a_0 - \left(\mathbf{1}^\top a^*\right)^2 \leq \frac{C_1}{p}\|a^*\|_2^2$ for some constants $C_1, A \geq 0$. For any $\delta \in (0,1)$, if we choose $\eta \leq C\left(k^4 p^6 \max\{1, \log\frac{1}{\delta}\}\right)^{-1}$, then with at least probability $1-\delta$, we have*

$$-A - 2\left(\mathbf{1}^\top a^*\right)^2 \leq \mathbf{1}^\top a^* \mathbf{1}^\top a_t - \left(\mathbf{1}^\top a^*\right)^2 \leq \frac{C_2}{p}\|a^*\|_2^2,$$

for $\forall t \leq T = \widetilde{O}\left(\frac{1}{\eta^2}\right)$ and some constant $C_2 > C_1$.

Proof. We only give the prove for the right side, and the left side follows similar lines.

We start with

$$\begin{aligned}
\mathbb{E}[\mathbf{1}^\top a^* \mathbf{1}^\top a_{t+1} | \mathcal{F}_t] &= \left(1 - \frac{\eta\left(k+\pi-1\right)}{2\pi}\right)\mathbf{1}^\top a^* \mathbf{1}^\top a_t + \frac{\eta\left(k + \mathbb{E}g\left(\phi\right) - 1\right)}{2\pi}\left(\mathbf{1}^\top a^*\right)^2 \\
&\leq \left(1 - \frac{\eta\left(k+\pi-1\right)}{2\pi}\right)\mathbf{1}^\top a^* \mathbf{1}^\top a_t + \frac{\eta\left(k+\pi-1\right)}{2\pi}\left(\mathbf{1}^\top a^*\right)^2.
\end{aligned}$$

Denote $G_t = \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)^{-t}\left(\mathbf{1}^\top a^* \mathbf{1}^\top a_t - \left(\mathbf{1}^\top a^*\right)^2\right)$. Thus, we have

$$\mathbb{E}[G_{t+1} | \mathcal{F}_t] \leq G_t.$$

Denote $\mathscr{E}_t = \{\forall \tau \leq t, \mathbf{1}^\top a^* \mathbf{1}^\top a_\tau - \left(\mathbf{1}^\top a^*\right)^2 \leq \frac{C_2}{p}\|a^*\|_2^2\} \subset \mathcal{F}_t$. We have

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{E}_t} | \mathcal{F}_t] \leq G_t \mathbb{1}_{\mathscr{E}_t} \leq G_t \mathbb{1}_{\mathscr{E}_{t-1}}.$$

Thus, $G_t \mathbb{1}_{\mathscr{E}_{t-1}}$ is a supermartingale with initial value G_0.

We have the following bound.

$$\begin{aligned}
d_t &\triangleq |G_t \mathbb{1}_{\mathscr{E}_{t-1}} - \mathbb{E}[G_t \mathbb{1}_{\mathscr{E}_{t-1}} | \mathcal{F}_{t-1}]| \\
&= \left(1 - \frac{\eta\left(k+\pi-1\right)}{2\pi}\right)^{-t}\eta\left|-\frac{k+\pi-1}{2\pi}\mathbf{1}^\top a^* \mathbf{1}^\top \epsilon_{t-1} + \frac{\mathbb{E}g\left(\phi_{\xi_{t-1}}\right) - g\left(\phi_{\xi_{t-1}}\right)}{2\pi}\left(\mathbf{1}^\top a^*\right)^2\right| \\
&\leq \left(1 - \frac{\eta\left(k+\pi-1\right)}{2\pi}\right)^{-t}\eta\left(\frac{(k+\pi-1)k\rho_a}{2\pi}|\mathbf{1}^\top a^*| + \frac{\left(\mathbf{1}^\top a^*\right)^2}{2}\right) \\
&= \left(1-\lambda\right)^{-t} M,
\end{aligned}$$

where $\lambda = \frac{\eta(k+\pi-1)}{2\pi}$ and $M = \eta\left(\frac{(k+\pi-1)k\rho_a}{2\pi}|\mathbf{1}^\top a^*| + \frac{\left(\mathbf{1}^\top a^*\right)^2}{2}\right)$. Denote $r_t = \sqrt{\sum_{i=1}^t d_i^2}$. Then Azuma's Inequality can be applied, and we have

$$\mathbb{P}\left(G_t \mathbb{1}_{\mathscr{E}_{t-1}} - G_0 \geq \widetilde{O}\left(1\right) r_t \log^{1/2}\left(\frac{1}{\eta\delta}\right)\right) = \exp\left(-\frac{\widetilde{O}\left(1\right) r_t^2 \log\left(\frac{1}{\eta\delta}\right)}{2\sum_{i=1}^t d_i^2}\right) = \exp\left(-\frac{\widetilde{O}\left(1\right) r_t^2 \log\left(\frac{1}{(\eta\delta)^2}\right)}{2\sum_{i=1}^t d_i^2}\right) = \widetilde{O}\left(\eta^2\delta\right).$$

Therefore, with at least probability $1-\widetilde{O}\left(\eta^2\delta\right)$, we have

$$G_t \mathbb{1}_{\mathscr{E}_{t-1}} \leq G_0 + \widetilde{O}\left(1\right) r_t \log^{1/2}\left(\frac{1}{\eta\delta}\right).$$

We next prove that conditioning on $\mathscr{E}_{t-1}$, we have $\mathscr{E}_t$ with at least probability $1-\widetilde{O}\left(\eta^2\delta\right)$. Thus, $\mathscr{E}_T$ holds with at least probability $1-\delta$, when $T = \widetilde{O}\left(\frac{1}{\eta^2}\right)$.

From $G_t \mathbb{1}_{\mathscr{E}_{t-1}} \leq G_0 + \widetilde{O}\left(1\right) r_t \log^{1/2}\left(\frac{1}{\eta\delta}\right)$, we know

$$\begin{aligned}\mathbf{1}^\top a^* \mathbf{1}^\top a_t - \left(\mathbf{1}^\top a^*\right)^2 &\leq \left(1-\lambda\right)^t \left(\mathbf{1}^\top a^* \mathbf{1}^\top a_0 - \left(\mathbf{1}^\top a^*\right)^2 + \widetilde{O}\left(1\right) r_t \log^{1/2}\left(\frac{1}{\eta\delta}\right)\right)\\ &\leq \mathbf{1}^\top a^* \mathbf{1}^\top a_0 - \left(\mathbf{1}^\top a^*\right)^2 + \widetilde{O}\left(1\right)\left(1-\lambda\right)^t r_t \log^{1/2}\left(\frac{1}{\eta\delta}\right).\end{aligned}$$

We have

$$\begin{aligned}\left(1-\lambda\right)^t r_t = \left(1-\lambda\right)^t M\sqrt{\sum_{i=1}^t \left(1-\lambda\right)^{-2i}} &= M\sqrt{\sum_{i=0}^{t-1}\left(1-\lambda\right)^{2i}}\\ &\leq M\sqrt{\frac{1}{1-\left(1-\lambda\right)^2}} \leq \frac{M}{\sqrt{\lambda}} \qquad (16)\\ &= \left(\frac{k+\pi-1}{2\pi}k\rho_a|\mathbf{1}^\top a^*| + \frac{\left(\mathbf{1}^\top a^*\right)^2}{2}\right)\sqrt{\frac{2\pi\eta}{k+\pi-1}}.\end{aligned}$$

By carefully choosing $\eta_{\max} = \widetilde{O}\left(\frac{1}{k^4p^6}\right)$ and let $\eta = \frac{\eta_{\max}}{\max\{1, \log\frac{1}{\delta}\}}$, we have

$$\begin{aligned}\mathbf{1}^\top a^* \mathbf{1}^\top a_t - \left(\mathbf{1}^\top a^*\right)^2 &= \frac{C_1}{p}\|a^*\|_2^2 + \widetilde{O}\left(1\right)\log^{1/2}\left(\frac{1}{\eta\delta}\right)\left(\frac{k+\pi-1}{2\pi}k\rho_a|\mathbf{1}^\top a^*| + \frac{\left(\mathbf{1}^\top a^*\right)^2}{2}\right)\sqrt{\frac{2\pi\eta}{k+\pi-1}}\\ &\leq \frac{C_2}{p}\|a^*\|_2^2.\end{aligned}$$

□

Lemma 17 (Bound on $a_t^\top a^*$). *Suppose* $\frac{C_1}{p}\|a^*\|_2^2 \leq a_0^\top a^* \leq M_a$, $\mathbb{E}_{\xi_{t-1}} g\left(\phi_t\right) \geq 1 + \frac{C_2}{p}$ *and*

$$-A - 2\left(\mathbf{1}^\top a^*\right)^2 \leq \mathbf{1}^\top a^* \mathbf{1}^\top a_t - \left(\mathbf{1}^\top a^*\right)^2 \leq \frac{C_3}{p}\|a^*\|_2^2$$

holds for $\forall t \leq T = \widetilde{O}\left(\frac{1}{\eta^2}\right)$ *and some positive constants* $M_a, C_1, C_2 > C_3$. *If we take* $\eta \leq C\left(k^4p^6\max\{1, \log\frac{1}{\delta}\}\right)^{-1}$, *then with at least probability* $1-\delta$, *we have*

$$\frac{C_4}{p}\|a^*\|_2^2 \leq a_t^\top a^* \leq A + M_a + \left(1 + \frac{C_5}{p}\right)\|a^*\|_2^2 + 2\left(\mathbf{1}^\top a^*\right)^2$$

for $\forall t \leq T = \widetilde{O}\left(\frac{1}{\eta^2}\right)$ *and some positive constants* C_4, C_5.

Proof. We only give the proof for the left side, the right side follows similar lines.

$$\begin{aligned}\mathbb{E}[a_{t+1}^\top a^* | \mathcal{F}_t] &= \left(1 - \frac{\eta\left(\pi-1\right)}{2\pi}\right) a_t^\top a^* + \frac{\eta\left(\mathbb{E}g\left(\phi_t\right)-1\right)}{2\pi}\|a^*\|_2^2 + \frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top a^*\right)^2 - \mathbf{1}^\top a^* \mathbf{1}^\top a_t\right) \\ &\geq \left(1 - \frac{\eta\left(\pi-1\right)}{2\pi}\right) a_t^\top a^* + \eta \frac{C_2 - C_3}{2\pi p}\|a^*\|_2^2.\end{aligned}$$

Denote $C = \frac{C_2 - C_3}{2\pi p}\|a^*\|_2^2$ and $G_t = \left(1 - \frac{\eta(\pi-1)}{2\pi}\right)^{-t}\left(a_t^\top a^* - \frac{2\pi}{\pi-1}C\right)$. The above inequality changes to

$$\mathbb{E}[G_{t+1}|\mathcal{F}_t] \geq G_t.$$

Denote $\mathscr{E}_t = \{\forall \tau \leq t, a_\tau^\top a^* \geq \frac{C_4}{p}\|a^*\|_2^2\}$ for some constant $C_4 = \min\{\frac{C_1}{2}, \frac{C_2 - C_3}{2(\pi-1)}\}$. Then, for all t, $G_{t+1}\mathbb{1}_{\mathscr{E}_t}$ satisfies

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{E}_t}|\mathcal{F}_t] \geq G_t \mathbb{1}_{\mathscr{E}_t} \geq G_t \mathbb{1}_{\mathscr{E}_{t-1}}.$$

Thus, $\{G_{t+1}\mathbb{1}_{\mathscr{E}_t}\}$ is submartingale.

