A. Proof Details

From now on, we always implicitly require that Assumption 1 and Assumption 2 hold if not specified. Since the INITIALIZE algorithm also depends on REFINE, we will start our analysis with the latter.

First of all, we recall a few important notations. Given $w \in \mathbb{R}^d$, the sampling region for the unlabeled example is a band, which is given by

$$X_{\hat{w},b} := \{x \in \mathbb{R}^d : 0 < \hat{w} \cdot x \leq b\}, \text{ where } \hat{w} := \frac{w}{\|w\|}. \tag{4}$$

We denoted by $D_{X|\hat{w},b}$ the distribution of D_X conditioned on $x \in X_{\hat{w},b}$, and by $D_{\hat{w},b}$ the distribution of D conditioned on $x \in X_{\hat{w},b}$.

Recall that given $w \in \mathbb{R}^d$ and $x \sim D_{X|\hat{w},b}$, our prediction $\hat{y} = \operatorname{sign}(w \cdot x) = 1$ and we set the gradient in REFINE as

$$g = x \cdot \mathbf{1}_{\{y=-1\}},$$

where y is the label returned by the adversary. Also recall that we defined the potential function

$$f_{u,b}(w) = \mathbb{E}_{x \sim D_{X|\hat{w},b}} \left[|u \cdot x| \cdot \mathbf{1}_{\{u \cdot x < 0\}} \right].$$

Finally, we note that the capital letters C and K, and their subscript variants such as C_1 and K_1, are used to denote absolute constants whose values may differ from appearance to appearance. However, we reserve c_0, c_1, and $c_2 > 0$ for specific absolute constants: c_0 is a sufficiently small constant such that the noise rate $\nu \leq c_0 \epsilon$, c_1 and c_2 are specified in Lemma 28 and Lemma 29 respectively.

A.1. Analysis of REFINE

Intuitively, since we are performing gradient descent, we would hope that the negative gradient has a nontrivial correlation with the underlying halfspace u. The following lemma formalizes the intuition.

Lemma 13. *Given $w \in \mathbb{R}^d$ with $\|w - u\| \leq r$ and $b > 0$, let $g = x \cdot \mathbf{1}_{\{y=-1\}}$ where $(x, y) \sim D_{\hat{w},b}$. If the noise rate $\nu \leq c_0 b$ for some absolute constant $c_0 > 0$, then*

$$\mathbb{E}\left[u \cdot (-g)\right] \geq f_{u,b}(w) - \sqrt{\frac{c_0 c_1}{c_2}} \cdot (b + r),$$

where the expectation is taken over the random draw of (x, y).

Proof. By the definition of g, it follows that

$$\mathbb{E}\left[u \cdot (-g)\right] = \mathbb{E}\left[-(u \cdot x) \cdot \mathbf{1}_{\{y=-1\}}\right]. \tag{5}$$

As we can rewrite the indicator function $\mathbf{1}_{\{y=-1\}}$ in an equivalent form as follows:

$$\mathbf{1}_{\{y=-1\}} = \mathbf{1}_{\{u \cdot x < 0\}} + \mathbf{1}_{\{u \cdot x > 0, y=-1\}} - \mathbf{1}_{\{y=1, u \cdot x < 0\}}, \tag{6}$$

(5) can be written as

$$\mathbb{E}\left[u \cdot (-g)\right] = \underbrace{\mathbb{E}\left[-(u \cdot x) \cdot \mathbf{1}_{\{u \cdot x < 0\}}\right]}_{E_1} + \underbrace{\mathbb{E}\left[-(u \cdot x) \cdot \left(\mathbf{1}_{\{u \cdot x > 0, y=-1\}} - \mathbf{1}_{\{y=1, u \cdot x < 0\}}\right)\right]}_{E_2}.$$

First, we argue that $E_1 = f_{u,b}(w)$. In fact, when $u \cdot x \geq 0$, $E_1 = 0 = f_{u,b}(w)$; when $u \cdot x < 0$, $E_1 = \mathbb{E}[|u \cdot x|] = f_{u,b}(w)$.

Let us now consider the term E_2, whose absolute value can be bounded by

$$|E_2| \leq \mathbb{E}\left[|u \cdot x| \cdot \mathbf{1}_{\{\operatorname{sign}(u \cdot x) \neq y\}}\right] \leq \sqrt{\mathbb{E}\left[(u \cdot x)^2\right] \cdot \mathbb{E}\left[\mathbf{1}_{\{\operatorname{sign}(u \cdot x) \neq y\}}\right]}.$$

By Lemma 28, $\mathbb{E}\left[(u \cdot x)^2\right] \le c_1(b^2 + r^2)$. On the other side, from the definition of ν-adversarial noise, we have

$$\mathbb{E}\left[\mathbf{1}_{\{\operatorname{sign}(u\cdot x)\neq y\}}\right] = \Pr_{(x,y)\sim D_{\hat{w},b}}(y \neq \operatorname{sign}(u\cdot x)) \le \frac{\Pr_{(x,y)\sim D}(y\neq \operatorname{sign}(u\cdot x))}{\Pr_{x\sim D_X}(x\in X_{\hat{w},b})} \le \frac{\nu}{c_2\cdot b}.$$

In the first inequality of the above expression, we use the fact that for an event A, $\Pr_{(x,y)\sim D_{\hat{w},b}}(A) = \Pr_{(x,y)\sim D}(A \mid x \in X_{\hat{w},b}) \le \frac{\Pr_{(x,y)\sim D}(A)}{\Pr_{x\sim D_X}(x\in X_{\hat{w},b})}$. In the second inequality, we use Lemma 29 to bound the denominator from below.

