A Proof of Theorem 1

A.1 Intermediate Results

Before we present the proof of the Theorem, we present useful intermediate results which we require in our proof. The following Lemmas present some monotonicity properties of the logistic loss.

Lemma 1. *Let y be a discrete random variable such that*

$$y = \begin{cases} 1, & \text{with probability} \geq \frac{1}{2} + \gamma \\ -1, & \text{with probability} \leq \frac{1}{2} - \gamma \end{cases},$$

for some $\gamma > 0$. Let $\xi = \log \frac{1+2\gamma}{1-2\gamma}$ and let $z < \xi$ be a constant. Define $h(u)$ as follows

$$h(u) = \mathbb{E}_y[\log(1 + e^{-y((1-u)z+u\xi)})].$$

Then $h(u)$ is a strictly decreasing function over the domain $[0, 1)$.

Proof. Let $p = P(y = 1)$. The derviative of $h(u)$, w.r.t u is given by

$$h'(u) = p \times \frac{(z-\xi)}{1 + e^{(1-u)z+u\xi}} + (1-p) \times \frac{(\xi - z)e^{(1-u)z+u\xi}}{1 + e^{(1-u)z+u\xi}}.$$

We will now show that $h'(u) < 0$. Suppose $p < 1$ (otherwise it is easy to see that $h'(u) < 0$). Then

$$\begin{aligned} \left(\tfrac{1+e^{(1-u)z+u\xi}}{\xi - z}\right) \times h'(u) &= -p + (1-p)e^{(1-u)z+u\xi} \\ &= (1-p)\left(e^{(1-u)z+u\xi} - \tfrac{p}{1-p}\right) \\ &\leq (1-p)\left(e^{(1-u)z+u\xi} - e^{\xi}\right) \\ &= (1-p)e^{\xi}\left(e^{(1-u)(z-\xi)} - 1\right) \\ &< 0. \end{aligned}$$

□

Lemma 2. *Let u, ξ be such that $\xi > 0, u \in [0, 1)$. Define functions $h_1(z), h_2(z)$ as follows*

$$h_1(z) = \log(1 + e^{-(1-u)z-u\xi}) - \log(1 + e^{-z}).$$

$$h_2(z) = \log(1 + e^{+(1-u)z+u\xi}) - \log(1 + e^{z}).$$

Then $h_1(z)$ is an increasing function over the domain $(-\infty, \xi)$ and $h_2(z)$ is a decreasing function over $(-\infty, \xi)$.

Proof. The derivative of $h_1(z)$ w.r.t z is given by

$$h_1'(z) = -\frac{1-u}{1 + e^{(1-u)z+u\xi}} + \frac{1}{1+e^{z}}.$$

We will now show that $h_1'(z) \geq 0$.

$$\begin{aligned} h_1'(z) &= -\tfrac{1-u}{1+e^{(1-u)z+u\xi}} + \tfrac{1}{1+e^{z}} \\ &\geq -\tfrac{1}{1+e^{(1-u)z+u\xi}} + \tfrac{1}{1+e^{z}} \\ &\geq -\tfrac{1}{1+e^{z}} + \tfrac{1}{1+e^{z}} \\ &= 0 \end{aligned},$$

where the first inequality follows from the fact that $u \in [0, 1)$ and the second inequality follows from the fact that $z < \xi$. This shows that h_1 is increasing over $(-\infty, \xi)$.

We use a similar argument to show that $h_2'(z)$ is a decreasing function. Consider the derivative of $h_2(z)$ w.r.t z

$$h_2'(z) = \frac{1-u}{1 + e^{-(1-u)z-u\xi}} - \frac{1}{1+e^{-z}}.$$

We will now show that $h_2'(z) \leq 0$.

$$\begin{aligned} h_2'(z) &= \frac{1-u}{1+e^{-(1-u)z-u\xi}} - \frac{1}{1+e^{-z}} \\ &\leq \frac{1}{1+e^{-(1-u)z-u\xi}} - \frac{1}{1+e^{-z}} \\ &\leq \frac{1}{1+e^{-z}} - \frac{1}{1+e^{-z}} \\ &= 0 \end{aligned},$$

This shows that h_2 is decreasing over $(-\infty, \xi)$. □

A.2 Main Argument

0/1 loss. We first prove the Theorem for 0/1 loss; that is, we show that any minimizer of $R_{0-1}(f)+\lambda R_{\text{adv},0-1}(f)$ is a Bayes optimal classifier. We prove the result by contradiction. Let f^* be a Bayes optimal classifier such that $\text{sign}(f^*(\mathbf{x})) = g(\mathbf{x})$ a.e. Suppose $\hat{f}$ is a minimizer of the joint objective. Let $\text{sign}(\hat{f}(\mathbf{x}))$ disagree with $\text{sign}(f^*(\mathbf{x}))$ over a set X of non-zero measure. We show that the joint risk of $\hat{f}$ is strictly larger than f^*.