We have the following bound of the difference between $G_{t+1}\mathbb{1}_{\mathscr{E}_t}$ and $\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{E}_t}]$.

$$\begin{aligned}d_{t+1} &\triangleq |G_{t+1}\mathbb{1}_{\mathscr{E}_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{E}_t}|\mathcal{F}_t]| \\ &= \eta\left(1 - \frac{\eta\left(\pi-1\right)}{2\pi}\right)^{-t-1}\left|-\frac{\pi-1}{2\pi}\epsilon_t^\top a^* + \frac{\mathbb{E}g\left(\phi_{\xi_t}\right) - g\left(\phi_{\xi_t}\right)}{2\pi}\|a^*\|_2^2 - \frac{\mathbf{1}^\top a^* \mathbf{1}^\top \epsilon_t}{2\pi}\right| \\ &\leq \eta\left(1 - \frac{\eta\left(\pi-1\right)}{2\pi}\right)^{-t-1}\left(\frac{\pi-1}{2\pi}\rho_a\|a^*\|_2 + \frac{\|a^*\|_2^2}{2\pi} + \frac{|\mathbf{1}^\top a^*|k\rho_a}{2\pi}\right) \\ &= \left(1-\lambda\right)^{-t-1}M,\end{aligned}$$

where $\lambda = \frac{\eta(\pi-1)}{2\pi}$ and $M = \eta\left(\frac{\pi-1}{2\pi}\rho_a\|a^*\|_2 + \frac{\|a^*\|_2^2}{2\pi} + \frac{|\mathbf{1}^\top a^*|k\rho_a}{2\pi}\right)$.

Denote $r_t = \sqrt{\sum_{i=0}^t d_i^2}$. By Azuma's Inequality again, we have

$$\mathbb{P}\left(G_t\mathbb{1}_{\mathscr{E}_{t-1}} - G_0 \leq -\widetilde{O}\left(1\right)r_t \log^{\frac{1}{2}}\left(\frac{1}{\eta\delta}\right)\right) \leq \exp\left(-\frac{\widetilde{O}\left(1\right)r_t^2 \log\left(\frac{1}{\eta\delta}\right)}{2\sum_{i=0}^t d_i^2}\right) = \widetilde{O}\left(\eta^2\delta\right).$$

Therefore, with at least probability $1 - \widetilde{O}\left(\eta^2\delta\right)$, we have

$$G_t\mathbb{1}_{\mathscr{E}_{t-1}} \geq G_0 - \widetilde{O}\left(1\right)r_t \log^{\frac{1}{2}}\left(\frac{1}{\eta\delta}\right).$$

This means that when $\mathscr{E}_{t-1}$ holds, with at least probability $1 - \widetilde{O}\left(\eta^2\delta\right)$,

$$a_t^\top a^* \geq \left(1-\lambda\right)^t\left(a_0^\top a^* - \frac{2\pi}{\pi-1}C - \widetilde{O}\left(1\right)r_t \log^{\frac{1}{2}}\left(\frac{1}{\eta\delta}\right)\right) + \frac{2\pi}{\pi-1}C \geq C_6 - \widetilde{O}\left(1\right)\left(1-\lambda\right)^t r_t \log^{\frac{1}{2}}\left(\frac{1}{\eta\delta}\right),$$

where $C_6 = \min\{\frac{C_1}{p}\|a^*\|_2^2, \frac{2\pi}{\pi-1}C\}$. Following similar lines to (16), we have

$$\left(1-\lambda\right)^t r_t \leq \frac{M}{\sqrt{\lambda}} = \left(\frac{\pi-1}{2\pi}\rho_a\|a^*\|_2 + \frac{\|a^*\|_2^2}{2\pi} + \frac{|\mathbf{1}^\top a^*|k\rho_a}{2\pi}\right)\sqrt{\frac{2\pi\eta}{\pi-1}}.$$

With a proper step size $\eta \leq \left(k^4p^6\max\{1, \log\frac{1}{\delta}\}\right)^{-1}$, we then have

$$a_t^\top a^* \geq C_6 - \widetilde{O}\left(1\right)\log^{\frac{1}{2}}\left(\frac{1}{\eta\delta}\right)\left(\frac{\pi-1}{2\pi}\rho_a\|a^*\|_2 + \frac{\|a^*\|_2^2}{2\pi} + \frac{|\mathbf{1}^\top a^*|k\rho_a}{2\pi}\right)\sqrt{\frac{2\pi\eta}{\pi-1}} \geq \frac{C_6}{2} = \frac{C_4}{p}\|a^*\|_2^2.$$

□

C.4. Detailed Proof of Theorem 6

C.4.1. Proof of Lemma 11

Proof. For any $a \in \mathbb{B}_0\left(\frac{|\mathbf{1}^\top a^*|}{\sqrt{k}}\right)$, we have

$$\mathbf{1}^\top a \le \|a\|_1 \le \sqrt{k}\|a\|_2 \le |\mathbf{1}^\top a^*|,$$
$$\mathbf{1}^\top a \ge -\|a\|_1 \ge -\sqrt{k}\|a\|_2 \ge -|\mathbf{1}^\top a^*|.$$

Thus, we have

$$|\mathbf{1}^\top a| \le |\mathbf{1}^\top a^*|,$$

which is equivalent to the following inequality.

$$-2\left(\mathbf{1}^\top a^*\right)^2 \le \mathbf{1}^\top a^* \mathbf{1}^\top a - \left(\mathbf{1}^\top a^*\right)^2 \le 0.$$

By this inequality, we prove that $a \in \mathcal{A}_{C_3}$. □

C.4.2. Proof of Lemma 12

Proof. For calculation simplicity, we rescale w and ξ by ρ_w. Specifically, For any $w \in \mathbb{S}_0(1)$, define $\nu = \rho_w^{-1} w$ and $r_\nu \triangleq \|\nu\|_2 = \|w\|_2/\rho_w = 1/\rho_w$, where $\rho_w = \Omega\left(p^2\right)$ and $r_\nu = O\left(p^{-2}\right)$. Moreover, let $\zeta \triangleq \xi/\rho_w \sim$ unif$(\mathbb{B}_0(1))$. Then we have $\angle(v+\zeta, w^*) = \angle(w+\zeta, w^*)$.

Without loss of generality, we assume $w^* = (1, 0, ..., 0)^\top$. To calculate the expectation, we need to rewrite ν in the p-dimension polar coordinate system as discussed in Section B. Specifically, ν can be written as $\nu = (r, \theta_1, \theta_2, ..., \theta_{p-1})$, where $r \ge 0$, $\theta_i \in [0, \pi]$, $i = 1, 2, ..., p-2$, $\theta_{p-1} \in [0, 2\pi]$ and $\theta_1 = \arccos(\nu_1/\|\nu\|) = \angle(w, w^*)$. Moreover, under the polar coordinate, $\nu + \zeta$ is expressed as $(r^\zeta, \theta_1^\zeta, \theta_2^\zeta, ..., \theta_{p-1}^\zeta)$. We then have

$$\theta_1^\zeta = \arccos\frac{\nu_1 + \zeta_1}{\|\nu+\zeta\|_2} = \angle(v+\zeta, w^*) = \phi_\xi,$$

where $\zeta = (\zeta_1, \zeta_2, \cdots, \zeta_p)$.

Therefore, for sufficiently large ρ_w we have

$$\begin{aligned}
&\mathbb{E}_\xi\left(\pi - \phi_\xi\right)\cos\phi_\xi + \sin\phi_\xi = \mathbb{E}_\zeta\left(\pi - \theta_1^\zeta\right)\cos\theta_1^\zeta + \sin\theta_1^\zeta \\
&= \int_{\mathbb{B}_0(1)}\left[\left(\pi - \arccos\frac{\nu_1 + x_1}{\|\nu+x\|_2}\right)\frac{\nu_1+x_1}{\|\nu+x\|_2} + \sin\arccos\frac{\nu_1+x_1}{\|\nu+x\|_2}\right]\frac{1}{V(1)}dx \\
&= \int_{\mathbb{B}_\nu(1)}\left[\left(\pi - \arccos\frac{x_1}{\|x\|_2}\right)\frac{x_1}{\|x\|_2} + \sin\arccos\frac{x_1}{\|x\|_2}\right]\frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}}dx \\
&\ge \int_{\mathbb{B}_0(1-r_\nu)}\left[\left(\pi - \arccos\frac{x_1}{\|x\|_2}\right)\frac{x_1}{\|x\|_2} + \sin\arccos\frac{x_1}{\|x\|_2}\right]\frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}}dx \\
&= \int_0^{1-r_\nu} r^{p-1}dr\frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}}\int_0^\pi\left(\left(\pi-\theta_1\right)\cos\theta_1 + \sin\theta_1\right)\sin^{p-2}\theta_1 d\theta_1 \\
&\quad \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \sin^{p-3}\theta_2 \cdots \sin\theta_{p-2} d\theta_2 \cdots d\theta_{p-1}
\end{aligned}$$

We then apply Lemma 10 by taking $f(\theta) = (\pi - \theta)\cos\theta + \sin\theta$ and get the following result.

$$\begin{aligned}
&\mathbb{E}_\xi\left(\pi - \phi_\xi\right)\cos\phi_\xi + \sin\phi_\xi \\
&= L_p\left(r_\nu\right)A_p\left(f\right) = M_p A_p\left(f\right) + \left(L_p\left(r_\nu\right) - M_p\right)A_p\left(f\right) \\
&= \frac{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{p+2}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)^2} - O\left(\frac{1}{p}\right) = 1 + \frac{1}{2p} + \frac{1}{8p^2} + \cdots - O\left(\frac{1}{p}\right) = 1 + \Omega\left(\frac{1}{p}\right),
\end{aligned}$$

where the last equality is due to the Taylor expansion of first term at $p = +\infty$, i.e.

$$\frac{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{p+2}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)^2} = 1 + \frac{1}{2p} + \frac{1}{8p^2} + o\left(\frac{1}{p^2}\right).$$

Similarly, we have

$$\begin{aligned}\mathbb{E}_\xi\phi_\xi = \mathbb{E}_\zeta\theta_1^\zeta &= \int_{\mathbb{B}_\nu(1)}\left(\arccos\frac{x_1}{\|x\|_2}\right)\frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}}dx \le \int_{\mathbb{B}_0(1+r_\nu)}\left(\arccos\frac{x_1}{\|x\|_2}\right)\frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}}dx \\ &= \left(\int_0^{1+r_\nu} r^{p-1}dr\frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}}\int_0^\pi \theta_1\sin^{p-2}(\theta_1)\,d\theta_1\right)\left(\int_0^\pi\cdots\int_0^\pi\int_0^{2\pi}\sin^{d-3}\theta_2\cdots\sin\theta_{d-2}d\theta_1\cdots d\theta_{d-1}\right).\end{aligned}$$

We then apply Lemma 10 by taking $g(\theta) = \theta$ and get the following result.

$$\begin{aligned}\mathbb{E}_\xi\phi_\xi &= H_p(r_\nu)A_p(g) = M_pA_p(g) + (H_p(r_\nu) - M_p)A_p(g) \\ &= \frac{\pi}{2} + O\left(\frac{1}{p}\right) \le \frac{3\pi}{4}.\end{aligned}$$

□

C.4.3. Proof of Lemma 13

Proof. Denote $\mathscr{E}_t = \{\forall\tau \le t, a_\tau \in \mathcal{A}_{C_2,C_3}\}$. Then, by Theorem 5, when $(w, a) \in \mathcal{A}_{C_2,C_3}$, we have

$$\langle -\mathbb{E}_{\xi,\epsilon}\nabla_aL(w_t+\xi, a_t+\epsilon), a^* - a_t\rangle \ge \frac{C}{p}\|a_t - a^*\|_2^2,$$

for some constant C.

