

A Missing Proofs Related to Strategic Behavior

The following statement characterizes action x as a function of z and the prediction rule, where x is the observed attributes vector by the algorithm.

Lemma 3 (Strategic Action). *Consider the following utility structure for agents, where $\|v_t\| = 1$. Each agent receives a value of 1 if classified as positive and 0 otherwise, and pays a cost of c per unit of movement (manipulation). The agent's utility is defined as the value received minus the cost incurred. Under the prediction rule defined above:*

1. If $z_t \cdot v_t < 0$, the agent does not move and is classified negative.
2. If $0 \leq z_t \cdot v_t < 1/c$, the agent moves in the direction of v_t to a point where $x_t \cdot v_t = 1/c$, and is classified positive.
3. If $1/c \leq z_t \cdot v_t$, the agent does not move and is classified positive.

Proof. Using the prediction rule in [Algorithm 2](#), an agent with observed vector x is classified as positive iff $v_t \cdot x \geq 1/c$. We analyze the utility-maximizing behavior of agents under the given utility structure in [Section 2](#). The utility of an agent is defined as:

$$\text{Utility}(x) = \begin{cases} 1 - c\|x - z_t\| & \text{if } v_t \cdot x \geq \frac{1}{c}, \\ -c\|x_t - z_t\| & \text{otherwise.} \end{cases}$$

As implied from Observation 1 in [Ahmadi et al. \[2021\]](#), the agent with attributes vector z_t only manipulates iff by movement cost at most c can be classified as positive. We now analyze the three cases: (i) If $z_t \cdot v_t < 0$, any movement incurs negative utility, so the agent does not move. (ii) If $0 \leq z_t \cdot v_t < 1/c$, the agent maximizes utility by moving to $x \cdot v_t = 1/c$ in direction v_t , i.e.,

$$x = z_t + (1/c - z_t \cdot v_t) \cdot v_t.$$

This is the minimum manipulation cost to be classified as positive. (iii) Finally, if $z_t \cdot v_t \geq 1/c$, the agent already achieves maximum utility without moving. \square

Corollary 4. *The prediction rule $\mathbb{1}[x \cdot v \geq 1/c]$ classifies all z such that $z \cdot v \geq 0$ as +1 and all z such that $z \cdot v < 0$ as -1.*

In a nonstrategic setting, the difference in the error rate of two prediction rules can be related to the angles of their normal vectors.

Lemma 5 ([Yan and Zhang \[2017\]](#)). *In the nonstrategic setting, for any unit-sized vector v_1 and v_2 and prediction rules $\mathbb{1}[z \cdot v_i \geq 0]$,*

$$|\text{err}(v_1) - \text{err}(v_2)| \leq \mathbb{P}[\mathbb{1}[z \cdot v_1 \geq 0] \neq \mathbb{1}[z \cdot v_2 \geq 0]] = \frac{\theta(v_1, v_2)}{\pi}.$$

We extend the above argument to the strategic setting.

Lemma 6. *Consider the prediction rule of [Algorithm 2](#) for unit-sized vectors v_1 and v_2 . $|\text{err}(h_{v_1}) - \text{err}(h_{v_2})| \leq \frac{\theta(v_1, v_2)}{\pi}$.*

Proof. The prediction rule for v_i is 1 iff $v_i \cdot x - 1/c \geq 0$. Consider an arbitrary example z with its utility maximizing action x with respect to prediction rule v_i . By [Corollary 4](#), the set of z classified as positive or negative under the rule $v_i \cdot z \geq 0$ in the non-strategic setting exactly matches the set classified as positive or negative under $v_i \cdot x \geq 1/c$ in the strategic setting, respectively. Invoking [Lemma 5](#) for the nonstrategic setting gives the difference in error in the strategic setting as well as $\theta(v_1, v_2)/\pi$. \square

B Proof of [Theorem 2](#)

We provide the proof of our main result, [Theorem 2](#) in this section.

As shown in [Corollary 4](#), the prediction rule identified by v_i in the strategic setting using adjusted threshold $1/c$, predicts exactly the same as v_i in the nonstrategic setting. Therefore, intuitively, the

key remaining part to extend the guarantees is to show that the *update rule* in our setting (strategic agents and arbitrary-sized vectors \mathbf{z} from the uniform distribution).

The update rule of our algorithm only uses negatively classified observed examples. Using [Lemma 3](#) for these examples $\mathbf{z} = \mathbf{x}$.

