Supplementary Material

A. Omitted Details

The matrix $M_{\widehat{x}\leftrightarrow\widehat{z}}$ in the definition of WDC (Definition 3.7) is defined as

$$M_{\widehat{x}\leftrightarrow\widehat{z}} := U^\top \begin{bmatrix} \cos\angle(x,z) & \sin\angle(x,z) & 0 \\ \sin\angle(x,z) & -\cos\angle(x,z) & 0 \\ 0 & 0 & \mathbf{0}_{(p-2)\times(p-2)} \end{bmatrix} U,$$

where the matrix U denotes a rotation matrix such that $U\widehat{x} = e_1$ and $U\widehat{z} = \cos\angle(x,z)\cdot e_1 + \sin\angle(x,z)\cdot e_2$ with $e_1 = [1, 0, \cdots, 0]^\top$ and $e_2 = [0, 1, 0, \cdots, 0]^\top$. Moreover, if $\angle(x,z) = 0$ or $\angle(x,z) = \pi$, then we have $M_{\widehat{x}\leftrightarrow\widehat{z}} = \widehat{x}\widehat{x}^\top$ or $M_{\widehat{x}\leftrightarrow\widehat{z}} = -\widehat{x}\widehat{x}^\top$ respectively.

To verify Theorems 3.8 and 3.10, we illustrate the landscape of $L(x)$ in Figure 1. The simulation is based on a large sample number $m \to +\infty$, which intends to show the expectation of the risk $L(x)$. We are more interested in Case 1 of Theorem 3.8, where x_0 can be potentially recovered. By letting $x_0 = [1, 1]^\top$ which is sufficiently far away from the origin, Figure 1 shows that there are no stationary points outside the neighbors of x_0 and its negative multiple and the directional derivatives along any directions at the origin are negative, which matches the Case 1 of Theorem 3.8. In addition, the function values at the neighbor of x_0 is lower than that of its negative multiple, which therefore verifies the result in Theorem 3.10. The landscape will further inspire us to design efficient algorithms to solve the ERM in (4).

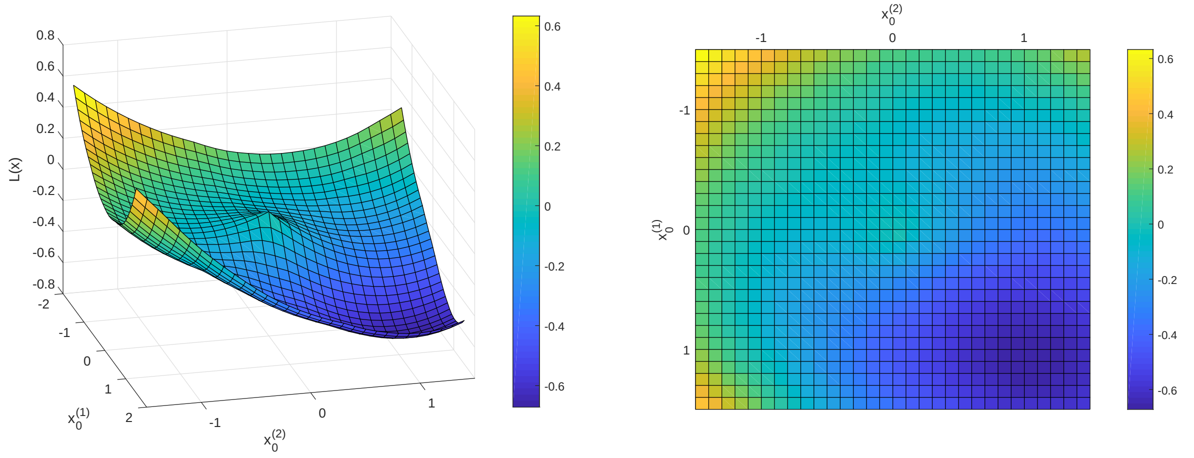

Figure 1. Illustration of landscape for $L(x)$. We build a two-layer ReLU network $G(\cdot)$ with input x_0 where $x_0 = [1, 1]^\top$, Gaussian weights $W_1 \in \mathbb{R}^{64\times 2}$ and $W_2 \in \mathbb{R}^{1024\times 64}$ such that $k = 2$ and $d = 1024$. The samples $\{(a_i, y_i)\}_{i=1}^m$ are generated via standard Guassian vector a_i and $y_i = \text{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$ with noise $\xi_i \sim \mathcal{N}(0, 0.01)$, dithering $\tau_i \sim \text{Unif}[-10, 10]$, and a large sample number $m \to \infty$.

B. Proof of Theorem 3.2

In this section, we provide the proofs of the two key lemmas, i.e., Lemma 4.1 and Lemma 4.2 as well as other supporting lemmas. The proof of Theorem 3.4 is immediately obtained by following Lemma 4.1 and Lemma 4.2 as shown in Section §4.1.

B.1. Bias of the Expected Risk

We prove Lemma 4.1 in this subsection.

Lemma B.1 (Lemma 4.1). *There exists an absolute constant $c_1 > 0$ such that the following holds:*

$$\left|\mathbb{E}[y_i\langle a_i, G(x) - G(x_0)\rangle] - \frac{1}{\lambda}\langle G(x_0), G(x) - G(x_0)\rangle\right|$$
$$\leq \sqrt{c_1(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}(\sqrt{2(\lambda+1)} + 2)e^{-\lambda/2(\|a\|_{\psi_1}R+\|\xi\|_{\psi_1})}\|G(x) - G(x_0)\|_2.$$

Furthermore, for any $\varepsilon \in (0,1)$, if $\lambda \geq 4C_{a,\xi,R} \cdot \log(64C_{a,\xi,R} \cdot \varepsilon^{-1})$ where $C_{a,\xi,R} = \max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}$, and $\|G(x) - G(x_0)\|_2 > \varepsilon$, then, we have

$$\|G(x)\|_2^2 - \|G(x_0)\|_2^2 - 2\lambda\mathbb{E}[y_i\langle a_i, G(x) - G(x_0)\rangle] \geq \frac{1}{2}\|G(x) - G(x_0)\|_2^2.$$

Proof of Lemma 4.1. Recall that $y_i = \operatorname{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$. For simplicity of notations, we set $V_i = \langle a_i, G(x_0)\rangle + \xi_i$ and $Z_i = \langle a_i, G(x) - G(x_0)\rangle$. Note first that due to the independence between V_i and τ_i, we have

$$\begin{aligned}\mathbb{E}[\operatorname{sign}(V_i + \tau_i)|V_i] =& \frac{V_i}{\lambda}\mathbf{1}_{\{|V_i|\leq\lambda\}} + \mathbf{1}_{\{V_i>\lambda\}} - \mathbf{1}_{\{V_i<-\lambda\}}\\ =& \frac{V_i}{\lambda} - \frac{V_i}{\lambda}\mathbf{1}_{\{|V_i|>\lambda\}} + \mathbf{1}_{\{V_i>\lambda\}} - \mathbf{1}_{\{V_i<-\lambda\}}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\left|\mathbb{E}[Z_i\operatorname{sign}(V_i+\tau_i)] - \frac{\mathbb{E}[Z_iV_i]}{\lambda}\right| &= \left|-\mathbb{E}\left[\frac{Z_iV_i}{\lambda}\mathbf{1}_{\{|V_i|>\lambda\}}\right] + \mathbb{E}\big[Z_i\mathbf{1}_{\{V_i>\lambda\}}\big] - \mathbb{E}\big[Z\mathbf{1}_{\{V_i>\lambda\}}\big]\right|\\ &\leq \left|\mathbb{E}\left[\frac{Z_iV_i}{\lambda}\mathbf{1}_{\{|V_i|>\lambda\}}\right]\right| + 2\left|\mathbb{E}\big[Z_i\mathbf{1}_{\{|V_i|>\lambda\}}\big]\right|\\ &\leq \frac{\|Z_i\|_{L_2}\cdot\|V_i\mathbf{1}_{\{|V_i|>\lambda\}}\|_{L_2}}{\lambda} + 2\|Z_i\|_{L_2}\Pr(|V_i|>\lambda)^{1/2},\end{aligned} \tag{12}$$

where the last line follows from Cauchy-Schwarz inequality. Now we bound these terms respectively. First of all, by the isotropic assumption of a_i, we have

$$\|Z_i\|_{L_2} = \big\{\mathbb{E}\big[|\langle a_i, G(x) - G(x_0)\rangle|^2\big]\big\}^{1/2} = \|G(x) - G(x_0)\|_2.$$

Next, we have

$$\begin{aligned}\|V_i\mathbf{1}_{\{|V_i|>\lambda\}}\|_{L_2} &= \mathbb{E}\big[V_i^2\mathbf{1}_{\{|V_i|>\lambda\}}\big]^{1/2} = \left(\int_\lambda^\infty w^2dP(w)\right)^{1/2}\\ &= \left(2\int_\lambda^\infty wP(|V_i|>w)dw\right)^{1/2} \leq \left(2c_1\int_\lambda^\infty we^{-w/\|\langle a_i,G(x_0)\rangle+\xi_i\|_{\psi_1}}dw\right)^{1/2}\\ &\leq \sqrt{2c_1(\lambda+1)\|\langle a_i, G(x_0)\rangle + \xi_i\|_{\psi_1}}e^{-\lambda/2\|\langle a_i,G(x_0)\rangle+\xi_i\|_{\psi_1}},\end{aligned}$$

where the second from the last inequality follows from sub-exponential assumption of $\langle a_i, G(x_0)\rangle + \xi_i$ and $c_1 > 0$ is an absolute constant. Note that

$$\|\langle a_i, G(x_0)\rangle + \xi_i\|_{\psi_1} \leq \|\langle a_i, G(x_0)\rangle\|_{\psi_1} + \|\xi_i\|_{\psi_1} \leq \|a\|_{\psi_1}\|G(x_0)\|_2 + \|\xi\|_{\psi_1} \leq \|a\|_{\psi_1}R + \|\xi\|_{\psi_1},$$

where we use the assumption that $\|G(x_0)\|_2 \leq R$. Substituting this bound into the previous one gives

$$\|V_i\mathbf{1}_{\{|V_i|>\lambda\}}\|_{L_2} \leq \sqrt{2c_1(\lambda+1)(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}e^{-\lambda/2(\|a\|_{\psi_1}R+\|\xi\|_{\psi_1})}.$$

Furthermore,

$$\Pr(|V_i| > \lambda)^{1/2} \leq \sqrt{c_1(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}e^{-\lambda/2(\|a\|_{\psi_1}R+\|\xi\|_{\psi_1})}.$$

Overall, substituting the previous computations into (12), we obtain

$$\begin{aligned}&\left|\mathbb{E}[Z_i \operatorname{sign}(V_i+\tau_i)] - \frac{\mathbb{E}[Z_iV_i]}{\lambda}\right| \\ &\qquad\le \sqrt{c_1(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}(\sqrt{2(\lambda+1)}/\lambda + 2)e^{-\lambda/2(\|a\|_{\psi_1}R+\|\xi\|_{\psi_1})}\|G(x)-G(x_0)\|_2,\end{aligned}$$

finishing the first part of the proof.

To prove the second part, we need to compute

$$\left|2\lambda\mathbb{E}[y_i\langle a_i, G(x)-G(x_0)\rangle] - 2\langle G(x_0), G(x)-G(x_0)\rangle\right| = 2\left|\lambda\mathbb{E}[Z_i \operatorname{sign}(V_i+\tau_i)] - \mathbb{E}[Z_iV_i]\right|.$$

Note that when $\varepsilon < 1$ and

$$\lambda \ge 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}\log(64\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon).$$

One can check that

$$\begin{aligned}&|\lambda\mathbb{E}[Z_i \operatorname{sign}(V_i+\tau_i)] - \mathbb{E}[Z_iV_i]| \\ &\qquad\le \sqrt{c_1(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}(\sqrt{2(\lambda+1)} + 2\lambda)e^{-\lambda/2(\|a\|_{\psi_1}R+\|\xi\|_{\psi_1})}\|G(x)-G(x_0)\|_2 \\ &\qquad\le \frac{1}{4}\varepsilon\|G(x)-G(x_0)\|_2.\end{aligned}$$

Thus, it follows

$$\begin{aligned}&\|G(x)\|_2^2 - \|G(x_0)\|_2^2 - 2\lambda\mathbb{E}[y_i\langle a_i, G(x)-G(x_0)\rangle] \\ &\qquad\ge \|G(x)\|_2^2 - \|G(x_0)\|_2^2 - 2\langle G(x_0), G(x)-G(x_0)\rangle - \frac{1}{2}\varepsilon\|G(x)-G(x_0)\|_2 \\ &\qquad= \|G(x)-G(x_0)\|_2^2 - \frac{1}{2}\varepsilon\|G(x)-G(x_0)\|_2.\end{aligned}$$

Thus, when $\|G(x)-G(x_0)\|_2 > \varepsilon$ the second claim holds. □

B.2. Analysis of Variances: Uniform Bounds of An Empirical Process

Our goal in this subsection is to prove Lemma 4.2. Note that one can equivalently write the $\{G(x_0) : \|G(x_0)\|_2 \le R,\ x_0 \in \mathbb{R}^k\}$ as $G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R)$, where $\mathbb{B}_2^d(R)$ denotes the ℓ_2-ball of radius R. The strategy of bounding this supremum is as follows: Consider a δ-covering net over the set $G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R)$, namely $\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$, and bounding the supremum over each individual covering ball. The δ value will be decided later.

B.2.1. Bounding Supremum Under Fixed Signs: A Covering Net Argument

First of all, since for any point $\theta \in G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R)$, there exists a $v \in \mathbb{R}^k$ such that $\theta = G(v)$, we use $G(v)$ to denote any point in the net $\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$. We replace each $\operatorname{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$ by $\operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)$ and have the following lemma regarding the supremum for each fixed $G(v)$:

Lemma B.2. *Let $c, c_1 > 0$ be some absolute constants. For any $u \ge 0$ and fixed $G(v)$, the following holds with probability at least $1 - 2\exp(-u - c_2kn\log ed)$,*

$$\begin{aligned}&\sup_{x\in\mathbb{R}^k,\ x_0\in\mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)\langle a_i, G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2} \\ &\qquad\le \sqrt{\frac{8(u + ckn\log(ed))}{m}} + \frac{4\|a\|_{\psi_1}(u + ckn\log(ed))}{m}.\end{aligned}$$

Proof of Lemma B.2. First of all, since v is fixed and ε_i is independent of $\operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)$, it follows the distribution of ε_i is the same as the distribution of $\varepsilon_i \operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)$. Thus, it is enough to work with the following supremum:

$$\sup_{x\in\mathbb{R}^k,\ x_0\in\mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\langle a_i, G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2}.$$

To this point, we will then use the piecewise linear structure of the ReLU function. Note that the ReLU network has n layers with each layer having at most d nodes, where each layer of the network is a linear transformation followed by at most d pointwise nonlinearities. Consider any node in the first layer, which can be written as $\max\{\langle w, x\rangle, 0\}$ with a weight vector w and an input vector x, splits the input space $\mathbb{R}^k$ into two disjoint pieces, namely $\mathcal{P}_1$ and $\mathcal{P}_2$, where for any input in $\mathcal{P}_1$, the node is a linear mapping $\langle w, x\rangle$ and for any input in $\mathcal{P}_2$ is the other linear mapping $\langle 0, x\rangle$.

Thus, each node in the first layer corresponds to a splitting hyperplane in $\mathbb{R}^k$. We have the following claim on the number of possible pieces split by d hyperplanes:

Claim 1: The maximum number of pieces when splitting $\mathbb{R}^k$ with d hyperplanes, denoted as $\mathcal{C}(d,k)$, is

$$\mathcal{C}(d,k) = \binom{d}{0} + \binom{d}{1} + \cdots + \binom{d}{k}.$$

The proof of this claim, which follows from, for example (Winder, 1966), is based on an induction argument on both d and k and omitted here for brevity. Note that $\mathcal{C}(d,k) \le d^k + 1$. For the second layer, we can consider each piece after the first layer, which is a subset of $\mathbb{R}^k$ and will then be further split into at most $d^k + 1$ pieces. Thus, we will get at most $(d^k+1)^2$ pieces after the second layer. Continuing this argument through all n layers and we have the input space $\mathbb{R}^k$ is split into at most $(d^k+1)^n \le (2d)^{kn}$ pieces, where within each piece the function $G(\cdot)$ is simply a linear transformation from $\mathbb{R}^k$ to $\mathbb{R}^d$.

Now, we consider any two pieces, namely $\mathcal{P}_1,\ \mathcal{P}_2 \subseteq \mathbb{R}^k$, from the aforementioned collection of pieces, and aim at bounding the following quantity:

$$\sup_{t_1\in\mathcal{P}_1, t_2\in\mathcal{P}_2} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, G(t_1) - G(t_2)\rangle\right|}{\|G(t_1) - G(t_2)\|_2}.$$

By the previous argument, we know that within $\mathcal{P}_1$ and $\mathcal{P}_2$, the function $G(\cdot)$ can simply be represented by some fixed linear maps W_1 and W_2, respectively. As a consequence, it suffices to bound

$$\begin{aligned}
&\sup_{t_1\in\mathcal{P}_1, t_2\in\mathcal{P}_2} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, W_1 t_1 - W_2 t_2\rangle\right|}{\|W_1t_1 - W_2t_2\|_2} \\
&\le \sup_{t_1,\ t_2\in\mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, W_1 t_1 - W_2 t_2\rangle\right|}{\|W_1t_1 - W_2t_2\|_2} \\
&\le \sup_{t\in\mathbb{R}^{2k}} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, W_0 t\rangle\right|}{\|W_0 t\|_2},
\end{aligned}$$

where $W_0 := [W_1,\ -W_2]$, and the last inequality follows from concatenating t_1 and t_2 to form a vector $t \in \mathbb{R}^{2k}$ and then expanding the set to take supremum over $t \in \mathbb{R}^{2k}$. Let $\mathcal{E}_{2k}$ be the subspace in $\mathbb{R}^d$ spanned by the $2k$ columns of W_0, then, the above supremum can be rewritten as

$$E_m := \sup_{b\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, b\rangle\right|.$$

To bound the supremum, we consider a $1/2$-covering net of the set $\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}$, namely, $\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)$. A simple volume argument shows that the cardinality $|\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)| \le 3^{2k}$.