We next bound the expectation of the norm of the perturbed gradient.

$$\begin{aligned}\mathbb{E}_{\xi,\epsilon}\|\nabla_aL(w_t+\xi, a_t+\epsilon)\|_2^2 &= \mathbb{E}_{\xi,\epsilon}\|\nabla_aL(w_t+\xi, a_t+\epsilon) - \nabla_aL(w^*, a^*)\|_2^2 \\ &= \mathbb{E}_{\xi,\epsilon}\left\|\frac{1}{2\pi}\left(\mathbb{1}\mathbb{1}^\top + (\pi-1)I\right)(a_t+\epsilon-a^*) - \frac{g(\phi_\xi)-\pi}{2\pi}a^*\right\|_2^2 \\ &\le \frac{1}{2\pi^2}\mathbb{E}_{\xi,\epsilon}\left\|\left(\mathbb{1}\mathbb{1}^\top + (\pi-1)I\right)(a_t+\epsilon-a^*)\right\|_2^2 + \frac{1}{2}\|a^*\|_2^2 \\ &\le \frac{(k+\pi-1)^2}{\pi^2}\left(\|a_t-a^*\|_2^2 + \rho_a^2\right) + \frac{1}{2}\|a^*\|_2^2.\end{aligned}$$

Therefore, the expectation $\mathbb{E}\left[\|a_{t+1}-a^*\|_2^2\mathbb{1}_{\mathscr{E}_t}\right]$ can be bounded as follows.

$$\begin{aligned}\mathbb{E}\left[\|a_{t+1}-a^*\|_2^2\mathbb{1}_{\mathscr{E}_t}\right] &= \mathbb{E}\left[\|a_t-a^*\|_2^2\mathbb{1}_{\mathscr{E}_t}\right] - 2\mathbb{E}\left[\langle -\eta\mathbb{E}_{\xi,\epsilon}\nabla_aL(w_t+\xi, a_t+\epsilon), a^*-a_t\rangle\mathbb{1}_{\mathscr{E}_t}\right] \\ &\quad + \mathbb{E}\left[\|\eta\nabla_aL(w_t+\xi, a_t+\epsilon)\|_2^2\mathbb{1}_{\mathscr{E}_t}\right] \\ &\le \left(1 - \eta\frac{C}{p} + \eta^2\frac{(k+\pi-1)^2}{\pi^2}\right)\left[\mathbb{E}\|a_t-a^*\|_2^2\mathbb{1}_{\mathscr{E}_t}\right] \\ &\quad + \frac{\eta^2(k+\pi-1)^2}{\pi^2}\rho_a^2 + \frac{\eta^2}{2}\|a^*\|_2^2 \\ &\le (1-\lambda_1)\mathbb{E}\|a_t-a^*\|_2^2\mathbb{1}_{\mathscr{E}_{t-1}} + b_1,\end{aligned}$$

where $\lambda_1 = \eta\frac{C}{p} - \eta^2\frac{(k+\pi-1)^2}{\pi^2}$ and $b_1 = \frac{\eta^2(k+\pi-1)^2}{\pi^2}\rho_a^2 + \frac{\eta^2}{2}\|a^*\|_2^2$.

Note that $\mathscr{E}_t$ implies that $\|a_t - a^*\|_2^2 \geq \frac{1}{4}\|a^*\|_2^2$, then we have

$$\frac{1}{4}\|a^*\|_2^2\mathbb{P}\left(\mathscr{E}_t\right) \leq \mathbb{E}\left[\|a_t - a^*\|_2^2\mathbb{1}_{\mathscr{E}_t}\right] \leq \mathbb{E}\left[\|a_t - a^*\|_2^2\mathbb{1}_{\mathscr{E}_{t-1}}\right] \leq (1-\lambda_1)^t\|a_0 - a^*\|_2^2 + \frac{b_1}{\lambda_1}. \tag{17}$$

With our choice of small η, we have $\lambda_1 = O\left(\eta/p\right) \in (0,1), \frac{b_1}{\lambda_1} \leq \frac{1}{8}\|a^*\|_2^2$. Thus when $t = \widetilde{O}\left(\frac{p}{\eta}\right)$, we have $\mathbb{P}\left(\mathscr{E}_t\right) \leq \frac{1}{2}$. We recursively apply the same procedure with $\log\frac{1}{\delta}$ times, and after $T_0 = \widetilde{O}\left(\frac{p}{\eta}\log\frac{1}{\delta}\right)$, we have $\mathbb{P}\left(\mathscr{E}_{T_0}\right) < \delta$, which implies that with probability at least $1-\delta$, there exists $\tau_{11} \leq T_0$ such that

$$\frac{C_2}{p}\|a^*\|_2^2 \leq a_{\tau_{11}}^\top a^* \text{ and } \|a_{\tau_{11}} - a^*/2\|_2^2 \leq \|a^*\|_2^2,$$

for some constant C_2. Moreover, $\|a_{\tau_{11}} - a^*/2\|_2^2 \leq \|a^*\|_2^2$ further implies $a_{\tau_{11}}^\top a^* \leq 2\|a^*\|_2^2$. Thus, we have

$$\frac{C_2}{p}\|a^*\|_2^2 \leq a_{\tau_{11}}^\top a^* \leq 2\|a^*\|_2^2.$$

Then by Lemma 16 and Lemma 17, we get the desired result. □

C.4.4. PROOF OF LEMMA 14

Proof. When ρ_w is sufficiently large, with probability $1-\delta$, the norm of perturbed gradient w.r.t. w, i.e., $\|\nabla_w L(w,a)\|$, is at least $O\left(\rho_w\right)$ once in $\log\frac{1}{\delta}$ steps. Thus, with at least probability $1-\delta$, there exists $t \leq \log\frac{1}{\delta}$ such that

$$w_t^\top w^* = -1 + O\left(\eta^2\rho_w^2\right).$$

We take this point as w_0 in the later proof. Recall that $\widetilde{w}_t = w_{t-1} - \eta\left(I - w_{t-1}w_{t-1}^\top\right)\nabla_w L$ and $w_t = \mathrm{Proj}_{\mathbb{S}_0(1)}\left(\widetilde{w}_t\right)$. Without loss of generality, we assume $\widetilde{w}_{t+1}^\top w^* \leq 0$, otherwise we already have $1 + w_{t+1}^\top w^* \geq 1$. Notice that $\|\widetilde{w}_{t+1}\|_2 \geq 1$, we then have

$$\begin{aligned}
1 + w_{t+1}^\top w^* =& 1 + \frac{\widetilde{w}_{t+1}^\top w^*}{\|\widetilde{w}_{t+1}\|_2} \geq 1 + \widetilde{w}_{t+1}^\top w^* \\
=& 1 + w_t^\top w^* - \eta {w^*}^\top\left(I - w_t w_t^\top\right)\nabla_w L(w_t + \xi, a_t + \epsilon) \\
=& 1 + w_t^\top w^* + \frac{\eta}{2\pi}(1 + w_t^\top w^*)(1 - w_t^\top w^*)(a_t + \epsilon)^\top a^*\left(\pi - \phi_\xi\right) - \eta {w^*}^\top\left(I - w_t w_t^\top\right)\Bigg(\frac{\|a_t + \epsilon\|_2^2}{2} \\
&+ \frac{\sum_{i\neq j}\left(a_{t,i} + \epsilon_i\right)\left(a_{t,j} + \epsilon_j\right)}{2\pi} - \frac{\left(a_t + \epsilon\right)^\top a^*\sin\phi_\xi}{2\pi}\frac{1}{\|w_t + \xi\|_2} - \frac{\sum_{i\neq j}\left(a_{t,i} + \epsilon_i\right)a_j^*}{2\pi}\frac{1}{\|w_t + \xi\|_2}\Bigg)\xi.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}\left(1 + w_{t+1}^\top w^* | \mathcal{F}_t\right) \geq& (1 + w_t^\top w^*)\left(1 + \frac{\eta}{2\pi}(1 - w_t^\top w^*)a_t^\top a^*\left(\pi - \mathbb{E}_\xi\phi_\xi\right)\right) \\
&+ \eta\mathbb{E}_\xi\left(w^* - w_t^\top w^* w_t\right)^T\left(\frac{a_t^\top a^*\sin\phi_\xi}{2\pi}\frac{1}{\|w_t + \xi\|_2} + \frac{\sum_{i\neq j}a_{t,i}a_j^*}{2\pi}\frac{1}{\|w_t + \xi\|_2}\right)\xi \\
=& (1 + w_t^\top w^*)\left(1 + \frac{\eta}{2\pi}(1 - w_t^\top w^*)a_t^\top a^*\left(\pi - \mathbb{E}_\xi\phi_\xi\right)\right) + \frac{\eta}{2\pi}a_t^\top a^*\mathbb{E}_\xi\frac{\sin\phi_\xi\left(w^* - w_t^\top w^* w_t\right)^T\xi}{\|w_t + \xi\|_2},
\end{aligned}$$

where the last line is due to (12).