Corollary 7. *For any negatively classified example at time t , i.e., $\mathbf{x}_t \cdot \mathbf{v}_t < 1/c$, the observed attributes are the same as the true attributes, i.e., $\mathbf{x}_t = \mathbf{z}_t$ and $\mathbf{z}_t \cdot \mathbf{v}_t < 0$. Furthermore, such examples are distributed uniformly according to D subject to $\mathbf{z}_t \cdot \mathbf{v}_t < 0$ and satisfy the ν bounded separability.*

The above lemma implies that the strategic aspect does not affect the prediction step.

Next, we show that using our update rule induces a coupling between examples inside the unit sphere and those on its surface.

Lemma 8. *Consider the update rule $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + 2(\mathbf{w}_t \cdot \hat{\mathbf{x}}_t)\hat{\mathbf{x}}_t$, where \mathbf{x}_t is sampled from distribution D conditioned on $-b \leq \mathbf{w}_t \cdot \hat{\mathbf{x}}_t \leq \frac{-b}{2}$. The distribution over updated vectors \mathbf{w}_{t+1} matches that of the update rule $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + 2(\mathbf{w}_t \cdot \mathbf{x}_t)\mathbf{x}_t$, where \mathbf{x}_t is sampled uniformly from the $\|\mathbf{x}\| = 1$ conditioned on $-b \leq \mathbf{w}_t \cdot \mathbf{x}_t \leq \frac{-b}{2}$.*

Proof. We construct a coupling between the samples used in the two update rules. Fix a unit vector $\hat{\mathbf{x}}$ with $\|\hat{\mathbf{x}}\| = 1$, and consider the set of points \mathbf{x} such that $\mathbf{x}/\|\mathbf{x}\| = \hat{\mathbf{x}}$. Under the distribution D (uniform over the unit ball), the density over such \mathbf{x} induces a uniform distribution over directions $\hat{\mathbf{x}}$ on the unit sphere.

Since the update rule using $\hat{\mathbf{x}}$ depends only on the direction of \mathbf{x} (and not its norm), and since every \mathbf{x} whose unit-sized scaled version is $\hat{\mathbf{x}}$ yields the same update vector, the two update distributions are equivalent. That is, sampling from D and normalizing before updating yields the same distribution over \mathbf{w}_{t+1} as sampling directly from the unit sphere and updating without normalization. \square

By [Corollary 7](#) and [Lemma 8](#), we can disregard the effects of strategic behavior and the arbitrary norms of the examples, and directly apply the guarantees established in the non-strategic setting:

Theorem 9 (Adapted from [\[Yan and Zhang, 2017\]](#), Theorem 3). *Suppose [Algorithm 1](#) has inputs satisfying the ν -bounded inseparability condition with respect to halfspace \mathbf{u} , initial halfspace \mathbf{v}_0 such that $\theta(\mathbf{v}_0, \mathbf{u}) \leq \pi/2$, target error ε , confidence δ , sample schedule $\{m_k\}$ where $m_k = \Theta(d(\ln d + \ln \frac{k}{\delta}))$, and band width $\{b_k\}$ where $b_k = \Theta(\frac{2^{-k}}{\sqrt{d \ln(km_k/\delta)}})$. Additionally, $\nu \leq \Theta(\frac{\varepsilon}{\ln d + \ln \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}})$. Then with probability at least $1 - \delta$:*

1. The output halfspace \mathbf{v} outputs a prediction different from \mathbf{u} with probability at most ε .
2. The number of label queries is $O(d \ln \frac{1}{\varepsilon} \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon}))$.
3. The number of unlabeled examples drawn is $O(d \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon})^2 \cdot \frac{1}{\varepsilon} \ln \frac{1}{\varepsilon})$.
4. The algorithm runs in time $O(d^2 \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon})^2 \cdot \frac{1}{\varepsilon} \ln \frac{1}{\varepsilon})$.

