A. Additional discussion on related work

The work of (Hardt et al., 2016) is most relevant to our work—they introduced a Stackelberg game framework to model the interaction between the learner and the test data. Our model can be viewed as a generalization of (Hardt et al., 2016) by allowing *heterogeneous* preferences over classification outcomes. (Hardt et al., 2016) assume a special class of *separably cost functions*, and prove that the optimal classifier is always a *threshold* classifier. Essentially, the assumption of separable cost functions reduces the feature space to a low dimension, which is also why the strategic VC dimension in this case is at most 2 as we proved. Despite this clean characterization, it appears a strong and somewhat unrealistic requirement. For example, one consequence of separable cost functions is that for *any* two features $\boldsymbol{x}, \boldsymbol{z}$, the manipulation cost from either $\boldsymbol{x}$ to $\boldsymbol{z}$ or from $\boldsymbol{z}$ to $\boldsymbol{x}$ must be 0.[4] This appears unrealistic in reality. For example, a high-school student with true average math grade 80 and true average literature grade 95 is likely to incur cost if she/he wants to appear as 95 for math and 80 for literature, and vice versa. This is because different students are good at different aspects. Our model imposes less assumptions on the cost functions. For example, in our study of strategic linear classification, the cost functions are induced by arbitrary semi-norms.

(Brückner & Scheffer, 2011) is one of the first to consider the Stackelberg game formulation of strategic classification, motivated by spam filtering; however they do not study generalization bounds. (Zhang & Conitzer, 2021) provide the sample complexity result for strategic PAC-learning under the homogeneous preference setting and in particular study the case under the incentive-compatibility constraints, i.e., subject to no data points will misreport features. These two works all assume the positive labels are always and equally preferred. There has also been work on understanding the social implications of strategically robust classification (Akyol et al., 2016; Milli et al., 2019; Hu et al., 2019b); these works show that improving the learner's performance may lead to increased social burden and unfairness. (Dong et al., 2018; Chen et al., 2020) extend strategic linear classification to an online setting where the input features are not known a-priori, but instead are revealed in an online manner. They both focused on the optimization problem of regret minimization. Our setting however is in the more canonical PAC-learning setup and our objective is to design statistically and computationally efficient learning algorithms. All these aforementioned works, including the present work, consider *gaming* behaviors. A relevant but quite different line of recent works study *strategic improvements* where the manipulation does really change the inherent quality and labels (Kleinberg & Raghavan, 2019; Miller et al., 2019; Ustun et al., 2019; Bechavod et al., 2020; Shavit et al., 2020). The question there is mainly to design incentive mechanisms to encourage agents' efforts or improvements.

Finally, going beyond classification, strategic behaviors in machine learning has received significant recent attentions, including in regression problems (Perote & Perote-Peña, 2004; Dekel et al., 2010; Chen et al., 2018), distinguising distributions (Zhang et al., 2019a;b), and learning for pricing (Amin et al., 2013; Mohri & Munoz, 2015; Vanunts & Drutsa, 2019). These are similar in spirit to us, but study a completely different set of problems using different techniques. Their results are not comparable to ours.

B. Omitted Proofs from Section 3

B.1. Proof of Theorem 1

Proof. Let $\mathcal{Y} = \{+1, -1\}$. Define another binary hypothesis class $\tilde{H} = \{\kappa_c(h) : h \in H\}$, where $\kappa_c : (\mathcal{X} \to \mathcal{Y}) \to (\mathcal{X} \times R \to \mathcal{Y})$ is a mapping such that $\kappa_c(h)(\boldsymbol{x}, r) = h(\Delta_c(\boldsymbol{x}, r; h)), \forall (\boldsymbol{x}, r) \in \mathcal{X} \times R$. Note that the input of classifier $\kappa_c(h)$ consists of both the feature vector $\boldsymbol{x}$ and the preference r. By the definition of SVC, we have VC$(\tilde{H})$ =SVC$(\mathcal{H}, R, c) = d$.

Given any distribution $\mathcal{D}$, cost function c, and $h \in \mathcal{H}$, the strategic 0-1 loss of h is $L_c(h, \mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},y,r)\sim\mathcal{D}}\Big[\mathbb{I}\big[\kappa_c(h)(\boldsymbol{x}, r) \neq y\big]\Big] = L(\kappa_c(h), \mathcal{D})$, where $L(\tilde{h}, \mathcal{D})$ is the standard expected risk of the newly defined $\tilde{h} \in \tilde{\mathcal{H}}$ under the distribution $\mathcal{D}$ in the non-strategic setting. Therefore, studying the PAC sample complexity upper bound for $\mathcal{H}$ under the strategic setting $\langle R, c\rangle$ is equivalent to studying the sample complexity for $\tilde{H}$ in the non-strategic setting. The latter problem can be addressed by employing the standard PAC learning analysis. From the Fundamental Theorem of Statistical Learning (Theorem 6.8 in (Shalev-Shwartz & Ben-David, 2014)), we know $\tilde{H}$ is agnostic PAC learnable with sample complexity $O(\epsilon^{-2}(\text{VC}(\tilde{H}) + \log\frac{1}{\delta}))$, meaning that there exists a constant C such that for any $(\delta, \epsilon) \in (0, 1)^2$ and any distribution $\mathcal{D}$

[4] A cost function $c(\boldsymbol{z}; \boldsymbol{x})$ is separable if there exists two functions $c_1, c_2 : \mathcal{X} \to \mathbb{R}$ such that $c(\boldsymbol{z}; \boldsymbol{x}) = \max\{c_2(\boldsymbol{z}) - c_1(\boldsymbol{x}), 0\}$. Since $c(\boldsymbol{x}; \boldsymbol{x}) = 0$, we have $c_2(\boldsymbol{x}) \leq c_1(\boldsymbol{x})$ for any $\boldsymbol{x}$. Therefore, $c_2(\boldsymbol{x})+c_2(\boldsymbol{z})-c_1(\boldsymbol{x})-c_1(\boldsymbol{z}) \leq 0$. Consequently, either $c_2(\boldsymbol{x})-c_1(\boldsymbol{z}) \leq 0$ or $c_2(\boldsymbol{z}) - c_1(\boldsymbol{x}) \leq 0$, yielding either $c(\boldsymbol{z}; \boldsymbol{x}) = 0$ or $c(\boldsymbol{x}; \boldsymbol{z}) = 0$.

for $(\boldsymbol{x}, y, r)$, as long as $n \geq C \cdot \epsilon^{-2}(\mathrm{VC}(\tilde{H}) + \log \frac{1}{\delta})$, with at least probability $1-\delta$, we have

$$L(\tilde{h}^*, \mathcal{D}) - \inf_{\tilde{h} \in \tilde{H}} L(\tilde{h}, \mathcal{D}) \leq \epsilon,$$

where $\tilde{h}^*$ is the solution of ERM with n i.i.d. samples from $\mathcal{D}$ as input. Let h^* be the solution of the corresponding SERM conditioned on the same n i.i.d. samples from $\mathcal{D}$. By the definition of $\tilde{H}$ and L_c, we have $L_c(h^*, \mathcal{D}) = L(\tilde{h}^*, \mathcal{D})$, and $\inf_{h \in \mathcal{H}} L_c(h, \mathcal{D}) = \inf_{\tilde{h} \in \tilde{H}} L(\tilde{h}, \mathcal{D})$. Therefore, with probability $1-\delta$, we have

$$L_c(h^*, \mathcal{D}) - \inf_{h \in \mathcal{H}} L_c(h, \mathcal{D}) \leq \epsilon,$$

which implies $\mathrm{STRAC}\langle \mathcal{H}, R, c\rangle$ is agnostic PAC learnable with sample complexity $O(\epsilon^{-2}[d + \log(\frac{1}{\delta})])$ by the SERM.

□

B.2. Proof of Proposition 1

Proof. The adversarial VC-dimension defined in (Cullina et al., 2018) relies on an auxiliary definition of *corrupted classifier* $\tilde{h} = \kappa_R(h)$ of any classifier h for the standard non-adversarial setting such that $\tilde{h}(\boldsymbol{x}) = h(\boldsymbol{x})$ if all the points in $N(\boldsymbol{x})$ have the same label as $\boldsymbol{x}$ and otherwise, $\tilde{h}(\boldsymbol{x}) = \perp$. Recall that $N(\boldsymbol{x}) = \{\boldsymbol{z} \in \mathcal{X} : (\boldsymbol{z}; \boldsymbol{x}) \in \mathcal{B}\} = \{\boldsymbol{z} \in \mathcal{X} : c(\boldsymbol{z}; \boldsymbol{x}) \leq r\}$ denotes the set of all possible adversarial features $\boldsymbol{x}$ can move to. Given this auxiliary definition, the adversarial VC-dimension is defined as $\mathrm{AVC}(\mathcal{H}, \mathcal{B}) = \sup\{n : \sigma_n(\mathcal{F}, \mathcal{B}) = 2^n\}$, where

$$\sigma_n(\mathcal{F}, \mathcal{B}) = \max_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \{+1,-1\}^n} |\{(f(\boldsymbol{x}_1, y_1; h), \ldots, f(\boldsymbol{x}_n, y_n; h)) : h \in \mathcal{H}\}| \tag{7}$$

is the shattering coefficient, and $f(\boldsymbol{x}_i, y_i) = \mathbb{I}(\tilde{h}(\boldsymbol{x}_i) \neq y_i)$ is the *loss function* of the corrupted classifier $\tilde{h} = \kappa_R(h)$.