Therefore,

$$|E_2| \le \sqrt{c_1(b^2+r^2)\cdot \frac{\nu}{c_2\cdot b}} \le \sqrt{\frac{c_0c_1}{c_2}(b^2+r^2)} \le \sqrt{\frac{c_0c_1}{c_2}}\cdot(b+r).$$

Now combining the above estimate and that of E_1, we prove the lemma. □

Lemma 14. *There exists an absolute constant $C>0$ such that the following holds. Suppose the algorithm* Refine *is run with initialization w_0, step size $\alpha>0$, bandwidth $b>0$, convex constraint set $\mathcal{K}$, regularizer $\Phi(w)=\frac{1}{2(p-1)}\|w-w_0\|_p^2$, number of iterations T, where the following are satisfied:*

1. *$\|w_0-u\|_1 \le \rho$;*
2. *$w_0\in\mathcal{K}$ and $u\in\mathcal{K}$;*
3. *for all $w\in\mathcal{K}$, $\|w-u\|\le r$.*

Then, with probability $1-\delta$,

$$C\cdot\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \le (b+r)\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{\log(1/\delta)}{T} + C\sqrt{\frac{c_0c_1}{c_2}}\right) + \frac{\rho^2\log d}{\alpha T} + \alpha\cdot\log^2\frac{Td}{b\delta}.$$

Proof. By standard analysis of online mirror descent (see, e.g. Theorem 6.8 of Orabona (2019)), we have

$$\alpha\sum_{t=1}^{T} w_{t-1}\cdot g_t - \alpha\sum_{t=1}^{T} u\cdot g_t \le \mathcal{B}_\Phi(u,w_0) + \sum_{t=1}^{T}\|\alpha g_t\|_q^2.$$

Since $w_{t-1}\cdot g_t = w_{t-1}\cdot x_t\cdot \mathbf{1}_{\{y_t=-1\}}$ and x_t is such that $\hat{w}_{t-1}\cdot x_t > 0$, we have $w_{t-1}\cdot g_t \ge 0$ for all t. Using this observation and dividing both sides by α, we obtain

$$\sum_{t=1}^{T} u\cdot(-g_t) \le \frac{\mathcal{B}_\Phi(u,w_0)}{\alpha} + \alpha\sum_{t=1}^{T}\|g_t\|_q^2. \tag{7}$$

We first present upper bounds for the right-hand side. In particular, note that

$$\mathcal{B}_\Phi(u,w_0) = \frac{1}{2(p-1)}\|u-w_0\|_p^2 \le \frac{\ln(8d)-1}{2}\rho^2 \le \frac{\rho^2\ln(8d)}{2}, \tag{8}$$

where in the first inequality we use the fact that for any $p>1, \|u-w_0\|_p \le \|u-w_0\|_1$.

For the ℓ_q-norm of g_t, denote $g_t^{(j)}$ the jth coordinate of g_t. We have

$$\|g_t\|_q = \left(\sum_{j=1}^{d}\left|g_t^{(j)}\right|^q\right)^{1/q} \le \left(d\,\|g_t\|_\infty^q\right)^{1/q} \le 2\,\|g_t\|_\infty \le 2\,\|x_t\|_\infty, \tag{9}$$

where the first inequality makes use of the definition of ℓ_∞-norm, the second inequality applies the setting $q=\ln(8d)$, and the last step follows from the setting of g_t. On the other side, it is known that for any $x_t \sim D_{X|\hat{w}_{t-1},b}$, with probability

$1 - \frac{\delta}{2T}$, we have $\|x_t\|_\infty \leq K_1 \cdot \log \frac{Td}{b\delta}$ for some absolute constant $K_1 > 0$; see Lemma 30. Hence, the union bound implies that with probability $1 - \frac{\delta}{2}$, $\max_{1 \leq t \leq T} \|x_t\|_\infty \leq K \cdot \log \frac{Td}{b\delta}$, which further gives

$$\max_{1 \leq t \leq T} \|g_t\|_q \leq 2K_1 \cdot \log \frac{Td}{b\delta}. \tag{10}$$

Now we consider a lower bound of $\sum_{t=1}^T u \cdot (-g_t)$. Define the filtration $\mathcal{F}_t := \sigma(w_0, x_1, y_1, w_1, \dots, x_{t-1}, y_{t-1}, w_t)$, and denote by $\mathbb{E}_{t-1}[\cdot]$ the expectation over $(x_t, y_t) \sim D_{\hat{w}_{t-1}, b}$ conditioning on the past filtration $\mathcal{F}_{t-1}$; likewise for $\Pr_{t-1}(\cdot)$.

By existing tail bound of one-dimensional isotropic log-concave distributions in the band $X_{\hat{w}_{t-1}, b}$ (see, e.g. Lemma 3.3 of Awasthi et al. (2017)), and the fact that $\|u - \hat{w}_{t-1}\| \leq 2\|u - w_{t-1}\| \leq 2r$, we have

$$\Pr_{t-1}\left(|u \cdot x_t| \geq a \right) \leq K \exp\left(-K' \cdot \frac{a}{2r + b} \right),$$

for some constants $K, K' > 0$, implying that

$$\Pr_{t-1}\left(|u \cdot (-g_t)| \geq a \right) \leq K \exp\left(-K' \cdot \frac{a}{2r + b} \right).$$

Now applying Lemma 31 with $Z_t = u \cdot (-g_t)$ therein gives that with probability $1 - \frac{\delta}{2}$,

$$\left| \sum_{t=1}^T u \cdot (-g_t) - \mathbb{E}_{t-1}[u \cdot (-g_t)] \right| \leq K_2 (b + r) \left(\sqrt{T \log \frac{1}{\delta}} + \log \frac{1}{\delta} \right). \tag{11}$$