Lemma 10 (Adapted from [\[Yan and Zhang, 2017\]](#), Lemma 3). *Suppose [Algorithm 2](#) has inputs satisfying the ν -bounded inseparability condition with respect to halfspace \mathbf{u} , initial halfspace \mathbf{w}_0 and angle upper bound $\theta \in (0, \pi/2]$ such that $\theta(\mathbf{w}_0, \mathbf{u}) \leq \theta$, confidence δ , number of iterations $m = \Theta(d(\ln d + \ln \frac{1}{\delta}))$, and band width $b = \Theta(\frac{\theta}{\sqrt{d \ln(m/\delta)}})$. Additionally, $\nu = O(\frac{\theta}{\ln(m/\delta)})$. Then with probability at least $1 - \delta$:*

1. The output \mathbf{w}_m is such that $\theta(\mathbf{w}_m, \mathbf{u}) \leq \frac{\theta}{2}$.
2. The number of label queries is $O(d \cdot (\ln d + \ln \frac{1}{\delta}))$.
3. The number of unlabeled examples drawn is $O(d \cdot (\ln d + \ln \frac{1}{\delta})^2 \cdot \frac{1}{\theta})$.
4. The algorithm runs in time $O(d^2 \cdot (\ln d + \ln \frac{1}{\delta})^2 \cdot \frac{1}{\theta})$.

The only remaining part is proving the mistake bound of the algorithm.

Lemma 11. *Suppose [Algorithm 1](#) has inputs satisfying the ν -bounded inseparability condition with respect to halfspace \mathbf{u} , initial halfspace \mathbf{v}_0 such that $\theta(\mathbf{v}_0, \mathbf{u}) \leq \pi/2$, target error ε ,*

confidence δ , sample schedule $\{m_k\}$ where $m_k = \Theta(d(\ln d + \ln \frac{k}{\delta}))$, and band width $\{b_k\}$ where $b_k = \Theta(\frac{2^{-k}}{\sqrt{d \ln(km_k/\delta)}})$. Additionally, $\nu \leq \Theta(\frac{\varepsilon}{\ln d + \ln \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}})$. Then with probability at least $1 - \delta$, The additional number of mistakes that the algorithm makes compared to \mathbf{u} is $O(d \cdot \ln \frac{1}{\varepsilon} \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon})^2)$.

Proof. Similar to the proof of Theorem 3 in [Yan and Zhang, 2017], we define, for each iteration k , a corresponding event with high individual success probability. Specifically, by Lemma 10, for every k , there exists an event E_k such that $\Pr(E_k) \geq 1 - \frac{\delta}{k(k+1)}$. Moreover, on event E_k , items 1 through 4 of Lemma 10 hold for the input $\mathbf{w}_0 = \mathbf{v}_k$.

The excess error of \mathbf{w}_t relative to \mathbf{u} is θ_t/π , where θ_t is the angle between \mathbf{w}_t and \mathbf{u} . Consider Algorithm 2 with initial halfspace \mathbf{w}_0 and angle bound $\theta \in [0, \frac{\pi}{2}]$ such that $\theta(\mathbf{w}_0, \mathbf{u}) \leq \theta$. Let the number of iterations be $m = \Theta(\frac{d}{(1-2\eta)^2} (\ln \frac{d}{(1-2\eta)^2} + \ln \frac{k}{\delta}))$. Then, with probability at least $1 - \frac{\delta}{k(k+1)}$, the output halfspace \mathbf{w}_m satisfies $\theta(\mathbf{w}_m, \mathbf{u}) \leq \frac{\theta}{2}$.

By items 2 and 3 in Lemma 10, with probability at least $1 - \frac{\delta}{k(k+1)}$, the total number of examples seen during epoch k is: $O(d \cdot (\ln d + \ln \frac{k}{\delta})^2 \cdot \frac{1}{\theta_t})$. Since each example differs in classification from \mathbf{u} with probability $\frac{\theta}{\pi}$, the number of additional misclassified examples compared to \mathbf{u} is $O(d \cdot (\ln d + \ln \frac{k}{\delta})^2)$. The number of total time epochs, $k_0 = \lceil \log_2(1/\varepsilon) \rceil$ and $k \leq k_0$. The total number of epochs is $k_0 = \lceil \log_2(1/\varepsilon) \rceil$, and we have $k \leq k_0$. Therefore, by a union bound over all epochs, with probability at least $1 - \delta$, the total number of additional mistakes compared to \mathbf{u} is $O(d \cdot \ln \frac{1}{\varepsilon} \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon})^2)$.