Since $\mathcal{B}$ and c are r-consistent, we have $\mathcal{B} = \{(\boldsymbol{z}; \boldsymbol{x}) : c(\boldsymbol{z}; \boldsymbol{x}) \leq r\}$. Let $R = \{+r, -r\}$. We now prove the proposition by arguing

$$\sup\{n \in \mathbb{N} : \sigma_n(\mathcal{H}, R, c) = 2^n\} = \sup\{n : \sigma_n(\mathcal{F}, \mathcal{B}) = 2^n\}. \tag{8}$$

1. If $\sup\{n \in \mathbb{N} : \sigma_n(\mathcal{H}, R, c) = 2^n\} = n$, by Definition 1, there exists $(\boldsymbol{x}'_i, r'_i) \in \mathcal{X} \times R, i = 1, \cdots, n$ such that $|\{(h(\Delta_c(\boldsymbol{x}'_1, r'_1; h)), \cdots, h(\Delta_c(\boldsymbol{x}'_n, r'_n; h)) : h \in \mathcal{H}\}| = 2^n$. Since Definition 1 does not rely on the true labels of $\boldsymbol{x}'_i$, we may let the true labels of $\boldsymbol{x}'_i$ be $y'_i = -r'_i/r$ for any i. In this case, each $\boldsymbol{x}'_i$'s strategic preference is against its true label, which corresponds to the loss function f in Equation (7) for the adversarial setting. Therefore, taking $(\boldsymbol{x}_i, y_i) = (\boldsymbol{x}'_i, y'_i)$ in Equation (7) gives $\sigma_n(\mathcal{F}, \mathcal{B}) = 2^n$. This implies $\sup\{n \in \mathbb{N} : \sigma_n(\mathcal{H}, R, c) = 2^n\} \leq \sup\{n : \sigma_n(\mathcal{F}, \mathcal{B}) = 2^n\}$.
2. Conversely, if $\sup\{n : \sigma_n(\mathcal{F}, \mathcal{B}) = 2^n\} = n$, from Equation (7), there exists $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times R, i = 1, \cdots, n$ such that $|\{(f(\boldsymbol{x}_1, y_1), \ldots, f(\boldsymbol{x}_n, y_n)) : f \in \mathcal{F}\}| = 2^n$. Similarly, taking $r_i = -ry_i \in R$ gives $\sigma_n(\mathcal{H}, R, c) = 2^n$, which implies $\sup\{n \in \mathbb{N} : \sigma_n(\mathcal{H}, R, c) = 2^n\} \geq \sup\{n : \sigma_n(\mathcal{F}, \mathcal{B}) = 2^n\}$.

Therefore, we have $\mathrm{AVC}(\mathcal{H}, \mathcal{B})$ =$\mathrm{SVC}(\mathcal{H}, \{+r, -r\}, c)$ for any r-consistent pair $(\mathcal{B}, c)$. □

B.3. Proof of Corollary 1.1

Proof. Since $\{+r, -r\} \subseteq \mathcal{B}$, we have $\sigma_n(\mathcal{H}, R, c) \geq \sigma_n(\mathcal{H}, \{+r, -r\}, c)$ by Definition 1. As a result, $\mathrm{SVC}(\mathcal{H}, R, c) \geq$$\mathrm{SVC}(\mathcal{H}, \{+r, -r\}, c)$. Then by applying Proposition 1 we have
$\mathrm{SVC}(\mathcal{H}, R, c) \geq$$\mathrm{SVC}(\mathcal{H}, \{+r, -r\}, c)$ =$\mathrm{AVC}(\mathcal{H}, \mathcal{B})$. □

B.4. Proof of Proposition 2

Given any positive integer n, let $[n]$ denotes $\{1, 2, \cdots, n\}$, and $\mathcal{S}$ be the power set of $[n]$, i.e., the set that contains all the subsets of $[n]$. Let $\mathcal{X} = [n] \cup \mathcal{S}$ be the sample space of size $n + 2^n$, and the hypothesis class $\mathcal{H}$ is the set of all the point classifiers with points from $\mathcal{S}$, i.e., $\mathcal{H} = \{h_s : s \in \mathcal{S}\}$, where point classifier h_s only classifies the point $s \in \mathcal{S}$ as positive.

The cost function $c(z;x)$ is a metric defined as follows. Since metric is symmetric, i.e., $c(z;x) = c(x;z)$, we will use the notation $c(x,z)$ instead throughout this proof.

$$c(x,z) = \begin{cases} x, & \text{if} \quad x \in [n], z \in \mathcal{S}, x \in z \\ x+1, & \text{if} \quad x \in [n], z \in \mathcal{S}, x \notin z \\ c(z,x), & \text{if} \quad x \in \mathcal{S}, z \in [n] \\ x+z, & \text{if} \quad x, z \in [n], x \neq z \\ 1, & \text{if} \quad x, z \in \mathcal{S}, x \neq z \\ 0, & \text{if} \quad x = z, \end{cases} \tag{9}$$

and R is set to be $[-n,-1] \cup [1,n]$.

First, we verify that $c(\cdot,\cdot)$ is indeed a metric. Given the Definition (9), it is easy to see that $c(x,z) = 0$ iff $x = z$, and $c(x,z) = c(z,x), \forall x, z \in \mathcal{X}$. It remains to check the triangle inequality, i.e., for any $x, y, z \in \mathcal{X}$, $c(x,y) + c(y,z) \geq c(x,z)$. Consider the case when x, y, z are different elements in $\mathcal{X}$. By enumerating all the possibility that whether each x, y, z is in $[n]$ or $\mathcal{S}$, it suffices to discuss the following $8 (= 2^3)$ cases:

1. if $x, y, z \in [n]$, $c(x,y) + c(y,z) = x + y + y + z > x + z = c(x,z)$.
2. if $x, y, z \in \mathcal{S}$, $c(x,y) + c(y,z) = 2 > 1 = c(x,z)$.
3. if $x, z \in [n], y \in \mathcal{S}$, then $c(x,y) \geq x, c(y,z) \geq z. \Longrightarrow c(x,y) + c(y,z) \geq x + z = c(x,z)$.
4. if $x, y \in [n], z \in \mathcal{S}$, we need to show that $c(x,y) \geq c(x,z) - c(y,z)$. Conditioned on the relationship between x, y and set z, the maximum value of $c(x,z) - c(y,z)$ is $x - y + 1$ when $y \in z, x \notin z$. Therefore, $c(x,y) = x + y \geq x - y + 1 \geq c(x,z) - c(y,z)$.
5. if $x, z \in \mathcal{S}, y \in [n]$, then $c(x,y) + c(y,z) \geq y + y > 1 \geq c(x,z)$.
6. if $x, y \in \mathcal{S}, z \in [n]$, then the maximum value for $c(x,z) - c(y,z)$ is $z + 1 - z = 1$ when $z \notin x, z \in y$. Therefore, $c(x,y) \geq 1 \geq c(x,z) - c(y,z)$.
7. if $x \in \mathcal{S}, y, z \in [n]$, it is equivalent to case 4.
8. if $y, z \in \mathcal{S}, x \in [n]$, it is equivalent to case 6.

Next, we show VC$(\mathcal{H}) = 1$, AVC$(\mathcal{H}, \mathcal{B}_c(r)) = 1$, and SVC$(\mathcal{H}, R, c) \geq n$. Observe that VC$(\mathcal{H}) = 1$ follows easily since no point classifier $h_s \in \mathcal{H}$ can generate the label pattern $(+1, +1)$ for any pair of distinct data points.

Next we prove AVC$(\mathcal{H}, \mathcal{B}_c(r)) = 1$. We first show AVC$(\mathcal{H}, \mathcal{B}_c(r)) \leq 1$ by arguing that under binary nearness relation $\mathcal{B}_c(r) = \{(z;x) : c(z,x) \leq r\}$ with $r \geq 1$, any two elements x_1, x_2 in $\mathcal{X}$ cannot be shattered by $\mathcal{H}$.