The above concentration of martingales, in allusion to Lemma 13, gives

$$\sum_{t=1}^T u \cdot (-g_t) \geq \sum_{t=1}^T f_{u,b}(w_{t-1}) - T \cdot \sqrt{\frac{c_0 c_1}{c_2}} \cdot (b + r) - K_2 (b + r) \left(\sqrt{T \log \frac{1}{\delta}} + \log \frac{1}{\delta} \right). \tag{12}$$

Combining (7), (8), (10), and (12), we obtain

$$\begin{aligned} C \cdot \frac{1}{T} \sum_{t=1}^T f_{u,b}(w_{t-1}) \leq (b + r) &\left(\frac{\sqrt{\log(1/\delta')}}{\sqrt{T}} + \frac{\log(1/\delta')}{T} + C \sqrt{\frac{c_0 c_1}{c_2}} \right) \\ &+ \frac{\rho^2 \log d}{\alpha T} + \alpha \cdot \log^2 \frac{Td}{b\delta} \end{aligned}$$

for some absolute constant $C > 0$. □

The following proposition is an immediate result of Lemma 14 by specifying the involved hyper-parameters and showing that u stays in the convex constraint set $\mathcal{K}$.

Proposition 15. *Suppose that the adversarial noise rate $\nu \leq c_0 \epsilon$ for some sufficiently small absolute constant $c_0 > 0$. Consider running the* Refine *algorithm with step size $\alpha = \tilde{\Theta}\left(\theta \cdot \log^{-2} \frac{d}{\delta\theta} \right)$, bandwidth $b = \Theta(\theta)$, convex constraint set $\mathcal{K} = \{ w \in \mathbb{R}^d : \|w - w_0\| \leq \theta, \|w\| \leq 1 \}$, regularizer $\Phi(w) = \frac{1}{2(p-1)} \|w - w_0\|_p^2$, number of iterations $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta\theta})$. If the initial iterate w_0 is such that $\|w_0\| = 1$, $\|w_0\|_0 \leq s$, and $\|w_0 - u\| \leq \theta$ for some $\theta \leq \frac{\pi}{16}$, then with probability $1 - \delta$,*

$$\frac{1}{T} \sum_{t=1}^T f_{u,b}(w_{t-1}) \leq \frac{\theta}{50 \cdot 3^4 \cdot 2^{33}}.$$

Proof. We will first verify that the premises of Lemma 14 are satisfied. First of all, it is easy to see that $\|w_0 - u\|_1 \leq \sqrt{\|w_0 - u\|_0} \cdot \|w_0 - u\| \leq \sqrt{2s} \cdot \theta$. Hence we can choose $\rho = \sqrt{2s} \cdot \theta$ in Lemma 14. Next, we have $\|u - w_0\| \leq \theta$ in view of our condition on w_0, which together with the fact that $\|u\| = 1$ implies $u \in \mathcal{K}$. Last, for all $w \in \mathcal{K}$, by triangle inequality $\|w - u\| \leq \|w - w_0\| + \|w_0 - u\| \leq 2\theta$. Hence we can choose $r = 2\theta$ in Lemma 14.

Now with $\rho = \sqrt{2s}\cdot\theta$ and $r = 2\theta$, Lemma 14 indicates that with probability $1-\delta$,

$$C\cdot\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \le (b+2\theta)\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{\log(1/\delta)}{T} + C\sqrt{\frac{c_0 c_1}{c_2}}\right) + \frac{\theta^2\cdot 2s\log d}{\alpha T} + \alpha\cdot\log^2\frac{Td}{b\delta}.$$

We need each term on the right-hand side is upper bounded by $\frac{C\cdot\theta}{3\cdot 50\cdot 3^4\cdot 2^{33}}$. First, we choose $b = \Theta(\theta)$. Then for the first term, it suffices to choose $T \ge \log(1/\delta)$ and set c_0 to be a sufficiently small absolute constant (this is possible since C, c_1, and c_2 are all fixed absolute constants). The second and the last terms require $\alpha T \ge \Omega(\theta\cdot s\log d)$ and $\alpha\cdot\log^2\frac{Td}{b\delta} = O(\theta)$ respectively. The latter implies $\alpha = O(\theta\cdot\log^{-2}\frac{Td}{b\delta})$, thus combining it with the former we need

$$\frac{\theta\cdot s\log d}{T} \le \theta\cdot\log^{-2}\frac{Td}{b\delta}.$$

This can be satisfied if we choose $T = \tilde{O}(s\log d\cdot\log^2\frac{d}{\delta\theta})$. Finally, we have $\alpha = \tilde{\Theta}(\theta\cdot\log^{-2}\frac{d}{\delta\theta})$. □

Remark 16. In the above proof, we note that $C\sqrt{\frac{c_0 c_1}{c_2}}$ is an extra term introduced by the adversarial noise model. It is important to observe that c_0 is chosen as a very small *absolute constant*. Thus the noise tolerance still reads as $\nu = \Omega(\epsilon)$. The term $C\sqrt{\frac{c_0 c_1}{c_2}}$ does not appear in the bounded noise analysis though; see Lemma 8 of Zhang et al. (2020).