We next show that

$$\mathbb{E}_\xi\frac{\sin\phi_\xi\left(w^* - w^\top w^* w\right)^T\xi}{\|w + \xi\|_2} \geq 0 \tag{18}$$

Let $\phi_x = \angle(x, w^*)$.

$$\mathbb{E}_\xi \frac{\sin\phi_\xi \left(w^* - w^\top w^* w\right)^T \xi}{\|w+\xi\|_2} = \int_{\mathbb{B}_w(\rho_w)} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx. \tag{19}$$

$$= \int_{\mathbb{B}_w(\rho_w)\cap\mathbb{B}_{-w}(\rho_w)} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \tag{20}$$

$$+ \int_{\mathbb{B}_w(\rho_w)\backslash\mathbb{B}_{-w}(\rho_w)} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx. \tag{21}$$

Let's calculate these two integrals separately. For the first integral,

$$\begin{aligned}
&\int_{\mathbb{B}_w(\rho_w)\cap\mathbb{B}_{-w}(\rho_w)} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
=& \int_{\mathbb{B}_w(\rho_w)\cap\mathbb{B}_{-w}(\rho_w),(w^*-w^\top w^* w)^\top x>0} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&+ \int_{\mathbb{B}_w(\rho_w)\cap\mathbb{B}_{-w}(\rho_w),(w^*-w^\top w^* w)^\top x<0} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx
\end{aligned}$$

By the symmetric property with respect to the origin, we have

$$\int_{\mathbb{B}_w(\rho_w)\cap\mathbb{B}_{-w}(\rho_w)} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx = 0.$$

For the second integral, let's consider the the symmetric point of x with respect to the vector w, i.e., $\widetilde{x} = 2w^\top x w - x$. We have the following properties:

$$\|\widetilde{x}\|_2 = \|x\|_2, \ \|w+x\|_2 = \|w+\widetilde{x}\|_2, \text{ and } \left(w^* - w^\top w^* w\right)^\top \widetilde{x} = -\left(w^* - w^\top w^* w\right)^\top x.$$

We further have

$$\begin{aligned}
\sin\phi_{\widetilde{x}} = \sqrt{1-\cos^2\phi_{\widetilde{x}}} = \sqrt{1-\left(\frac{(w^*)^\top\widetilde{x}}{\|\widetilde{x}\|_2}\right)^2} &= \sqrt{1-\left(\frac{(2w^\top x(w^*)^\top w - (w^*)^\top x)}{\|x\|_2}\right)^2} \\
&= \sqrt{1-\left(\frac{(w^*)^\top x}{\|x\|_2}\right)^2 + \left(\frac{(4w^\top x(w^*)^\top w\left[(w^*)^\top x - w^\top x(w^*)^\top w\right]}{\|x\|_2^2}\right)}.
\end{aligned}$$

Since $x \in \mathbb{B}_w\left(\rho_w\right)\backslash\mathbb{B}_{-w}\left(\rho_w\right)$, we have $\|x+w\|_2^2 \geq \rho_w^2 \geq \|x-w\|_2^2$, which implies $w^\top x \geq 0$. Moreover, when , $\left(w^* - w^\top w^* w\right)^\top x > 0$, we have $(w^*)^\top x - w^\top x (w^*)^\top w \geq 0$. Together with $(w^*)^\top w \leq 0$, a we have

$$\sin\phi_{\widetilde{x}} \leq \sin\phi_x.$$

Then the second integral can be estimated as follows.

$$\begin{aligned}
&\int_{\mathbb{B}_w(\rho_w)\backslash\mathbb{B}_{-w}(\rho_w)} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
=& \int_{\mathbb{B}_w(\rho_w)\backslash\mathbb{B}_{-w}(\rho_w),(w^*-w^\top w^* w)^\top x>0} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
&+ \int_{\mathbb{B}_w(\rho_w)\backslash\mathbb{B}_{-w}(\rho_w),(w^*-w^\top w^* w)^\top x<0} \frac{1}{V\left(\rho_w\right)} \frac{\sin(\phi_x)\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} dx \\
=& \int_{\mathbb{B}_w(\rho_w)\backslash\mathbb{B}_{-w}(\rho_w),(w^*-w^\top w^* w)^\top x>0} \frac{1}{V\left(\rho_w\right)} \frac{\left(w^* - w^\top w^* w\right)^\top x}{\|x\|_2} (\sin(\phi_x) - \sin(\phi_{\widetilde{x}})) dx \\
\geq& 0
\end{aligned}$$

Thus, combining the above calculations for two integrals, (18) is proved. Then with the fact that $a_t^\top a^* \geq m_a > 0$ for all t, we have

$$\begin{aligned}\mathbb{E}\left(1 + w_{t+1}^\top w^* | \mathcal{F}_t\right) &\geq (1 + w_t^\top w^*)\left(1 + \frac{\eta}{2\pi}(1 - w_t^\top w^*) a_t^\top a^* \left(\pi - \mathbb{E}_\xi \phi_\xi\right)\right) \\ &\geq (1 + C_1 \frac{\eta}{p})(1 + w_t^\top w^*),\end{aligned}$$

for some positive constant C_1.

Thus,

$$\mathbb{E}\left(1 + w_t^\top w^*\right) \geq (1 + C_1 \frac{\eta}{p})^t (1 + w_0^\top w^*).$$

When $t = \widetilde{O}(\frac{p}{\eta} \log \frac{1}{\eta})$, we have $\mathbb{E}\left(1 + w_t^\top w^*\right) \geq C_2$ for some constant $C_2 \in (-1, 0)$. Thus, with constant probability we have $1 + w_t^\top w^* \geq C_2$. And We could have with at least probability $1 - \delta$, $1 + w_{\tau_{12}}^\top w^* \geq C_2$ for some $\tau_{12} = \widetilde{O}\left(\frac{p}{\eta} \log \frac{1}{\eta} \log \frac{1}{\delta}\right)$. □

C.4.5. Proof of Lemma 15

Proof. Recall that we have $-1 < C_4 \leq w_0^\top w^* \leq 0$ and $m_a \leq a_t^\top a^* \leq M_a$ for all t.

Our proof has two steps.

Step 1: We show that $w_t^\top w^*$ have a lower bound $\frac{C_4 - 1}{2}$with probability $1 - \delta$ for $\forall t \leq \widetilde{O}\left(\frac{1}{\eta^2}\right)$.

Denote $\mathscr{E}_t = \{\forall \tau \leq t, w_\tau^\top w^* \geq \frac{C_4 - 1}{2}\} \subset \mathcal{F}_t$. Then if $\mathscr{E}_t$ holds, we have $(w_\tau, a_\tau) \in \mathcal{K}_{(C_4-1)/2, m_a, M_a}$ for $\forall \tau \leq t$. Recall that $\widetilde{w}_{t+1}$ is defined as

$$\widetilde{w}_{t+1} = w_t - \left(I - w_t w_t^\top\right) \nabla_w L\left(w_t + \xi_t, a_t + \epsilon_t\right),$$

and

$$w_{t+1} = \text{Proj}_{\mathbb{S}_0}\left(\widetilde{w}_{t+1}\right).$$

By Theorem 5, when $(w, a) \in \mathcal{K}_{(C_4-1)/2, m_a, M_a}$, we have

$$\langle -\mathbb{E}_{\xi,\epsilon}\left(I - ww^\top\right) \nabla_w L\left(w + \xi, a + \epsilon\right), w^* - w\rangle \geq \frac{m_a\left(1 + C_4\right)}{32}\|w - w^*\|_2^2 - \gamma,$$

where $\gamma = O\left(k/\rho_w\right)$.
Moreover, we have the bound on expectation of the norm of the perturbed (manifold) gradient.

$$\begin{aligned}\mathbb{E}_{\xi,\epsilon}\left\|\left(I - ww^\top\right) \nabla_w L\left(w + \xi, a + \epsilon\right)\right\|_2^2 =& \frac{\mathbb{E}_\epsilon\left(\left(a + \epsilon\right)^\top a^*\right)^2 \mathbb{E}_\xi\left(\pi - \phi\right)^2}{2\pi^2} w^{*\top}\left(I - ww^\top\right) w^* \\ &+ 2\mathbb{E}_{\xi,\epsilon}\Big(\frac{\|a + \epsilon\|_2^2}{2} + \frac{\sum_{i \neq j}\left(a_i + \epsilon_i\right)\left(a_j + \epsilon_j\right)}{2\pi} - \frac{\left(a + \epsilon\right)^\top a^* \sin \phi_\xi}{2\pi} \frac{1}{\|w + \xi\|_2} \\ &- \frac{\sum_{i \neq j}\left(a_i + \epsilon_i\right) a_j^*}{2\pi} \frac{1}{\|w + \xi\|_2}\Big)^2 \xi^\top\left(I - ww^T\right)\xi \\ &\leq \frac{M_a^2 + \rho_a^2\|a^*\|_2^2}{2} + C_1 k^2 \rho_w^2 \rho_a^2,\end{aligned}$$

where C_1 is a constant.

Combine the above two inequalities, we get

$$\begin{aligned}\mathbb{E}[\|\widetilde{w}_{t+1}-w^*\|_2^2\mathbb{1}_{\mathscr{E}_t}|\mathcal{F}_t] =&\|w_t-w^*\|_2^2\mathbb{1}_{\mathscr{E}_t} - 2\langle -\eta\mathbb{E}_{\xi_t,\epsilon_t}\nabla_w L\left(w_t+\xi_t, a_t+\epsilon_t\right), w^*-w_t\rangle\mathbb{1}_{\mathscr{E}_t}\\ &+\mathbb{E}_{\xi_t,\epsilon_t}\left\|\eta\left(I-w_tw_t^\top\right)\nabla_w L\left(w_t+\xi_t,a_t+\epsilon_t\right)\right\|_2^2\mathbb{1}_{\mathscr{E}_t}\\ \le& \left(1-\frac{\eta m_a\left(1+C_4\right)}{16}\right)\|w_t-w^*\|_2^2\mathbb{1}_{\mathscr{E}_t} + \left(\eta\gamma + \frac{\eta^2\left(M_a^2+\rho_a^2\|a^*\|_2^2\right)}{2} + \eta^2C_1k^2\rho_w^2\rho_a^2\right)\mathbb{1}_{\mathscr{E}_t}\\ =&\left(1-\lambda_2\right)\|w_t-w^*\|_2^2\mathbb{1}_{\mathscr{E}_t}+b_2\mathbb{1}_{\mathscr{E}_t},\end{aligned}$$

where $\lambda_2 = O\left(\eta/p\right)$, $b_2 = O\left(\eta k/\rho_w\right)$ and $\frac{b_2}{\lambda_2} \le \min\{\frac{1+C_4}{2}, \frac{1}{4}\}$ by proper choice of small η and large ρ_w.

We next show that $\|w_{t+1}-w^*\|_2^2 \le \|\widetilde{w}_{t+1}-w^*\|_2^2$. We first have the following inequality.

$$\|\widetilde{w}_{t+1}\|_2^2 = \|w_t\|_2^2 + \left\|\eta\left(I-w_tw_t^\top\right)\nabla_w L\left(w_t+\xi, a_t+\epsilon\right)\right\|_2^2 \ge 1.$$

Since we have $w_{t+1}^\top w^* \le 1$, we obtain

$$\begin{aligned}\|\widetilde{w}_{t+1}-w^*\|_2^2 &= 1 + \|\widetilde{w}_{t+1}\|_2^2 - 2\|\widetilde{w}_{t+1}\|_2 w_{t+1}^\top w^*\\ &\ge 1+1-2w_{t+1}^\top w^* = \|w_{t+1}-w^*\|_2^2.\end{aligned}$$

The above inequality comes from $a^2-2ab+1\ge 2-2b \Leftrightarrow a+1\ge 2b$ for $a\ge 1$. Therefore, we have

$$\mathbb{E}[\|w_{t+1}-w^*\|_2^2\mathbb{1}_{\mathscr{E}_t}|\mathcal{F}_t] \le \left(1-\lambda_2\right)\|w_t-w^*\|_2^2\mathbb{1}_{\mathscr{E}_t}+b_2\mathbb{1}_{\mathscr{E}_t}.$$

Denote $G_t = \left(1-\lambda_2\right)^{-t}\left(\|w_t-w^*\|_2^2 - \frac{b_2}{\lambda_2}\right)$, the above recursive relation becomes

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{E}_t}|\mathcal{F}_t] \le G_t\mathbb{1}_{\mathscr{E}_t} \le G_t\mathbb{1}_{\mathscr{E}_{t-1}}.$$

Thus, $\{G_t\mathbb{1}_{\mathscr{E}_t}\}$ is a supermartingale.