1. If at least one of r_1, r_2 equals $-r$, e.g., $r_1 = -r$, we show that x_1 can never be classified as $+1$ by contradiction. Suppose some $h_s \in \mathcal{H}$ classifies $(x_1, -r)$ as $+1$: if $x_1 \neq s$, since $r_1 = -r < 0$, x_1 will not manipulate its feature and be classified as -1; if $x_1 = s$, x_1 can move to any $z \in \mathcal{S}$ with cost $1 \leq r$, and will also be classified as -1. Therefore, (x_1, x_2) can not be shattered.
2. If $r_1 = r_2 = r$, consider the following two cases:
   (a) If at least one of x_1, x_2 belongs to $\mathcal{S}$, e.g., $x_1 \in \mathcal{S}$, then x_1 can move to any $s \in \mathcal{S}$ as $c(x_1, s) = 1 \leq r$ for any $s \in \mathcal{S}$. Therefore x_1 can never be classified as -1 by any point classifier in $\mathcal{H}$.
   (b) if $x_1, x_2 \in [n]$, we may w.l.o.g. assume $x_1 < x_2$, i.e., $x_1 + 1 \leq x_2$. Observe that when $r < x_1$, any $h_s \in \mathcal{H}$ will classify x_1 as -1 because $c(x_1, s) = x_1 > r, \forall s \in \mathcal{S}$; when $r \geq x_1 + 1$, any $h_s \in \mathcal{H}$ will classify x_1 as +1 because $c(x_1, s) = x_1 + 1 \leq r, \forall s \in \mathcal{S}$. Therefore, in order to shatter (x_1, x_2), r must lie in the interval $[x_1, x_1 + 1) \cap [x_2, x_2 + 1) = \emptyset$, which draws the contradiction.

To see that $\text{AVC}(\mathcal{H}, \mathcal{B}_c(r)) \geq 1$, for any $x \in [n]$ with $r > 0$, it can be classified as either $+1$ or -1 as long as $r \in [x, x+1)$. We thus have $\text{AVC}(\mathcal{H}, c) = 1$.

Finally, we prove that $\text{SVC}(\mathcal{H}, R, c) = n$. Consider the subset $[n] \subset \mathcal{X}$ of size n, with each element i equipped with a strategic preference $r_i = i$. For any label pattern $\mathcal{L} \in \{+1, -1\}^n$, let $s_{\mathcal{L}} = \{i \in [n] : \mathcal{L}_i = +1\}$ be an element in $\mathcal{S}$. We claim that $h_{s_{\mathcal{L}}} \in \mathcal{H}$ gives exactly the label pattern $\mathcal{L}$ on $[n]$. To see this, consider any $i \in [n]$:

1. If $i \in s_{\mathcal{L}}$, i will move to $s_{\mathcal{L}} \in \mathcal{S}$ and be classified as $+1$, as the cost $c(i, s_{\mathcal{L}}) = i \leq r_i = i$.
2. If $i \notin s_{\mathcal{L}}$, i will not move to $s_{\mathcal{L}} \in \mathcal{S}$ and be classified as -1, as the cost $c(i, s_{\mathcal{L}}) = i + 1 > r_i = i$.

Therefore, any label pattern $\mathcal{L} \in \{+1, -1\}^n$ can be achieved by some $h_{s_{\mathcal{L}}} \in \mathcal{H}$. This implies $\text{SVC}(\mathcal{H}, R, c) \geq n$. On the other hand, it's easy to see $\mathcal{H}$ cannot shatter $n+1$ points, because any subset of size $n+1$ must contain an element s_0 in $\mathcal{S}$, and no matter what strategic preference s_0 has, it will either be classified as $+1$ by all $h_s \in \mathcal{H}$, or be classified as $+1$ by only one classifier in $\mathcal{H}$, i.e., h_{s_0}. Either case renders the shattering for $n+1$ points impossible.

B.5. Proof of Proposition 3

Proof. Define the adversarial region for an adversary $(\boldsymbol{x}, r)$ as $N(\boldsymbol{x}, r) = \{\boldsymbol{z} \in \mathcal{X} : c_2(\boldsymbol{z}) \leq c_1(\boldsymbol{x}) + |r|\} \supseteq \{\boldsymbol{x}\}$. Since staying with the same feature has no cost, this implies $c(\boldsymbol{x}; \boldsymbol{x}) = 0$ or equivalently $c_2(\boldsymbol{x}) \leq c_1(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$. Then, the best response function for $(\boldsymbol{x}, r)$ can be characterized by

1. if $h(\boldsymbol{x}) = \text{sgn}(r)$, then $h(\Delta(\boldsymbol{x}, r; h)) = \text{sgn}(r)$;
2. if $h(\boldsymbol{x}) = -\text{sgn}(r)$, then

$$h(\Delta(\boldsymbol{x}, r; h)) = \begin{cases} -\text{sgn}(r), & \forall \boldsymbol{z} \in N(\boldsymbol{x}, r) : h(\boldsymbol{z}) = -\text{sgn}(r) \\ \text{sgn}(r), & \exists \boldsymbol{z} \in N(\boldsymbol{x}, r) : h(\boldsymbol{z}) = \text{sgn}(r) \end{cases} \tag{10}$$

Suppose there are three points $\{(x_i, r_i)\}_{i=1}^3$ that can be shattered by $\mathcal{H}$. Let $b_i = c_1(\boldsymbol{x}_i) + r_i$ and w.l.o.g. let $b_1 \leq b_2 \leq b_3$. From $b_1 \leq b_2 \leq b_3$, we have $N(\boldsymbol{x}_1, r_1) \subseteq N(\boldsymbol{x}_2, r_2) \subseteq N(\boldsymbol{x}_3, r_3)$.

By Pigeonhole principle, there must exists two elements in $\{r_1, r_2, r_3\}$ which have the same sign. Suppose these two elements are r_1, r_2 and consider the following two cases:

1. $r_1 > 0, r_2 > 0$. From Equation 10, for any $h \in \mathcal{H}$, $h(\Delta(\boldsymbol{x}_2, r_2; h)) = -1$ means $h(\boldsymbol{z}) = -1, \forall \boldsymbol{z} \in N(\boldsymbol{x}_2, r_2)$. Note that $N(\boldsymbol{x}_1, r_1) \subseteq N(\boldsymbol{x}_2, r_2)$, we also have $h(\boldsymbol{z}) = -1, \forall \boldsymbol{z} \in N(\boldsymbol{x}_1, r_1)$. As a result, $h(\Delta(\boldsymbol{x}_1, r_1; h)) = -1$, meaning the sign pattern $\{+, -\}$ cannot be achieved by any $h \in \mathcal{H}$ for $\{(\boldsymbol{x}_1, r_1), (\boldsymbol{x}_2, r_2)\}$.
2. $r_1 < 0, r_2 < 0$. From Equation 10, for any $h \in \mathcal{H}$, $h(\Delta(\boldsymbol{x}_2, r_2; h)) = 1$ means $h(\boldsymbol{z}) = 1, \forall \boldsymbol{z} \in N(\boldsymbol{x}_2, r_2)$. Similarly, from $N(\boldsymbol{x}_1, r_1) \subseteq N(\boldsymbol{x}_2, r_2)$ we conclude $h(\boldsymbol{z}) = 1, \forall \boldsymbol{z} \in N(\boldsymbol{x}_1, r_1)$ and $h(\Delta(\boldsymbol{x}_1, r_1; h)) = 1$, meaning the sign pattern $\{-, +\}$ cannot be achieved by any $h \in \mathcal{H}$ for $\{(\boldsymbol{x}_1, r_1), (\boldsymbol{x}_2, r_2)\}$.

Therefore, $\{(x_i, r_i)\}_{i=1}^3$ cannot be shattered by $\mathcal{H}$, which implies $\text{SVC}(\mathcal{H}, R, c) \leq 2$.

□

C. Proof of Theorem 2

Proof. Let $\mathcal{X} = \mathbb{R}^2$, and consider the linear hypothesis class on $\mathcal{X}$: $\mathcal{H} = \{h = \text{sgn}(\boldsymbol{w} \cdot \boldsymbol{x} + b) : (\boldsymbol{w}, b) \in \mathbb{R}^3, \boldsymbol{x} \in \mathcal{X}\}$. We show that for any $n \in \mathbb{Z}^+$ and $R = \{+1\}$, there exist n points $\{\boldsymbol{x}_i\}_{i=1}^n \in \mathcal{X}^n$ and corresponding cost functions $\{c_i\}_{i=1}^n$, such that the n'th shattering coefficients $\sigma_n(\mathcal{H}, R, \{c_i\}_{i=1}^n) = 2^n$ (see Definition 1 for σ_n). Note that the cost function is instance-wise. For convenience, here we equivalently think of it as each data point i has a different cost function c_i.