Lemma 17. *Let $\theta \in [0, \frac{\pi}{16}]$ be a given scalar. Let $w_0, \dots w_{T-1}$ be a sequence of vectors such that for all $1 \le t \le T$, $\|w_{t-1} - u\| \le 2\theta$ and $\|w_{t-1}\| \le 1$. Further assume that $\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \le \frac{\theta}{50\cdot 3^4\cdot 2^{33}}$. Let $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_{t-1}$ and $v = \frac{\mathcal{H}_s(\bar{w})}{\|\mathcal{H}_s(\bar{w})\|}$. Then $\theta(v,u) \le \frac{\theta}{2}$.*

Proof. Define the index set $S = \left\{t \in [T] : f_{u,b}(w_{t-1}) \ge \frac{\theta}{5\cdot 3^4\cdot 2^{21}}\right\}$. It is easy to show that $\frac{|S|}{T} \le \frac{1}{10\cdot 2^{12}}$ as otherwise the average of $f_{u,b}(w_{t-1})$ will exceed the assumed upper bound. Therefore, $\frac{|\bar{S}|}{T} \ge 1 - \frac{1}{10\cdot 2^{12}}$. For all $t \in \bar{S}$ we have $f_{u,b}(w_{t-1}) \le \frac{\theta}{50\cdot 3^4\cdot 2^{21}}$; by Lemma 26, we have $\theta(w_{t-1}, u) \le \frac{\theta}{5}$ for these t.

Now consider $\theta(w_{t-1}, u)$ for $t \in S$. As we showed in the proof of Proposition 15, we have $\|w_{t-1} - u\| \le 2\theta$. Since $\|u\| = 1$ and $\|w_{t-1}\| \le 1$, we use the basic fact that $\theta(w_{t-1}, u) \le \pi\|w_{t-1} - u\| < 8\theta$.

Now we translate these bounds on the angles to those of the cosine distance, and obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\cos\theta(w_{t-1},u) &\ge \cos\frac{\theta}{5}\cdot\left(1 - \frac{1}{20\cdot 2^{12}}\right) + \cos(8\theta)\cdot\frac{1}{20\cdot 2^{12}}\\ &\ge \left(1 - \frac{\theta^2}{50}\right)\left(1 - \frac{1}{20\cdot 2^{12}}\right) + \left(1 - \frac{(8\theta)^2}{2}\right)\frac{1}{20\cdot 2^{12}}\\ &\ge 1 - \frac{1}{5}\left(\frac{\theta}{32}\right)^2 \ge \cos\frac{\theta}{32}.\end{aligned}$$

where in the second inequality we use the fact $\cos\theta \ge 1 - \frac{\theta^2}{2}$ for any $\theta \in [0,\pi]$, and in the last inequality we use the fact that $\cos\theta \le 1 - \frac{\theta^2}{5}$.

The above inequality, in combination with Lemma 32 yields the following guarantee for $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_{t-1}$:

$$\cos\theta(\bar{w},u) \ge \frac{1}{T}\sum_{t=1}^{T}\cos\theta(w_{t-1},u) \ge \cos\frac{\theta}{32}.$$

Finally, we use Lemma 33 to show that $\theta(v,u) \le \pi\|v-u\| \le 4\pi\|\bar{w}-u\| \le 16\cdot\theta(\bar{w},u) \le \frac{\theta}{2}$, which concludes the proof. □

Theorem 18 (Restatement of Theorem 12)**.** *Suppose that the adversarial noise rate $\nu \le c_0\epsilon$ for some sufficiently small absolute constant $c_0 > 0$. Consider running the* Refine *algorithm with step size $\alpha = \tilde{\Theta}\left(\theta\cdot\log^{-2}\frac{d}{\delta'\theta}\right)$, bandwidth*

$b = \Theta(\theta)$, convex constraint set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w - w_0\| \leq \theta, \|w\| \leq 1\}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, number of iterations $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta'\theta})$. If the initial iterate w_0 is such that $\|w_0\| = 1$, $\|w_0\|_0 \leq s$, and $\|w_0 - u\| \leq \theta$ for some $\theta \leq \frac{\pi}{16}$, then the output of the REFINE *algorithm, $\tilde{w}$, satisfies $\theta(\tilde{w}, u) \leq \frac{\theta}{2}$ with probability $1 - \delta'$. In addition, the label complexity of* REFINE *is T, and the sample complexity is $O(T/b + T \log \frac{T}{\delta})$.*

Proof. The result of $\theta(\tilde{w}, u) \leq \frac{\theta}{2}$ follows from combining Proposition 15 and Lemma 17. In particular, all the required conditions in Proposition 15 are assumed here. The condition $\|w_{t-1} - u\|$ appearing in Lemma 17 can easily be verified by observing $\|w_{t-1} - w_0\| \leq \theta$ and $\|w_0 - u\| \leq \theta$.

The label complexity bound is exactly T since REFINE runs in T iterations and requests one label per iteration. Since the marginal distribution is assumed to be isotropic log-concave, Lemma 29 shows that $\Pr_{x_t \sim D_X}(x_t \in X_{\hat{w}_{t-1}, b}) \geq c_2 b$ for some absolute constant $c_2 > 0$. Thus, by Chernoff bound, we need to call EX for $O(b^{-1} + \log \frac{T}{\delta})$ times in order to obtain one x_t with probability $1 - \frac{\delta}{2T}$. Thus, by union bound over the T iterations in REFINE, with probability $1 - \frac{\delta}{2}$, the total number of calls to EX is $O(T/b + T \log \frac{T}{\delta})$. □

A.2. Analysis of INITIALIZE

In this subsection, we use $\mathbb{E}[\cdot]$ denote the expectation $\mathbb{E}_{(x,y)\sim D}[\cdot]$ and likewise for $\Pr(\cdot)$.