We then have the following bound.

$$\begin{aligned}d_t &\triangleq |G_t\mathbb{1}_{\mathscr{E}_{t-1}} - \mathbb{E}[G_t\mathbb{1}_{\mathscr{E}_{t-1}}|\mathcal{F}_{t-1}]| = \left(1-\lambda_2\right)^{-t}|\|w_t-w^*\|_2^2 - \mathbb{E}[\|w_t-w^*\|_2^2|\mathcal{F}_{t-1}]|\\ &\le \left(1-\lambda_2\right)^{-t}\left(\eta\left(M_a+\rho_a\|a^*\|_2\right)+C_2\eta k\rho_w\right)\\ &= \left(1-\lambda_2\right)^{-t}M_2,\end{aligned}$$

where $M_2 = O\left(\eta k\rho_w\right)$.

Denote $r_t = \sqrt{\sum_{i=0}^t d_i^2}$. By Azuma's Inequality,

$$\mathbb{P}\left(G_t\mathbb{1}_{\mathscr{E}_{t-1}} - G_0 \ge \widetilde{O}\left(1\right)r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \le \exp\left(-\frac{\widetilde{O}\left(1\right)r_t^2\log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^t d_i^2}\right) = \widetilde{O}\left(\eta^2\delta\right)$$

Therefore, with at least probability $1-\widetilde{O}\left(\eta^2\delta\right)$, we have

$$\begin{aligned}\|w_t-w^*\|_2^2 &\le \left(1-\lambda_2\right)^t\left(\|w_0-w^*\|_2^2 - \frac{b_2}{\lambda_2}\right) + \widetilde{O}\left(1\right)\left(1-\lambda_2\right)^t r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right) + \frac{b_2}{\lambda_2}\\ &\le \|w_0-w^*\|_2^2 + \widetilde{O}\left(1\right)\frac{M}{\sqrt{\lambda_2}}\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right) + \frac{b_2}{\lambda_2}\\ &\le 3-C_4,\end{aligned}$$

where the last line is true by our choice of small η.

The above inequality shows that $w_t^\top w^* \geq \frac{C_4-1}{2}$ holds with at least probability $1-O\left(\eta^2\delta\right)$, which implies that $\mathscr{E}_t$ holds with at least probability $1-O\left(\eta^2\delta\right)$ when $\mathscr{E}_{t-1}$ holds. Hence, with at least probability $1-\delta$, we have $w_t^\top w^* \geq \frac{C_4-1}{2}$ for all $t \leq \widetilde{O}\left(\frac{1}{\eta^2}\right)$.

Step 2: We show that if the result in Step 1 holds, there exists $\tau_{13} = \widetilde{O}\left(\frac{p}{\eta}\log\frac{1}{\delta}\right)$ such that $\phi_{\tau_{13}} \leq \frac{\pi}{3}$ and ϕ_t stays in the region $\left\{\phi\middle|\phi \leq \frac{5\pi}{12}\right\}$ with probability $1-\delta$ during the later $\widetilde{O}\left(\frac{1}{\eta^2}\right)$ steps.

Following similar lines to Step 1, we have

$$\mathbb{E}[G_{t+1}|\mathcal{F}_t] \leq G_t,$$

where $G_t = (1-\lambda_2)^{-t}\left(\|w_t - w^*\|_2^2 - \frac{b_2}{\lambda_2}\right)$.

Thus, recall that $\lambda_2 = O(\eta/p)$, $\frac{b_2}{\lambda_2} \leq \frac{1}{4}$, and let $t = \widetilde{O}\left(\frac{p}{\eta}\right)$, and we know that

$$\mathbb{E}[\|w_t - w^*\|_2^2|\mathcal{F}_{t-1}] \leq (1-\lambda_2)^t\|w_0 - w^*\|_2^2 + \frac{b_2}{\lambda_2} \leq \frac{1}{2}.$$

By Markov Inequality, we know

$$\mathbb{P}\left(\|w_t - w^*\|_2^2 \geq 1\right) \leq \frac{1}{2}.$$

We recursively apply the above inequality with $\log\frac{1}{\delta}$ times. Then, with at most $\tau_{13} = \widetilde{O}\left(\frac{1}{\eta}\log\frac{1}{\delta}\right)$,we have

$$\mathbb{P}\left(\phi_{\tau_{13}} \leq \frac{\pi}{3}\right) = \mathbb{P}\left(\|w_{\tau_{13}} - w^*\|_2^2 \leq 1\right) \geq 1-\delta.$$

For notational simplify, we assume $\phi_0 \leq \frac{\pi}{3}$ in the later proof. We will show that ϕ_t stays in the region $\left\{\phi\middle|\phi \leq \frac{5\pi}{12}\right\}$ with high probability during the later $\widetilde{O}\left(\frac{1}{\eta^2}\right)$ steps.

Denote $\mathscr{H}_t = \{\forall\tau \leq t, \phi_\tau \leq \frac{5\pi}{12}\}$. With the similar argument in Step 1, when $\mathscr{H}_{t-1}$ holds, with at least probability $1-\widetilde{O}\left(\eta^2\delta\right)$, we have

$$\begin{aligned}\|w_t - w^*\|_2^2 &\leq (1-\lambda_2)^t\left(\|w_0 - w^*\|_2^2 - \frac{b_2}{\lambda_2}\right) + \frac{b_2}{\lambda_2} + \widetilde{O}\left(1\right)(1-\lambda_2)^t r_t \log^{\frac{1}{2}}\left(\frac{1}{\eta\delta}\right)\\ &\leq \|w_0 - w^*\|_2^2 + \frac{b_2}{\lambda_2} + \widetilde{O}\left(1\right)\frac{M}{\sqrt{\lambda_2}}\log^{\frac{1}{2}}\left(\frac{1}{\eta\delta}\right) \leq 1.4,\end{aligned}$$

which implies that $\phi_t \leq \frac{5\pi}{12}$, i.e., $\mathscr{H}_t$ holds.

Hence, for all $t \leq T = \widetilde{O}\left(\frac{1}{\eta^2}\right)$, we have $\phi_t \leq \frac{5\pi}{12}$ with at least probability $1-\delta$.

Combining the above two steps, with probability $1-\delta$, we have $\phi_t \leq 5\pi/12$ for all t's such that $\tau_{13} \leq t \leq T = \widetilde{O}(\eta^{-2})$, where $\tau_{13} = \widetilde{O}\left(\frac{p}{\eta}\log\frac{1}{\delta}\right)$.

□

D. Proof for Phase II

D.1. Technical Lemma

The next lemma shows that perturbed GD imitates the behavior of GD, when the noise is small enough. Thus, it can finally converge to the global optimum.

Lemma 18. *Denote $g(\phi) = (\pi - \phi)\cos\phi + \sin\phi$, $\xi \sim \text{unif}(\mathbb{B}_0(1)) \in \mathbb{R}^d$ and $w \in \mathbb{R}^d$, $\|w\|_2 = 1$. Define*

$$\phi_\xi = \arccos \frac{v^\top (w + \rho\xi)}{\|v\|_2 \|w + \rho\xi\|_2}, \quad \phi = \arccos \frac{v^\top w}{\|v\|_2 \|w\|_2}.$$

Suppose $\phi \leq \frac{\pi}{2}$ and $\rho < 1$. Then we have

$$\mathbb{E}_\xi \phi_\xi \leq U_\phi^{(1)}(\rho), \quad \mathbb{E}_\xi (\pi - g(\phi_\xi))^2 \leq U_\phi^{(2)}(\rho), \quad \mathbb{E}_\xi g(\phi_\xi) \geq U_\phi^{(3)}(\rho).$$

where $\lim_{\rho\to 0} U_\phi^{(1)}(\rho) = \phi$, $\lim_{\rho\to 0} U_\phi^{(2)}(\rho) = (\pi - g(\phi))^2$, $\lim_{\rho\to 0} U_\phi^{(3)}(\rho) = g(\phi)$ and $U_\phi^{(1)}(\rho)$, $U_\phi^{(2)}(\rho)$ is non-decreasing, $U_\phi^{(3)}(\rho)$ is non-increasing. Moreover, we have

$$|\mathbb{E}_\xi \phi_\xi - \phi| = O(\rho), \quad |\mathbb{E}_\xi (\pi - g(\phi_\xi))^2 - (\pi - g(\phi))^2| = O(\rho), \quad |\mathbb{E}_\xi g(\phi_\xi) - g(\phi)| = O(\rho).$$

Proof. Without loss of generality, let $v = (1, 0, ..., 0)^\top$. Then since $\phi \leq \frac{\pi}{2}$, we have $w_1 \geq 0$.

We find the upper bound of $\phi_\xi = \arccos \frac{w_1 + \rho\xi_1}{\|w+\rho\xi\|_2}$, when ρ and w are fixed. ϕ_ξ could be explained as the angle between X and v, where $X = w + \rho\xi \in \mathbb{B}_w(\rho)$. Thus, ϕ_ξ achieves the maximum when X is tangent to $\mathbb{B}_w(\rho)$. This means that $(w + \rho\xi)^\top \rho\xi = 0$ and $\|\xi\|_2 = 1$, which is equivalent to $w^\top \xi + \rho = 0$ and $\|\xi\|_2 = 1$. This leads to $\|w + \rho\xi\|_2 = \sqrt{1 - \rho^2}$. Therefore, to get the upper bound of ϕ_ξ, we need the lower bound of ξ_1. This is formulated as following,

$$\min \xi_1 \text{ s.t. } \sum_i w_i \xi_i + \rho = 0, \; \sum_i w_i^2 = 1, \; \sum_i \xi_i^2 = 1.$$

By the Lagrange multiplier method, we have $\xi_1^* = -\sqrt{(1-\rho^2)(1-w_1^2)} - \rho w_1$. Thus,

$$\phi_\xi \leq \arccos\left(w_1 \sqrt{1-\rho^2} - \rho\sqrt{1 - w_1^2}\right) \triangleq U_\phi^{(1)}(\rho).$$

Moreover, with the same argument above, we have

$$\phi_\xi \geq \arccos\left(w_1 \sqrt{1-\rho^2} + \rho\sqrt{1 - w_1^2}\right).$$

Therefore, we have $|\mathbb{E}_\xi \phi_\xi - \phi| = |\mathbb{E}_\xi \phi_\xi - \arccos w_1| \leq C_1 \rho$.