Let $\boldsymbol{x}_i = (0, 0), \forall i \in [n]$ be the set of data points. The main challenge of the proof is a very careful construction of the cost function for each data point. To do so, we first pick a set of 2^n different points $S = \{s_j\}_{j=1}^{2^n}$ lying on the *unit circle*, i.e.,

$S \subset \{(x, y) : x^2 + y^2 = 1\}$. The number 2^n is *not* arbitrarily chosen — indeed, we will map each point s_j to one of the 2^n subsets of $[n]$ in a *bijective* manner so that each s_j corresponds to a unique subset of $[n]$. What are these 2^n different points will not matter to our construction neither it matters which point is mapped to which subset so long as it is a bijection. Moreover, let $\bar{S} = \{(-x, -y) : (x, y) \in S\}$ be the set that is origin-symmetric to S such that $\bar{S} \cap S = \emptyset$. $\bar{S}$ is chosen to "symmetrize" our construction to obtain a norm and needs not to have any interpretation. For any $\boldsymbol{x}_i$, we now define its cost function c_i through the following steps :

1. Let $S_i = \{s \subseteq [n] : i \in s\} \subset S$ contains all the 2^{n-1} subsets of $[n]$ that include the element i.
2. Let $\bar{S}_i = \{(-x, -y) : (x, y) \in S_i\} \subseteq \bar{S}$ be the set that is origin-symmetric to S_i.
3. Let G_i be the convex, origin-symmetric polygon with the vertex set being $S_i \cup \bar{S}_i$.
4. The cost function of $\boldsymbol{x}_i$ is defined as $c_i(\boldsymbol{z}; \boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{z}\|_{G_i}$, where $\| \cdot \|_{G_i} = \inf\{\epsilon \in \mathbb{R}_{\geq 0} : x \in \epsilon G_i\}$ is a norm derived from polygon G_i (note the origin-symmetry of $S_i \cup \bar{S}_i$ and thus G_i).

Next we show that for any label pattern $\mathcal{L} \in \{+1, -1\}^n$, there exists some linear classifier $h \in \mathcal{H}_2$ such that $(h(\Delta_{c_1}(\boldsymbol{x}_1, +1; h)), \cdots, h(\Delta_{c_n}(\boldsymbol{x}_n, +1; h)) = \mathcal{L}$.

With slight abuse of notation, let $s_{\mathcal{L}} = \{i \in [n] : \mathcal{L}_i = +1\} \in S$ be the point in S that corresponds to the set of the indexes of $\mathcal{L}$ with $\mathcal{L}_i = 1$. Let $h_{\mathcal{L}}$ be any *linear classifier* whose decision boundary intersects the unit circle centered at $\boldsymbol{x}_i$ and *strictly* separates $s_{\mathcal{L}}$ from all the other elements in $S \cup \bar{S}$. We will use $h_{\mathcal{L}}$ to denote both the linear classifier and its decision boundary (i.e., a line in $\mathbb{R}^2$) interchangeably. Due to the convexity of G_i, such $h_{\mathcal{L}}$ must exist. We further let $h_{\mathcal{L}}$ give prediction result $+1$ for the half plane that contains $s_{\mathcal{L}}$ and -1 for the other half plane. Figure 3 illustrates the geometry of this example.

We now argue that $h_{\mathcal{L}}$ induces the given label pattern $\mathcal{L}$ for instances $\{(\boldsymbol{x}_i, 1, c_i)\}_{i=1}^n$. To see this, we examine $h_{\mathcal{L}}(\Delta_{c_i}(\boldsymbol{x}_i, 1; h))$ for each i:

1. If $i \in s_{\mathcal{L}}$, then $s_{\mathcal{L}} \in S_i$ and $\boldsymbol{x}_i$ can move to $s_{\mathcal{L}}$ with cost $c_i(s_{\mathcal{L}}; \boldsymbol{x}_i) < 1$. This is because G_i is convex and there exist a point $\boldsymbol{x}_i'$ on $h_{\mathcal{L}}$ such that $c_i(\boldsymbol{x}_i'; \boldsymbol{x}_i) < c_i(s_{\mathcal{L}}; x_i) = 1 = r_i$ (e.g., choose $\boldsymbol{x}_i'$ as the intersection point of the segment $[\boldsymbol{x}_i, s_{\mathcal{L}}]$ and $h_{\mathcal{L}}$). Therefore, $h_{\mathcal{L}}$ will classify $\boldsymbol{x}_i$ as positive. This case is shown in the left panel of Figure 3.
2. If $i \notin s_{\mathcal{L}}$, then $s_{\mathcal{L}} \notin S_i$ and G_i does not intersect $h_{\mathcal{L}}$. In this case, $h_{\mathcal{L}}(\boldsymbol{x}) = -1$, and moving across $h_{\mathcal{L}}$ always induces a cost strictly larger than 1. Therefore, the best response for $\boldsymbol{x}_i$ is to stay put and $h_{\mathcal{L}}$ will classify $\boldsymbol{x}_i$ as negative. This case is shown in the right panel of Figure 3.

Now we have shown that the n'th shattering coefficients $\sigma_n(\mathcal{H}, \{+1, -1\}, \{c_i\}_{i=1}^n) = 2^n$. Since n can take any integer, we conclude the strategic VC-dimension in this case is $+\infty$.

□

D. Proof of Theorem 3

The following lemma from advanced linear algebra is widely known and will be useful for our analysis.

Lemma 1. *For any seminorm $l : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, and the cost function $c(\boldsymbol{z}; \boldsymbol{x}) = l(\boldsymbol{z} - \boldsymbol{x})$ induced by l, the minimum manipulation cost for $\boldsymbol{x}$ to move to the hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ is given by the following:*

$$\min_{\boldsymbol{x}'}\{c(\boldsymbol{x}'; \boldsymbol{x}) : \boldsymbol{w} \cdot \boldsymbol{x}' + b = 0\} = \frac{|\boldsymbol{w} \cdot \boldsymbol{x} + b|}{l^*(\boldsymbol{w})}$$

where $l^*(\boldsymbol{w}) = \sup_{\boldsymbol{z} \in B}\{\boldsymbol{w} \cdot \boldsymbol{z}\} \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$, and $\mathcal{B} = \{\boldsymbol{z} : l(\boldsymbol{z}) \leq 1\}$ is the unit ball induced by l.

The proof is divided into the following two parts. The first part is the more involved one.

Proof of SVC$(\mathcal{H}_d, R, c) \leq d + 1 - \dim(V_l)$**:**

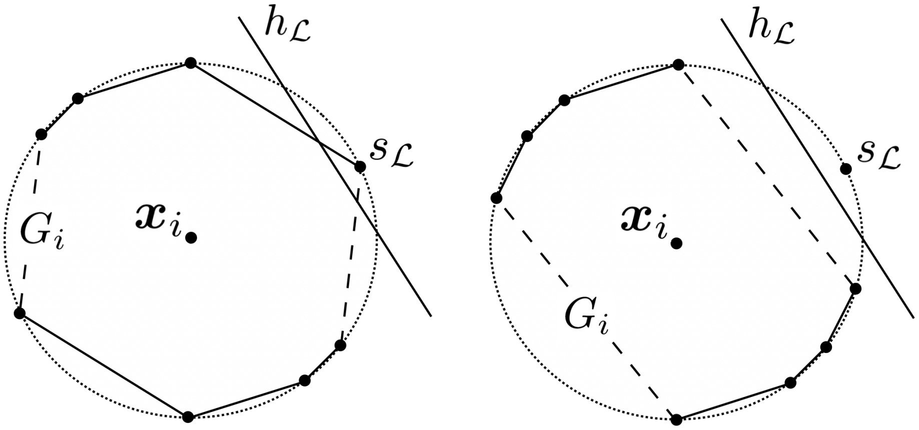

Figure 3. Left: If $i \in s_{\mathcal{L}}$, $h_{\mathcal{L}}$ intersects with G_i, and $\boldsymbol{x}_i$ can manipulate its feature within G_i to cross $h_{\mathcal{L}}$. Right: If $i \notin s_{\mathcal{L}}$, $h_{\mathcal{L}}$ and G_i are disjoint; $\boldsymbol{x}_i$ cannot manipulate its feature within G_i to cross $h_{\mathcal{L}}$. Given any label pattern $\mathcal{L} \in \{+1, -1\}^n$, G_i is the convex, origin-symmetric polygon associated with $\boldsymbol{x}_i$'s cost function. The linear classifier $h_{\mathcal{L}}$ is chosen to separate $s_{\mathcal{L}}$ from all other elements in $\bar{S} \cup S$ and classifies $s_{\mathcal{L}}$ as $+1$. The left/right panel shows the two situations, depending on $i \in s_{\mathcal{L}}$ or $i \notin s_{\mathcal{L}}$.