By Bernstein's inequality (Lemma B.6), we have for any fixed $b \in \mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)$,

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, b\rangle\right| \ge \sqrt{\frac{2u'}{m}} + \frac{\|a\|_{\psi_1} u'}{m}\right) \le 2e^{-u'}.$$

Taking $u' = u + ckn\log(ed)$ for some $c > 6$, we have with probability at least $1 - 2\exp(-u - ckn\log(ed))$,

$$\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, b\rangle\right| \le \sqrt{\frac{2(u + ckn\log(ed))}{m}} + \frac{\|a\|_{\psi_1}(u + ckn\log(ed))}{m}.$$

Taking a union bound over all $b \in \mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)$, we have with probability at least $1-2\exp(-u-ckn\log(ed))\cdot 3^{2k} \geq 1-2\exp(-u-c_1kn\log(ed))$ for some absolute constant $c_1 > 2$.

$$\sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\langle a_i, b\rangle\right| \leq \sqrt{\frac{2(u+ckn\log(ed))}{m}} + \frac{\|a\|_{\psi_1}(u+ckn\log(ed))}{m}. \tag{13}$$

Let $P_{\mathcal{N}}(\cdot)$ be the projection of any point in $\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}$ onto $\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}, 1/2)$. we have

$$\begin{aligned} E_m &\leq \sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\langle a_i, b\rangle\right| + \sup_{b\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\langle a_i, b-P_{\mathcal{N}}(b)\rangle\right| \\ &\leq \sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\langle a_i, b\rangle\right| + \frac{1}{2}\sup_{b\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \frac{\varepsilon_i\langle a_i, b-P_{\mathcal{N}}(b)\rangle}{\|b-P_{\mathcal{N}}(b)\|_2}\right| \\ &\leq \sup_{b\in\mathcal{N}(\mathcal{E}^{2k}\cap\mathcal{S}^{d-1},1/2)} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\langle a_i, b\rangle\right| + \frac{1}{2}E_m, \end{aligned} \tag{14}$$

where the second inequality follows from the homogeneity of the set $\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}$ under constant scaling. Combining (13) and (14) gives

$$\sup_{b\in\mathcal{E}^{2k}\cap\mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\langle a_i, b\rangle\right| \leq 2\sqrt{\frac{2(u+ckn\log(ed))}{m}} + \frac{2\|a\|_{\psi_1}(u+ckn\log(ed))}{m}.$$

Taking a further union bound over at most $(2d)^{kn}$ different pair of subspaces $\mathcal{P}_1,\ \mathcal{P}_2$ finishes the proof. □

B.2.2. Counting the Sign Differences: A VC-Dimension Bound

In this section, we consider all possible sign changes replacing each $\operatorname{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$ by $\operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)$, where we recall $G(v)$ is a nearest point to $G(x_0)$ in $\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$.

First of all, since $\tau_i \sim \text{Unif}[-\lambda, +\lambda]$, for any $\eta > 0$, defining a new random variable $X_i := \langle a_i, G(v)\rangle + \xi_i$ which is thus independent of τ_i, for all $i = 1, 2, \cdots, m$, we have

$$\Pr(|\langle a_i, G(v)\rangle + \xi_i + \tau_i| \leq \eta) = \Pr(-\eta \leq X_i + \tau_i \leq \eta) \leq \frac{\eta}{\lambda},$$

by computing the integral of the probability density functions of X_i and τ_i in $-\eta \leq X_i + \tau_i \leq \eta$. Using Chernoff bound (Lemma B.8), one has with probability at least $1-\exp(-\eta m/3\lambda)$,

$$\sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, G(v)\rangle + \xi_i + \tau_i| \geq \eta\}} \geq \left(1 - \frac{2\eta}{\lambda}\right) m. \tag{15}$$

Next, we prove the following lemma:

Lemma B.3. *Let $\eta, \delta > 0$ be chosen parameters. For any $u \geq 0$ and fixed $G(v)$, the following holds with probability at least $1-2\exp(-u)$,*

$$\sup_{x_0\in\mathbb{R}^k, \|G(x_0)-G(v)\|_2\leq\delta} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, G(x_0)-G(v)\rangle| \geq \eta\}} \leq m\cdot\Pr(|\langle a_i, z\rangle| \geq \eta/\delta) + L\sqrt{(kn\log(ed)+u)m},$$

where z is any fixed vector in $\mathbb{B}_2^d(1)$ and $L > 1$ is an absolute constant.

This lemma implies that the counting process $\{\mathbf{1}_{\{|\langle a_i, G(x_0)-G(v)\rangle| \geq \eta\}}\}_{i=1}^m$ enjoys a tight sub-Gaussian uniform concentration. The proof relies on a book-keeping VC dimension argument.

Proof of Lemma B.3. First of all, let $T = G(\mathbb{R}^k)$, and it suffices to bound the following supremum:

$$\sup_{t\in(T-T)\cap\mathbb{B}_2^d(\delta)} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta\}}.$$

Let $\mathcal{T}$ be the set of all distinctive pieces split by $G(\cdot)$. By the same argument as that of Lemma B.2, the cardinality of $\mathcal{T}$ is at most $(d^k+1)^n \leq (2d)^{kn}$, and we have

$$\begin{aligned}
&\sup_{t\in(T-T)\cap\mathbb{B}_2^d(\delta)} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta\}}\\
&\leq \sup_{\mathcal{P}_1,\ \mathcal{P}_2\in\mathcal{T}, t\in(\mathcal{P}_1-\mathcal{P}_2)\cap\mathbb{B}_2^d(\delta)} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta\}}\\
&\leq \sup_{\mathcal{P}_1,\ \mathcal{P}_2\in\mathcal{T},\ t\in\text{affine}(\mathcal{P}_1-\mathcal{P}_2)\cap\mathbb{B}_2^d(\delta)} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta\}}\\
&= \sup_{\mathcal{P}_1,\ \mathcal{P}_2\in\mathcal{T},\ t\in\text{affine}(\mathcal{P}_1-\mathcal{P}_2)\cap\mathbb{B}_2^d(1)} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta/\delta\}},
\end{aligned}$$

where affine$(\mathcal{P}_1-\mathcal{P}_2)$ denotes the affine subspace spanned by $\mathcal{P}_1-\mathcal{P}_2$, which is of dimension at most $2k$. To this point, define $a_1^m := \{a_i\}_{i=1}^m$, define the set

$$\mathcal{C} := \{t : t\in \text{affine}(\mathcal{P}_1-\mathcal{P}_2)\cap\mathbb{B}_2^d(1), \mathcal{P}_1,\ \mathcal{P}_2\in\mathcal{T}\}, \tag{16}$$

and define an empirical process

$$\mathcal{R}(a_1^m,t) := \frac{1}{m}\sum_{i=1}^m \left(\mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta/\delta\}} - \mathbb{E}\big[\mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta/\delta\}}\big]\right).$$

Our goal is to bound

$$\sup_{t\in\mathcal{C}} |\mathcal{R}(a_1^m,t)|.$$

By symmetrization inequality (Lemma B.7) it suffices to bound

$$\sup_{t\in\mathcal{C}}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta/\delta\}}\right|,$$

where $\{\varepsilon\}_{i=1}^m$ are i.i.d. Rademacher random variables. Define the set of indicator functions:

$$\mathcal{F} := \{\mathbf{1}_{\{|\langle\cdot,t\rangle|\geq\eta/\delta\}} :\ t\in\mathcal{C}\}.$$

By Hoeffding's inequality, the stochastic process $m^{-1/2}\sum_{i=1}^m \varepsilon_i\mathbf{1}_{\{|\langle a_i,t\rangle|\geq\eta/\delta\}}$ parametrized by $\mathcal{F}$ when fixing a_1^m is a sub-Gaussian process with respect to the empirical L_2 metric:

$$\|f-g\|_{L_2(\mu_m)} := \sqrt{\frac{1}{m}\sum_{i=1}^m (f(a_i)-g(a_i))^2},\ \forall f,g\in\mathcal{F}.$$

By Lemma B.9, one can easily derive the following bound:

$$\mathbb{E}\left[\sup_{t\in\mathcal{C}}|\mathcal{R}(a_1^m,t)|\right] \leq \frac{C_0}{\sqrt{m}}\int_0^2 \sqrt{\log|\mathcal{N}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_{L_2(\mu_m)})|}d\varepsilon, \tag{17}$$

where $\mathcal{N}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_{L_2(\mu_m)})$ is the ε-covering net of $\mathcal{F}$ under the empirical L_2-metric. By Haussler's inequality (Theorem 2.6.4 of (Wellner et al., 2013)),

$$|\mathcal{N}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_{L_2(\mu_m)})| \leq C_1 V(\mathcal{F})(4e)^{V(\mathcal{F})}\left(\frac{1}{\varepsilon}\right)^{2V(\mathcal{F})},$$

where $V(\mathcal{F})$ is the VC dimension of the class $\mathcal{F}$ and C_1 is an absolute constant. To compute $V(\mathcal{F})$, note first that for any fixed $\mathcal{P}_1,\ \mathcal{P}_2 \in \mathcal{T}$ and any fixed constant c, the VC dimension of the class of half-spaces defined as

$$\mathcal{H}' := \{\langle \cdot, t\rangle \geq c:\ t \in \text{affine}(\mathcal{P}_1 - \mathcal{P}_2)\}$$

is bounded by $2k$. Thus, for any p points on $\mathbb{R}^k$ and the number of different subsets of these points picked by $\mathcal{H}'$ is bounded by $(p+1)^{2k}$. Next, note that any element in the class

$$\mathcal{H} := \{|\langle \cdot, t\rangle| \geq c:\ t \in \text{affine}(\mathcal{P}_1 - \mathcal{P}_2)\}$$

is the intersection of two halfspaces in $\mathcal{H}'$. Thus, the number of different subsets of p points picked by $\mathcal{H}$ is bounded by

$$\binom{(p+1)^{2k}}{2} \leq e^2(p+1)^{4k}/4 \leq 2(p+1)^{4k}.$$

Taking into account that the class $\mathcal{F}$ is the union of at most $(2d)^{2kn}$ different classes of the form

$$\{\mathbf{1}_{\{|\langle \cdot, t\rangle| \geq \eta/\delta\}}:\ t \in \text{affine}(\mathcal{P}_1 - \mathcal{P}_2)\},$$

we arrive at the conclusion that the number of distinctive mappings in $\mathcal{F}$ from any p points in $\mathbb{R}^k$ to $\{0,1\}^p$ is bounded by $2d^{2kn}(p+1)^{4k}$. To get the VC dimension of $\mathcal{F}$, we try to find the smallest p such that

$$2d^{2kn}(p+1)^{4k} < 2^p.$$

A sufficient condition is to have $2kn\log_2(d) + 4k\log_2(p+1) + 1 < p$, which holds when $p > c_0 kn\log(ed) - 1$ for some absolute constant c_0 large enough. Thus, $V(\mathcal{F}) \leq c_0 kn\log(ed)$. Thus, it follows

$$\begin{aligned}\log|\mathcal{N}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_{L_2(\mu_m)})| &\leq \log C_1 + \log V(\mathcal{F}) + V(\mathcal{F})\log(4e) + 2V(\mathcal{F})\log(1/\varepsilon)\\ &\leq c_1 kn\log(ed)(\log(1/\varepsilon)+1),\end{aligned}$$

for some absolute constant $c_1 > 0$. Substituting this bound into (17), and we obtain

$$\mathbb{E}\left[\sup_{t\in\mathcal{C}}|\mathcal{R}(a_1^m, t)|\right] \leq c_2\sqrt{\frac{kn\log(ed)}{m}},$$

for some absolute constant c_2. Finally, by bounded difference inequality, we obtain with probability at least $1 - 2e^{-u}$,

$$\sup_{t\in\mathcal{C}}|\mathcal{R}(a_1^m, t)| \leq \mathbb{E}\left[\sup_{t\in\mathcal{C}}|\mathcal{R}(a_1^m, t)|\right] + \sqrt{\frac{u}{m}} \leq L\sqrt{\frac{kn\log(ed)+u}{m}},$$

finishing the proof. □

Combining Lemma B.3 and (15) we have the following bound on the number of sign differences:

Lemma B.4. *Let $u > 0$ be any constant. Suppose $m \geq c_2\lambda^2(kn\log(ed) + k\log(2R) + u)$ for some absolute constant c_2 large enough and $\lambda \geq 1$. Define the following parameters:*

$$\delta = \frac{\eta}{\|a\|_{\psi_1}}\log(c_1\lambda/\eta), \tag{18}$$

$$\eta = (\lambda + \|a\|_{\psi_1})L\sqrt{\frac{kn\log(ed)+u'}{m}}, \tag{19}$$

and $u' > 0$ satisfying

$$u' = u + kn\log(ed) + k\log(2R) + Ck\log\left(\frac{m}{k\log(ed)+u'}\right). \tag{20}$$

We have with probability at least $1 - \exp(-cu) - 2\exp(-u)$,

$$\sup\ \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{\text{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i) \neq \text{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)\}} \leq \frac{4\eta}{\lambda},$$

where the supremum is taken over $x_0 \in \mathbb{R}^k$, $\|G(x_0) - G(v)\|_2 \leq \delta, G(v) \in \mathcal{N}(G(\mathbb{R}^k) \cap \mathbb{B}_2^d(R),\ \delta)$ and $c, c_1, C, L > 0$ are absolute constants.

Proof of Lemma B.4. We compute $\Pr(|\langle a_i, z\rangle| \geq \eta/\delta)$. By the fact that $\langle a_i, z\rangle$ is a sub-exponential random variable,

$$\Pr(|\langle a_i, z\rangle| \geq \eta/\delta) \leq c_1 \exp\left(-\frac{\eta}{\delta\|a\|_{\psi_1}}\right),$$

where $c_1 > 0$ is an absolute constant. We choose δ according to (18), which implies

$$\Pr(|\langle a_i, z\rangle| \geq \eta/\delta) \leq \frac{\eta}{\lambda}.$$

From Lemma B.3, we readily obtain with probability at least $1 - 2\exp(-u')$,

$$\sup_{x_0\in\mathbb{R}^k, \|G(x_0)-G(v)\|_2\leq\delta} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, G(x_0)-G(v)\rangle|\geq\eta\}} \leq \left(\frac{\eta}{\lambda} + L\sqrt{\frac{kn\log(ed)+u'}{m}}\right)m. \tag{21}$$

We will then take a further supremum over all $G(v) \in \mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$. Note that by a simple volume argument, $\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$ satisfies

$$\log|\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)| \leq kn\log(ed) + k\log(2R/\delta).$$

Choose η according to (19). Then, By the aforementioned choices of η and δ in (19) and (18), we obtain

$$\log(1/\delta) \leq C\log\left(\frac{m}{k\log(ed)+u'}\right),$$

where C is an absolute constant. Thus,

$$\log|\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)| \leq kn\log(ed) + k\log(2R) + Ck\log\left(\frac{m}{k\log(ed)+u'}\right). \tag{22}$$

Finally, for any $u > 0$, take u' so that it satisfies (20). By (21), we obtain that, with probability at least

$$1 - 2\exp\left(-u - kn\log(ed) - k\log(2R) - Ck\log\left(\frac{m}{k\log(ed)+u'}\right)\right),$$

the following holds

$$\sup_{x_0\in\mathbb{R}^k, \|G(x_0)-G(v)\|_2\leq\delta} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, G(x_0)-G(v)\rangle|\geq\eta\}} \leq \left(\frac{\eta}{\lambda} + L\sqrt{\frac{kn\log(ed)+u'}{m}}\right)m.$$

Taking a union bound over all $G(v) \in \mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$, we get with probability at least $1 - 2\exp(-u)$,

$$\sup_{x_0\in\mathbb{R}^k, \|G(x_0)-G(v)\|_2\leq\delta, G(v)\in\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, G(x_0)-G(v)\rangle|\geq\eta\}} \leq \left(\frac{\eta}{\lambda} + L\sqrt{\frac{kn\log(ed)+u'}{m}}\right)m.$$

Note that by definition of η in (19), $L\sqrt{(kn\log(ed)+u')/m} \leq \eta/\lambda$, and this readily implies with probability at least $1 - 2\exp(-u)$,

$$\sup_{x_0\in\mathbb{R}^k, \|G(x_0)-G(v)\|_2\leq\delta, G(v)\in\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i, G(x_0)-G(v)\rangle|\geq\eta\}} \leq \frac{2\eta}{\lambda}m. \tag{23}$$

Moreover, taking a union bound over all $G(v) \in \mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)$ in (15), we have with probability at least

$$\begin{aligned}
&1 - \exp(\log|\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)| - \eta m/3\lambda) \\
&\quad \geq 1 - \exp\left(kn\log(ed) + k\log(2R) + Ck\log\left(\frac{m}{k\log(ed)+u'}\right) - \eta m/3\lambda\right),
\end{aligned}$$

one has

$$\inf_{G(v)\in\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \quad \sum_{i=1}^m \mathbf{1}_{\{|\langle a_i,G(v)\rangle+\xi_i+\tau_i|\geq\eta\}} \geq (1-\frac{2\eta}{\lambda})m. \tag{24}$$

Note that by assumption, we have $m \geq c_2\lambda^2(kn\log(ed)+k\log(2R)+u)/\varepsilon^2$ for some $\varepsilon<1$ and some absolute constant c_2 large enough. Thus, it follows

$$\begin{aligned}\frac{\eta m}{3\lambda} =& \frac{L}{3}\sqrt{(kn\log(ed)+u')m}\\ \geq& \frac{L}{3}\sqrt{\left(u+kn\log(ed)+k\log(2R)+Ck\log\frac{m}{kn\log(ed)+u'}\right)m}\\ \geq& C'\left(u+kn\log(ed)+k\log(2R)+\sqrt{km\log\frac{m}{kn\log(ed)+u'}}\right)\\ \geq& C'\left(u+kn\log(ed)+k\log(2R)+k\log\frac{m}{k\log(ed)+u'}\right),\end{aligned}$$

where C' is an absolute constant related to $L,\ c_2,\ C$, and the last inequality follows from the assumption that $m \geq \sqrt{km} \geq \sqrt{k\log m}$ for any $m\geq 1$. Overall, when c_2 is large enough so that $C' > C$, we have (24) holds with probability at least $1-\exp(-C'u)$. Overall, combining (23) and (24) we finish the proof. □

B.2.3. PUTTING BOUNDS TOGETHER: PROOF OF LEMMA 4.2

Lemma B.5 (Lemma 4.2)**.** *Suppose Assumption 3.1 holds and*

$$m \geq c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed)+k\log(2R)+k\log m+u)/\varepsilon^2, \tag{25}$$

for some absolute constant c_2 large enough, then, with probability at least $1-c\exp(-u)$,

$$\sup_{x_0\in\mathbb{R}^k,\ \|G(x_0)\|_2\leq R,\ x\in\mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\,\mathrm{sign}(\langle a_i,G(x_0)\rangle+\xi_i+\tau_i)\langle a_i,G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2} \leq \frac{\varepsilon}{16\lambda},$$

where $\{\varepsilon_i\}_{i=1}^m$ are i.i.d. Rademacher random variables and $c>0$ is an absolute constant.