Since $g(\phi)$ is decreasing, $(\pi - g(\phi))^2$ is increasing and both of them are Lipschitz continuous, we have

$$\begin{aligned} &\mathbb{E}_\xi (\pi - g(\phi_\xi))^2 \leq \left(\pi - g\left(U_\phi^{(1)}(\rho)\right)\right)^2 \triangleq U_\phi^{(2)}(\rho), \quad \mathbb{E}_\xi g(\phi_\xi) \geq g\left(U_\phi^{(1)}(\rho)\right) \triangleq U_\phi^{(3)}(\rho), \\ &|\mathbb{E}_\xi (\pi - g(\phi_\xi))^2 - (\pi - g(\phi))^2| \leq C_2 \rho, \quad |\mathbb{E}_\xi g(\phi_\xi) - g(\phi)| \leq C_3 \rho. \end{aligned}$$

By simple manipulation, one can easily verify that $\lim_{\rho\to 0} U_\phi^1(\rho) = \phi$, $\lim_{\rho\to 0} U_\phi^2(\rho) = (\pi - g(\phi))^2$, $\lim_{\rho\to 0} U_\phi^3(\rho) = g(\phi)$ and $U_\phi^{(1)'}(\rho) \geq 0$, $U_\phi^{(2)'}(\rho) \geq 0$, $U_\phi^{(3)'}(\rho) \leq 0$. □

D.2. Proof for Theorem 7

Proof. When $\rho_w < 1$, we have

$$\mathbb{E}_\xi \frac{\sin\phi_\xi \left(w^* - w^\top w^* w\right)^\top \xi}{\|w + \xi\|_2} \leq C_1 \sqrt{1 - (w^\top w^*)^2} \rho_w = O(\rho_w)$$

Taking $\rho_w = O\left(\frac{\gamma}{kp}\right)$ to be small enough and combining (12), we have

$$\begin{aligned}\left\langle -\mathbb{E}_{\xi,\epsilon}\left(I - ww^\top\right)\nabla_w L\left(w+\xi, a+\epsilon\right), w^* - w\right\rangle &\geq \frac{a^\top a^*\left(\pi - \mathbb{E}_\xi \phi_\xi\right)}{2\pi}\left(1 - (w^\top w^*)^2\right) - \gamma \\ &= \frac{a^\top a^*\left(\pi - \mathbb{E}_\xi \phi_\xi\right)}{4\pi}\left(1 + w^\top w^*\right)\|w - w^*\|_2^2 - \gamma \\ &\geq \frac{m\left(1 + C_9\right)}{16}\|w - w^*\|_2^2 - \gamma.\end{aligned}$$

Given small enough ρ_w, by Lemma 18 and $\|w - w^*\|_2^2 \leq C_{10}\gamma$, we have

$$\begin{aligned}\left\langle -\mathbb{E}_{\xi,\epsilon}\nabla_a L\left(w+\xi, a+\epsilon\right), a^* - a\right\rangle &= \frac{1}{2\pi}\left(\mathbf{1}^\top a - \mathbf{1}^\top a^*\right)^2 + \frac{1}{2\pi}\left(\left(\pi - 1\right)a - \left(\mathbb{E}_\xi g\left(\phi_\xi\right) - 1\right)a^*\right)^\top\left(a - a^*\right) \\ &= \frac{1}{2\pi}\left(\mathbf{1}^\top a - \mathbf{1}^\top a^*\right)^2 + \frac{1}{2\pi}\left(\pi - \mathbb{E}_\xi g\left(\phi_\xi\right)\right)a^{*\top}\left(a - a^*\right) + \frac{\pi - 1}{2\pi}\|a - a^*\|_2^2 \\ &\geq \frac{\pi - 1}{2\pi}\|a - a^*\|_2^2 - \gamma.\end{aligned}$$

□

D.3. Proof Sketch for Theorem 8

Proof Sketch. The perturbed GD is already in the solution set of Phase I, which is actually in the dissipative region $\mathcal{K}_{C_9,m,M}$. The first lemma shows that even if the noise is reduced, our proposed algorithm never escape this set.

Lemma 19. *Define $\phi_t(\xi) = \angle(x_t + \xi, x^*)$. Assume there exists some constant C_8 such that $1 + C_8/p \leq \mathbb{E}_\xi g(\phi_t(\xi)) \leq \pi$ and $\mathbb{E}_\xi \phi_t(\xi) \leq \frac{3\pi}{4}$ for all t. Suppose*

$$0 < m_a \leq a_0^\top a^* \leq M_a \quad \textit{and} \quad \phi_0 \leq \frac{5}{12}\pi.$$

For any $\delta \in (0,1)$, we choose step size

$$\eta = C_{11}\Big(\max\Big\{k^4 p^6, \frac{k^2 p}{\gamma}\Big\}\max\Big\{1, p\log\frac{1}{\gamma}\log\frac{1}{\delta}\Big\}\Big)^{-1}$$

for some constant C_{11}. Then with at least probability at lease $1 - \delta/3$, we have for all $t \leq T = \widetilde{O}(\eta^{-2})$,

$$0 < m_a' \leq a_t^\top a^* \leq M_a' \quad \textit{and} \quad \phi_t \leq \frac{11}{24}\pi,$$

where $m_a' = m_a/2$, $M_a' = 3M_a$.

Lemma 19 shows that throughout sufficiently many iterations of Phase II, (w_t, a_t)'s are at least as accurate as the initial solution with high probability. Thus, we can guarantee that the perturbed GD algorithm stays away from the spurious local optimum, and the benign optimization landscape in Theorem 7 holds.

The next lemma characterizes the convergence properties of the perturbed GD algorithm for w.

Lemma 20. *Suppose $\phi_t \leq \frac{11}{24}\pi$ and $0 < m_a' \leq a_t^\top a^* \leq M_a'$ hold for all t. For any $\gamma > 0$, we choose $\rho_w^1 \leq C_w^1\frac{\gamma}{kp} < 1$ and $\rho_a \leq C_a^1$ for small enough constant C_w^1 and C_a^1. For any $\delta \in (0,1)$, we choose step size*

$$\eta = C_{11}\Big(\max\Big\{k^4 p^6, \frac{k^2 p}{\gamma}\Big\}\max\Big\{1, p\log\frac{1}{\gamma}\log\frac{1}{\delta}\Big\}\Big)^{-1}$$

for some constant C_{11}. Then with at least probability at least $1 - \delta/3$, we have

$$\|w_t - w^*\|_2^2 \leq C_{12}\gamma$$

for all t's such that $\tau_{21} \leq t \leq \widetilde{O}(\eta^{-2})$, where C_{12} is a constant and

$$\tau_{21} = \widetilde{O}\Big(\frac{p}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\Big).$$

Lemma 20 shows that at τ_{21} iterations, the perturbed GD algorithm enters $\mathcal{R}_{m_a',M_a',C_{12}}$. Then we can characterize its convergence properties for a, as shown in the next lemma.

Lemma 21. *Suppose $(w_t, a_t) \in \mathcal{R}_{m_a',M_a',C_{12}}$ holds for all t. For any $\gamma > 0$, we choose $\rho_w^1 \leq C_w^1 \frac{\gamma}{kp} < 1$ and $\rho_a \leq C_a^1$ for small enough constant C_w^1 and C_a^1. For any $\delta \in (0, 1)$, we choose step size*

$$\eta = C_{11}\Big(\max\Big\{k^4p^6, \frac{k^2p}{\gamma}\Big\}\max\Big\{1, p\log\frac{1}{\gamma}\log\frac{1}{\delta}\Big\}\Big)^{-1}$$

for some constant C_{11}. Then with at least probability $1-\delta/3$, we have

$$\|a_t - a^*\|_2^2 \leq \gamma$$

for all t's such that $\tau_{22} \leq t \leq \widetilde{O}(\eta^{-2})$, where

$$\tau_{22} = \widetilde{O}\Big(\frac{p}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\Big).$$

Similar to Lemmas 13–15, the proof of Lemmas 19–21 also requires supermartingale-based analysis. See more details in Appendix D.

Combining the above lemmas together, we take $T_2 = \tau_{21} + \tau_{22}$, and complete the proof of Theorem 8. □

D.4. Detailed Proof of Theorem 8

D.4.1. Proof of Lemma 19

Proof. Since $\phi_0 \leq \frac{5\pi}{12}$, we have $g(\phi_0) > 1.4$. By Lemma 18, conditions $\mathbb{E}_\xi \phi_\xi \leq \frac{3\pi}{4}$ and $1 + O\left(\frac{1}{p}\right) \leq \mathbb{E}_\xi g(\phi_\xi) \leq \pi$ are satisfied with our choice of small noise level ρ_w. Recall that $\mathbf{1}^\top a_0 \mathbf{1}^\top a^* - \left(\mathbf{1}^\top a^*\right)^2$ is bounded. Then, using Lemma 16, we have $\mathbf{1}^\top a_t \mathbf{1}^\top a^* - \left(\mathbf{1}^\top a^*\right)^2$ is still bounded in the same order for $\forall t \leq \widetilde{O}\left(\eta^{-2}\right)$ with probability $1-\delta/3$. Combined with Lemma 17, with probability $1-\delta/3$ we have $m_a' \leq a_t^\top a^* \leq M_a'$ in the following $\widetilde{O}\left(\eta^{-2}\right)$ steps, where $m_a' = m_a/2$, $M_a' = 3M_a$. Then, following the same arguments in Step 2 of the proof of Lemma 6, we have $\phi_t \leq \frac{1}{2}\left(\frac{\pi}{2} + \frac{5\pi}{12}\right) = \frac{11\pi}{24}$ in the following $\widetilde{O}\left(\eta^{-2}\right)$ steps with probability $1-\delta/3$. Therefore, combining the above results together, we have the desired results. □

D.4.2. Proof of Lemma 20

Proof. Recall that we have $\phi_t \leq \frac{11\pi}{24}$ and $m_a' \leq a_t^\top a^* \leq M_a'$ for all t. This implies that $w_t^\top w^* \geq 0.1$. Thus, we have $(w_t, a_t) \in \mathcal{K}_{0.1,m_a',M_a'}$ for all t.

The following two steps proof is similar with Lemma 15.

Step 1: We show that there exists $\tau_{21} = \widetilde{O}\left(\frac{1}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\right)$ such that $\|w_{\tau_{21}} - w^*\|_2^2 \leq \gamma/2$ holds with at least probability $1-\delta$.

By Theorem 7, for $(w, a) \in \mathcal{K}_{0.1,m_a',M_a'}$, we have

$$\left\langle -\mathbb{E}_{\xi,\epsilon}\left(I - ww^\top\right)\nabla_w L\left(w+\xi, a+\epsilon\right), w^* - w\right\rangle \geq \frac{11m_a}{160}\|w - w^*\|_2^2 - M_a\rho_w,$$

where $M_a\rho_w = O(\gamma/p)$ is a small constant by our choice of ρ_w.

Also, we have the bound on the expectation of the norm of the perturbed (manifold) gradient.

$$\mathbb{E}_{\xi,\epsilon}\left\|\left(I-ww^{\top}\right)\nabla_w L\left(w+\xi,a+\epsilon\right)\right\|_2^2 \le \frac{M_a^2+\rho_a^2\|a^*\|_2^2}{2}+C_1k^2\rho_w^2,$$

for some constant C_1.