It suffices to show that for any $n > d + 1 - \dim(V_l)$ and n data points $(\boldsymbol{x}_i, r_i) \in \mathbb{R}^d \times R, \forall i = 1, \cdots, n$, there exists a label pattern $\mathcal{L} \in \{+1, -1\}^n$, such that for any $h \in \mathcal{H}_d$ cannot induce $\mathcal{L}$, i.e.,

$$(h(\Delta_c(\boldsymbol{x}_1, r_1; h), \cdots, h(\Delta_c(\boldsymbol{x}_n, r_n; h))) \neq \mathcal{L}.$$

The first step of our proof derives a succinct characterization about the classification outcome for a set of data points. For any seminorm l, it is known the set $\mathcal{B} = \{x : l(x) \leq 1\}$ is nonempty, closed, convex, and origin-symmetric. Let $l^*(\boldsymbol{w}) = \sup_{\boldsymbol{z} \in B}\{\boldsymbol{w} \cdot \boldsymbol{z}\}$. We have $l^*(\boldsymbol{w}) > 0$ for all $\boldsymbol{w} \neq \boldsymbol{0}$ since $\boldsymbol{0}$ is an interior point of $\mathcal{B}$. According to Lemma 1, for any $\boldsymbol{x} \in \mathbb{R}^d$ and any linear classifier $h = (\boldsymbol{w}, b) \in \mathcal{H}_d$, the minimum manipulation cost for $\boldsymbol{x}$ to move to the decision boundary of h is $|\boldsymbol{w} \cdot \boldsymbol{x} + b| / l^*(\boldsymbol{w})$. Note that we may w.l.o.g. restrict to $\boldsymbol{w}$'s such that $l^*(\boldsymbol{w}) = 1$ since the sign function sgn$(\boldsymbol{w} \cdot \boldsymbol{x} + b)$ does not change after re-scaling. For any data point $(\boldsymbol{x}, r) \in \mathcal{X} \times R$ and linear classifier $h \in \mathcal{H}_d$, we define the *signed* manipulation cost to the classification boundary as

$$\delta(h, \boldsymbol{x}) = h(\boldsymbol{x}) \cdot \frac{|\boldsymbol{w} \cdot \boldsymbol{x} + b|}{l^*(\boldsymbol{w})} = \boldsymbol{w} \cdot \boldsymbol{x} + b,$$

using the condition $l^*(\boldsymbol{w}) = 1$. We claim that $h(\Delta_c(\boldsymbol{x}, r; h)) = 2\mathbb{I}(\boldsymbol{w} \cdot \boldsymbol{x} + b \geq -r) - 1$. This follows a case analysis:

1. If $r \leq 0$, then $h(\Delta_c(\boldsymbol{x}, r, h)) = 1$ if and only if $h(\boldsymbol{x}) = 1$ and $\boldsymbol{x}$ *cannot* move across the decision boundary of h within cost $|r| = -r$. This implies $h(\Delta_c(\boldsymbol{x}, r; h)) = 2\mathbb{I}(\boldsymbol{w} \cdot \boldsymbol{x} + b \geq -r) - 1$.

2. If $r > 0$, then $h(\Delta_c(\boldsymbol{x}, r, h)) = -1$ if and only if $h(\boldsymbol{x}) = -1$ and $\boldsymbol{x}$ cannot move across the decision boundary of h within cost r. In this case, $h(\Delta_c(\boldsymbol{x}, r; h)) = -(2\mathbb{I}(-(\boldsymbol{w} \cdot \boldsymbol{x} + b) > r) - 1) = 2\mathbb{I}(\boldsymbol{w} \cdot \boldsymbol{x} + b \geq -r) - 1$. Note that the first inequality holds strictly because we assume h always gives $+1$ for those $\boldsymbol{x}$ on the decision boundary.

For a set of samples $(\boldsymbol{X}, \boldsymbol{r})$ where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n), \boldsymbol{r} = (r_1, \cdots, r_n)$, define the set of all possible vectors (over the choice of linear classifiers $(\boldsymbol{w}, b) \in \mathcal{H}_d$) of signed manipulation costs as

$$\mathcal{D}(\mathcal{H}_d, \boldsymbol{X}) = \{(\boldsymbol{w} \cdot \boldsymbol{x}_1 + b, \cdots, \boldsymbol{w} \cdot \boldsymbol{x}_n + b) : h \in \mathcal{H}_d\}, \tag{11}$$

there is a $h \in \mathcal{H}_d$ that achieves a label pattern $\mathcal{L}$ on $(\boldsymbol{X}, \boldsymbol{r})$ if and only if there exist an element in $D(\mathcal{H}_d, \boldsymbol{x}) + \boldsymbol{r}$ with the corresponding sign pattern $\mathcal{L}$.

Recall that a linear classifier is described by $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$. The second step of our proof rules out "trivial" linear classifiers under strategic behaviors, and consequently allows us to work with only linear classifiers in a linear space of smaller dimension. Let $\mathcal{B} = \{\boldsymbol{x} : l(\boldsymbol{x}) \leq 1\}$ and V_l be the largest linear space contained in $\mathcal{B}$. We argue that it suffice to consider only linear classifiers $(\boldsymbol{w}, b)$ with $\boldsymbol{w} \perp V_l$. This is because for any $\boldsymbol{w}$ that is not orthogonal to the subspace V_l, we can find $\bar{\boldsymbol{z}} \in V_l$ such that $c(\bar{\boldsymbol{z}}; \boldsymbol{x}) = 0$ and $\boldsymbol{w} \cdot \bar{\boldsymbol{z}} \to \infty$ since V_l is a linear subspace. This means any data point can induce its preferred label $\text{sgn}(r)$ with 0 cost, by moving to $\bar{\boldsymbol{z}}$ if $\text{sgn}(r) = +$ and $-\bar{\boldsymbol{z}}$ otherwise. Any such linear classifier will result in the same label pattern, simply specified by $\text{sgn}(r)$. As a consequence, we only need to focus on linear classifiers $(\boldsymbol{w}, b)$ with $\boldsymbol{w} \perp V_l$. Let $\tilde{H}_d = \{(\boldsymbol{w}, b) : \boldsymbol{w} \perp V_l\}$ denote all such linear classifiers.

Next, we argue that when restricting to the non-trivial class of linear classifiers $\tilde{\mathcal{H}}_d$, the $\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})$ defined in Equation (11) lies in a linear subspace with dimension at most $d + 1 - \dim(V_l)$. Consider the linear mapping $\mathcal{G}_{\boldsymbol{X}} : \tilde{\mathcal{H}}_d \to \mathbb{R}^n$ determined by the data features $\boldsymbol{X}$, defined as

$$\mathcal{G}_{\boldsymbol{X}}(\boldsymbol{w}, b) = (\boldsymbol{w} \cdot \boldsymbol{x}_1 + b, \cdots, \boldsymbol{w} \cdot \boldsymbol{x}_n + b), \quad \forall (\boldsymbol{w}, b) \in \tilde{\mathcal{H}}_d.$$

Since $\boldsymbol{w} \perp V_l$, $\boldsymbol{w}$ is from a linear subspace of $d - \dim(V_l)$. Linear mapping will not increase the dimension of the image space, therefore $\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})$ lies in a space with dimension at most $d + 1 - \dim(V_l)$.

Finally, we prove that there must exist label patterns that cannot be induced by linear classifiers whenever the number of data points $n > d + 1 - \dim(V_l)$. Let $\text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big)$ denote the smallest linear space that contains $\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})$. Since $\text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big)$ has dimension at most $d + 1 - \dim(V_l) < n$ but $\text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big) \subset \mathbb{R}^n$, there must exist a non-zero vector $\bar{\boldsymbol{u}} \in \mathbb{R}^n$ such that: (1) $\bar{\boldsymbol{u}} \neq \boldsymbol{0}$; (2) $\bar{\boldsymbol{u}} \perp \text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big)$ (i.e., $\bar{\boldsymbol{u}} \cdot \boldsymbol{v} = 0, \forall \boldsymbol{v} \in \text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big)$); and (3) $\bar{\boldsymbol{u}} \cdot \boldsymbol{r} \leq 0$ (if $\bar{\boldsymbol{u}} \cdot \boldsymbol{r} \geq 0$, simply takes its negation). Note that this implies $\bar{\boldsymbol{u}} \cdot \boldsymbol{v} \leq 0, \forall \boldsymbol{v} \in \text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big) + \boldsymbol{r}$.