Proof of Lemma 4.2. Let $\mathcal{I}$ be the set of indices such that $\mathrm{sign}(\langle a_i,G(v)\rangle+\xi_i+\tau_i)\neq\mathrm{sign}(\langle a_i,G(x_0)\rangle+\xi_i+\tau_i)$. By Lemma B.4, we know that $|\mathcal{I}|\leq 4\eta/\lambda$. Now, we have with probability at least $1-\exp(-cu)-2\exp(-u)$,

$$\begin{aligned}&\sup_{x_0\in\mathbb{R}^k,\ \|G(x_0)\|_2\leq R,\ x\in\mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\,\mathrm{sign}(\langle a_i,G(x_0)\rangle+\xi_i+\tau_i)\langle a_i,G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2}\\ &\leq \sup_{x\in\mathbb{R}^k,\ x_0\in\mathbb{R}^k,G(v)\in\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i^v\langle a_i,G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2}\\ &\quad+ \sup_{x\in\mathbb{R}^k,x_0\in\mathbb{R}^k,\|G(x_0)-G(v)\|_2\leq\delta,G(v)\in\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i(y_i-y_i^v)\langle a_i,G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2}\\ &\leq \underbrace{\sup_{x\in\mathbb{R}^k,\ x_0\in\mathbb{R}^k,G(v)\in\mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\,\mathrm{sign}(\langle a_i,G(v)\rangle+\xi_i+\tau_i)\langle a_i,G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2}}_{\text{(I)}}\\ &\quad+\underbrace{\sup_{x\in\mathbb{R}^k,x_0\in\mathbb{R}^k}\ \max_{|\mathcal{I}|\leq 4\eta/\lambda}\frac{2}{m}\sum_{i\in\mathcal{I}}\frac{|\langle a_i,G(x)-G(x_0\rangle|}{\|G(x)-G(x_0\|_2}}_{\text{(II)}},\end{aligned}$$

where, for simplicity, we let $y_i^v := \mathrm{sign}(\langle a_i,G(v)\rangle+\xi_i+\tau_i)$ be the sign function associated with $G(v)$ in the net in the first inequality, and the second inequality is according to Lemma B.4.

For the rest of the proof, we will bound (I) and (II) respectively. To bound (I), take u in Lemma B.2 to be $kn\log(ed) + k\log(2R) + Ck\log m + u$, we have with probability at $1 - 2\exp(-c_2 kn\log(ed) - k\log(2R) - Ck\log m - u)$, for a fixed $G(v)$, any $x \in \mathbb{R}^k,\ x_0 \in \mathbb{R}^k$,

$$\frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)\langle a_i, G(x) - G(x_0)\rangle\right|}{\|G(x) - G(x_0)\|_2}$$
$$\leq \sqrt{\frac{8(ckn\log(ed) + k\log(2R) + Ck\log m + u)}{m}} + \frac{4\|a\|_{\psi_1}(ckn\log(ed) + k\log(2R) + Ck\log m + u)}{m},$$

where $c, c_2, C > 0$ are absolute constants. Take a further union bound over all $G(v) \in \mathcal{N}(G(\mathbb{R}^k) \cap \mathbb{B}_2^d(R),\ \delta)$ with the net size satisfying (22), we have with probability at least $1 - 2\exp(-u)$,

$$\text{(I)} \leq \sqrt{\frac{8(ckn\log(ed) + k\log(2R) + Ck\log m + u)}{m}} + \frac{4\|a\|_{\psi_1}(ckn\log(ed) + k\log(2R) + Ck\log m + u)}{m}. \tag{26}$$

Next, we will bound the term (II). Let $t = (G(x) - G(x_0))/\|G(x) - G(x_0)\|_2$ and it is enough to bound

$$\sup_{x_0 \in \mathbb{R}^k, x_0 \in \mathbb{R}^k} \max_{|\mathcal{I}| \leq 4\eta/\lambda} \frac{1}{m}\sum_{i\in\mathcal{I}} |\langle a_i, t\rangle| - \mathbb{E}[|\langle a_i, t\rangle|] + \mathbb{E}[|\langle a_i, t\rangle|].$$

It is obvious that $|\langle a_i, t\rangle| - \mathbb{E}[|\langle a_i, t\rangle|]$ is also a sub-exponential random variable with sub-exponential norm bounded by $2\|a\|_{\psi_1}$, and $\mathbb{E}[|\langle a_i, t\rangle|] \leq 1$. Thus, by Bernstein's inequality,

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} |\langle a_i, t\rangle| - \mathbb{E}[|\langle a_i, t\rangle|] \leq \frac{2\sqrt{u_2}}{\sqrt{|\mathcal{I}|}} + \frac{2\|a\|_{\psi_1} u_2}{|\mathcal{I}|},$$

with probability at least $1 - 2\exp(-u_2)$. Thus,

$$\frac{1}{m}\sum_{i\in\mathcal{I}} |\langle a_i, t\rangle| - \mathbb{E}[|\langle a_i, t\rangle|] \leq \frac{1}{m}(2\sqrt{u_2|\mathcal{I}|} + 2\|a\|_{\psi_1} u_2).$$

Here we take

$$u_2 = C_1 \log\left(\frac{\lambda m}{kn\log(ed) + u'}\right)\left(u + 2kn\log(ed) + k\log(2R) + k\log\frac{m}{kn\log(ed) + u'}\right)^{1/2}\sqrt{m},$$

where C_1 is an absolute constant large enough and u' satisfies (20). Using the fact that

$$|\mathcal{I}| \leq \frac{4\eta}{\lambda} \leq 2L\Big(u + 2kn\log(ed) + k\log(2R) + k\log\frac{m}{kn\log(ed) + u'}\Big)^{1/2}\sqrt{m},$$

we have with probability at least

$$1 - 2\exp\left(-C_1\log\left(\frac{\lambda m}{kn\log(ed) + u'}\right)\left(u + 2kn\log(ed) + k\log(2R) + k\log\frac{m}{kn\log(ed) + u'}\right)^{1/2}\sqrt{m}\right),$$

the following holds:

$$\left|\frac{1}{m}\sum_{i\in\mathcal{I}} |\langle a_i, t\rangle| - \mathbb{E}[|\langle a_i, t\rangle|]\right|$$
$$\leq C_1\|a\|_{\psi_1}\log\left(\frac{\lambda m}{kn\log(ed) + u'}\right)\sqrt{\frac{u + 2kn\log(ed) + k\log(2R) + k\log\frac{m}{kn\log(ed)+u'}}{m}}. \tag{27}$$

To bound the maximum over $|\mathcal{I}| \leq 4\eta/\lambda$, we take a union bound over all $\binom{m}{4\eta m/\lambda}$ possibilities, where

$$\binom{m}{4\eta m/\lambda} \leq \left(\frac{em}{4\eta m/\lambda}\right)^{4\eta m/\lambda} = \left(\frac{\lambda}{\eta}\right)^{4\eta m/\lambda}.$$

Thus, it follows from the definition of η in terms of λ in Lemma B.4,

$$\log\binom{m}{4\eta m/\lambda} \leq \frac{4\eta m}{\lambda}\log\frac{\lambda}{\eta}$$
$$\leq L\Big(u + 2kn\log(ed) + k\log(2R) + \log\frac{m}{kn\log(ed)+u'}\Big)^{1/2}\cdot\sqrt{m}\log\Big(\frac{\lambda m}{kn\log(ed)+u'}\Big),$$

and when $C_1 > L$, the union bound gives, with probability at least

$$1 - 2\exp\left(-C_2\log\left(\frac{\lambda m}{kn\log(ed)+u'}\right)\left(u + 2kn\log(ed) + k\log(2R) + \log\frac{m}{kn\log(ed)+u'}\right)^{1/2}\sqrt{m}\right),$$

the quantity

$$\max_{|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}|\langle a_i,t\rangle| - \mathbb{E}[|\langle a_i,t\rangle|]\right|$$

is also bounded by the right hand side of (27) with a possibly different constant C_1, where $t = (G(x)-G(x_0))/\|G(x)-G(x_0)\|_2$. Now, using the same trick as that of Lemma B.2, we obtain

$$\sup_{x\in\mathbb{R}^k, x_0\in\mathbb{R}^k}\max_{|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}|\langle a_i,t\rangle| - \mathbb{E}[|\langle a_i,t\rangle|]\right|$$

is bounded by the right hand side of (27) with a possibly different constant C_1 and with probability

$$1 - 2\cdot 3^{2k}(2d)^{kn}\exp\left(-C_2\log\left(\frac{\lambda m}{kn\log(ed)+u'}\right)\left(u + 2kn\log(ed) + k\log(2R) + \log\frac{m}{kn\log(ed)+u'}\right)^{1/2}\sqrt{m}\right),$$

where C_2 is another absolute constant. Note that by assumption in Theorem 3.2,

$$m \geq c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon^2,$$

for some absolute constant c_2 large enough. This implies

$$\text{(II)} \leq 2C_1\|a\|_{\psi_1}\log\left(\frac{\lambda m}{kn\log(ed)+u'}\right)\sqrt{\frac{u + 2kn\log(ed) + k\log(2R) + k\log\frac{m}{kn\log(ed)+u'}}{m}},$$

with probability at least $1 - c_3\exp(-u)$, where $c_3 \geq 1$ is an absolute constant. Combining this bound with (26) and using (25), we obtain with probability $1 - c_3\exp(-u) - \exp(-cu) - 2\exp(-u)$,

$$\sup_{x_0\in\mathbb{R}^k,\ \|G(x_0)\|_2\leq R,\ x\in\mathbb{R}^k}\frac{\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\,\mathrm{sign}(\langle a_i,G(x_0)\rangle + \xi_i + \tau_i)\langle a_i, G(x)-G(x_0)\rangle\right|}{\|G(x)-G(x_0)\|_2} \leq \frac{\varepsilon}{16\lambda}.$$

This finishes the proof. □

B.3. Useful Probability Bounds for Proving Theorem 3.2

We recall the following well-known concentration inequality.

Lemma B.6 (Bernstein's inequality). *Let $X_1,\cdots,X_m$ be a sequence of independent centered random variables. Assume that there exist positive constants f and D such that for all integers $p\geq 2$*

$$\frac{1}{m}\sum_{i=1}^m\mathbb{E}[|X_i|^p] \leq \frac{p!}{2}f^2D^{p-2},$$

then

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^m X_i\right| \geq \frac{f}{\sqrt{m}}\sqrt{2u} + \frac{D}{m}u\right) \leq 2\exp(-u).$$

In particular, if $X_1,\cdots,X_m$ are all sub-exponential random variables, then f and D can be chosen as $f = \frac{1}{m}\sum_{i=1}^m\|X_i\|_{\psi_1}$ and $D = \max_{i=1\ldots m}\|X_i\|_{\psi_1}$.

The following version of Symmetrization inequality can be found, for example, in (Wellner et al., 2013).

Lemma B.7 (Symmetrization inequality)**.** *Let $\{Z_t(i)\}_{i=1}^m$ be i.i.d. copies of a mean 0 stochastic process $\{Z_t :\ t \in T\}$. For every $1 \leq i \leq m$, let $g_t(i) :\ T \to \mathbb{R}$ be an arbitrary function. Let $\{\varepsilon_i\}_{i=1}^m$ be a sequence of independent Rademacher random variables. Then, for every $x > 0$,*

$$\left(1 - \frac{4m}{x^2} \sup_{t\in T} var(Z_t)\right) \cdot \Pr\left(\sup_{t\in T}\left|\sum_{i=1}^m Z_t(i)\right| > x\right) \leq 2\Pr\left(\sup_{t\in T}\left|\sum_{i=1}^m \varepsilon_i(Z_t(i) - g_t(i))\right| > \frac{x}{4}\right),$$

where $var(Z_t) = \mathbb{E}\big[(Z_t - \mathbb{E}[Z_t])^2\big]$.

The following classical bound can be found, for example in Proposition 2.4 of Angluin & Valiant (1979).

Lemma B.8 (Chernoff bound)**.** *Let $X_1, \ldots, X_n$ be a sequence of i.i.d. copies of X such that $\Pr(X = 1) = 1 - \Pr(X = 0) = p \in (0, 1)$, and define $S_n := \sum_{i=1}^n X_i$. Then*

$$\Pr\Big(\frac{S_n}{n} \geq (1+\tau)p\Big) \leq \inf_{\theta>0}\left[e^{-\theta np(1+\tau)}\mathbb{E}e^{\theta S_n}\right] \leq \begin{cases} e^{-\frac{\tau^2 np}{2+\tau}}, & \tau > 1, \\ e^{-\frac{\tau^2 np}{3}}, & 0 < \tau \leq 1. \end{cases}$$

The following bound is the well-known Dudley's entropy estimate which can be found, for example, in Corollary 2.2.8 of (Wellner et al., 2013).

Lemma B.9 (Dudley's entropy bound)**.** *Let (T, d) be an arbitrary semi-metric space, and let $\{X_t,\ t \in T\}$ be a separable sub-Gaussian stochastic process with* [5]

$$\|X_s - X_t\|_{\psi_2} \leq Cd(s,t),\ \forall s,t \in T,$$

for some constant $C > 0$. Then, for every $r > 0$,

$$\mathbb{E}\left[\sup_{d(s,t)\leq r} |X_s - X_t|\right] \leq C_0 \int_0^r \sqrt{\log \mathcal{N}(\varepsilon, d)}d\varepsilon,$$

where $\mathcal{N}(\varepsilon, d)$ is the ε covering number of the set T and C_0 is an absolute constant.

C. Proof of Theorem 3.4

We provide detailed proofs of Proposition 4.4 and Lemma 4.5 in this section. As shown in Section §4.2, Theorem 3.4 can be proved immediately following Proposition 4.4 and Lemma 4.5.

Definition C.1. *A vector $v \in \mathbb{R}^d$ is k-group sparse if, when dividing v into k blocks of sub-vectors of size d/k,*[6] *each block has exactly one non-zero entry.*

Proposition C.2 (Proposition 4.4)**.** *Any nonnegative k-group sparse vector in $\mathbb{B}_2^d(1)$ can be generated by a ReLU network of the form (3) with a $k+1$ dimensional input and and depth $n = 3$.*

Proof of Proposition 4.4. Consider an $k+1$ dimensional input of a network. The idea is to map each of the first k entries of the input into a block in $\mathbb{R}^d$ of length d/k, respectively, and use one another input entry to construct proper offsets.

We first construct a single hidden layer ReLU network (i.e. $n = 2$) *with offsets* and k dimensional input $[x_1,\ \cdots, x_k]^T$ that can generate all positive k-group sparse signals. For each entry x_i of $x \in \mathbb{R}^k$, we consider a sequence of functions of the form:

$$\widetilde{\Gamma}_r(x_i) := \sigma(\sigma(x_i - 2r) - 2\sigma(x_i - 2r - 1)),\ \ r \in \left\{1,\ 2,\ \cdots,\ \frac{d}{k}\right\}. \tag{28}$$

Graphically, it is a sequence of d/k non-overlapping triangle functions on the positive real line with width 2 and height 1. We use outputs of $\widetilde{\Gamma}_r(x_i)$ over all r as the output of the i-th block in $\mathbb{R}^d$. It then follows that for any $x_i \in \mathbb{R}$, there is only

[5] For a sub-Gaussian random variable X, the ψ_2-norm is defined as $\sup_{p\geq 1} p^{-1/2}\|X\|_{L_p}$.