Lemma 19. *Suppose that $\nu \leq \frac{1}{4}$. Then $\mathbb{E}[y(u \cdot x)] \geq \frac{1}{9 \cdot 2^{17}}$.*

Proof. We have

$$\begin{aligned}
\mathbb{E}[y(u \cdot x)] &= \mathbb{E}[y(u \cdot x) \mid y = \operatorname{sign}(u \cdot x)] \cdot \Pr(y = \operatorname{sign}(u \cdot x)) \\
&\quad + \mathbb{E}[y(u \cdot x) \mid y \neq \operatorname{sign}(u \cdot x)] \cdot \Pr(y \neq \operatorname{sign}(u \cdot x)) \\
&\geq \mathbb{E}[|u \cdot x|] \cdot (1 - \nu) - \mathbb{E}[|u \cdot x|] \cdot \nu \\
&= (1 - 2\nu)\, \mathbb{E}[|u \cdot x|].
\end{aligned}$$

Since $u \cdot x$ is an isotropic log-concave random variable in $\mathbb{R}$, its density function is lower bounded by 2^{-16} when $|u \cdot x| \leq \frac{1}{9}$ in view of Lemma 29. Thus $\mathbb{E}[|u \cdot x|] \geq \frac{1}{9 \cdot 2^{16}}$. On the other side, we assumed $\nu \leq \frac{1}{4}$. Together, we obtain $\mathbb{E}[y(u \cdot x)] \geq \frac{1}{9 \cdot 2^{17}}$. □

Lemma 20. *Let $m = O(\log \frac{1}{\delta})$ and let $(x_1, y_1), \ldots, (x_m, y_m)$ be m i.i.d. samples drawn from D. Then with probability $1 - \delta$,*

$$w_{\mathrm{avg}} \cdot u \geq \frac{1}{9 \cdot 2^{18}},$$

where $w_{\mathrm{avg}} := \frac{1}{m}\sum_{i=1}^m y_i x_i$.

Proof. First, Lemma 29 shows that $u \cdot x$ is isotropic log-concave, and hence $y(u \cdot x)$ is a $(32, 16)$-subexponential random variable by Lemma 34 of Zhang et al. (2020). Therefore, the standard concentration bound implies that there is an absolute constant $K_1 > 0$, such that if $m = O(\log \frac{1}{\delta})$, with probability $1 - \frac{\delta}{2}$,

$$\left| \frac{1}{m}\sum_{i=1}^m y_i(u \cdot x_i) - \mathbb{E}[y(u \cdot x)] \right| \leq K_1 \left(\sqrt{\frac{\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{m} \right) \leq \frac{1}{9 \cdot 2^{18}}.$$

This in allusion to Lemma 19 gives $\frac{1}{m}\sum_{i=1}^m y_i(u \cdot x_i) \geq \frac{1}{9 \cdot 2^{18}}$, namely

$$w_{\mathrm{avg}} \cdot u \geq \frac{1}{9 \cdot 2^{18}},$$

which is the desired lower bound. □

Lemma 21. *Let $\tilde{s} \leq d$ be a positive integer, and set $m = O(\tilde{s} \log \frac{d}{\delta})$. Let $(x_1, y_1), \ldots, (x_m, y_m)$ be m i.i.d. samples drawn from D. Then with probability $1 - \delta$,*

$$w^\sharp \cdot u \geq \frac{1}{9 \cdot 2^{20}}.$$

Proof. Denote $w' = \mathcal{H}_{\tilde{s}}(w_{\text{avg}})$. Using Lemma 17 of Zhang et al. (2020) we know that with probability $1 - \frac{\delta}{2}$, $\|w'\| \leq 2$. From the choice of m and Lemma 20, we have $w_{\text{avg}} \cdot u \geq \frac{1}{9 \cdot 2^{18}}$ with probability $1 - \frac{\delta}{2}$. We hence condition on both events happening.

Now Lemma 16 of Zhang et al. (2020) implies that

$$\left|w' \cdot u - w_{\text{avg}} \cdot u\right| \leq \sqrt{\frac{s}{\tilde{s}}} \left\|w'\right\| \leq 2\sqrt{\frac{s}{\tilde{s}}}.$$

Therefore,

$$w' \cdot u \geq w_{\text{avg}} \cdot u - 2\sqrt{\frac{s}{\tilde{s}}} \geq \frac{1}{9 \cdot 2^{18}} - 2\sqrt{\frac{s}{\tilde{s}}}.$$

Now taking $\tilde{s} = 81 \cdot 2^{40} s$ gives us $w' \cdot u \geq \frac{1}{9 \cdot 2^{19}}$. Finally, by algebra

$$w^\sharp \cdot u = \frac{1}{\|w'\|}(w' \cdot u) \geq \frac{1}{2} \cdot \frac{1}{9 \cdot 2^{19}} = \frac{1}{9 \cdot 2^{20}}.$$

The proof is complete. □

Proposition 22. *Suppose that the adversarial noise rate $\nu \leq c_0 \epsilon$ for some sufficiently small absolute constant $c_0 > 0$. Let $\zeta = \frac{1}{9 \cdot 2^{20}}$. Consider running the* INITIALIZE *algorithm with step size $\alpha = \tilde{\Theta}\left(\log^{-2} \frac{d}{\delta}\right)$, bandwidth $b = \Theta(1)$, convex constraint set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq 1,\ w^\sharp \cdot w \geq \zeta\}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\|w - w_0\|_p^2$, number of iterations $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta})$, where w_0 is an arbitrary point in $\mathcal{K} \cap \{w \in \mathbb{R}^d : \|w\|_1 \leq \sqrt{s}\}$ that can be found in polynomial time. Then with probability $1 - \delta$,*

$$\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\zeta}{20 \cdot 3^4 \cdot 2^{33}}.$$

Proof. We will first verify that the premises of Lemma 14 are satisfied. First of all, it is easy to see that $\|w_0 - u\|_1 \leq 2\sqrt{s}$. Hence we can choose $\rho = 2\sqrt{s}$ in Lemma 14. Next, we have $\|u\| = 1$, which together with Lemma 21 implies $u \in \mathcal{K}$ with probability $1 - \frac{\delta}{2}$. Last, for all $w \in \mathcal{K}$, by triangle inequality $\|w - u\| \leq \|w\| + \|u\| \leq 2$. Hence we can choose $r = 2$ in Lemma 14.