583 □

Theorem 2. Suppose Algorithm 1 has inputs satisfying the ν -bounded inseparability condition with respect to halfspace \mathbf{u} , initial halfspace \mathbf{v}_0 such that $\theta(\mathbf{v}_0, \mathbf{u}) \leq \pi/2$, target error ε , confidence δ , sample schedule $\{m_k\}$ where $m_k = \Theta(d(\ln d + \ln \frac{k}{\delta}))$, and band width $\{b_k\}$ where $b_k = \Theta(\frac{2^{-k}}{\sqrt{d \ln(km_k/\delta)}})$. Additionally, $\nu \leq \Theta(\frac{\varepsilon}{\ln d + \ln \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}})$. Then with probability at least $1 - \delta$:

1. The output halfspace \mathbf{v} outputs a prediction different from \mathbf{u} with probability at most ε .
2. The number of label queries is $O(d \ln \frac{1}{\varepsilon} \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon}))$.
3. The number of unlabeled examples drawn is $O(d \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon})^2 \cdot \frac{1}{\varepsilon} \ln \frac{1}{\varepsilon})$.
4. The additional number of mistakes that the algorithm makes compared to \mathbf{u} is $O(d \cdot \ln \frac{1}{\varepsilon} \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon})^2)$.
5. The algorithm runs in time $O(d^2 \cdot (\ln d + \ln \frac{1}{\delta} + \ln \ln \frac{1}{\varepsilon})^2 \cdot \frac{1}{\varepsilon} \ln \frac{1}{\varepsilon})$.

Proof. Corollary 7, Lemma 8 and Theorem 9 imply items 1, 2, 3, and 5. Lemma 11 implies item 4. □

596 C Initialization Step

Algorithm 1 assumes that the initial vector \mathbf{v}_0 forms an angle less than $\pi/2$ with the true separator \mathbf{u} . To obtain such a vector, we adopt an approach similar to that of [Yan and Zhang, 2017]. Intuitively, Algorithm 3 begins with two opposite vectors and, after a small number of trials, selects the one with lower classification error. This initialization step incurs only a constant overhead in terms of label, time, and mistake complexities.

Algorithm 3: Master Algorithm in Adversarial Noise Setting, (adapted from [Yan and Zhang \[2017\]](#))

Input: Labeling oracle \mathcal{O} , confidence δ .

Output: A halfspace \mathbf{v} such that $\theta(\hat{\mathbf{v}}, \mathbf{u}) \leq \pi/4$.

$\mathbf{v}_0 \leftarrow (1, 0, \dots, 0)$.

$\mathbf{v}_+ \leftarrow \text{Active-Strategic-Perceptron}(\mathcal{O}, \mathbf{v}_0, \frac{1}{16}, \frac{\delta}{3}, \{m_k\}, \{b_k\}, \{c\})$.

602 $\mathbf{v}_- \leftarrow \text{Active-Strategic-Perceptron}(\mathcal{O}, -\mathbf{v}_0, \frac{1}{16}, \frac{\delta}{3}, \{m_k\}, \{b_k\}, \{c\})$.

Define region $R := \{\mathbf{x} : \text{sgn}(\mathbf{v}_+ \cdot \mathbf{x} - 1/c) \neq \text{sgn}(\mathbf{v}_- \cdot \mathbf{x} - 1/c)\}$.

$S \leftarrow \text{Observe } 8 \ln \frac{6}{\delta} \text{ examples in } R \text{ and query their labels.}$

if $\text{err}_S(h_{\mathbf{v}_+}) \leq \text{err}_S(h_{\mathbf{v}_-})$ **then**

| **return** \mathbf{v}_+ .

| **else**

| **return** \mathbf{v}_- .

603 By [Corollary 7](#) and [Theorem 2](#), we can effectively disregard the strategic behavior of agents and
 604 apply Theorem 12 from [Yan and Zhang \[2017\]](#), which implies the following.

605 **Theorem 12.** Suppose [Algorithm 3](#) has inputs labeling oracle \mathcal{O} that satisfies ν -inseparability
 606 condition with respect to \mathbf{u} , confidence δ , and sample schedule $\{m_k\}$ where

$$m_k = \Theta \left(d \left(\ln d + \ln \frac{k}{\delta} \right) \right),$$

607 and band width $\{b_k\}$ where

$$b_k = \tilde{\Theta} \left(\frac{2^{-k}}{\sqrt{d}} \right).$$

608 Then, with probability at least $1 - \delta$, the output $\hat{\mathbf{v}}$ is such that $\theta(\hat{\mathbf{v}}, \mathbf{u}) \leq \frac{\pi}{4}$. Furthermore:

- 609 1. the total number of label queries to oracle \mathcal{O} is at most $\tilde{O}(d)$;
- 610 2. the total number of unlabeled examples drawn is $\tilde{O}(d)$;
- 611 3. the total number of additional mistakes compared to \mathbf{u} is $\tilde{O}(d)$;
- 612 4. the algorithm runs in time $\tilde{O}(d^2)$.