[6] We assume WLOG that d/k is an integer.

one of $\widetilde{\Gamma}_r(x_i)$ that can be nonzero. Furthermore, the nonzero entry can take any value in $[0, 1]$. Thus, lining up all k blocks constructed in such a way, we have any positive k-group sparse vector in $\mathbb{B}_\infty^d(1)$ can be generated by this network, and so does any vector in $\mathbb{B}_2^d(1)$.

To represent such a network above using a ReLU network *with no offset*, we add another hidden layer of width $(k + 2d/k)$ before passing to $\widetilde{\Gamma}_r(\cdot)$ and make use of the additional $k+1$ entries. The proposed network with a $k+1$ dimensional input of the form: $[x_1, \cdots, x_k, z]^T$ can be constructed as follows. The first k nodes are:

$$\sigma(x_i), \;\; i \in \{1, 2, \cdots, k\}.$$

The next $2d/k$ nodes are used to construct the offsets:

$$\sigma(r \cdot z), \;\; r \in \left\{1, 2, \cdots, \frac{2d}{k}\right\}.$$

The second and the third hidden layers are almost the same as (28) mapping each $\sigma(x_i)$ into a block in $\mathbb{R}^d$ of length d/k, except that we replace the offsets $2r$ and $2r+1$ by the output computed in the first hidden layer, i.e., $\sigma(r \cdot z)$. Then, we construct the second layer that can output the following results for all $i \in \{1, 2, ..., k\}$ and $r \in \{1, 2, ..., d/k\}$:

$$\Upsilon_r(x_i, z) = \sigma(\sigma(x_i) - 2\sigma(r \cdot z)) \quad \text{and} \quad \Upsilon'_r(x_i, z) = \sigma(\sigma(x_i) - 2\sigma(r \cdot z) - \sigma(z)).$$

Finally, by constructing the third layer, we have for all $i \in \{1, 2, ..., k\}$ and $r \in \{1, 2, ..., d/k\}$

$$\Gamma_r(x_i, z) := \sigma\big(\Upsilon_r(x_i, z) - 2\Upsilon'_r(x_i, z)\big). \tag{29}$$

Note that (28) fires only when $x_i \geq 0$, on which case we have $\sigma(x_i) = x_i$. Finally, we take z always equal to 1 and obtain $\widetilde{\Gamma}_r(x_i) = \Gamma_r(x_i, 1)$. Thus, the proposed network (29) can generate all nonnegative k-group sparse signals in $\mathbb{B}_2^d(1)$. □

Furthermore, based on the next two lemmas, we give the proof of Lemma 4.5.

Lemma C.3 (Theorem 4.2 of Plan et al. (2016)). *Assume that $\theta_0 \in K$ where $K \subseteq \mathbb{R}^d$ satisfies $\lambda v \in K$ for any $v \in K$ and $\lambda \in [0, 1)$. Assume that $\check{y} = \langle a, \theta_0 \rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$ and $a \sim \mathcal{N}(0, \mathbf{I}_d)$. Let*

$$\delta_* := \inf_{t>0} \left\{ t + \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_t}\right) \right\},$$

where P_t with $t > 0$ is the packing number of $K \cap \mathbb{B}_2^d(t)$ with balls of radius $t/10$. Then, there exists an absolute constant $c > 0$ such that any estimator $\widehat{\theta}$ which depends only on m observations of $(a, \check{y})$ satisfies

$$\sup_{\theta_0 \in K} \mathbb{E}\Big[\|\widehat{\theta} - \theta_0\|_2\Big] \geq c \min\{\delta_*, \mathit{diam}(K)\}.$$

Lemma C.4. *When $k \leq d/4$, for any $t \leq 1$, we have $P_t \geq \exp(ck \log d/k)$, where P_t is defined as in Lemma C.3 with letting $K \subseteq \mathbb{B}_2^d(1)$ being a set containing all k group sparse vectors in $\mathbb{B}_2^d(1)$. Here $c > 0$ is an absolute constant.*

Proof of Lemma C.4. The proof of this lemma follows from the idea of randomized packing construction in Section 4.3 of (Plan et al., 2016). For any t, since P_t is defined as the packing number with balls of radius scaling as t, which is the radius of the set $K \cap \mathbb{B}_2^d(t)$, then we have $P_t = P_1$. Thus, we only need to consider the lower bound of P_1. Furthermore, since $\mathcal{S}^{d-1} \subseteq \mathbb{B}_2^d(1)$, where $\mathcal{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$ of radius 1, the packing number of $K \cap \mathbb{B}_2^d(1)$ is larger $K \cap \mathcal{S}^{d-1}$. Thus, we consider $1/10$ packing of the set $K \cap \mathcal{S}^{d-1}$ to obtain the lower bound of P_1. Consider a subset $K \cap \mathcal{S}^{d-1}$ such that it contains all nonnegative k group sparse signals in $\mathbb{R}^d$ where each non-zero entry equals $1/\sqrt{k}$. This is possible due to Proposition C.2. Then, we have $|\mathcal{C}| = (d/k)^k$. We will show that there exists a large enough subset $\mathcal{X} \subseteq \mathcal{C}$ such that $\forall x, y \in \mathcal{X}$, $\|x - y\|_2 > 1/10$. Consider picking vectors $x, y \in \mathcal{C}$ uniformly at random and computing the probability of the event $\|x - y\|_2^2 \leq 1/100$. When the event happens, it requires x and y to have at least $0.99k$ matching non-zero coordinates. Assume without loss of generality that $0.01k$ is an integer, this event happens with probability

$$\binom{k}{0.99k}\binom{d - 0.99k}{0.01k} \Big/ (d/k)^k.$$

Using Stirling's approximation and $k \leq d/4$, we have $\Pr(\|x - y\|_2^2 \leq 1/100) \leq \exp(-c' k \log(d/k))$, where $c' > 0$ is an absolute constant. This implies the claim that when choosing $\mathcal{X}$ to have $\exp(ck \log(d/k))$ uniformly chosen vectors from $\mathcal{C}$, which satisfies $\forall x, y \in \mathcal{X}$, $\|x - y\|_2 > 1/10$ with a constant probability. □

Lemma C.5 (Lemma 4.5)**.** *Assume that $\theta_0 \in K \subseteq \mathbb{B}_2^d(1)$ where K is a set containing any k-group sparse vectors in $\mathbb{B}_2^d(1)$, and K satisfies that $\forall v \in K$ then $\lambda v \in K, \forall \lambda \in [0,1)$. Assume that $\check{y} = \langle a, \theta_0 \rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$ and $a \sim \mathcal{N}(0, \mathbf{I}_d)$. Then, there exist absolute constants $c_1, c_2 > 0$ such that any estimator $\widehat{\theta}$ which depends only on m observations of $(a, \check{y})$ satisfies that when $m \geq c_1 k \log(d/k)$, there is*

$$\sup_{\theta_0 \in K} \mathbb{E}\|\widehat{\theta} - \theta_0\|_2 \geq c_2 \sqrt{\frac{k \log(d/k)}{m}}.$$

Proof of Lemma 4.5. Since K satisfies $\lambda v \in K$ for any $v \in K$ and $\lambda \in [0,1)$. Thus, by Lemma C.3, we have

$$\delta_* = \inf_{t>0} \left\{ t + \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_t}\right) \right\}.$$

Consider that for any $t > 1$, then we can observe that

$$t + \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_t}\right) > 1.$$

On the other hand, for any $t \leq 1$, then we have

$$\begin{aligned}
&\inf_{0<t\leq 1} \left\{ t + \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_t}\right) \right\} \\
&= \inf_{0<t\leq 1} \left\{ t + \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_1}\right) \right\} \\
&= \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_1}\right),
\end{aligned}$$

where the first equality is due to Lemma C.4, the second equality is by taking $\inf$ over t. If $m \geq \sigma^2(1 + \sqrt{\log P_1})^2$, we have

$$\inf_{0<t\leq 1} \{t + \tfrac{\sigma}{\sqrt{m}}(1 + \sqrt{\log P_t})\} \leq 1.$$

Comparing the cases $t > 1$ and $t \leq 1$, we get that, if $m \geq \sigma^2(1 + \sqrt{\log P_1})^2$, then

$$\delta_0 = \inf_{t>0} \left\{ t + \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_t}\right) \right\} = \inf_{0<t\leq 1} \left\{ t + \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_t}\right) \right\} = \frac{\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_1}\right).$$

Moreover, since the $\operatorname{diam}(K) \leq 1$, then by Lemma C.3, we have

$$\sup_{\theta_0 \in K} \mathbb{E}\|\widehat{\theta} - \theta_0\|_2 \geq c \min\{\delta_*, \operatorname{diam}(K)\} = \frac{c\sigma}{\sqrt{m}} \left(1 + \sqrt{\log P_1}\right) \geq \frac{c\sigma}{\sqrt{m}} \sqrt{\log P_1},$$

by letting $m \geq \sigma^2(1 + \sqrt{\log P_1})^2$. Furthermore, according to Lemma C.4, we know $\log P_1 \geq c' k \log(d/k)$ with c' being an absolute constant. Then, there exists a sufficient large absolute constant c_1 such that when $m \geq c_1 k \log(d/k)$, we have

$$\sup_{\theta_0 \in K} \mathbb{E}[\|\widehat{\theta} - \theta_0\|_2] \geq c_2 \sqrt{\frac{k \log(d/k)}{m}}.$$

□

D. Proof of Theorem 3.8

Before presenting the proof of Theorem 3.8, we first introduce some notations and definitions used hereafter. These notations and definitions will also be used in the proof of Theorem 3.10 in Section §E. According to the definition of $W_{i,+,x}$ in the paper, we can know that $G(x)$ can be represented as

$$G(x) = \left(\prod_{i=1}^{n} W_{i,+,x} \right) x = (W_{n,+,x} W_{n-1,+,x} \cdots W_{1,+,x}) x.$$

We therefore further define a more general form $H_x(z)$ as follows,

$$H_x(z) := \left(\prod_{i=1}^{n} W_{i,+,x}\right) z = (W_{n,+,x} W_{n-1,+,x} \cdots W_{1,+,x}) z,$$

by which we can see that $H_x(x) = G(x)$.

Recall that as shown in the main body of the paper, for any x such that $L(x)$ is differentiable, we can write the gradient of $L(x)$ w.r.t. x as follows

$$\nabla L(x) = 2\left(\prod_{j=1}^{n} W_{j,+,x}\right)^{\top}\left(\prod_{j=1}^{n} W_{j,+,x}\right) x - \frac{2\lambda}{m}\sum_{i=1}^{m} y_i \left(\prod_{j=1}^{n} W_{j,+,x}\right)^{\top} a_i,$$

by which we further have

$$\langle \nabla L(x), z\rangle = 2\langle G(x), H_x(z)\rangle - \frac{2\lambda}{m}\sum_{i=1}^{m} y_i \langle a_i, H_x(z)\rangle,$$

for any x and z.

We then let

$$h_{x,x_0} := \frac{1}{2^n}x - \frac{1}{2^n}\left[\left(\prod_{i=0}^{n-1}\frac{\pi-\overline{\varrho}_i}{\pi}\right)x_0 + \sum_{i=0}^{n-1}\frac{\sin\overline{\varrho}_i}{\pi}\left(\prod_{j=i+1}^{d-1}\frac{\pi-\overline{\theta}_j}{\pi}\right)\frac{\|x_0\|_2}{\|x\|_2}x\right], \tag{30}$$

$$S_{\varepsilon,x_0} := \{x \neq 0 : \|h_{x,x_0}\|_2 \le \frac{1}{2^n}\varepsilon \max(\|x\|_2, \|x_0\|_2)\}. \tag{31}$$

where $\overline{\varrho}_0 = \angle(x, x_0)$ and $\overline{\varrho}_i = g(\overline{\varrho}_{i-1})$, and $g(\varrho) := \cos^{-1}\left(\frac{(\pi-\varrho)\cos\varrho + \sin\varrho}{\pi}\right)$ as defined in Lemma D.3. In the following subsections, we provides key lemmas for the proof of Theorem 3.8, and then a proof sketch of this theorem, followed by a detailed proof.

D.1. Lemmas for Theorem 3.8

Lemma D.1. *Define $H_x(z) = \prod_{j=1}^{n} W_{j,+,x} z$. Suppose that $G(x_0)$ satisfies $|G(x_0)| \le R$. There exists an absolute constant $c_1 > 0$ such that for any z and any x, when*

$$\lambda \ge 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}\log(64\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon),$$

the following holds:

$$|\lambda\mathbb{E}[y_i\langle a_i, H_x(z)\rangle] - \langle G(x_0), H_x(z)\rangle| \le \frac{1}{4}\varepsilon\|H_x(z)\|_2.$$

Proof of Lemma D.1. Recall that $y_i = \operatorname{sign}(\langle a_i, G(x_0)\rangle + \xi_i + \tau_i)$. We let $V_i = \langle a_i, G(x_0)\rangle + \xi_i$ and $Z_i = \langle a_i, H_z(x)\rangle$. Still, we assume V_i and τ_i are independent. Thus, there is

$$\mathbb{E}[\operatorname{sign}(V_i + \tau_i)|V_i] = \frac{V_i}{\lambda} - \frac{V_i}{\lambda}\mathbf{1}_{\{|V_i|>\lambda\}} + \mathbf{1}_{\{V_i>\lambda\}} - \mathbf{1}_{\{V_i<-\lambda\}}.$$

Therefore, we have

$$\begin{aligned}
&\left|\mathbb{E}[Z_i \operatorname{sign}(V_i + \tau_i)] - \frac{\mathbb{E}[Z_i V_i]}{\lambda}\right| \\
&\quad = \left|-\mathbb{E}\left[\frac{Z_i V_i}{\lambda}\mathbf{1}_{\{|V_i|>\lambda\}}\right] + \mathbb{E}\big[Z_i \mathbf{1}_{\{V_i>\lambda\}}\big] - \mathbb{E}\big[Z\mathbf{1}_{\{V_i>\lambda\}}\big]\right| \\
&\quad \le \frac{\|Z_i\|_{L_2}\cdot\|V_i\mathbf{1}_{\{|V_i|>\lambda\}}\|_{L_2}}{\lambda} + 2\|Z_i\|_{L_2}\Pr(|V_i| > \lambda)^{1/2},
\end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality.

First, by the isotropic assumption of a_i, we have

$$\|Z_i\|_{L_2} = \left\{\mathbb{E}\left[|\langle a_i, H_x(z)\rangle|^2\right]\right\}^{1/2} = \|H_x(z)\|_2.$$

Next, same to Lemma 4.1, we have

$$\begin{aligned}\|V_i \mathbf{1}_{\{|V_i|>\lambda\}}\|_{L_2} \leq & \sqrt{2c_1(\lambda+1)\|\langle a_i, G(x_0)\rangle + \xi_i\|_{\psi_1}} e^{-\lambda/2\|\langle a_i, G(x_0)\rangle + \xi_i\|_{\psi_1}} \\ \leq & \sqrt{2c_1(\lambda+1)(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})} e^{-\lambda/2(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}.\end{aligned}$$

due to our assumption that $\|G(x_0)\|_2 \leq R$ and V_i is sub-gaussian. Moreover, we also have

$$\Pr(|V_i| > \lambda)^{1/2} \leq \sqrt{c_1(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})} e^{-\lambda/2(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}.$$

Overall, we can obtain

$$|\lambda\mathbb{E}[Z_i \operatorname{sign}(V_i + \tau_i)] - \mathbb{E}[Z_i V_i]| \leq \sqrt{c_1(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}(\sqrt{2(\lambda+1)} + 2\lambda) e^{-\lambda/2(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}\|H_x(z)\|_2.$$

When

$$\lambda \geq 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}\log(64\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon),$$

it is immediate that

$$\sqrt{c_1(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})}(\sqrt{2(\lambda+1)} + 2\lambda) e^{-\lambda/2(\|a\|_{\psi_1}R + \|\xi\|_{\psi_1})} \leq \frac{1}{4}\varepsilon.$$

As a consequence, we have

$$|\lambda\mathbb{E}[y_i\langle a_i, H_x(z)\rangle] - \langle G(x_0), H_x(z)\rangle| \leq \frac{1}{4}\varepsilon\|H_x(z)\|_2,$$

which finishes the proof. □

Lemma D.2. *Define $H_x(z) := \prod_{j=1}^n W_{j,+,x}z$. Suppose that $G(x_0)$ satisfies $|G(x_0)| \leq R$. Then, with probability at least $1 - c_4\exp(-u)$ where $c_4 > 0$ is an absolute constant,*

$$\sup_{x\in\mathbb{R}^k, z\in\mathbb{R}^k, x_0\in\mathbb{R}^k, |G(x_0)|\leq R} \frac{\left|\frac{\lambda}{m}\sum_{i=1}^m y_i\langle a_i, H_x(z)\rangle - \lambda\mathbb{E}[y_i\langle a_i, H_x(z)\rangle]\right|}{\|H_x(z)\|_2} \leq \frac{\varepsilon}{8},$$

where the sample complexity is

$$m \geq c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon^2,$$

for some absolute constant c_2 large enough.

Proof of Lemma D.2. The proof of Lemma D.2 is very similar to the proofs shown in the previous subsection. Therefore, we only outline the main proof steps here but ignore detailed calculation for some inequalities. We aim to bound the following term

$$\sup_{x\in\mathbb{R}^k, z\in\mathbb{R}^k, x_0\in\mathbb{R}^k, |G(x_0)|\leq R} \frac{\left|\frac{1}{m}\sum_{i=1}^m y_i\langle a_i, H_x(z)\rangle - \mathbb{E}[y_i\langle a_i, H_x(z)\rangle]\right|}{\|H_x(z)\|_2}.$$

By Symmetrization inequality in Lemma B.7, it suffices to bound

$$\sup_{x\in\mathbb{R}^k, z\in\mathbb{R}^k, x_0\in\mathbb{R}^k, |G(x_0)|\leq R} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i\langle a_i, H_x(z)\rangle\right|}{\|H_x(z)\|_2}$$

where $\{\varepsilon_i\}$ are i.i.d. Rademacher random variables that are independent of other random variables.