We argue that the sign pattern of the vector $\bar{\boldsymbol{u}}$, denoted as $\text{sgn}(\bar{\boldsymbol{u}})$, and the sign pattern of all negatives ($\mathcal{L} = (-1, \cdots, -1)$) cannot be achieved simultaneously by $\tilde{\mathcal{H}}_d$. Suppose $\text{sgn}(\bar{\boldsymbol{u}})$ can be achieved by $\tilde{\mathcal{H}}_d$, then there must exist $\boldsymbol{v}^1 \in \text{span}(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})) + \boldsymbol{r}$ such that $\text{sgn}(\bar{\boldsymbol{u}}) = \text{sgn}(\boldsymbol{v}^1)$ and $\bar{\boldsymbol{u}} \cdot \boldsymbol{v}^1 \leq 0$. Since $\text{sgn}(\bar{\boldsymbol{u}}) = \text{sgn}(\boldsymbol{v}^1)$ also implies $\bar{\boldsymbol{u}} \cdot \boldsymbol{v}^1 \geq 0$, we thus have $\bar{\boldsymbol{u}} \cdot \boldsymbol{v}^1 = \sum_{j=1} \bar{u}_j v^1_j = 0$. We claim that there must exist j such that $\bar{u}_j > 0$. First of all, we cannot have $\bar{u}_j < 0$ for any j since that implies $v^1_j < 0$ (only strictly less v^1_j's will be assigned -1 pattern due to our tie breaking rule) and consequently, $\bar{\boldsymbol{u}} \cdot \boldsymbol{v}^1 < 0$, a contradiction. Also note that $\bar{\boldsymbol{u}} \neq \boldsymbol{0}$, so there exist $j \in [n]$ such that $\bar{u}_j > 0$.

Utilizing the above property of $\bar{\boldsymbol{u}}$, we show that the sign pattern $\mathcal{L} = (-1, \cdots, -1)$ cannot be achieved by $\tilde{\mathcal{H}}_d$. Suppose, for the sake of contradiction, that this is not true. Then there exists another $\boldsymbol{v}^2 = (v^2_1, \cdots, v^2_n) \in \text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big) + \boldsymbol{r}$ with all its elements being strictly negative. Now consider $\boldsymbol{v} = \boldsymbol{v}^1 - \boldsymbol{v}^2 \in \text{span}(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X}))$, we have $\bar{u} \cdot \boldsymbol{v} = \bar{u} \cdot \boldsymbol{v}^1 - \bar{u} \cdot \boldsymbol{v}^2 = 0 - \bar{u} \cdot \boldsymbol{v}^2 > 0$. Here the inequality holds because $\bar{u}_j \geq 0, v^2_j < 0$ for all j and there exists some j such that $\bar{u}_j > 0$. Therefore, we draw a contradiction to the fact that $\bar{\boldsymbol{u}} \cdot \boldsymbol{v} = 0$ for any $\boldsymbol{v} \in \text{span}(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X}))$.

Now we proved that $\text{sgn}(\bar{\boldsymbol{u}})$ and $\mathcal{L} = (-1, \cdots, -1)$ cannot be achieved simultaneously by non-trivial classifiers $\tilde{\mathcal{H}}_d$, and the only achievable sign pattern for trivial classifiers is $\text{sgn}(\boldsymbol{r})$. Note that $\boldsymbol{r} \in \text{span}\big(\mathcal{D}(\tilde{\mathcal{H}}_d, \boldsymbol{X})\big) + \boldsymbol{r}$, $\text{sgn}(\boldsymbol{r})$ is thus also achievable by $\tilde{\mathcal{H}}_d$. Therefore, the trivial classifiers has no contributions to the shattering coefficient, and we conclude at least one of $\text{sgn}(\bar{\boldsymbol{u}})$ and $\mathcal{L} = (-1, \cdots, -1)$ cannot be achieved by $\mathcal{H}_d$.

Proof of SVC$(\mathcal{H}_d, R, c) \geq d + 1 - \dim(V_l)$:

The second step of the proof shows $\text{SVC}(\mathcal{H}, R, c) \geq d + 1 - \dim(V_l)$ by giving an explicit construction of $(\boldsymbol{X}, \boldsymbol{r})$ that can be shattered by $\mathcal{H}_d$. Let $\boldsymbol{x}_0 = \boldsymbol{0}$, and $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)$ be a basis of the subspace orthogonal to V_l, $(\boldsymbol{x}_{t+1}, \cdots, \boldsymbol{x}_d)$ be a basis of the subspace V_l, where $t = d - \dim(V_l)$.

We claim that the $t + 1 = d + 1 - \dim(V_l)$ data points in $\{0, 1, \cdots, t\}$ can be shattered by $\mathcal{H}_d$. In particular, for any given subset $S \subseteq \{0, 1, \cdots, t\}$, consider the linear system

$$\begin{cases} \boldsymbol{x}_i \cdot \boldsymbol{w}_S + b_S = 1, & \text{if} \quad i \in S \\ \boldsymbol{x}_i \cdot \boldsymbol{w}_S + b_S = -1, & \text{if} \quad i \leq t, \text{ and } i \notin S \\ \boldsymbol{x}_i \cdot \boldsymbol{w}_S = 0, & t + 1 \leq i \leq d. \end{cases}$$

Because $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_d)$ has full rank, the solution $(\boldsymbol{w}_S, b_S)$ must exist. Therefore, the half-plane $h = \boldsymbol{w}_S \cdot \boldsymbol{x} + b_S$ separates S and $\{\boldsymbol{x}_0, \cdots, \boldsymbol{x}_d\}/S$. Now consider the case when each $\boldsymbol{x}_i$ has a strategic preference $r_i \in R$. Since $\boldsymbol{w}_S$ is chosen to be orthogonal to V_l, $\boldsymbol{w}_S \cdot \boldsymbol{x}_i$ is bounded when $\boldsymbol{x}_i \in \{\boldsymbol{z} : c(\boldsymbol{z}; \boldsymbol{x}_i) \leq r_i\}$. Let $\delta_S = \max_{0 \leq i \leq t}\{\sup\{\boldsymbol{w}_S \cdot (\boldsymbol{z} - \boldsymbol{x}_i) : c(\boldsymbol{z}; \boldsymbol{x}_i) \leq r_i\}\}$, and $\delta = \max(1, 2\delta_S)$. Then the data set $\{\delta \boldsymbol{x}_0, \cdots, \delta \boldsymbol{x}_t\}$ can be shattered by $\mathcal{H}_d$ for any given c, R, because the classifier $(\delta \boldsymbol{w}_S, \delta b_S)$ separates the subset S and the other points regardless their strategic responses.

E. Proof of Theorem 4

Proof of Theorem 4. For any data point $(\boldsymbol{x}, y, r)$, let the manipulation cost for the data point be $c(\boldsymbol{z}; \boldsymbol{x}) = l_{\boldsymbol{x}}(\boldsymbol{z} - \boldsymbol{x})$ where $l_{\boldsymbol{x}}$ is any seminorm. Since the instance is separable, there exists a hyperplane $h : \boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ that separates the given n training points $(\boldsymbol{x}_1, y_1, r_1), \cdots, (\boldsymbol{x}_n, y_n, r_n)$ under strategic behaviors. The SERM problem is thus a feasibility problem, which we now formulate. Utilizing Lemma 1 about the signed distance from $\boldsymbol{x}_i$ to hyperplane h under cost function $c(\boldsymbol{z}; \boldsymbol{x}_i) = l_{\boldsymbol{x}_i}(\boldsymbol{z} - \boldsymbol{x}_i)$, we can formulate the SERM problem under the separability assumption. Concretely, we would like to find a hyperplane $h : \boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ such that it satisfies the following for any $(\boldsymbol{x}_i, y_i, r_i)$:

1. If $y_i = 1$ and $r_i \geq 0$, we must have either $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq 0$ or $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq 0$ and $\frac{-(\boldsymbol{w} \cdot \boldsymbol{x} + b)}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} \leq r_i$;
2. If $y_i = 1$ and $r_i \leq 0$, we must have $\frac{\boldsymbol{w} \cdot \boldsymbol{x} + b}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} \geq -r_i$ (this implies $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq 0$);
3. If $y_i = -1$ and $r_i \leq 0$, we must have either $\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq 0$ or $\boldsymbol{w} \cdot \boldsymbol{x}_i + b > 0$ and $\frac{\boldsymbol{w} \cdot \boldsymbol{x} + b}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} < -r_i$;
4. If $y_i = -1$ and $r_i \geq 0$, we must have $\frac{-(\boldsymbol{w} \cdot \boldsymbol{x} + b)}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} > r_i$ (this implies $\boldsymbol{w} \cdot \boldsymbol{x}_i + b < 0$);