Now with $\rho = \sqrt{s}$ and $r = 2$, Lemma 14 indicates that with probability $1 - \frac{\delta}{2}$,

$$C \cdot \frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq (b+2)\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{\log(1/\delta)}{T} + C\sqrt{\frac{c_0 c_1}{c_2}}\right) + \frac{s \log d}{\alpha T} + \alpha \cdot \log^2 \frac{Td}{b\delta}.$$

We need each term on the right-hand side is upper bounded by $\frac{C \cdot \zeta}{20 \cdot 3^5 \cdot 2^{33}}$. First, we choose $b = \Theta(1)$. Then for the first term, it suffices to choose $T \geq \log(1/\delta)$ and set c_0 to be a sufficiently small absolute constant (this is possible since C, c_1, and c_2 are all fixed absolute constants). The second and the last terms require $\alpha T \geq \Omega(s \log d)$ and $\alpha \cdot \log^2 \frac{Td}{b\delta} = O(1)$ respectively. The latter implies $\alpha = O(\log^{-2} \frac{Td}{b\delta})$, thus combining it with the former we need

$$\frac{s \log d}{T} \leq \cdot \log^{-2} \frac{Td}{b\delta}.$$

This can be satisfied if we choose $T = \tilde{O}(s \log d \cdot \log^2 \frac{d}{\delta})$, which results in $\alpha = \tilde{\Theta}(\log^{-2} \frac{d}{\delta})$. □

Lemma 23. *Set $b = \frac{1}{81 \cdot 2^{22}}$ and let $\zeta = \frac{1}{9 \cdot 2^{20}}$. Let $w_0, \dots w_{T-1}$ be a sequence of vectors such that for all $1 \leq t \leq T$, $\|w_{t-1}\| \leq 1$. Further assume that $\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\zeta}{20 \cdot 3^4 \cdot 2^{33}}$. Let $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_{t-1}$ and $v = \frac{\mathcal{H}_s(\bar{w})}{\|\mathcal{H}_s(\bar{w})\|}$. Then $\theta(v, u) \leq \frac{\pi}{8}$.*

Proof. In light of Lemma 21, we know that $u \in \mathcal{K}$. Also, our choices of b and ζ implies that $b \leq \frac{\zeta}{36}$. Define $S = \{1 \leq t \leq T : f_{u,b}(w_{t-1}) \geq \frac{\zeta}{3^4 \cdot 2^{21}}\}$. Then by the second part of Lemma 26, for all $t \in \bar{S}$, we have $\theta(w_{t-1}, u) < \zeta$. For all $t \in S$, we have a trivial estimate of $\theta(w_{t-1}, u) \in [0, \pi]$.

Next we bound the size of S. Using the condition $\frac{1}{T}\sum_{t=1}^{T} f_{u,b}(w_{t-1}) \leq \frac{\zeta}{20\cdot 3^4\cdot 2^{33}}$, it is possible to show that $\frac{|S|}{T} \leq \frac{1}{20\cdot 2^{12}}$ and thus $\frac{|\bar{S}|}{T} \geq 1 - \frac{1}{20\cdot 2^{12}}$.

Now we translate these bounds on the angles to those of the cosine distance, and obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\cos\theta(w_{t-1},u) &\geq \cos\zeta\cdot\left(1-\frac{1}{20\cdot 2^{12}}\right)+(\cos\pi)\cdot\frac{1}{20\cdot 2^{12}}\\ &\geq 1-\frac{1}{5}\left(\frac{\pi}{128}\right)^2 \geq \cos\frac{\pi}{128}.\end{aligned}$$

where in the last inequality we use the fact that $\cos\theta \leq 1-\frac{\theta^2}{5}$.

The above inequality, in combination with Lemma 32 yields the following guarantee for $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_{t-1}$:

$$\cos\theta(\bar{w},u) \geq \frac{1}{T}\sum_{t=1}^{T}\cos\theta(w_{t-1},u) \geq \cos\frac{\pi}{128}.$$

Finally, we use Lemma 33 to show that $\theta(v,u) \leq \pi\left\|v-u\right\| \leq 4\pi\left\|\bar{w}-u\right\| \leq 16\cdot\theta(\bar{w},u) \leq \frac{\pi}{8}$, which concludes the proof. □

Theorem 24 (Restatement of Theorem 11)**.** *Suppose that the adversarial noise rate $\nu \leq c_0\epsilon$ for some sufficiently small absolute constant $c_0 > 0$. Let $\zeta = \frac{1}{9\cdot 2^{20}}$. Consider running the* INITIALIZE *algorithm with step size $\alpha = \tilde{\Theta}\left(\log^{-2}\frac{d}{\delta}\right)$, bandwidth $b = \Theta(1)$, convex constraint set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq 1,\ w^\sharp\cdot w \geq \zeta\}$, regularizer $\Phi(w) = \frac{1}{2(p-1)}\left\|w-w_0\right\|_p^2$, number of iterations $T = \tilde{O}(s\log d\cdot\log^2\frac{d}{\delta})$, where w_0 is an arbitrary point in $\mathcal{K}\cap\{w\in\mathbb{R}^d : \|w\|_1 \leq \sqrt{s}\}$ that can be found in polynomial time. Then with probability $1-\delta$, the output of* INITIALIZE*, v_0, is such that $\theta(v_0,u) \leq \frac{\pi}{8}$.*

Proof. This is an immediate result by combining Proposition 22 and Lemma 23. □

B. The Structure of $f_{u,b}(w)$

Our definition of the potential function $f_{u,b}(w)$ slightly differs from that of Zhang et al. (2020): in this work, the expectation is taken over D conditioned on $\{x\in\mathbb{R}^d : 0 < w\cdot x \leq b\}$ while in Zhang et al. (2020) it is conditioned on $\{x\in\mathbb{R}^d : -b \leq w\cdot x \leq b\}$. It can be seen that our function value is always less than that of Zhang et al. (2020). However, we note that the difference in sampling region does not lead to significant difference in the structure of the potential function. In particular, we are still able to show that under certain conditions, $f_{u,b}(w)$ serves as an upper bound of $\theta(w,u)$ – a crucial observation made in Zhang et al. (2020).