We rewrite the set $\{G(x_0) : \|G(x_0)\|_2 \leq R,\ x_0 \in \mathbb{R}^k\}$ as $G(\mathbb{R}^k) \cap \mathbb{B}_2^d(R)$. To bound the supremum above is based on building a δ-covering net over the set $G(\mathbb{R}^k) \cap \mathbb{B}_2^d(R)$, namely $\mathcal{N}(G(\mathbb{R}^k) \cap \mathbb{B}_2^d(R),\ \delta)$. The δ value should be carefully chosen. For a simply notation, we let $y_i^v := \operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i)$ be the sign function associated with $G(v)$ in the net. We begin our proof by bounding the supremum term as follows, with probability at least $1 - \exp(-cu) - 2\exp(-u)$,

$$
\begin{aligned}
&\sup_{x,z,x_0 \in \mathbb{R}^k,\ \|G(x_0)\|_2 \leq R} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i \langle a_i, H_x(z)\rangle\right|}{\|H_x(z)\|_2} \\
&\quad\leq \sup_{x,z \in \mathbb{R}^k, G(v) \in \mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i^v \langle a_i, H_x(z)\rangle\right|}{\|H_x(z)\|_2} \\
&\qquad + \sup_{x,z,x_0 \in \mathbb{R}^k, \|G(x_0)-G(v)\|_2 \leq \delta, G(v) \in \mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i (y_i - y_i^v) \langle a_i, H_x(z)\rangle\right|}{\|H_x(z)\|_2} \\
&\quad\leq \underbrace{\sup_{x,z \in \mathbb{R}^k, G(v) \in \mathcal{N}(G(\mathbb{R}^k)\cap\mathbb{B}_2^d(R),\ \delta)} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i) \langle a_i, H_x(z)\rangle\right|}{\|H_x(z)\|_2}}_{\text{(I)}} \\
&\qquad + \underbrace{\sup_{x,z \in \mathbb{R}^k} \max_{|\mathcal{I}| \leq 4\eta/\lambda} \frac{2}{m} \sum_{i \in \mathcal{I}} \frac{|\langle a_i, H_x(z)\rangle|}{\|H_x(z)\|_2}}_{\text{(II)}},
\end{aligned}
$$

where the first inequality is due to decomposition of the supremum term and the second inequality is by Lemma B.4, which bounds the number of difference between $\{y_i\}$ and $\{y_i^v\}$ with high probability.

Bounding Term (I): We first show the bound based on fixed $G(v)$. Then we give a uniform bound for any $G(v)$ in the δ-net. For a fixed $G(v)$, we have

$$
\sup_{x,z \in \mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\langle a_i, G(v)\rangle + \xi_i + \tau_i) \langle a_i, H_x(z)\rangle\right|}{\|H_x(z)\|_2} = \sup_{x,z \in \mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, H_x(z)\rangle\right|}{\|H_x(z)\|_2}.
$$

For the function $H_x(z)$, we can see that as x varies, $H_x(z)$ can be different linear functions, which constructs at most $[\mathcal{C}(d,k)]^n = [\binom{d}{0} + \binom{d}{1} + \cdots + \binom{d}{k}]^n \leq (d^k + 1)^n \leq (2d)^{kn}$ hyperplanes that split the whole $\mathbb{R}^k$ space.

Now, we consider any one piece $H_{\widetilde{x}}$ where $\widetilde{x} \in \mathcal{P} \subseteq \mathbb{R}^k$ and bound the following quantity:

$$
\begin{aligned}
\sup_{z \in \mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, H_{\widetilde{x}} z)\rangle\right|}{\|H_{\widetilde{x}}(z)\|_2} &\leq \sup_{z \in \mathbb{R}^k} \frac{\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, W_{\widetilde{x}} z\rangle\right|}{\|W_{\widetilde{x}} z\|_2} \\
&\leq \sup_{b \in \mathcal{E}^k \cap \mathcal{S}^{d-1}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, b\rangle\right|,
\end{aligned}
$$

where we let $W_{\widetilde{x}} = \prod_{j=1}^n W_{j,+,\widetilde{x}}$ be the linear function at $\widetilde{x}$ such that $H_{\widetilde{x}}(z) = \left(\prod_{j=1}^n W_{j,+,\widetilde{x}}\right) z$. $\mathcal{E}_k$ be the subspace in $\mathbb{R}^d$ spanned by the k columns of $W_{\widetilde{z}}$. We also define $b = W_{\widetilde{x}} z / \|W_{\widetilde{x}} z\|$ in the above formulation.

It suffices to bound the last term in the above formulation. We consider a $1/2$-covering net of the set $\mathcal{E}^k \cap \mathcal{S}^{d-1}$, namely, $\mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)$. A simple volume argument shows that the cardinality $|\mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)| \leq 3^k$.

By Bernstein's inequality in Lemma B.6, we have for any fixed $v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)$,

$$
\Pr\left(\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, b\rangle\right| \geq \sqrt{\frac{2u'}{m}} + \frac{\|a\|_{\psi_1} u'}{m}\right) \leq 2e^{-u'}.
$$

Taking $u' = u + ckn\log(ed)$ for some $c > 6$, we have with probability at least $1 - 2\exp(-u - ckn\log(ed))$,

$$
\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \langle a_i, v\rangle\right| \leq \sqrt{\frac{2(u + ckn\log(ed))}{m}} + \frac{\|a\|_{\psi_1}(u + ckn\log(ed))}{m}.
$$

Taking a union bound over all $v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)$, we have with probability at least $1 - 2\exp(-u - ckn\log(ed)) \cdot 3^k \geq 1 - 2\exp(-u - c_1 kn\log(ed))$ for some absolute constant $c_1 > 2$.

$$\sup_{v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, b \rangle \right| \leq \sqrt{\frac{2(u + ckn\log(ed))}{m}} + \frac{\|a\|_{\psi_1}(u + ckn\log(ed))}{m}. \tag{32}$$

Therefore, we will have

$$\begin{aligned}
&\sup_{b \in \mathcal{E}^k \cap \mathcal{S}^{d-1}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, b \rangle \right| \\
&\quad \leq \sup_{v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, v \rangle \right| + \sup_{b \in \mathcal{E}^k \cap \mathcal{S}^{d-1}, v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2), \|b - v\|_2 \leq 1/2} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, b - v \rangle \right| \\
&\quad \leq \sup_{v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, v \rangle \right| + \frac{1}{2} \sup_{b \in \mathcal{E}^k \cap \mathcal{S}^{d-1}, v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2), \|b - v\|_2 \leq 1/2} \left| \frac{1}{m} \sum_{i=1}^m \frac{\varepsilon_i \langle a_i, b - v \rangle}{\|b - v\|_2} \right| \\
&\quad \leq \sup_{v \in \mathcal{N}(\mathcal{E}^k \cap \mathcal{S}^{d-1}, 1/2)} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, v \rangle \right| + \frac{1}{2} \sup_{b \in \mathcal{E}^k \cap \mathcal{S}^{d-1}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, b \rangle \right|.
\end{aligned}$$

Now we can obtain

$$\begin{aligned}
\sup_{z \in \mathbb{R}^k} \frac{\left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, H_{\widetilde{x}}(z) \rangle \right|}{\|H_{\widetilde{x}}(z)\|_2} &\leq \sup_{b \in \mathcal{E}^k \cap \mathcal{S}^{d-1}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle a_i, b \rangle \right| \\
&\leq 2\sqrt{\frac{2(u + ckn\log(ed))}{m}} + \frac{2\|a\|_{\psi_1}(u + ckn\log(ed))}{m}.
\end{aligned}$$

Taking a further union bound over at most $(2d)^{kn}$ linear functions, we have

$$\text{(I)} \leq 2\sqrt{\frac{2(u + ckn\log(ed))}{m}} + \frac{2\|a\|_{\psi_1}(u + ckn\log(ed))}{m}$$

with probability at least $1 - 2\exp(-u - c_1 kn\log(ed)) \cdot (2d)^{kn} \geq 1 - 2\exp(-u - c_2 kn\log(ed))$ where $c_2 > 1$.

Bounding Term (II): Now we bound the term

$$\text{(II)} = \sup_{x, z \in \mathbb{R}^k} \max_{|\mathcal{I}| \leq 4m\eta/\lambda} \frac{2}{m} \sum_{i \in \mathcal{I}} \frac{|\langle a_i, H_x(z) \rangle|}{\|H_x(z)\|_2}$$

Let $t = H_x(z)/\|H_x(z)\|_2$ and it is enough to bound

$$\sup_{t \in \mathcal{E}^k \cap \mathcal{S}^{d-1}} \max_{|\mathcal{I}| \leq 4m\eta/\lambda} \frac{1}{m} \sum_{i \in \mathcal{I}} \left(|\langle a_i, t \rangle| - \mathbb{E}[|\langle a_i, t \rangle|] + \mathbb{E}[|\langle a_i, t \rangle|] \right).$$

Note that $|\langle a_i, t \rangle| - \mathbb{E}[|\langle a_i, t \rangle|]$ is also a sub-exponential random variable with sub-exponential norm bounded by $2\|a\|_{\psi_1}$, and $\mathbb{E}[|\langle a_i, t \rangle|] \leq 1$. Given x, $H_x(z)$ is a linear function and there are at most $(2d)^{kn}$ different linear function for different x.

For the extra expectation term, we have

$$\sup_{t \in \mathcal{E}^k \cap \mathcal{S}^{d-1}} \max_{|\mathcal{I}| \leq 4m\eta/\lambda} \frac{1}{m} \sum_{i \in \mathcal{I}} \mathbb{E}[|\langle a_i, t \rangle|] \leq \max_{|\mathcal{I}| \leq 4m\eta/\lambda} \frac{|\mathcal{I}|}{m}$$

Next, we bound the term $\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1}}\max_{|\mathcal{I}|\leq 4m\eta/\lambda}\frac{1}{m}\sum_{i\in\mathcal{I}}\left(|\langle a_i,t\rangle|-\mathbb{E}[|\langle a_i,t\rangle|]\right)$. We have

$$\begin{aligned}
&\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1}}\max_{|\mathcal{I}|\leq 4m\eta/\lambda}\frac{1}{m}\sum_{i\in\mathcal{I}}\left(|\langle a_i,t\rangle|-\mathbb{E}[|\langle a_i,t\rangle|]\right)\\
&\quad=\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1}}\max_{|\mathcal{I}|\leq 4m\eta/\lambda}\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{\{i\in\mathcal{I}\}}\left(|\langle a_i,t\rangle|-\mathbb{E}[|\langle a_i,t\rangle|]\right)\\
&\quad\leq\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1}}\max_{|\mathcal{I}|\leq 4m\eta/\lambda}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{\{i\in\mathcal{I}\}}\left(|\langle a_i,t\rangle|-\mathbb{E}[|\langle a_i,t\rangle|]\right)\right|
\end{aligned}$$

By Symmetrization inequality, it suffices to bound

$$\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1}}\max_{|\mathcal{I}|\leq 4m\eta/\lambda}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\mathbf{1}_{\{i\in\mathcal{I}\}}|\langle a_i,t\rangle|\right|=\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1},|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,t\rangle|\right|$$

We consider a $1/2$-covering net of the set $\mathcal{E}^k\cap\mathcal{S}^{d-1}$, namely, $\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2)$. A simple volume argument shows that the cardinality $|\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2)|\leq 3^k$. Therefore, we will have

$$\begin{aligned}
&\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1},|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,t\rangle|\right|\\
&\quad\leq\sup_{v\in\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2),|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,v\rangle|\right|+\sup_{\substack{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1},v\in\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2),\\ \|t-v\|_2\leq 1/2,|\mathcal{I}|\leq 4\eta/\lambda}}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,t-v\rangle|\right|\\
&\quad\leq\sup_{v\in\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2),|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,v\rangle|\right|+\frac{1}{2}\sup_{\substack{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1},v\in\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2),\\ \|t-v\|_2\leq 1/2,|\mathcal{I}|\leq 4\eta/\lambda}}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\frac{\varepsilon_i|\langle a_i,t-v\rangle|}{\|t-v\|_2}\right|\\
&\quad\leq\sup_{v\in\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2),|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,v\rangle|\right|+\frac{1}{2}\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1},|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,t\rangle|\right|,
\end{aligned}$$

which implies

$$\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1},|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,t\rangle|\right|\leq 2\sup_{v\in\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2),|\mathcal{I}|\leq 4\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,v\rangle|\right|.$$

For any fixed v in the $1/2$-net and a fixed $\mathcal{I}$, by Bernstein's inequality, we have

$$\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,v\rangle|\right|=\frac{|\mathcal{I}|}{m}\left|\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\varepsilon_i\langle a_i,v\rangle\right|\leq\frac{1}{m}(2\sqrt{u_2|\mathcal{I}|}+2\|a\|_{\psi_1}u_2).$$

We take

$$u_2=C_1\log\left(\frac{\lambda m}{kn\log(ed)+u'}\right)\left(u+2kn\log(ed)+k\log(2R)+k\log\frac{m}{kn\log(ed)+u'}\right)^{1/2}\sqrt{m},$$

where C_1 is an absolute constant large enough and u' satisfies (20). Using the fact that

$$|\mathcal{I}|\leq\frac{4\eta}{\lambda}m\leq 2L\Big(u+2kn\log(ed)+k\log(2R)+k\log\frac{m}{kn\log(ed)+u'}\Big)^{1/2}\sqrt{m},$$

we have with probability at least

$$1-2\exp\left(-C_1\log\left(\frac{\lambda m}{kn\log(ed)+u'}\right)\left(u+2kn\log(ed)+k\log(2R)+k\log\frac{m}{kn\log(ed)+u'}\right)^{1/2}\sqrt{m}\right),$$

the following holds:

$$\left|\frac{1}{m}\sum_{i\in\mathcal{I}}\varepsilon_i|\langle a_i,v\rangle|\right| \le C_1\|a\|_{\psi_1}\log\Big(\frac{\lambda m}{kn\log(ed)+u'}\Big)\sqrt{\frac{u+2kn\log(ed)+k\log(2R)+k\log\frac{m}{kn\log(ed)+u'}}{m}}.$$

To bound the maximum over $|\mathcal{I}|\le 4\eta/\lambda$, we take a union bound over all $\binom{m}{4\eta m/\lambda}$ possibilities,

$$\binom{m}{4\eta m/\lambda}\le\left(\frac{em}{4\eta m/\lambda}\right)^{4\eta m/\lambda}=\left(\frac{\lambda}{\eta}\right)^{4\eta m/\lambda}.$$

Thus, it follows

$$\begin{aligned}\log\binom{m}{4\eta m/\lambda} &\le \frac{4\eta m}{\lambda}\log\frac{\lambda}{\eta}\\ &\le L\Big(u+2kn\log(ed)+k\log(2R)+\log\frac{m}{kn\log(ed)+u'}\Big)^{1/2}\sqrt{m}\log\Big(\frac{\lambda m}{kn\log(ed)+u'}\Big),\end{aligned}$$

and when $C_1>L$, taking the union bound gives, with probability at least

$$1-2\exp\left(-C_2\log\left(\frac{\lambda m}{kn\log(ed)+u'}\right)\Big(u+2kn\log(ed)+k\log(2R)+\log\frac{m}{kn\log(ed)+u'}\Big)^{1/2}\sqrt{m}\right),$$

we have

$$\begin{aligned}&\max_{|\mathcal{I}|\le 4m\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}|\langle a_i,v\rangle|\right|\\ &\quad\le C_1\|a\|_{\psi_1}\log\Big(\frac{\lambda m}{kn\log(ed)+u'}\Big)\sqrt{\frac{u+2kn\log(ed)+k\log(2R)+k\log\frac{m}{kn\log(ed)+u'}}{m}}.\end{aligned}$$

Furthermore, taking the union bound on all the $1/2$-net, we obtain

$$\begin{aligned}&\sup_{t\in\mathcal{E}^k\cap\mathcal{S}^{d-1}}\max_{|\mathcal{I}|\le 4m\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}|\langle a_i,t\rangle|\right|\\ &\quad\le 2\sup_{v\in\mathcal{N}(\mathcal{E}^k\cap\mathcal{S}^{d-1},1/2)}\max_{|\mathcal{I}|\le 4m\eta/\lambda}\left|\frac{1}{m}\sum_{i\in\mathcal{I}}|\langle a_i,v\rangle|\right|\\ &\quad\le 2C_1\|a\|_{\psi_1}\log\Big(\frac{\lambda m}{kn\log(ed)+u'}\Big)\sqrt{\frac{u+2kn\log(ed)+k\log(2R)+k\log\frac{m}{kn\log(ed)+u'}}{m}}.\end{aligned}$$

with probability

$$1-2\cdot 3^k\cdot(2d)^{kn}\exp\left(-C_2\log\left(\frac{\lambda m}{kn\log(ed)+u'}\right)\Big(u+2kn\log(ed)+k\log(2R)+\log\frac{m}{kn\log(ed)+u'}\Big)^{1/2}\sqrt{m}\right),$$

where C_2 is an absolute constant. Particularly, if we set

$$m\ge c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed)+k\log(2R)+k\log m+u)/\varepsilon^2,$$

for some absolute constant c_2 large enough, we have

$$\text{(II)}\le 2C_1\|a\|_{\psi_1}\log\Big(\frac{\lambda m}{kn\log(ed)+u'}\Big)\sqrt{\frac{u+2kn\log(ed)+k\log(2R)+k\log\frac{m}{kn\log(ed)+u'}}{m}},$$

with probability at least $1 - c_3 \exp(-u)$, where $c_3 \geq 1$ is an absolute constant.