Note that we classify any point on the hyperplane as $+1$ as well, which is why the strict inequality for Case 3 and 4. Case 1 can be summarized as $\frac{\boldsymbol{w} \cdot \boldsymbol{x} + b}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} \geq -r_i$. Similarly, Case 3 can be summarized as $\frac{\boldsymbol{w} \cdot \boldsymbol{x} + b}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} < -r_i$. To impose the strict inequality for Case 3 and 4, we may introduce an ϵ slack variable. These observations lead to the following formulation of the SERM problem.

$$\begin{array}{lll} \text{find} & \boldsymbol{w}, b, \epsilon > 0 & \\ \text{subject to} & \frac{\boldsymbol{w} \cdot \boldsymbol{x}_i + b}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} \geq -r_i, & \text{for points } (\boldsymbol{x}_i, y_i, r_i) \text{ with } y_i = 1. \\ & \frac{\boldsymbol{w} \cdot \boldsymbol{x}_i + b}{l^*_{\boldsymbol{x}_i}(\boldsymbol{w})} \leq -r_i - \epsilon, & \text{for points } (\boldsymbol{x}_i, y_i, r_i) \text{ with } y_i = -1. \end{array} \tag{12}$$

We now consider the two settings as described in the theorem statement. We first consider **Situation 1**, i.e., the essentially adversarial case with $\min^- \geq \max^+$ and an instance-invariant cost function induced by the same seminorm l, i.e., $c(\boldsymbol{z}; \boldsymbol{x}) = l(\boldsymbol{x} - \boldsymbol{z})$ for any $\boldsymbol{x}$. In this case, System (12) is equivalent to the following

$$\begin{array}{lll} \text{find} & \boldsymbol{w}, b, \epsilon > 0 & \\ \text{subject to} & \boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq -r_i, & \text{for points } (\boldsymbol{x}_i, y_i, r_i) \text{ with } y_i = 1. \\ & \boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -(r_i + \epsilon), & \text{for points } (\boldsymbol{x}_i, y_i, r_i) \text{ with } y_i = -1. \\ & l^*(\boldsymbol{w}) = 1 & \end{array} \tag{13}$$

This system is unfortunately not a convex feasibility problem. To solve System (13), we consider the following optimization program (OP), which is a relaxation of System (13) by relaxing the non-convex constraint $l^*(\boldsymbol{w}) = 1$ to the convex constraint $l^*(\boldsymbol{w}) \leq 1$.

$$\begin{array}{lll} \text{maximize} & \epsilon & \\ \text{subject to} & \boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq -r_i, & \text{for points } (\boldsymbol{x}_i, r_i) \text{ with label 1.} \\ & \boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -r_i - \epsilon, & \text{for points } (\boldsymbol{x}_i, r_i) \text{ with label -1.} \\ & l^*(\boldsymbol{w}) \leq 1 & \end{array} \tag{14}$$

Note that OP (14) is a convex program because the objective and constraints are either linear or convex. Therefore, OP (14) can be efficiently solved in polynomial time.[5] Note that this relaxation is not tight in general as we will show later that solving System (13) is NP-hard in general.

Our main insight is that under the assumption of $\min^- \geq \max^+$, the above relaxation is tight — i.e., there always exists an optimal solution to the above problem with $l^*(\boldsymbol{w}) = 1$. This solution is then a feasible solution to System (13) as well,

[5] Note that convex programs can only be solved to be within precision ϵ in poly$(1/\epsilon)$ time sine it may have irrational solutions. In this case, we simply say it can be "solved" efficiently.

thus completing our proof. Concretely, given any optimal solution $(\boldsymbol{w}^*, b^*, \epsilon^*)$ to OP (14), we construct another solution $(\bar{\boldsymbol{w}}, \bar{b}, \bar{\epsilon})$ as follows:

$$\bar{\boldsymbol{w}} = \frac{\boldsymbol{w}^*}{\alpha}, \quad \bar{b} = \frac{b^*}{\alpha} + (\frac{1}{\alpha} - 1)\frac{\min^- + \max^+}{2}, \quad \bar{\epsilon} = \frac{\epsilon^*}{\alpha}, \quad \text{where } \alpha = l^*(\boldsymbol{w}^*) \leq 1.$$

We claim that the constructed solution above remains feasible to OP (14). Note that for data point with label 1, we have: (1) $\frac{\min^- + \max^+}{2} \geq r_i$ by assumption $r_i \leq \max^+ \leq \min^-$; (2) $\boldsymbol{x}_i \cdot \boldsymbol{w}^* + b^* \geq -r_i$ by the feasibility of $(\boldsymbol{w}^*, b^*, \epsilon^*)$. Therefore

$$\begin{aligned}
& \boldsymbol{x}_i \cdot \frac{\boldsymbol{w}^*}{\alpha} + \frac{b^*}{\alpha} \geq -\frac{r_i}{\alpha} \\
\Rightarrow \quad & \boldsymbol{x}_i \cdot \frac{\boldsymbol{w}^*}{\alpha} + \frac{b^*}{\alpha} + (\frac{1}{\alpha} - 1)\frac{\min^- + \max^+}{2} \geq -\frac{r_i}{\alpha} + (\frac{1}{\alpha} - 1)r_i \\
\Leftrightarrow \quad & \boldsymbol{x}_i \cdot \bar{\boldsymbol{w}} + \bar{b} \geq -r_i
\end{aligned}$$

This proves that the constructed solution is feasible for data points with label 1. Similar argument using the inequality $\frac{\min^- + \max^+}{2} \leq r_i$ for any negative label data point shows that it is also feasible for negative data points. It is easy to see that the solution quality is as good as the optimal solution ϵ^* since $\alpha \leq 1$. This proves the optimality of the constructed solution.

Finally, we consider the **Situation 2** where the instance is adversarial, i.e, $\min^- \geq 0 \geq \max^+$. In this case, r_i in the first constraint of System (12) is always non-positive whereas r_i in the second constraint is always non-negative. After basic algebraic manipulations, the SERM problem becomes the following optimization problem.

$$\begin{array}{lll}
\text{find} & \boldsymbol{w}, b, \epsilon > 0 & \\
\text{subject to} & \boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq (-r_i) \cdot l^*_{\boldsymbol{x}_i}(\boldsymbol{w}), & \text{for points } (\boldsymbol{x}_i, y_i, r_i) \text{ with } r_i \leq 0. \\
& -(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq (r_i + \epsilon) \cdot l^*_{\boldsymbol{x}_i}(\boldsymbol{w}), & \text{for points } (\boldsymbol{x}_i, y_i, r_i) \text{ with } r_i \geq 0.
\end{array} \tag{15}$$

This is again not a convex feasibility problem due to the non-convex term $(r_i + \epsilon) \cdot l^*_{\boldsymbol{x}_i}(\boldsymbol{w})$, however for any fixed $\epsilon > 0$ both constraints are convex. Moreover, if the system is feasible for some $\epsilon_0 > 0$ and it is feasible for any $0 < \epsilon \leq \epsilon_0$. Therefore, we can determine the feasibility of the (convex) system for any fixed ϵ and then binary search for the feasible ϵ. Therefore, the feasibility problem in System (12) can be solved in polynomial time. □

F. Proof of Theorem 5

Proof. We start with **Situation 1**, i.e., the preferences are arbitrary but the cost function is $c(\boldsymbol{z}; \boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{z}\|_2^2$. We will show later that the second situation can be reduced from the first. In the first situation, the feasibility problem is System (13) with l as the l_2 norm. Our reduction starts by reducing this system to the following optimization problem (OP)

$$\begin{array}{lll}
\text{maximize} & ||\boldsymbol{w}||_2^2 & \\
\text{subject to} & \boldsymbol{x}_i \cdot \boldsymbol{w} + b \geq -r_i, & \text{for points } (\boldsymbol{x}_i, r_i) \text{ with label 1.} \\
& \boldsymbol{x}_i \cdot \boldsymbol{w} + b \leq -r_i - \epsilon, & \text{for points } (\boldsymbol{x}_i, r_i) \text{ with label -1.} \\
& ||\boldsymbol{w}||_2^2 \leq 1 &
\end{array} \tag{16}$$

Formally, we claim that for any fixed ϵ, system (13) is feasible if and only if OP (16) has optimal objective value 1. The "if" direction is simple. That is, if OP (16) has optimal objective value 1, then the optimal solution $(\boldsymbol{w}^*, b^*)$ is automatically a feasible solution to System (13) because $||\boldsymbol{w}^*||_2 = 1$. For the "only if" direction, let $(\bar{\boldsymbol{w}}, \bar{b})$ be any feasible solution to System (13), then it is easy to verify $\boldsymbol{w}^* = \frac{\bar{\boldsymbol{w}}}{||\boldsymbol{w}||_2}$ and $b^* = \frac{\bar{b}}{||\boldsymbol{w}||_2}$ must also be feasible to System (13). Moreover, it is an optimal solution to OP (16) with objective value 1, as desired.