Thus, denote $\lambda_3 = \frac{11m_a\eta}{160}$ and $b_3 = \frac{11m_a}{1280}\eta\gamma$. We have

$$\begin{aligned}\mathbb{E}[\|\widetilde{w}_{t+1}-w^*\|_2^2|\mathcal{F}_t] &= \|w_t-w^*\|_2^2 - 2\langle -\eta\mathbb{E}_{\xi_t,\epsilon_t}\left(I-w_tw_t^{\top}\right)\nabla_w L\left(w_t+\xi_t,a_t+\epsilon_t\right), w^*-w_t\rangle \\ &\quad + \mathbb{E}_{\xi_t,\epsilon_t}\left\|\eta\left(I-w_tw_t^{\top}\right)\nabla_w L\left(w_t+\xi_t,a_t+\epsilon_t\right)\right\|_2^2 \\ &\le \left(1-\frac{11m_a\eta}{160}\right)\|w_t-w^*\|_2^2 + \eta M_a\rho_w + \frac{\eta^2\left(M_a^2+\rho_a^2\|a^*\|_2^2\right)}{2} + C_1\eta^2k^2\rho_w^2 \\ &\le \left(1-\lambda_3\right)\|w_t-w^*\|_2^2 + b_3,\end{aligned}$$

where the last line is due to our choice of small parameters ρ_a, ρ_w and η.

With same argument in the proof of Lemma 15, and we have

$$\|\widetilde{w}_{t+1}-w^*\|_2^2 \ge \|w_{t+1}-w^*\|_2^2.$$

Hence, denote $\mathscr{E}_t = \{\forall\tau\le t, \|w_\tau-w^*\|_2^2 \ge \frac{\gamma}{2}\}$, with our choice of η and $t = \widetilde{O}\left(\frac{p}{\eta}\log\frac{1}{\gamma}\right)$, we have

$$\begin{aligned}\frac{\gamma}{2}\mathbb{P}\left(\mathscr{E}_t\right) \le \mathbb{E}\|w_t-w^*\|_2^2 &= \left(1-\lambda_3\right)\mathbb{E}\|w_{t-1}-w^*\|_2^2 + b_3 \\ &\le \left(1-\lambda_3\right)^t\|w_0-w^*\|_2^2 + \frac{b_3}{\lambda_3} \le \frac{1}{4}\gamma.\end{aligned}$$

Thus, we have $\mathbb{P}\left(\mathscr{E}_t\right) \le 0.5$ and recursively apply the above lines for $\log\frac{1}{\delta}$ times, we know there exists $\tau_{21} = \widetilde{O}\left(\frac{p}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\right)$ such that $\|w_{\tau_{21}}-w^*\|_2^2 \le \frac{\gamma}{2}$ with at least probability $1-\delta$.

Step 2: We show that if $\|w_0-w^*\|_2^2 \le \frac{\gamma}{2}$, w_t stays in the region $\left\{w\middle|\|w-w^*\|_2^2 \le \gamma\right\}$ in the following $\widetilde{O}\left(\frac{1}{\eta^2}\right)$ steps with at least probability $1-\delta$.

Denote $G_t = \left(1-\lambda_3\right)^{-t}\left(\|w_t-w^*\|_2^2 - \frac{b_3}{\lambda_3}\right)$ and $\mathscr{H}_t = \{\forall\tau\le t, \|w_\tau-w^*\|_2^2 \le \gamma\} \subset \mathcal{F}_t$. From Step 1, we have

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{H}_t}|\mathcal{F}_t] \le G_t\mathbb{1}_{\mathscr{H}_t} \le G_t\mathbb{1}_{\mathscr{H}_{t-1}}.$$

Thus, $\{G_t\mathbb{1}_{\mathscr{H}_t}\}$ is a supermartingale.

To apply Azuma's Inequality, we first have to bound the difference between $G_{t+1}\mathbb{1}_{\mathscr{H}_t}$ and $\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{H}_t}|\mathcal{F}_t]$.

$$\begin{aligned}d_{t+1} &\triangleq |G_{t+1}\mathbb{1}_{\mathscr{H}_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{H}_t}|\mathcal{F}_t]| \le \left(1-\lambda_3\right)^{-t-1}|\|w_{t+1}-w^*\|_2^2 - \mathbb{E}[\|w_{t+1}-w^*\|_2^2|\mathcal{F}_t]| \\ &\le \left(1-\lambda_3\right)^{-t-1}C_2\eta\gamma^{\frac{1}{2}}k = \left(1-\lambda_3\right)^{-t-1}M_3,\end{aligned}$$

where $\lambda_3 = \widetilde{O}\left(\eta/p\right)$, $M_3 = \widetilde{O}\left(\eta\gamma^{\frac{1}{2}}k\right)$.

Denote $r_t = \sqrt{\sum_{i=0}^t d_i^2}$. By Azuma's Inequality, we get

$$\mathbb{P}\left(G_t\mathbb{1}_{\mathscr{H}_{t-1}} - G_0 \ge \widetilde{O}\left(1\right)r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \le \exp\left(-\frac{\widetilde{O}\left(1\right)r_t^2\log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^t d_i^2}\right) = \widetilde{O}\left(\eta^2\delta\right).$$

Therefore, with at least probability $1-\widetilde{O}\left(\eta^2\delta\right)$, we have

$$\begin{aligned}\|w_t-w^*\|_2^2 &\le (1-\lambda_3)^t\left(\|w_0-w^*\|_2^2-\frac{b_3}{\lambda_3}\right)+\widetilde{O}\left(1\right)\left(1-\lambda_3\right)^t r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)+\frac{b_3}{\lambda_3}\\ &\le \|w_0-w^*\|_2^2+\widetilde{O}\left(1\right)\frac{M_3}{\sqrt{\lambda_3}}\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)+\frac{b_3}{\lambda_3}\le\gamma,\end{aligned}$$

where the last line holds, since we can always find $\eta\le\eta_{\max}=\widetilde{O}\left(\frac{\gamma}{k^2p}\right)$ to satisfy the condition.

The above inequality shows that if $\mathscr{E}_t$ holds, then $\mathscr{E}_{t+1}$ holds with at least probability $1-\widetilde{O}\left(\eta^2\delta\right)$. Hence, with at least probability $1-\delta$, we have $\|w_t-w^*\|_2^2\le\gamma$ for all $t\le T=\widetilde{O}\left(\frac{1}{\eta^2}\right)$.

Combining the above two steps, with probability $1-\delta$, we have $\|w_t-w^*\|_2^2\le C_{12}\gamma$ for all t's such that $\tau_{21}\le t\le\widetilde{O}(\eta^{-2})$, where C_{12} is a constant and $\tau_{21}=\widetilde{O}\left(\frac{p}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\right)$. □

D.4.3. PROOF OF LEMMA 21

Proof. Our proof has two steps.
Step 1: We show that with probability at least $1-\delta$, there exists $\tau_{22}=\widetilde{O}\left(\frac{1}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\right)$ such that $\|a_{\tau_{22}}-a^*\|_2^2\le\gamma/2$.

Recall that $(w_t,a_t)\in\mathcal{R}_{m_a',M_a',C_{12}}$ holds for all t. Then by Theorem 7, we have

$$\left\langle-\mathbb{E}_{\xi,\epsilon}\nabla_a L\left(w+\xi,a+\epsilon\right),a^*-a\right\rangle\ge\frac{\pi-1}{2\pi}\|a-a^*\|_2^2-\gamma.$$

Following similar lines to the proof of Lemma 13, we have the bound on the expectation of the norm of the perturbed gradient.

$$\begin{aligned}\mathbb{E}_{\xi,\epsilon}\|\nabla_a L\left(w+\xi,a+\epsilon\right)\|_2^2&=\mathbb{E}_{\xi,\epsilon}\|\nabla_a L\left(w+\xi,a+\epsilon\right)-\nabla_a L\left(w^*,a^*\right)\|_2^2\\ &=\mathbb{E}_{\xi,\epsilon}\left\|\frac{1}{2\pi}\left(\mathbb{1}\mathbf{1}^\top+\left(\pi-1\right)I\right)\left(a+\epsilon-a^*\right)-\frac{g\left(\phi_\xi\right)-\pi}{2\pi}a^*\right\|_2^2\\ &\le\frac{1}{2\pi^2}\mathbb{E}_{\xi,\epsilon}\left\|\left(\mathbb{1}\mathbf{1}^\top+\left(\pi-1\right)I\right)\left(a+\epsilon-a^*\right)\right\|_2^2+\gamma\le\frac{\left(k+\pi-1\right)^2}{\pi^2}\left(\|a-a^*\|_2^2+\rho_a^2\right)+\gamma.\end{aligned}$$

Combined the above two, with $\eta_{\max}=\widetilde{O}\left(\gamma/k^2p\right)$ and $\eta\le\eta_{\max}$, we have

$$\begin{aligned}\mathbb{E}[\|a_{t+1}-a^*\|_2^2|\mathcal{F}_t]&=\|a_t-a^*\|_2^2-2\langle-\eta\mathbb{E}_{\xi_t,\epsilon_t}\nabla_a L\left(w_t+\xi_t,a_t+\epsilon_t\right),a^*-a_t\rangle+\mathbb{E}_{\xi_t,\epsilon_t}\|\eta\nabla_a L\left(w_t+\xi_t,a_t+\epsilon_t\right)\|_2^2\\ &\le\left(1-\frac{\left(\pi-1\right)\eta}{\pi}+\eta^2\frac{\left(k+\pi-1\right)^2}{\pi^2}\right)\|a_t-a^*\|_2^2+2\eta\gamma+\eta^2\gamma+\frac{\left(k+\pi-1\right)^2}{\pi^2}\eta^2\rho_a^2\\ &\le\left(1-\frac{\left(\pi-1\right)\eta}{\pi}+\eta^2\frac{\left(k+\pi-1\right)^2}{\pi^2}\right)\|a_t-a^*\|_2^2+3\eta\gamma\end{aligned}$$

Thus, when $\eta\le\frac{\pi^2}{25(k+\pi-1)^2}$, we have

$$\begin{aligned}\mathbb{E}[\|a_{t+1}-a^*\|_2^2-5\gamma|\mathcal{F}_t]&\le\left(1-\frac{\left(\pi-1\right)\eta}{\pi}+\eta^2\frac{\left(k+\pi-1\right)^2}{\pi^2}\right)\left(\|a_t-a^*\|_2^2-5\gamma\right)-0.2\eta\gamma+5\frac{\left(k+\pi-1\right)^2}{\pi^2}\eta^2\gamma\\ &\le\left(1-\frac{\left(\pi-1\right)\eta}{\pi}+\eta^2\frac{\left(k+\pi-1\right)^2}{\pi^2}\right)\left(\|a_t-a^*\|_2^2-5\gamma\right)\\ &=\left(1-\lambda_4\right)\left(\|a_t-a^*\|_2^2-5\gamma\right).\end{aligned}$$

Denote $\mathscr{E}_t = \{\forall \tau \le t, \|a_\tau - a^*\|_2^2 \ge 12\gamma\}$. When $t \ge \frac{\log\left(\frac{\|a_0 - a^*\|_2^2}{\gamma}\right)}{\lambda_4} = \widetilde{O}\left(\frac{1}{\eta}\log\frac{1}{\gamma}\right)$, we have

$$12\gamma\mathbb{P}\left(\mathscr{E}_t\right) \le \mathbb{E}\|a_t - a^*\|_2^2 = \left(1-\lambda_4\right)^t\left(\|a_0 - a^*\|_2^2 - 5\gamma\right) + 5\gamma \le \left(1-\lambda_4\right)^t\|a_0 - a^*\|_2^2 + 5\gamma \le 6\gamma.$$

Therefore, $\mathbb{P}\left(\mathscr{E}_t\right) \le 0.5$. Recursively applying the above lines with $\log\frac{1}{\delta}$ times, we know that with at least probability $1-\delta$, there exists $\tau_{22} = \widetilde{O}\left(\frac{1}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\right)$ such that $\|a_{\tau_{22}} - a^*\|_2^2 \le 12\gamma$. Rescaling γ, we get the desired result.