Combining (I) and (II): Combining all the results above, we obtain with probability $1 - c_3 \exp(-u) - \exp(-cu) - 2\exp(-u)$,

$$\sup_{x \in \mathbb{R}^k, z \in \mathbb{R}^k, x_0 \in \mathbb{R}^k, |G(x_0)| \leq R} \frac{\left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i y_i \langle a_i, H_x(z) \rangle \right|}{\|H_x(z)\|_2} \leq \frac{\varepsilon}{16\lambda}.$$

which thus means, for any x, z, x_0, by Symmetrization, we have

$$\left| \frac{\lambda}{m} \sum_{i=1}^m y_i \langle a_i, H_x(z) \rangle - \lambda \mathbb{E}[y_i \langle a_i, H_x(z) \rangle] \right| \leq \frac{\varepsilon}{8} \|H_x(z)\|_2.$$

with probability at least $1 - c_4 \exp(-u)$. □

The following lemmas are some useful lemmas from previous papers. We rewrite them here for integrity.

Lemma D.3 ((Hand & Voroninski, 2018)). *Suppose* $8\pi n^6 \sqrt{\varepsilon} \leq 1$. *Let*

$$S_{\varepsilon, x_0} := \{x \neq 0 \in \mathbb{R}^k | \|h_{x,x_0}\|_2 \leq \frac{1}{2^n} \varepsilon \max(\|x\|_2, \|x_0\|_2)\},$$

where n *is an integer greater than 1 and let* h_{x,x_0} *be defined by*

$$h_{x,x_0} := \frac{1}{2^n} x - \frac{1}{2^n} \left[\left(\prod_{i=0}^{n-1} \frac{\pi - \overline{\varrho}_i}{\pi} \right) x_0 + \sum_{i=0}^{n-1} \frac{\sin \overline{\varrho}_i}{\pi} \left(\prod_{j=i+1}^{d-1} \frac{\pi - \overline{\varrho}_j}{\pi} \right) \frac{\|x_0\|_2}{\|x\|_2} x \right]$$

where $\overline{\varrho}_0 = \angle(x, x_0)$ *and* $\overline{\varrho}_i = g(\overline{\varrho}_{i-1})$. *Particularly, we define*

$$g(\varrho) := \cos^{-1} \left(\frac{(\pi - \varrho) \cos \varrho + \sin \varrho}{\pi} \right).$$

If $x \in S_{\varepsilon, x_0}$, *then we have*

$$S_{\varepsilon, x_0} \subset \mathcal{B}(x_0, 56n\sqrt{\varepsilon}\|x_0\|_2) \cup \mathcal{B}(-\rho_n x_0, 500n^{11}\sqrt{\varepsilon}\|x_0\|_2),$$

where ρ_n *is defined as*

$$\rho_n := \sum_{i=0}^{n-1} \frac{\sin \check{\varrho}_i}{\pi} \left(\prod_{j=i+1}^{n-1} \frac{\pi - \check{\varrho}_j}{\pi} \right) \leq 1,$$

and $\check{\varrho}_0 = \pi$ *and* $\check{\varrho}_i = g(\check{\varrho}_{i-1})$.

Lemma D.4 ((Hand & Voroninski, 2018)). *Fix* $0 < 16\pi n^2 \sqrt{\varepsilon_{\text{wdc}}} < 1$ *and* $n \geq 2$. *Suppose that* W_i *satisfies the WDC with constant* ε_{wdc} *for* $i = 1, \ldots, n$. *Define*

$$\widetilde{h}_{x,z} = \frac{1}{2^n} \left[\left(\prod_{i=0}^{n-1} \frac{\pi - \overline{\varrho}_i}{\pi} \right) z + \sum_{i=0}^{n-1} \frac{\sin \overline{\varrho}_i}{\pi} \left(\prod_{j=i+1}^{n-1} \frac{\pi - \overline{\varrho}_j}{\pi} \right) \frac{\|z\|_2}{\|x\|_2} x \right],$$

where $\overline{\varrho}_i = g(\overline{\varrho}_{i-1})$ *for* g *in Lemma D.3 and* $\overline{\varrho}_0 = \angle(x, z)$. *For all* $x \neq 0$ *and* $y \neq 0$,

$$\left\| \left(\prod_{i=1}^n W_{i,+,x} \right)^\top G(z) - \widetilde{h}_{x,z} \right\|_2 \leq 24 \frac{n^3 \sqrt{\varepsilon_{\text{wdc}}}}{2^n} \|z\|_2, \tag{33}$$

$$\langle G(x), G(z) \rangle \geq \frac{1}{4\pi} \frac{1}{2^n} \|x\|_2 \|z\|_2, \tag{34}$$

$$\|W_{i,+,x}\|_2 \leq \left(\frac{1}{2} + \varepsilon_{\text{wdc}} \right)^{1/2}. \tag{35}$$

D.2. Proof Sketch of Theorem 3.8

Under the conditions of Theorem 3.8, our proof is sketched as follows:

- The key to proving Theorem 3.8 lies in understanding the concentration of $L(x)$ and $\nabla L(x)$. Here we prove two critical lemmas, Lemmas D.1 and D.2 in this section, combining which we can show that for any x, z and $|G(x_0)| \leq R$, when λ and m are sufficiently large, the following holds with high probability

  $$\Big|\frac{\lambda}{m}\sum_{i=1}^{m} y_i\langle a_i, H_x(z)\rangle - \langle G(x_0), H_x(z)\rangle\Big| \leq \varepsilon\|H_x(z)\|_2.$$

  which further implies

  $$\frac{\lambda}{m}\sum_{i=1}^{m} y_i\langle a_i, H_x(z)\rangle \approx \langle G(x_0), H_x(z)\rangle,$$

  for any x, z.

  Therefore, we have $\forall z$ and $\forall x$ such that $L(x)$ is differentiable, we can approximate $\nabla L(x)$ as follows:

  $$\langle \nabla L(x), z\rangle \approx 2\langle G(x), H_x(z)\rangle - 2\langle G(x_0), H_x(z)\rangle.$$

- On the other hand, we can show that $\forall x, z$,

  $$\langle G(x), H_x(z)\rangle - \langle G(x_0), H_x(z)\rangle \approx \langle h_{x,x_0}, z\rangle,$$

  which therefore leads to

  $$\langle \nabla L(x), z\rangle \approx 2\langle h_{x,x_0}, z\rangle.$$

- Following the previous step, with v_x being defined in Theorem 3.8, the directional derivative is approximated as

  $$D_{-v_x}L(x)\cdot\|v_x\|_2 \approx -4\|h_{x,x_0}\|_2^2.$$

- We consider the error of approximating $D_{-v_x}L(x)\cdot\|v_x\|_2$ by $-4\|h_{x,x_0}\|_2^2$ in the following two cases:

  Case 1: When $\|x_0\|_2$ is not small and $x \neq 0$, one can show the error is negligible compared to $-4\|h_{x,x_0}\|_2^2$, so that $D_{-v_x}L(x) < 0$ as $-4\|h_{x,x_0}\|_2^2$.

  Case 2: When $\|x_0\|_2$ approaches 0, such an error is decaying slower than $-4\|h_{x,x_0}\|_2^2$ itself and eventually dominates it. As a consequence, one can only conclude that $\widehat{x}_m$ is around the origin.

- To characterize the directional derivative at 0 in Case 1, one can show

  $$D_wL(0)\cdot\|w\|_2 \leq \left|\langle G(x_0), H_{x_N}(w)\rangle - \frac{\lambda}{m}\sum_{i=1}^{m} y_i\langle a_i, H_{x_N}(w)\rangle\right| - \langle G(x_0), H_{x_N}(w)\rangle$$

  with $x_N \to 0$. By showing that the second term dominates according to (9) and Lemma D.4, we obtain

  $$D_wL(0) < 0, \forall w \neq 0.$$

D.3. Detailed Proof of Theorem 3.8

Proof of Theorem 3.8. According to Theorem 3.8, we define a non-zero direction as follows:

$$v_x := \begin{cases} \nabla L(x), & \text{if } L(x) \text{ is differentiable at } x,\\ \lim_{N\to+\infty}\nabla L(x_N), & \text{otherwise,}\end{cases}$$

where $\{x_N\}$ is a sequence such that $\nabla L(x)$ is differentiable at all point x_N in the sequence because of the piecewise linearity of $G(x)$.

On the other hand, by our definition of directional derivative, we have

$$D_{-v_x} L(x) = \begin{cases} \langle \nabla L(x), -\frac{v_x}{\|v_x\|_2} \rangle, & \text{if } L(x) \text{ is differentiable at } x, \\ \lim_{N\to+\infty} \langle \nabla L(\widetilde{x}_N), -v_x/\|v_x\|_2 \rangle, & \text{otherwise,} \end{cases}$$

where $\{\widetilde{x}_N\}$ is also a sequence with $\nabla L(\widetilde{x}_N)$ existing for all $\widetilde{x}_N$. Here we use $\widetilde{x}_N$ only in order to distinguish from the sequence of x_N in the definition of v_x above. We give the proof as follows:

Approximation of $\langle \nabla L(x), z \rangle$: The proof is mainly based on the two critical lemmas, i.e., Lemma D.1 and Lemma D.2.

First by (35) in Lemma D.4, we can have

$$\|G(x)\|_2 = (\prod_{i=1}^{n} W_{i,+,x})x \le (1/2 + \varepsilon_{\text{wdc}})^{n/2} \|x\|_2, \tag{36}$$

for any x. Thus, due to the assumption $\|x_0\|_2 \le R(1/2+\varepsilon_{\text{wdc}})^{-n/2}$ in Theorem 3.8 and $\|G(x_0)\|_2 \le (1/2+\varepsilon_{\text{wdc}})^{n/2}\|x_0\|_2$, we further have

$$\|G(x_0)\|_2 \le R.$$

By Lemma D.1 and $\|G(x_0)\|_2 \le R$, setting

$$\lambda \ge 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\} \log(64 \max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon),$$

the following holds for any x:

$$|\lambda \mathbb{E}[y_i \langle a_i, G(x)\rangle] - \langle G(x_0), G(x)\rangle| \le \frac{1}{4}\varepsilon \|G(x)\|_2, \tag{37}$$

if we let $z = x$ in Lemma D.1 such that $H_x(x) = G(x)$.

On the other hand, according to Lemma D.2 and $|G(x_0)| \le R$, we have that with probability at least $1 - c_4 \exp(-u)$, for any x, the following holds:

$$\left| \frac{\lambda}{m} \sum_{i=1}^{m} y_i \langle a_i, G(x)\rangle - \lambda \mathbb{E}[y_i \langle a_i, G(x)\rangle] \right| \le \frac{\varepsilon}{8} \|G(x)\|_2, \tag{38}$$

with sample complexity being

$$m \ge c_2 \|a\|_{\psi_1}^2 \lambda^2 \log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon^2,$$

where we set $z = x$ in Lemma D.2 with $H_x(x) = G(x)$.

Combining (37) and (38), we will have that with probability at least $1 - c_4\exp(-u)$, for any x, setting

$$\begin{aligned} \lambda \ge& 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\} \log(64 \max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon), \text{ and} \\ m \ge& c_2 \|a\|_{\psi_1}^2 \lambda^2 \log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon^2, \end{aligned}$$

the following holds

$$\left| \frac{\lambda}{m} \sum_{i=1}^{m} y_i \langle a_i, G(x)\rangle - \langle G(x_0), G(x)\rangle \right| \le \varepsilon \|G(x)\|_2, \tag{39}$$

which leads to

$$\left| \frac{1}{2} \langle \nabla L(x), z\rangle - (\langle G(x), H_x(z)\rangle - \langle G(x_0), H_x(z)\rangle) \right| \le \varepsilon \|G(x)\|_2. \tag{40}$$

Approximating $D_{-v_x}L(x)\cdot\|v_x\|_2$ and Bounding Errors: Without loss of generality, we directly prove the case where $L(x)$ is not differentiable at x. Then there exists a sequence $\widetilde{x}_N \to x$ such that $\nabla L(\widetilde{x}_N)$ exists for all $\widetilde{x}_N$. The proof is the same when x is the point such that $L(x)$ is differentiable. Therefore, we consider

$$D_{-v_x}L(x)\|v_x\|_2 = \lim_{\widetilde{x}_N\to x}\big\langle \nabla L(\widetilde{x}_N), -v_x\big\rangle. \tag{41}$$

When $L(x)$ is not differentiable, v_x is defined as $\lim_{x_M\to x}\nabla L(x_M)$, where $\{x_M\}$ could be another sequence such that $\nabla L(x_M)$ exists for all x_M. We decompose $D_{-v_x}L(x)\|v_x\|_2$ as follows

$$\begin{aligned}
D_{-v_x}L(x)\|v_x\|_2 =& \lim_{\widetilde{x}_N\to x}\big\langle \nabla L(\widetilde{x}_N), -v_x\big\rangle = \lim_{\widetilde{x}_N\to x}\lim_{x_M\to x} -\big\langle \nabla L(\widetilde{x}_N), L(x_M)\big\rangle\\
=& \lim_{\widetilde{x}_N\to x}\lim_{x_M\to x} -\big[4\langle h_{\widetilde{x}_N,x_0}, h_{x_M,x_0}\rangle + \langle \nabla L(\widetilde{x}_N) - 2h_{\widetilde{x}_N,x_0}, 2h_{x_M,x_0}\rangle\\
&+ \langle 2h_{\widetilde{x}_N,x_0}, \nabla L(\widetilde{x}_N) - 2h_{x_M,x_0}\rangle + \langle \nabla L(\widetilde{x}_N) - 2h_{\widetilde{x}_N,x_0}, \nabla L(\widetilde{x}_N) - 2h_{x_M,x_0}\rangle\big]\\
=& -4\|h_{x,x_0}\|_2^2 - \lim_{\widetilde{x}_N\to x}\lim_{x_M\to x}\big[\langle \nabla L(\widetilde{x}_N) - 2h_{\widetilde{x}_N,x_0}, 2h_{x_M,x_0}\rangle\\
&+ \langle 2h_{\widetilde{x}_N,x_0}, \nabla L(\widetilde{x}_N) - 2h_{x_M,x_0}\rangle + \langle \nabla L(\widetilde{x}_N) - 2h_{\widetilde{x}_N,x_0}, \nabla L(\widetilde{x}_N) - 2h_{x_M,x_0}\rangle\big],
\end{aligned}$$

where we regard the last term inside the limitation as approximation error term. It is equivalent to analyze

$$\begin{aligned}
\frac{1}{4}D_{-v_x}L(x)\|v_x\|_2 =& -\|h_{x,x_0}\|_2^2 - \lim_{\widetilde{x}_N\to x}\lim_{x_M\to x}\Big[\langle \frac{1}{2}\nabla L(\widetilde{x}_N) - h_{\widetilde{x}_N,x_0}, h_{x_M,x_0}\rangle\\
&+ \langle h_{\widetilde{x}_N,x_0}, \frac{1}{2}\nabla L(x_M) - h_{x_M,x_0}\rangle\\
&+ \langle \frac{1}{2}\nabla L(\widetilde{x}_N) - h_{\widetilde{x}_N,x_0}, \frac{1}{2}\nabla L(x_M) - h_{x_M,x_0}\rangle\Big].
\end{aligned} \tag{42}$$

For simply notation, we let

$$\overline{v}_{x,x_0} = \left(\prod_{i=1}^n W_{i,+,x}\right)^\top \left(\prod_{i=1}^n W_{i,+,x}\right)x - \left(\prod_{i=1}^n W_{i,+,x}\right)^\top \left(\prod_{i=1}^n W_{i,+,x_0}\right)x_0.$$

Thus we have

$$\langle \overline{v}_{x,x_0}, z\rangle = \langle G(x), H_x(z)\rangle - \langle G(x_0), H_x(z)\rangle.$$

For the term $\langle \frac{1}{2}\nabla L(\widetilde{x}_N) - h_{\widetilde{x}_N,x_0}, h_{x_M,x_0}\rangle$ in (42), we have that , setting λ and m sufficiently large as shown above, with probability at least $1 - c_4\exp(-u)$,

$$\begin{aligned}
&\left\langle \frac{1}{2}\nabla L(\widetilde{x}_N) - h_{\widetilde{x}_N,x_0}, h_{x_M,x_0}\right\rangle\\
&= \left\langle \frac{1}{2}\nabla L(\widetilde{x}_N) - \overline{v}_{\widetilde{x}_N,x_0}, h_{x_M,x_0}\right\rangle + \langle \overline{v}_{\widetilde{x}_N,x_0} - h_{\widetilde{x}_N,x_0}, h_{x_M,x_0}\rangle\\
&\geq -\varepsilon\|H_{\widetilde{x}_N}(h_{x_M,x_0})\|_2 - \|\overline{v}_{\widetilde{x}_N,x_0} - h_{\widetilde{x}_N,x_0}\|_2\|h_{x_M,x_0}\|_2\\
&\geq -\varepsilon\|H_{\widetilde{x}_N}(h_{x_M,x_0})\|_2 - 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|\widetilde{x}_N\|_2, \|x_0\|_2)\|h_{x_M,x_0}\|_2\\
&\geq -\varepsilon\left(\frac{1}{2}+\varepsilon_{\mathrm{wdc}}\right)^{n/2}\|h_{x_M,x_0}\|_2 - 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|\widetilde{x}_N\|_2, \|x_0\|_2)\|h_{x_M,x_0}\|_2,
\end{aligned}$$

where the first inequality is by (40) and Cauchy-Schwarz inequality,and the third inequality is by (33) in Lemma D.4. The second inequality above is due to

$$
\begin{aligned}
&\|\overline{v}_{\widetilde{x}_N,x_0} - h_{\widetilde{x}_N,x_0}\|_2 \\
&\leq \left\| \left(\prod_{i=1}^{n} W_{i,+,\widetilde{x}_N}\right)^\top \left(\prod_{i=1}^{n} W_{i,+,\widetilde{x}_N}\right) \widetilde{x}_N - \frac{1}{2^n}\widetilde{x}_N \right\|_2 \\
&\quad + \left\| \left(\prod_{i=1}^{n} W_{i,+,\widetilde{x}_N}\right)^\top \left(\prod_{i=1}^{n} W_{i,+,x_0}\right) x_0 - \frac{1}{2^n} \left[\left(\prod_{i=0}^{n-1} \frac{\pi - \overline{\varrho}_i}{\pi}\right) x_0 + \sum_{i=0}^{n-1} \frac{\sin \overline{\varrho}_i}{\pi} \left(\prod_{j=i+1}^{d-1} \frac{\pi - \overline{\varrho}_j}{\pi}\right) \frac{\|x_0\|_2}{\|\widetilde{x}_N\|_2} \widetilde{x}_N \right] \right\|_2 \\
&\leq 24 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \|\widetilde{x}_N\|_2 + 24 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \|x_0\|_2 \\
&\leq 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|\widetilde{x}_N\|_2, \|x_0\|_2),
\end{aligned}
$$

where the second inequality is by (33) in Lemma D.4.