Lemma 25. *Let w and u be two unit vectors. Suppose $b \in \left[0, \frac{\pi}{72}\right]$. We have*

1. *If $\theta(w,u) \in [36b, \frac{\pi}{2}]$, then $f_{u,b}(w) \geq \frac{\theta(w,u)}{3^4\cdot 2^{21}}$.*
2. *If $\theta(w,u) \in [\frac{\pi}{2}, \pi-36b]$, then $f_{u,b}(w) \geq \frac{\pi-\theta(w,u)}{3^4\cdot 2^{21}}$.*

Proof. The proof of the first part follows closely from [Lemma 22, Part 1] of Zhang et al. (2020). In particular, for the region

$$R_1 := \left\{x\in\mathbb{R}^d : w\cdot x\in[0,b],\ u\cdot x \in \left[-\frac{\sin\theta(w,u)}{18}, -\frac{\sin\theta(w,u)}{36}\right]\right\},$$

their analysis shows that

$$\Pr_{x\sim D_X}(x\in R_1) \geq \frac{b}{9\cdot 2^{18}}. \tag{13}$$

To see why it completes the proof of the first part, observe that

$$\begin{aligned}\mathbb{E}_{x\sim D_X}\left[\left|u\cdot x\right|\cdot \mathbf{1}_{\{0\le w\cdot x\le b\}}\cdot \mathbf{1}_{\{u\cdot x<0\}}\right] &\ge \mathbb{E}_{x\sim D_X}\left[\left|u\cdot x\right|\cdot \mathbf{1}_{\{x\in R_1\}}\right]\\ &\ge \frac{\sin\theta(w,u)}{36}\cdot \mathbb{E}_{x\sim D_X}\left[\mathbf{1}_{\{x\in R_1\}}\right]\\ &\ge \frac{\theta(w,u)}{72}\cdot \Pr_{x\sim D_X}\left(x\in R_1\right)\ge \frac{\theta(w,u)\cdot b}{3^4\cdot 2^{21}},\end{aligned}$$

where the first inequality uses the fact that R_1 is a subset of both sets $\left\{x\in\mathbb{R}^d : w\cdot x\in[0,b]\right\}$ and $\left\{x\in\mathbb{R}^d : u\cdot x<0\right\}$; the second inequality uses the fact that for all x in R_1, $|u\cdot x|\ge \frac{\sin\theta(w,u)}{36}$; the third inequality uses the elementary fact that $\sin\phi\ge\frac{\phi}{2}$ for any angle $\phi\in[0,\frac{\pi}{2}]$.

As $\Pr_{x\sim D_X}\left(w\cdot x\in[0,b]\right)\le b$ by Lemma 29, we have

$$f_{u,b}(w)=\frac{\mathbb{E}_{x\sim D_X}\left[\left|u\cdot x\right|\cdot \mathbf{1}_{\{0\le w\cdot x\le b\}}\cdot \mathbf{1}_{\{u\cdot x<0\}}\right]}{\Pr_{x\sim D_X}\left(w\cdot x\in[0,b]\right)}\ge \frac{\theta(w,u)\cdot b}{3^4\cdot 2^{21}}\cdot\frac{1}{b}=\frac{\theta(w,u)}{3^4\cdot 2^{21}}.$$

This completes the proof of the first part.

For the second part, we will define $\phi=\pi-\theta(w,u)$ and consider

$$R_2:=\left\{x\in\mathbb{R}^d : w\cdot x\in[0,b],\ u\cdot x\in\left[-\frac{\sin\phi}{18},-\frac{\sin\phi}{36}\right]\right\}.$$

Similar to the region R_1, we have $\Pr_{x\sim D_X}(x\in R_2)\ge\frac{b}{9\cdot 2^{18}}$. Hence, using the same induction with the proof of first part, we have $f_{u,b}(w)\ge\frac{\phi}{3^4\cdot 2^{21}}=\frac{\pi-\theta(w,u)}{3^4\cdot 2^{21}}$. □

The following lemma connects the potential function $f_{u,b}(w)$ to the angle $\theta(w,u)$.

Lemma 26. *Let w and u be two unit vectors. We have the following:*

1. *Suppose $\theta\in[0,\frac{\pi}{2}]$ is given. Set $b\le\frac{\theta}{5\cdot 36}$. If $f_{u,b}(w)\le\frac{\theta}{5\cdot 3^4\cdot 2^{21}}$, then $\theta(w,u)\le\frac{\theta}{5}$.*
2. *Let $w^\sharp\in\mathbb{R}^d$ be a unit vector with $w^\sharp\cdot u\ge\zeta$ for some $\zeta\in(0,1)$. Set $b\le\frac{\zeta}{36}$. If w is in $\mathcal{K}:=\{w\in\mathbb{R}^d : \|w\|\le 1,\ w\cdot w^\sharp\ge\zeta\}$ and $f_{u,b}(w)<\frac{\zeta}{3^4\cdot 2^{21}}$, then $\theta(w,u)<\zeta$.*

Proof. The first part was already set out in Claim 10 of Zhang et al. (2020). For the second part, first, we show that it is impossible for $\theta(w,u)\ge\frac{\pi}{2}$. Assume for contradiction that this holds. By Lemma 27, for all w in $\mathcal{K}$, we have that $\theta(w,u)\le\pi-\zeta$. By the choice of b, we know that $36b\le\zeta$, hence $\theta(w,u)\le\pi-36b$. Now using the second part of Lemma 25, we have

$$f_{u,b}(w)\ge\frac{\pi-\theta(w,u)}{3^4\cdot 2^{21}}\ge\frac{\zeta}{3^4\cdot 2^{21}},\tag{14}$$

which contradicts with the premise that $f_{u,b}(w)<\frac{\zeta}{3^4\cdot 2^{21}}$.