We now prove that determining whether the optimal objective value of OP (16) equals 1 or not is NP-complete. We reduce from the following well-known NP-complete problem called the *partition problem*:

Given d positive integers $c_1, \cdots, c_d$, decide whether there exists a subset $S \subset [d]$ such that $\sum_{i \in S} c_i = \sum_{i \notin S} c_i$

We now reduce the above partition problem to solving OP (16). Given any instance of partition problem, construct the following SERM instance.

The Constructed Hard SERM Instance for Situation 1: We will have $n = 2d + 3$ data points with feature vectors from $\mathbb{R}^d$. For convenience, we will use $\boldsymbol{e}_i$ to denote the basis vector in $\mathbb{R}^d$ whose entries are all 0 except that the i'th is 1. For each $i \in [d]$, there is a data point $(\boldsymbol{x}, y, r) = (2\sqrt{d} \cdot \boldsymbol{e}_i, 1, 4)$ as well as a data point $(\sqrt{d} \cdot \boldsymbol{e}_i, -1, 1 - \epsilon)$. The remaining three data points are $(\boldsymbol{c}, 1, 2)$, data point $(2\boldsymbol{c}, -1, 2 - \epsilon)$, and data point $(3\boldsymbol{c}, 1, 2)$.

We claim that OP (16) instantiated with the above constructed instance has an optimal objective value 1 if and only if the answer to the given partition problem is *Yes*. We first prove the "if" direction. If the partition problem is a Yes instance, then there exists an S such that $\sum_{i \in S} c_i - \sum_{i \notin S} c_i = 0$. We argue that the following construction is an optimal solution to OP (16) with optimal objective value 1:

$$b^* = -2, \ w_i = \frac{1}{\sqrt{d}} \ \forall i \in S, \ w_i = -\frac{1}{\sqrt{d}} \ \forall i \notin S.$$

Clearly, $||\boldsymbol{w}^*||_2^2 = 1$. We only need to prove feasibility of $(\boldsymbol{w}^*, b^*)$. For any label 1 point $(\boldsymbol{x}, r) = (2\sqrt{d} \cdot \boldsymbol{e}_i, 4)$, we have $\boldsymbol{x} \cdot \boldsymbol{w}^* + b^* = 2\sqrt{d}\boldsymbol{e}_i \cdot \boldsymbol{w}^* - 2 = -4 \geq -r$, as desired. Similarly, for any label -1 point $(\boldsymbol{x}, r) = (\sqrt{d} \cdot \boldsymbol{e}_i, 1 - \epsilon)$, we have $\boldsymbol{x} \cdot \boldsymbol{w}^* + b^* = \sqrt{d}\boldsymbol{e}_i \cdot \boldsymbol{w}^* - 2 = -1 \leq -r - \epsilon$. The feasibility of point $(\boldsymbol{c}, 2)$ with label 1 is argued as follows: $\boldsymbol{x} \cdot \boldsymbol{w}^* + b^* = \boldsymbol{c} \cdot \boldsymbol{w}^* - 2 = -r$. Feasibility of $(2\boldsymbol{c}, 2 - \epsilon)$ and $(3\boldsymbol{c}, 2)$ are similarly verified.

We now prove the "only if" direction. In particular, we prove that that if OP (16) has some optimal solution $(\boldsymbol{w}^*, b)$ with $||\boldsymbol{w}^*||_2^2 = 1$, then the partition instance must be Yes.

Let us first examine the feasibility of OP (16).

1. By the constraints with respect to positive-label data points $(2\sqrt{d} \cdot \boldsymbol{e}_i, 4)$, we have $2\sqrt{d}\boldsymbol{e}_i \cdot \boldsymbol{w} + b \geq -4$ or equivalently $w_i\sqrt{d} \geq -\frac{b}{2} - 2$.
2. By the constraints with respect to negative-label data points $(\sqrt{d} \cdot \boldsymbol{e}_i, 1 - \epsilon)$, we have $\sqrt{d}\boldsymbol{e}_i \cdot \boldsymbol{w} + b \leq -1$ or equivalently $w_i\sqrt{d} \leq -b - 1$.
3. By the constraints with respect to data point $(\boldsymbol{c}, 2)$ with label 1, we have $\boldsymbol{c} \cdot \boldsymbol{w} + b \geq -2$, or equivalently $-2 - b \leq \boldsymbol{c} \cdot \boldsymbol{w}$.
4. By the constraints with respect to data point $(2\boldsymbol{c}, 2 - \epsilon)$ with label -1, we have $2\boldsymbol{c} \cdot \boldsymbol{w} + b \leq -2$, or equivalently $-2 - b \geq 2\boldsymbol{c} \cdot \boldsymbol{w}$.
5. By the constraints with respect to data point $(3\boldsymbol{c}, 2)$ with label 1, we have $3\boldsymbol{c} \cdot \boldsymbol{w} + b \geq -2$, or equivalently $-2 - b \leq 3\boldsymbol{c} \cdot \boldsymbol{w}$.

Point 3–5 implies $2\boldsymbol{c} \cdot \boldsymbol{w} \leq -2 - b \leq \min\{\boldsymbol{c} \cdot \boldsymbol{w}, 3\boldsymbol{c} \cdot \boldsymbol{w}\}$. This must imply $\boldsymbol{c} \cdot \boldsymbol{w} = 0$ as any non-zero $\boldsymbol{c} \cdot \boldsymbol{w}$ cannot satisfy $2\boldsymbol{c} \cdot \boldsymbol{w} \leq \min\{\boldsymbol{c} \cdot \boldsymbol{w}, 3\boldsymbol{c} \cdot \boldsymbol{w}\}$. As a consequence, the only feasible b value is $b = -2$. Plugging $b = -2$ into Point 1 and 2, we have

$$-\frac{1}{\sqrt{d}} \leq w_i \leq \frac{1}{\sqrt{d}}.$$

Since the optimal objective value is $1 = \sum_{i=1}^{d} (w_i^*)^2$, it is easy to see that this optimal objective is achieved only when each w_i^* equals either $-\frac{1}{\sqrt{d}}$ or $\frac{1}{\sqrt{d}}$. Now define $S = \{i : w_i^* = \frac{1}{\sqrt{d}}\}$ to be the set of i such that w_i^* is positive. It is easy to verify that S will be a solution to the partition problem, implying that it is a *Yes* instance. This proves the NP-hardness for **Situation 1** stated in the theorem.

Finally, we consider **Situation 2** which can be reduced from the first situation. In particular, the constructed hard instance above has reward preferences all being positive (in fact, drawn from only three possible values $\{1, 2, 4\}$), but do not satisfy the essentially adversarial condition. However, if we are allowed to use instant-wise cost functions, we can simply scale down the reward preference for point with label 1 but propositionally scale down its cost function so that the right-hand-side of the first constraint in System (13) remains the same. Concretely, we now modify our constructed instance above to be the follows.

The Constructed Hard SERM Instance for Situation 2: We still have $n = 2d + 3$ data points with feature vectors from $\mathbb{R}^d$. For each $i \in [d]$, there is a data point $(\boldsymbol{x}, y, r) = (2\sqrt{d} \cdot \boldsymbol{e}_i, 1, 0.5)$ with cost function $c(\boldsymbol{z}; \boldsymbol{x}) = \frac{1}{8}|\boldsymbol{z} - \boldsymbol{x}||_2^2$ as well as a data point $(\sqrt{d} \cdot \boldsymbol{e}_i, -1, 1 - \epsilon)$ with cost function $c(\boldsymbol{z}; \boldsymbol{x}) = |\boldsymbol{z} - \boldsymbol{x}||_2^2$. The remaining three data points are: (1) data point $(\boldsymbol{c}, 1, 0.5)$ with cost function $c(\boldsymbol{z}; \boldsymbol{x}) = \frac{1}{4}|\boldsymbol{z} - \boldsymbol{x}||_2^2$; (2) data point $(2\boldsymbol{c}, -1, 2 - \epsilon)$ with cost function $c(\boldsymbol{z}; \boldsymbol{x}) = |\boldsymbol{z} - \boldsymbol{x}||_2^2$; (3) data point $(3\boldsymbol{c}, 1, 0.5)$ with cost function $c(\boldsymbol{z}; \boldsymbol{x}) = \frac{1}{4}|\boldsymbol{z} - \boldsymbol{x}||_2^2$.

It is easy to verify that the above instance satisfy situation 1 in the theorem statement and is equivalent to the instance we constructed for the second situation and thus is also NP-hard. □