Similarly, for the terms $\langle h_{\widetilde{x}_N,x_0}, \frac{1}{2}\nabla L(x_M) - h_{x_M,x_0}\rangle$ in (42), we have that, setting m and λ sufficiently large as above, with probability at least $1 - c_4 \exp(-u)$, the following holds:

$$
\left\langle h_{\widetilde{x}_N,x_0}, \frac{1}{2}\nabla L(x_M) - h_{x_M,x_0} \right\rangle \geq -\,\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}}\right)^{n/2} \|h_{\widetilde{x}_N,x_0}\|_2 - 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x_M\|_2, \|x_0\|_2) \|h_{\widetilde{x}_N,x_0}\|_2.
$$

For the terms $\langle \frac{1}{2}\nabla L(\widetilde{x}_N) - h_{\widetilde{x}_N,x_0}, \frac{1}{2}\nabla L(x_M) - h_{x_M,x_0}\rangle$ in (42), we have that, setting m and λ sufficiently large as above, with probability at least $1 - 2c_4 \exp(-u)$, the following holds:

$$
\begin{aligned}
&\left\langle \frac{1}{2}\nabla L(\widetilde{x}_N) - h_{\widetilde{x}_N,x_0}, \frac{1}{2}\nabla L(x_M) - h_{x_M,x_0} \right\rangle \\
&\quad \geq -\left[\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}}\right)^{n/2} + 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x_M\|_2, \|x_0\|_2) \right] \\
&\qquad \cdot \left[\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}}\right)^{n/2} + 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|\widetilde{x}_N\|_2, \|x_0\|_2) \right].
\end{aligned}
$$

Combining the above together, plugging in (42) and taking limit on both sides, we have

$$
\begin{aligned}
-\frac{1}{4} D_{-v_x} L(x) \|v_x\|_2 \geq \frac{1}{2}\|h_{x,x_0}\|_2 &\left[\|h_{x,x_0}\|_2 - 2\left(\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}}\right)^{n/2} + 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x\|_2, \|x_0\|_2) \right) \right] \\
&+ \frac{1}{2} \left[\|h_{x,x_0}\|_2^2 - 2\left(\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}}\right)^{n/2} + 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x\|_2, \|x_0\|_2) \right)^2 \right],
\end{aligned}
$$

with probability at least $1 - 4c_4 \exp(-u)$ by setting m and λ sufficiently large as above.

Discussion of Two Cases: We take our discussion from two aspects: $\|x_0\|_2 > 2^{n/2}\varepsilon_{\mathrm{wdc}}^{1/2}$ and $\|x_0\|_2 \leq 2^{n/2}\varepsilon_{\mathrm{wdc}}^{1/2}$.

Case 1: $\|x_0\|_2 > 2^{n/2}\varepsilon_{\mathrm{wdc}}^{1/2}$, or equivalently $\varepsilon_{\mathrm{wdc}} < 2^{-n}\|x_0\|_2^2$. This means $\|x\|_0$ is not close to 0. If we let $\varepsilon = \varepsilon_{\mathrm{wdc}}$,

$4\pi n\varepsilon_{\mathrm{wdc}} \leq 1$, then we have

$$\begin{aligned}
&\varepsilon\left(\frac{1}{2}+\varepsilon_{\mathrm{wdc}}\right)^{n/2} + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2)\\
&\leq \frac{\|x_0\|\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}(1+2\varepsilon_{\mathrm{wdc}})^{n/2} + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2)\\
&\leq \frac{\|x_0\|\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}(1+2n\varepsilon_{\mathrm{wdc}}) + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2)\\
&\leq \frac{3n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2) + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2)\\
&\leq 51\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2),
\end{aligned}$$

where the second inequality is due to $(1+2\varepsilon_{\mathrm{wdc}})^{n/2} \leq e^{n\varepsilon_{\mathrm{wdc}}} \leq 1+2n\varepsilon_{\mathrm{wdc}}$ when $\varepsilon_{\mathrm{wdc}}$ is sufficiently small satisfying the conditions of Theorem 3.8.

Recall the definition of $S_{121n^4\sqrt{\varepsilon_{\mathrm{wdc}}},x_0}$ in (31). We assume $x \neq 0$ and $x \notin S_{121n^4\sqrt{\varepsilon_{\mathrm{wdc}}},x_0}$, namely $\|h_{x,x_0}\|_2 > 121n^4/2^n\sqrt{\varepsilon_{\mathrm{wdc}}}\max(\|x\|_2,\|x_0\|_2)$. By Lemma D.3, if $x \in \mathcal{B}^c(x_0, 616n^3\varepsilon_{\mathrm{wdc}}^{-1/4}\|x_0\|_2) \cap \mathcal{B}^c(-\rho_n x_0, 5500n^{14}\varepsilon_{\mathrm{wdc}}^{-1/4}\|x_0\|_2)$, it is guaranteed that $x \notin S_{121n^3\sqrt{\varepsilon_{\mathrm{wdc}}},x_0}$ under the condition that $88\pi n^6\varepsilon_{\mathrm{wdc}}^{1/4} < 1$. Then we obtain

$$-\frac{1}{4}D_{-v_x}L(x)\|v_x\|_2 \geq \frac{9}{2}\|h_{x,x_0}\|_2\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2) + \frac{9439}{2}\frac{n^6\varepsilon_{\mathrm{wdc}}}{2^{2n}}[\max(\|x\|_2,\|x_0\|_2)]^2 > 0,$$

or equivalently,

$$D_{-v_x}L(x)\|v_x\|_2 < 0,$$

with probability at least $1-4c_4\exp(-u)$ when we set

$$\lambda \geq 4\max\{c_1(R\|a\|_{\psi_1}+\|\xi\|_{\psi_1}),1\}\log(64\max\{c_1(R\|a\|_{\psi_1}+\|\xi\|_{\psi_1}),1\}/\varepsilon_{\mathrm{wdc}}), \tag{43}$$

$$m \geq c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed)+k\log(2R)+k\log m+u)/\varepsilon_{\mathrm{wdc}}^2. \tag{44}$$

Next, we need to prove that $\forall w \neq 0$, $D_w L(0) < 0$. We compute the directional derivative as

$$\begin{aligned}
\frac{1}{2}D_w L(0)\cdot\|w\|_2 &= -\lim_{x_N\to 0}\frac{\lambda}{m}\sum_{i=1}^m y_i\langle a_i, H_{x_N}(w)\rangle\\
&= \lim_{x_N\to 0}\langle G(x_0), H_{x_N}(w)\rangle - \frac{\lambda}{m}\sum_{i=1}^m y_i\langle a_i, H_{x_N}(w)\rangle - \langle G(x_0), H_{x_N}(w)\rangle\\
&\leq \lim_{x_N\to 0}\left|\langle G(x_0), H_{x_N}(w)\rangle - \frac{\lambda}{m}\sum_{i=1}^m y_i\langle a_i, H_{x_N}(w)\rangle\right| - \langle G(x_0), H_{x_N}(w)\rangle\\
&\leq \varepsilon\left(\frac{1}{2}+\varepsilon_{\mathrm{wdc}}\right)^{n/2}\|w\|_2 - \frac{1}{4\pi}\frac{1}{2^n}\|w\|_2\|x_0\|_2\\
&\leq \frac{1}{2^{n/2}}\varepsilon(1+2n\varepsilon_{\mathrm{wdc}})\|w\|_2 - \frac{1}{4\pi}\frac{1}{2^n}\|w\|_2\|x_0\|_2,
\end{aligned}$$

where the first inequality is due to (40), and the second inequality is due to (34) in Lemma D.4. Now we still let $\varepsilon = \varepsilon_{\mathrm{wdc}}$, then $576\pi^2 n^6\varepsilon_{\mathrm{wdc}} \leq 1$ (which is guaranteed by the condition $88\pi n^6\varepsilon_{\mathrm{wdc}}^{1/4} < 1$). If $w \neq 0$, setting λ and m satisfying (43) and (44), the following holds with probability at least $1 - c_4 \exp(-u)$,

$$
\begin{aligned}
\frac{1}{2} D_w L(0) \cdot \|w\|_2 \leq & \frac{1}{2^{n/2}} \varepsilon_{\mathrm{wdc}} (1 + 2n\varepsilon_{\mathrm{wdc}}) \|w\|_2 - \frac{1}{4\pi} \frac{1}{2^n} \|w\|_2 \|x_0\|_2 \\
\leq & \frac{1}{2^n} \sqrt{\varepsilon_{\mathrm{wdc}}} (1 + 2n\varepsilon_{\mathrm{wdc}}) \|w\|_2 \|x_0\|_2 - \frac{1}{4\pi} \frac{1}{2^n} \|w\|_2 \|x_0\|_2 \\
\leq & \frac{1}{2^n} 3n^3 \sqrt{\varepsilon_{\mathrm{wdc}}} \|w\|_2 \|x_0\|_2 - \frac{1}{4\pi} \frac{1}{2^n} \|w\|_2 \|x_0\|_2 \\
\leq & - \frac{1}{8\pi} \frac{1}{2^n} \|w\|_2 \|x_0\|_2 < 0,
\end{aligned}
$$

where the first inequality is due to the condition that $\varepsilon_{\mathrm{wdc}} < 2^{-n} \|x_0\|_2^2$. This implies that

$$
D_w L(0) < 0, \forall w \neq 0.
$$

Summarizing the results in Case 1, we have that, if we let λ and m satisfying (43) and (44), the following holds with probability at least $1 - 5c_4 \exp(-u)$,

$$
\begin{aligned}
& D_{-v_x} L(x) < 0, \forall x \notin \mathcal{B}(x_0, 616 n^3 \varepsilon_{\mathrm{wdc}}^{1/4} \|x_0\|_2) \cup \mathcal{B}(-\rho_n x_0, 5500 n^{14} \varepsilon_{\mathrm{wdc}}^{1/4} \|x_0\|_2) \cup \{0\}, \\
& D_w L(0) < 0, \forall w \neq 0.
\end{aligned}
$$

Case 2: $\|x_0\|_2 \leq 2^{n/2} \varepsilon_{\mathrm{wdc}}^{1/2}$, or equivalently $\varepsilon_{\mathrm{wdc}} \geq 2^{-n} \|x_0\|_2^2$. This condition means $\|x_0\|$ is very small and close to 0. Then, for any z, we would similarly have

$$
\begin{aligned}
-\frac{1}{4} D_{-v_x} L(x) \|v_x\|_2^2 \geq & \frac{1}{2} \|h_{x,x_0}\|_2 \left[\|h_{x,x_0}\|_2 - 2 \left(\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}} \right)^{n/2} + 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x\|_2, \|x_0\|_2) \right) \right] \\
& + \frac{1}{2} \left[\|h_{x,x_0}\|_2^2 - 2 \left(\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}} \right)^{n/2} + 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x\|_2, \|x_0\|_2) \right)^2 \right].
\end{aligned}
$$

For any non-zero x satisfying $x \notin S_{121 n^4 \sqrt{\varepsilon_{\mathrm{wdc}}}, x_0}$, which can further imply that $\|h_{x,x_0}\|_2 > 121 n^4 2^{-n} \varepsilon_{\mathrm{wdc}} \max(\|x\|_2, \|x_0\|_2)$, we have

$$
\begin{aligned}
-\frac{1}{4} D_{-v_x} L(x) \|v_x\|_2^2 \geq & \frac{1}{2} \|h_{x,x_0}\|_2 \left[25 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x\|_2, \|x_0\|_2) - 2\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}} \right)^{n/2} \right] \\
& + \frac{1}{2} \left[53 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x\|_2, \|x_0\|_2) - \sqrt{2} \varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}} \right)^{n/2} \right] \\
& \cdot \left[\|h_{x,x_0}\|_2 + \sqrt{2} \left(\varepsilon \left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}} \right)^{n/2} + 48 \frac{n^3 \sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n} \max(\|x\|_2, \|x_0\|_2) \right) \right].
\end{aligned}
$$

Furthermore, for any x satisfying $\|x\|_2 \geq 2^{n/2} \sqrt{\varepsilon_{\mathrm{wdc}}}$, we have

$$
\|x\|_2 \geq 2^{n/2} \sqrt{\varepsilon_{\mathrm{wdc}}} \geq \|x_0\|, \text{ namely } x \notin \mathcal{B}(0, 2^{n/2} \varepsilon_{\mathrm{wdc}}^{1/2}),
$$

which leads to

$$\begin{aligned}
-\frac{1}{4}D_{-v_x}L(x)\|v_x\|_2^2 \geq & \frac{1}{2}\|h_{x,x_0}\|_2 \left[25\frac{n^3\sqrt{\varepsilon_{\text{wdc}}}}{2^n}2^{n/2}\sqrt{\varepsilon_{\text{wdc}}} - 2\varepsilon\left(\frac{1}{2}+\varepsilon_{\text{wdc}}\right)^{n/2}\right] \\
&+ \frac{1}{2}\left[53\frac{n^3\sqrt{\varepsilon_{\text{wdc}}}}{2^n}2^{n/2}\sqrt{\varepsilon_{\text{wdc}}} - \sqrt{2}\varepsilon\left(\frac{1}{2}+\varepsilon_{\text{wdc}}\right)^{n/2}\right] \\
&\cdot \left[\|h_{x,x_0}\|_2 + \sqrt{2}\left(\varepsilon\left(\frac{1}{2}+\varepsilon_{\text{wdc}}\right)^{n/2} + 48\frac{n^3\sqrt{\varepsilon_{\text{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2)\right)\right] \\
= & \frac{1}{2}\|h_{x,x_0}\|_2 \left[25\frac{n^3\varepsilon_{\text{wdc}}}{2^{n/2}} - 2\varepsilon\left(\frac{1}{2}+\varepsilon_{\text{wdc}}\right)^{n/2}\right] \\
&+ \frac{1}{2}\left[53\frac{n^3\varepsilon_{\text{wdc}}}{2^{n/2}} - \sqrt{2}\varepsilon\left(\frac{1}{2}+\varepsilon_{\text{wdc}}\right)^{n/2}\right] \\
&\cdot \left[\|h_{x,x_0}\|_2 + \sqrt{2}\left(\varepsilon\left(\frac{1}{2}+\varepsilon_{\text{wdc}}\right)^{n/2} + 48\frac{n^3\sqrt{\varepsilon_{\text{wdc}}}}{2^n}\max(\|x\|_2,\|x_0\|_2)\right)\right].
\end{aligned}$$

We let $\varepsilon = \varepsilon_{\text{wdc}}$. Then we have $\varepsilon(1/2+\varepsilon_{\text{wdc}})^{n/2} \leq 3n\varepsilon_{\text{wdc}}2^{-n/2}$, which consequently results in

$$-\frac{1}{4}D_{-v_x}L(x)\|v_x\|_2^2 > 0,$$

or equivalently,

$$D_{-v_x}L(x)\|v_x\|_2^2 < 0.$$

Note that in the above results, we also apply (40) in deriving the inequalities. Therefore, summarizing the above results in Case 2, we have that, if we let λ and m satisfying (43) and (44), the following holds with probability at least $1-4c_4\exp(-u)$,

$$D_{-v_x}L(x) < 0, \ \ \forall x \notin \mathcal{B}(x_0, 616n^3\varepsilon_{\text{wdc}}^{1/4}\|x_0\|_2) \cup \mathcal{B}(-\rho_n x_0, 5500n^{14}\varepsilon_{\text{wdc}}^{1/4}\|x_0\|_2) \cup \mathcal{B}(0, 2^{n/2}\varepsilon_{\text{wdc}}^{1/2}),$$

which completes the proof. □

E. Proof of Theorem 3.10

The proof of Theorem 3.10 is mainly based on Lemmas D.1 and D.2 proved in the last section and two additional lemmas in the previous literature (Huang et al., 2018) given as below.