Step 2: We show that, if $\|a_0 - a^*\|_2^2 \le \gamma/2$, then a_t stays in the region $\left\{a\middle|\|a - a^*\|_2^2 \le \gamma\right\}$ in the next $\widetilde{O}\left(\frac{1}{\eta^2}\right)$ steps with probability at least $1-\delta$.

Denote $G_t = \left(1-\lambda_4\right)^{-t}\left(\|a_t - a^*\|_2^2 - 5\gamma\right)$ and $\mathscr{H}_t = \{\forall\tau \le t, \|a_t - a^*\|_2^2 \le 6\gamma\}$. With the same argument in Step 1, we have

$$\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{H}_t}|\mathcal{F}_t] \le G_t\mathbb{1}_{\mathscr{H}_t} \le G_t\mathbb{1}_{\mathscr{H}_{t-1}}.$$

Thus, $\{G_t\mathbb{1}_{\mathscr{H}_t}\}$ is a supermartingale.

To use Azuma's Inequality, we first have to bound the difference between $G_{t+1}\mathbb{1}_{\mathscr{H}_t}$ and $\mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{H}_t}|\mathcal{F}_t]$.

$$\begin{aligned}
d_{t+1} &\triangleq |G_{t+1}\mathbb{1}_{\mathscr{H}_t} - \mathbb{E}[G_{t+1}\mathbb{1}_{\mathscr{H}_t}|\mathcal{F}_t]| \\
&= \left(1-\lambda_4\right)^{-t-1}\left|\|a_{t+1} - a^*\|_2^2 - \mathbb{E}[\|a_{t+1} - a^*\|_2^2|\mathcal{F}_t]\right|\mathbb{1}_{\mathscr{H}_t} \\
&\le \left(1-\lambda_4\right)^{-t-1}\left(C_1\eta\gamma^{\frac{1}{2}}k + C_2\eta^2k^2\right) = \left(1-\lambda_4\right)^{-t-1}M_4,
\end{aligned}$$

for some positive constant C_1 and C_2, where $\lambda_4 = \widetilde{O}\left(\eta\right)$, $M_4 = \widetilde{O}\left(\eta\gamma^{\frac{1}{2}}k\right)$.

Denote $r_t = \sqrt{\sum_{i=0}^t d_i^2}$. By Azuma's Inequality,

$$\mathbb{P}\left(G_t\mathbb{1}_{\mathscr{H}_{t-1}} - G_0 \ge \widetilde{O}\left(1\right)r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \le \exp\left(-\frac{\widetilde{O}\left(1\right)r_t^2\log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^t d_i^2}\right) = \widetilde{O}\left(\eta^2\delta\right).$$

Therefore, with at least probability $1 - \widetilde{O}\left(\eta^2\delta\right)$, we have

$$\begin{aligned}
\|a_t - a^*\|_2^2 &\le \left(1-\lambda_4\right)^t\left(\|a_0 - a^*\|_2^2 - 5\gamma\right) + \widetilde{O}\left(1\right)\left(1-\lambda_4\right)^t r_t\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right) + 5\gamma \\
&\le \|a_0 - a^*\|_2^2 + \widetilde{O}\left(1\right)\frac{M}{\sqrt{\lambda_4}}\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right) + 5\gamma \le 6\gamma,
\end{aligned}$$

where the last line holds, since we can always find $\eta \le \eta_{\max} = \widetilde{O}\left(\frac{\gamma}{k^2p}\right)$ to satisfy the condition. The above inequality shows that if $\mathscr{H}_t$ holds, then $\mathscr{H}_{t+1}$ holds with at least probability $1 - \widetilde{O}\left(\eta^2\delta\right)$. Hence, with at least probability $1-\delta$, we have $\|a_t - a^*\|_2^2 \le 6\gamma$ for all $t \le T = \widetilde{O}\left(\frac{1}{\eta^2}\right)$. Rescaling γ, we have the desired results for Step 2.

Combining the above two steps, with at least probability $1-\delta$, we have $\|a_t - a^*\|_2^2 \le \gamma$ for all t's such that $\tau_{22} \le t \le \widetilde{O}(\eta^{-2})$, where $\tau_{22} = \widetilde{O}\left(\frac{p}{\eta}\log\frac{1}{\gamma}\log\frac{1}{\delta}\right)$. □

E. An Additional Experiment for Training Overparameterized Neural Network

Our additional experiment still considers the regression problem under the realizable setting, where the response is generated by a noiseless **teacher** network

$$y = f(Z, w^*, a^*) = (a^*)^\top\sigma(Z^\top w^*).$$

The **student** network h, however, adopts a different architecture and contains two convolutional filters, i.e.,

$$h(Z, w, u, a, b) = a^\top \sigma(Z^\top w) + b^\top \sigma(Z^\top v),$$

where $v \in \mathbb{R}^p$ and $b \in \mathbb{R}^k$. Compared with the teach network, the student network is overparameterized. We then learn the overparameterized student network by solving the following optimization problem:

$$\min_{w,v,a,b} F(w, v, a, b) \quad \text{subject to} \quad w^\top w = 1 \text{ and } v^\top v = 1, \tag{22}$$

where $F(w, v, a, b) = \frac{1}{2}\mathbb{E}_Z(h(Z, w, v, a, b) - f(Z, w^*, a^*))^2$.

Unfortunately, $F(w, v, a, b)$ and $\nabla F(w, v, a, b)$ do not admit analytical forms. Therefore, we randomly sample n realizations of Z (denoted by $Z_i,\ i = 1, ...n$), and solve a finite sample approximation of (22),

$$\min_{w,v,a,b} F_n(w, v, a, b) \quad \text{subject to} \quad w^\top w = 1 \text{ and } v^\top v = 1, \tag{23}$$

where $F_n(w, v, a, b) = \frac{1}{2n}\sum_{i=1}^n (h(Z_i, w, v, a, b) - f(Z_i, w^*, a^*))^2$.

For our experiment, we choose $k = 10$ and $p = 15$. The first 5 entries of a^* all equal to $1/\sqrt{10}$ and the remaining entries of a^* all equal to $-1/\sqrt{10}$. w^* is randomly generated over the unit sphere. We choose $n = 10,000$, and expect (23) to have an optimization landscape to (22).

We run the gradient descent algorithm to solve (23). The initialization is chosen at

$$w = -w^*,\ v = -w^*,\ a_0 = (\mathbf{1}\mathbf{1}^\top + (\pi - 1)I)^{-1}(\mathbf{1}\mathbf{1}^\top - I)a^* \text{ and } b_0 = 0.$$

We choose the step size $\eta = 10^{-5}$ and run for 10^8 iterations. We eventually observe $\|\nabla F_n(w, v, a, b)\|_2 < 10^{-4}$ and $F_n(w, v, a, b) > 0.15$. We suspect that the gradient descent algorithm approaches some spurious local optimum.

F. Convergence Analysis for Perturbed-SGD

We then can characterize the estimation error of the stochastic gradient as follows.

Lemma 22. *Suppose that for any $\delta, \epsilon > 0$, $w \in \mathbb{S}_0(1)$ and $a \in \mathbb{B}_0(R)$, given a mini-batch size*

$$m = \mathrm{poly}\left(p, k, R, \frac{1}{\epsilon}, \log\frac{1}{\delta}\right),$$

with at least probability $1 - \delta$, we have

$$\left\|\nabla_w \widehat{\mathcal{L}}(w, a) - \nabla_w \mathcal{L}(w, a)\right\|_2^2 \leq \epsilon \quad \textit{and} \quad \left\|\nabla_a \widehat{\mathcal{L}}(w, a) - \nabla_a \mathcal{L}(w, a)\right\|_2^2 \leq \epsilon.$$

The proof of Lemma 22 is straightforward (by simple union bound and the concentration properties of sub-exponential random variable), and therefore omitted. Lemma 22 implies that as long as the batch size is sufficiently large, we can show the mini-batch stochastic gradient is sufficiently accurate with high probability. Then we can adapt the convergence analysis in Section 3, and show that P-SGD can avoid spurious local optimum with high probability in Phase I.

Theorem 23 (P-SGD escapes the spurious local optimum)**.** *Suppose $\|a^*\|_2 \leq R$, $\rho_w^0 = C_w^0 kp^2 \geq 1,\ \rho_a^0 = C_a^0,\ a_0 \in \mathbb{B}_0\left(\frac{|\mathbf{1}^\top a^*|}{\sqrt{k}}\right)$ and $w_0 \in \mathbb{S}_0(1)$. For any $\delta \in (0, 1)$, we choose a small enough step size*

$$\eta = \left(\mathrm{poly}\left(p, k, R, \log\frac{1}{\delta}\right)\right)^{-1}$$

and a large enough mini batch-size

$$m = \mathrm{poly}\left(p, k, R, \log\frac{1}{\delta}\right),$$

then with at least probability $1-\delta$*, we have*

$$m_a \leq a_t^\top a^* \leq M_a \quad \text{and} \quad \phi_t \leq \frac{5}{12}\pi$$

for all t*'s such that* $\widehat{T}_1 \leq t \leq \widetilde{O}(\eta^{-2})$*, where* m_a *and* M_a *are some constants and*

$$\widehat{T}_1 = \mathrm{poly}\Big(p, k, R, \log\frac{1}{\delta}\Big).$$

Similarly, for Phase II, we can show that P-SGD converges to the global optimum with high probability.

Theorem 24 (P-SGD converges to the global optimum)**.** *Suppose* $\|a^*\|_2 \leq R$*,* $\phi_0 \leq \frac{5}{12}\pi,\ 0 < m_a \leq a_0^\top a^* \leq M_a$*. For any* $\gamma > 0,$ *we choose* $\rho_w^1 \leq C_w^1 \frac{\gamma}{\sqrt{kp}} < 1$ *and* $\rho_a \leq C_a^1$ *for small enough constant* C_w^1 *and* C_a^1*. For any* $\delta \in (0,1)$*, we choose a small enough step size*

$$\eta = \left(\mathrm{poly}\left(p, k, R, \frac{1}{\gamma}, \log\frac{1}{\gamma}, \log\frac{1}{\delta}\right)\right)^{-1},$$

and a large enough batch size

$$m = \mathrm{poly}\left(p, k, R, \frac{1}{\gamma}, \log\frac{1}{\delta}\right),$$

then with at least probability $1-\delta$*, we have*

$$\|w_t - w^*\|_2^2 \leq C_{13}\gamma \quad \text{and} \quad \|a_t - a^*\|_2^2 \leq \gamma$$

for all t*'s such that* $\widehat{T}_2 \leq t \leq T = \widetilde{O}(\eta^{-2})$*, where* C_{13} *is a constant and*

$$\widehat{T}_2 = \mathrm{poly}\Big(p, k, R, \frac{1}{\gamma}, \log\frac{1}{\delta}\Big).$$

The proof of Lemma 22 is straightforward and therefore omitted, as the error of the mimi-batch stochastic gradient has been well controlled by a sufficiently large batch-size.