Therefore, $\theta(w,u)\in[0,\frac{\pi}{2}]$. We now conduct a case analysis. If $\theta(w,u)\le 36b$, then by the definition of b, we automatically have $\theta(w,u)<\zeta$. Otherwise, $\theta(w,u)\in[36b,\frac{\pi}{2}]$. In this case, the first part of Lemma 25 implies

$$f_{u,b}(w)\ge\frac{\theta(w,u)}{3^4\cdot 2^{21}}.$$

This inequality, in conjunction with the condition that $f_{u,b}(w)<\frac{\zeta}{3^4\cdot 2^{21}}$, implies that $\theta(w,u)\le\zeta$. In summary, in both cases, we have $\theta(w,u)\le\zeta$. This completes the proof. □

Lemma 27 (Lemma 19 of Zhang et al. (2020))**.** *Let $w^\sharp\in\mathbb{R}^d$ be a unit vector, and $\zeta\in(0,1)$. For any two vectors w and v in the set $\mathcal{K}=\{w\in\mathbb{R}^d : \|w\|\le 1,\ w\cdot w^\sharp\ge\zeta\}$, it holds that $\theta(w,v)\le\pi-\zeta$.*

C. Useful Lemmas

We collect a few useful results that are frequently invoked in our analysis.

Lemma 28 (Lemma 3.4 of Awasthi et al. (2017))**.** *There is an absolute constant $c_1 > 0$ such that the following holds. Let D_X be an isotropic log-concave distribution. Fix $w \in \mathbb{R}^d$. For all v with $\|v - w\| \le r$, $\mathbb{E}_{x \sim D_{X|\hat{w},b}}[(v \cdot x)^2] \le c_1(b^2 + r^2)$.*

Lemma 29 (Lovász & Vempala (2007))**.** *There exists an absolute constants $c_2, c_3 > 0$ such that the following holds. Let D_X be an isotropic log-concave distribution over $\mathbb{R}^d$.*

1. *Given any unit vector w, $w \cdot x$ is isotropic log-concave if x is drawn from D_X.*
2. *Given any unit vector $w \in \mathbb{R}^d$, $c_2 b \le \Pr_{x \sim D_X}(w \cdot x \in [0, b]) \le b$.*
3. *If $d = 1$ or $d = 2$, for all $x \in \mathbb{R}^d$ with $\|x\| \le \frac{1}{9}$, the density function $p(x) \ge 2^{-16}$.*

Lemma 30 (Lemma 16 of Shen & Zhang (2021))**.** *There exists an absolute constant $c_3 > 0$ such that the following holds for all isotropic log-concave distributions D_X. Let S be a set of i.i.d. instances drawn from $D_{X|\hat{w},b}$. Then*

$$\Pr_{S \sim D^n_{X|\hat{w},b}}\left(\max_{x \in S} \|x\|_\infty \ge c_3 \log \frac{|S|\, d}{b\delta}\right) \le \delta.$$

Lemma 31 (Lemma 36 of Zhang et al. (2020))**.** *Suppose $\{Z_t\}_{t=1}^T$ is sequence of random variables adapted to filtration $\{\mathcal{F}_t\}_{t=1}^T$. Denote by $\Pr_{t-1}(\cdot)$ and $\mathbb{E}_{t-1}[\cdot]$ the probability and expectation conditioned on $\mathcal{F}_{t-1}$, respectively. For every Z_t, suppose that $\Pr_{t-1}(|Z_t| > a) \le C \exp\left(-\frac{a}{\sigma}\right)$ for some absolute constant $C \ge 1$. Then, with probability $1 - \delta$,*

$$\left|\sum_{t=1}^T Z_t - \mathbb{E}_{t-1}[Z_t]\right| \le 16\sigma(\ln C + 1)\left(\sqrt{2T \ln \frac{2}{\delta}} + \ln \frac{2}{\delta}\right).$$

Lemma 32 (Lemma 24 of Zhang et al. (2020))**.** *Suppose we have a sequence of unit vectors $w_0, \dots, w_{T-1}$. Let $\bar{w} = \frac{1}{T}\sum_{t=1}^T w_{t-1}$ be their average. Suppose $\frac{1}{T}\sum_{t=1}^T \cos\theta(w_{t-1}, u) \ge 0$. Then, $\cos\theta(\bar{w}, u) \ge \frac{1}{T}\sum_{t=1}^T \cos\theta(w_{t-1}, u)$.*

Lemma 33. *Let w and v be two vectors in $\mathbb{R}^d$. The following holds:*

1. *If v is a unit vector, then $\|\hat{w} - v\| \le 2\|w - v\|$.*
2. *If v is s-sparse, then $\left\|\mathcal{H}_s(w) - v\right\| \le 2\|w - v\|$.*
3. *If v is a unit vector, then $\theta(w, v) \le \pi\|w - v\|$; if w is a unit vector as well, then we further have $\|w - v\| \le \theta(w, v)$.*

Proof. The first two parts are known expansion error of ℓ_2-normalization and hard thresholding, respectively. The proof can be found in, e.g. Shen & Li (2018) (which also presents a sharp bound for the second part). The last part can be derived by fundamental algebra. □