E.1. Lemmas for Theorem 3.10

Lemma E.1 ((Huang et al., 2018))**.** *Fix $0 < \psi \leq \frac{1}{4\pi}$. For any $\varphi, \zeta \in [\rho_n, 1]$, it holds that*

$$\begin{aligned}
\langle x, h_{x,x_0}\rangle - \frac{1}{2^{n+1}}\|x\|_2^2 &\leq \frac{1}{2^{n+1}}\left(\varphi^2 - 2\varphi + \frac{10\pi^2 n}{K_0^3}\psi\right)\|x_0\|_2^2, \forall x \in \mathcal{B}(\varphi x_0, \psi\|x_0\|_2) \\
\langle z, h_{z,x_0}\rangle - \frac{1}{2^{n+1}}\|z\|_2^2 &\geq \frac{1}{2^{n+1}}(\zeta^2 - 2\zeta\rho_n - 10\pi^2 n^3\psi)\|x_0\|_2^2, \forall z \in \mathcal{B}(-\zeta x_0, \psi\|x_0\|_2)
\end{aligned}$$

where $K_0 = \min_{n\geq 2}\rho_n$, and ρ_n is defined in Lemma D.3.

Lemma E.2 ((Huang et al., 2018))**.** *For all $n \geq 2$, there exists a constant K_1 such that*

$$\frac{1}{K_1(n+2)^2} \leq 1 - \rho_n.$$

E.2. Proof Sketches of Theorem 3.10

Our proof of Theorem 3.10 is sketched as follows:

- We first show that $L(x)$ can be approximated as $2\langle h_{x,x_0}, x\rangle - \|G(x)\|_2^2$ by the two critical lemmas, Lemma D.1 and Lemma D.2.
- Then we bound the approximation error $|L(x) - 2\langle(h_{x,x_0}, x\rangle - \|G(x)\|_2^2)|$, where h_{x,x_0} is defined in (30).
- By Lemmas E.1, E.2, we have that if x and z are around x_0 and $-\rho_n x_0$ respectively, by considering the approximation errors, the upper bound of $L(x)$ is smaller than the lower bound of $L(z)$, which further leads to $L(x) < L(z)$ with high probability.

E.3. Detailed Proof of Theorem 3.10

Proof of Theorem 3.10. By (35) in Lemma D.4, we have have

$$\|G(x)\|_2 \leq (1/2 + \varepsilon_{\mathrm{wdc}})^{n/2}\|x\|_2, \tag{45}$$

combining which and the assumption $\|x_0\|_2 \leq R(1/2 + \varepsilon_{\mathrm{wdc}})^{-n/2}$ in Theorem 3.10, we further have

$$\|G(x_0)\|_2 \leq R.$$

By Lemma D.1 and $\|G(x_0)\|_2 \leq R$, we set

$$\lambda \geq 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}\log(64\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon),$$

and $z = x$ in Lemma D.1 such that $H_x(x) = G(x)$, and the following holds for any x,

$$|\lambda\mathbb{E}[y_i\langle a_i, G(x)\rangle] - \langle G(x_0), G(x)\rangle| \leq \frac{1}{4}\varepsilon\|G(x)\|_2. \tag{46}$$

According to Lemma D.2 and $|G(x_0)| \leq R$, we have that with probability at least $1 - c_4\exp(-u)$, for any x, the following holds:

$$\left|\frac{\lambda}{m}\sum_{i=1}^{m} y_i\langle a_i, G(x)\rangle - \lambda\mathbb{E}[y_i\langle a_i, G(x)\rangle]\right| \leq \frac{\varepsilon}{8}\|G(x)\|_2, \tag{47}$$

with sample complexity being

$$m \geq c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon^2,$$

where we also set $z = x$ in Lemma D.2 such that $H_x(x) = G(x)$.

Combining (46) and (47), we will have that with probability at least $1 - c_4\exp(-u)$, for any x, setting

$$\lambda \geq 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}\log(64\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon),$$

and

$$m \geq c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon^2,$$

the following holds:

$$\left|\frac{\lambda}{m}\sum_{i=1}^{m} y_i\langle a_i, G(x)\rangle - \langle G(x_0), G(x)\rangle\right| \leq \varepsilon\|G(x)\|_2. \tag{48}$$

Bounding the error term: We next bound the term $|L(x) + \|G(x)\|_2^2 - 2\langle h_{x,x_0}, x\rangle|$ as follows. With λ, m satisfying the same conditions above, then with probability at least $1 - c_4\exp(-u)$, the following holds:

$$
\begin{aligned}
&\big|L(x) + \|G(x)\|_2^2 - 2\langle h_{x,x_0}, x\rangle\big| \\
&\quad = \left|2\|G(x)\|_2^2 - \frac{2\lambda}{m}\sum_{i=1}^m y_i\langle a_i, G(x)\rangle - 2\langle h_{x,x_0}, x\rangle\right| \\
&\quad = \left|2\langle G(x_0), G(x)\rangle - \frac{2\lambda}{m}\sum_{i=1}^m y_i\langle a_i, G(x)\rangle + 2\|G(x)\|_2^2 - 2\langle G(x_0), G(x)\rangle - 2\langle h_{x,x_0}, x\rangle\right|.
\end{aligned}
$$

Furthermore, we bound the above terms as follows

$$
\begin{aligned}
&\left|2\langle G(x_0), G(x)\rangle - \frac{2\lambda}{m}\sum_{i=1}^m y_i\langle a_i, G(x)\rangle\right| + \big|2\|G(x)\|_2^2 - 2\langle G(x_0), G(x)\rangle - 2\langle h_{x,x_0}, x\rangle\big| \\
&\qquad \le 2\varepsilon\|G(x)\|_2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2\|x\|_2 \\
&\qquad \le 2\varepsilon\left(\frac{1}{2} + \varepsilon_{\mathrm{wdc}}\right)^{n/2}\|x\|_2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2\|x\|_2 \\
&\qquad \le 2\varepsilon\frac{1 + 2n\varepsilon_{\mathrm{wdc}}}{2^{n/2}}\|x\|_2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2\|x\|_2,
\end{aligned}
$$

where the second inequality is due to (48) and (33) in Lemma D.4, the third inequality is due to (45), and the last inequality is due to $(1 + 2\varepsilon_{\mathrm{wdc}})^{n/2} \le e^{n\varepsilon_{\mathrm{wdc}}} \le 1 + 2n\varepsilon_{\mathrm{wdc}}$ if $\varepsilon_{\mathrm{wdc}}$ is sufficiently small satisfying the condition of Theorem 3.10. This result implies

$$
\big|L(x) + \|G(x)\|_2^2 - 2\langle h_{x,x_0}, x\rangle\big| \le 2\varepsilon\frac{1 + 2n\varepsilon_{\mathrm{wdc}}}{2^{n/2}}\|x\|_2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2\|x\|_2.
$$

Since we only consider the case that $\varepsilon_{\mathrm{wdc}} \le 2^{-n}\|x_0\|_2^2$. Letting $\varepsilon = \varepsilon_{\mathrm{wdc}}$, we have

$$
\begin{aligned}
&\big|L(x) + \|G(x)\|_2^2 - 2\langle h_{x,x_0}, x\rangle\big| \\
&\qquad \le 2\varepsilon_{\mathrm{wdc}}\frac{1 + 2n\varepsilon_{\mathrm{wdc}}}{2^{n/2}}\|x\|_2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2\|x\|_2 \\
&\qquad \le 2\sqrt{\varepsilon_{\mathrm{wdc}}}\frac{1 + 2n\varepsilon_{\mathrm{wdc}}}{2^n}\|x_0\|_2\|x\|_2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2 + 48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2\|x\|_2 \qquad (49)
\end{aligned}
$$

with probability $1 - c_4\exp(-u)$ if we set

$$
\lambda \ge 4\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}\log(64\max\{c_1(R\|a\|_{\psi_1} + \|\xi\|_{\psi_1}), 1\}/\varepsilon_{\mathrm{wdc}}), \text{ and} \qquad (50)
$$

$$
m \ge c_2\|a\|_{\psi_1}^2\lambda^2\log^2(\lambda m)(kn\log(ed) + k\log(2R) + k\log m + u)/\varepsilon_{\mathrm{wdc}}^2. \qquad (51)
$$

Upper bound of $L(x)$: For any $x \in \mathcal{B}(\varphi x_0, \psi\|x_0\|_2)$ with $0 < \psi \le 1/(4\pi)$ and any $\varphi \in [\rho_n, 1]$, we have

$$
\begin{aligned}
L(x) =& 2\langle x, h_{x,x_0}\rangle - \|G(x)\|_2^2 + \big(L(x) - 2\langle x, h_{x,x_0}\rangle + \|G(x)\|_2^2\big) \\
=& 2\langle x, h_{x,x_0}\rangle - \frac{1}{2^n}\|x\|_2^2 - \left(\|G(x)\|_2^2 - \frac{1}{2^n}\|x\|_2^2\right) + \big(L(x) - 2\langle x, h_{x,x_0}\rangle + \|G(x)\|_2^2\big) \\
\le& 2\langle x, h_{x,x_0}\rangle - \frac{1}{2^n}\|x\|_2^2 + \left|\|G(x)\|_2^2 - \frac{1}{2^n}\|x\|_2^2\right| + \big|L(x) - 2\langle x, h_{x,x_0}\rangle + \|G(x)\|_2^2\big| \\
\le& \frac{1}{2^n}(\varphi^2 - 2\varphi + \frac{10\pi^2 n}{K_0^3}\psi)\|x_0\|_2^2 + \left|\|G(x)\|_2^2 - \frac{1}{2^n}\|x\|_2^2\right| + \big|L(x) - 2\langle x, h_{x,x_0}\rangle + \|G(x)\|_2^2\big|,
\end{aligned}
$$

where the last inequality is due to Lemma E.1 and (45). In addition, we can also obtain

$$\begin{aligned}
&\left|L(x)-2\langle x,h_{x,x_0}\rangle+\|G(x)\|_2^2\right|\\
&\quad\le\sqrt{\varepsilon_{\mathrm{wdc}}}\frac{1+2n\varepsilon_{\mathrm{wdc}}}{2^n}\|x_0\|_2\|x\|_2+24\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2+24\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2\|x\|_2\\
&\quad\le 2\sqrt{\varepsilon_{\mathrm{wdc}}}\frac{1+2n\varepsilon_{\mathrm{wdc}}}{2^n}(\varphi+\psi)\|x_0\|_2^2+48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}(\varphi+\psi)^2\|x_0\|_2^2+48\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}(\varphi+\psi)\|x_0\|_2^2\\
&\quad\le 122\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2^2,
\end{aligned}$$

and

$$\left|\|G(x)\|_2^2-\frac{1}{2^n}\|x\|_2^2\right|\le 24\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x\|_2^2\le 30\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2,$$

due to $\|x\|_2\le(\varphi+\psi)\|x_0\|_2$ when $x\in\mathcal{B}(\varphi x_0,\psi\|x_0\|_2)$ and $\varphi+\psi\le 1+1/(4\pi)<1.1$, and (33) in Lemma D.4.

Combining the above results and letting λ and m satisfy (50) and (51), the following holds with probability at least $1-c_4\exp(-u)$,

$$L(x)\le\frac{1}{2^n}(\varphi^2-2\varphi+\frac{10\pi^2 n}{K_0^3}\psi+152n^3\sqrt{\varepsilon_{\mathrm{wdc}}})\|x_0\|_2^2,$$

for any $x\in\mathcal{B}(\varphi x_0,\psi\|x_0\|_2)$.

Lower bound of $L(z)$: Next, we should the lower bound of $L(z)$ when z is around $-\rho_n x_0$. Consider the situation for any $z\in\mathcal{B}(-\zeta x_0,\psi\|x_0\|_2)$ with $0<\psi\le 1/(4\pi)$ and any $\zeta\in[\rho_n,1]$. We can obtain

$$\begin{aligned}
L(z)=&2\langle z,h_{z,x_0}\rangle-\|G(z)\|_2^2+\left(L(z)-2\langle z,h_{z,x_0}\rangle+\|G(z)\|_2^2\right)\\
\ge&2\langle z,h_{z,x_0}\rangle-\|G(z)\|_2^2-\left|L(z)-2\langle z,h_{z,x_0}\rangle+\|G(z)\|_2^2\right|\\
=&2\langle z,h_{z,x_0}\rangle-\frac{1}{2^n}\|z\|_2^2-\left(\|G(z)\|_2^2-\frac{1}{2^n}\|z\|_2^2\right)-\left|L(z)-2\langle z,h_{z,x_0}\rangle+\|G(z)\|_2^2\right|\\
\ge&2\langle z,h_{z,x_0}\rangle-\frac{1}{2^n}\|z\|_2^2-\left|\|G(z)\|_2^2-\frac{1}{2^n}\|z\|_2^2\right|-\left|L(z)-2\langle z,h_{z,x_0}\rangle+\|G(z)\|_2^2\right|\\
\ge&\frac{1}{2^n}(\zeta^2-2\zeta\rho_n-10\pi^2n^3\psi)\|x_0\|_2^2-\left|\|G(x)\|_2^2-\frac{1}{2^n}\|x\|_2^2\right|-\left|L(x)-2\langle x,h_{x,x_0}\rangle+\|G(x)\|_2^2\right|,
\end{aligned}$$

where the last inequality is due to Lemma E.1. Furthermore, similar to the previous steps in the upper bound of $L(x)$, we have

$$\left|L(z)-2\langle z,h_{z,x_0}\rangle+\|G(z)\|_2^2\right|\le 122\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2^2,$$

and

$$\left|\|G(z)\|_2^2-\frac{1}{2^n}\|z\|_2^2\right|\le 30\frac{n^3\sqrt{\varepsilon_{\mathrm{wdc}}}}{2^n}\|x_0\|_2,$$

due to $\|z\|_2\le(\zeta+\psi)\|x_0\|_2$ when $z\in\mathcal{B}(-\zeta x_0,\psi\|x_0\|_2)$ and $\zeta+\psi\le 1+1/(4\pi)<1.1$.

Combining the above results, letting λ and m satisfy (50) and (51), the following holds with probability at least $1-c_4\exp(-u)$,

$$L(z)\ge\frac{1}{2^n}(\zeta^2-2\zeta\rho_n-10\pi^2n^3\psi-152n^3\sqrt{\varepsilon_{\mathrm{wdc}}})\|x_0\|_2^2,$$

for any $z\in\mathcal{B}(-\zeta x_0,\psi\|x_0\|_2)$.

Proving $L(x) < L(z)$: In order to have $L(x) < L(z)$, it is enough to ensure that

$$\min_{\zeta\in[\rho_n,1]} \frac{1}{2^n}(\zeta^2 - 2\zeta\rho_n - 10\pi^2 n^3\psi - 152n^3\sqrt{\varepsilon_{\mathrm{wdc}}})\|x_0\|_2^2$$
$$> \max_{\varphi\in[\rho_n,1]} \frac{1}{2^n}(\varphi^2 - 2\varphi + \frac{10\pi^2 n}{K_0^3}\psi + 152n^3\sqrt{\varepsilon_{\mathrm{wdc}}})\|x_0\|_2^2.$$

The minimizer for the left side of the above inequality is $\varphi = \rho_n$ while the maximizer for the right side is also $\zeta = \rho_n$. Then, to achieve the above inequality, we plug in the minimizer and maximizer for both sides and obtain

$$\rho_n^2 - 2\rho_n^2 - 10\pi^2 n^3\psi - 152n^3\sqrt{\varepsilon_{\mathrm{wdc}}} > \rho_n^2 - 2\rho_n + \frac{10\pi^2 n}{K_0^3}\psi + 152n^3\sqrt{\varepsilon_{\mathrm{wdc}}}.$$

Rearranging the terms, we would obtain

$$2\rho_n - 2\rho_n^2 > \left(10\pi^2 n^3 + \frac{10\pi^2 n}{K_0^3}\right)\psi + 304n^3\sqrt{\varepsilon_{\mathrm{wdc}}}.$$

To make the above inequality hold for all ρ_n, by computing the minimal value of the left-hand side according to Lemma E.2, we require $\varepsilon_{\mathrm{wdc}}$ to satisfy

$$\frac{2K_0}{K_1(n+2)^2} > \left(10\pi^2 n^3 + \frac{10\pi^2 n}{K_0^3}\right)\psi + 304n^3\sqrt{\varepsilon_{\mathrm{wdc}}}.$$

Due to $n+2 \leq 2n$ and $n \leq n^3$ (since we assume $n > 1$), it suffices to ensure

$$\frac{K_0}{4K_1 n^2} > \left(10\pi^2 n^3 + \frac{10\pi^2 n^3}{K_0^3}\right)\psi + 304n^3\sqrt{\varepsilon_{\mathrm{wdc}}},$$

which can be, therefore, guaranteed by the condition

$$35\sqrt{K_1/K_0}\,n^3\varepsilon_{\mathrm{wdc}}^{1/4} \leq 1 \text{ and } \psi \leq \frac{K_0}{50\pi^2 K_1(1+1/K_0^3)}n^{-5}.$$

Thus, under the condition of Theorem 3.10, for any $x \in \mathcal{B}(\varphi x_0, \psi\|x_0\|_2)$ and $z \in \mathcal{B}(-\zeta x_0, \psi\|x_0\|_2)$, letting λ and m satisfy (50) and (51), with probability at least $1 - 2c_4\exp(-u)$, we have

$$L(x) < L(z).$$

Note that the radius ψ satisfies $\psi < K_0 := \rho_n$, which means there are no overlap between $\mathcal{B}(\varphi x_0, \psi\|x_0\|_2)$ and $\mathcal{B}(-\zeta x_0, \psi\|x_0\|_2)$. This is because by Lemma E.2, we know that $1/K_1 \leq (n+2)^2 \leq 4n^2$. Therefore, $\psi \leq K_0 n^{-5}/(50\pi^2 K_1(1+1/K_0^3)) \leq K_0 n^{-3} < K_0$ when $n \geq 2$. This completes the proof. □