First, we show that the standard risk of $\hat{f}$ is strictly larger than f^*:

$$\begin{aligned} R_{0-1}(\hat{f}) - R_{0-1}(f^*) &= \mathbb{E}_{(\mathbf{x},y)}\left[\ell_{0-1}(\hat{f}(\mathbf{x}), y) - \ell_{0-1}(f^*(\mathbf{x}), y)\right] \\ &= \Pr(\mathbf{x} \in X) \times \mathbb{E}_{(\mathbf{x},y)}\left[\ell_{0-1}(\hat{f}(\mathbf{x}), y) - \ell_{0-1}(f^*(\mathbf{x}), y)\Big|\mathbf{x} \in X\right] \\ &= \Pr(\mathbf{x} \in X) \times \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_y\left[\ell_{0-1}(\hat{f}(\mathbf{x}), y) - \ell_{0-1}(f^*(\mathbf{x}), y)\Big|\mathbf{x}\right]\Big|\mathbf{x} \in X\right] \\ &= \Pr(\mathbf{x} \in X) \times \mathbb{E}_{\mathbf{x}}\left[P(y \neq \text{sign}(\hat{f}(\mathbf{x}))|\mathbf{x}) - P(y \neq \text{sign}(f^*(\mathbf{x}))|\mathbf{x})\Big|\mathbf{x} \in X\right] \\ &> 0, \end{aligned}$$

where the last inequality follows from the definition of Bayes optimal decision rule.

We now show that the adversarial risk of $\hat{f}$ is larger than f^*. Since the base classifier g agrees with f^* a.e. we have

$$R_{\text{adv},0-1}(f^*) = \mathbb{E}\left[\max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell_{0-1}\left(f^*(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \ell_{0-1}\left(f^*(\mathbf{x}), g(\mathbf{x})\right)\right] = 0.$$

Since $R_{\text{adv},0-1}$ of any classifier is always non-negative, this shows that $R_{\text{adv},0-1}(\hat{f}) \geq R_{\text{adv},0-1}(f^*)$. Combining this with the above result on classification risk we get

$$R_{0-1}(\hat{f}) + \lambda R_{\text{adv},0-1}(\hat{f}) > R_{0-1}(f^*) + \lambda R_{\text{adv},0-1}(f^*).$$

This shows that $\hat{f}$ can't be a minimizer of the joint objective and minimizer of joint objective should be a Bayes optimal classifier.

Logistic Loss. We now consider the logistic loss and show that any minimizer of $R(f) + \lambda R_{\text{adv}}(f)$ is a Bayes optimal classifier. We again prove the result by contradiction. Let $\xi = \log \frac{1+2\gamma}{1-2\gamma}$. Suppose $\hat{f}$ is a minimizer of the joint objective and is not Bayes optimal. Define set X as

$$X = \{\mathbf{x} : \hat{f}(\mathbf{x})g(\mathbf{x}) < \xi\}.$$

Note that, since $\hat{f}$ is not Bayes optimal, X is a set with non-zero measure. Construct a new classifier $\bar{f}$ as follows

$$\bar{f}(\mathbf{x}) = \begin{cases} \hat{f}(\mathbf{x}), & \text{if } \mathbf{x} \notin X \\ \hat{f}(\mathbf{x}) + \tau\left(\xi - \hat{f}(\mathbf{x})g(\mathbf{x})\right)g(\mathbf{x}), & \text{otherwise} \end{cases},$$

where $\tau \in (0, 1)$ is a constant. We now show that $\bar{f}$ has a strictly lower joint risk than $\hat{f}$. This will then contradict our assumption that $\hat{f}$ is a minimizer of the joint objective.

Let $\ell_{\text{adv}}(f, g, \mathbf{x})$ be the adversarial risk at point $\mathbf{x}$, computed w.r.t base classifier g

$$\ell_{\text{adv}}(f, g, \mathbf{x}) = \max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell\left(f(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \ell\left(f(\mathbf{x}), g(\mathbf{x})\right).$$

Define the conditional risk of f at $\mathbf{x}$ as

$$C(f, \mathbf{x}) = \mathbb{E}_y\left[\ell(f(\mathbf{x}), y)\Big|\mathbf{x}\right] + \lambda \ell_{\text{adv}}(f, g, \mathbf{x}).$$

Note that $\mathbb{E}_{\mathbf{x}}\left[C(f, \mathbf{x})\right]$ is equal to the joint risk $R(f) + \lambda R_{\text{adv}}(f)$. We now show that $C(\bar{f}, \mathbf{x}) - C(\hat{f}, \mathbf{x}) \leq 0, \forall \mathbf{x}$.

Case 1. Let $\mathbf{x} \notin X$. Then $\hat{f}(\mathbf{x}) = \bar{f}(\mathbf{x})$. So we have

$$\begin{aligned}
C(\bar{f}, \mathbf{x}) - C(\hat{f}, \mathbf{x}) &= \lambda\left(\ell_{\text{adv}}(\bar{f}, g, \mathbf{x}) - \ell_{\text{adv}}(\hat{f}, g, \mathbf{x})\right) \\
&\leq \lambda\left(\max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \ell\left(\hat{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right)\right) \\
&= \lambda \max\left\{0, \max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon, \mathbf{x}+\boldsymbol{\delta} \in X \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \ell\left(\hat{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right)\right\} \\
&= 0,
\end{aligned}$$

where the last equality follows from the observation that $g(\mathbf{x})\bar{f}(\mathbf{x}+\boldsymbol{\delta}) \geq g(\mathbf{x})\hat{f}(\mathbf{x}+\boldsymbol{\delta})$ and the logistic function $\ell(z) = \log(1+e^{-z})$ is a monotonically decreasing function.

Case 2. Let $\mathbf{x} \in X$. Then $\hat{f}(\mathbf{x}) \neq \bar{f}(\mathbf{x})$. Now, consider the difference $C(\bar{f}, \mathbf{x}) - C(\hat{f}, \mathbf{x})$:

$$C(\bar{f}, \mathbf{x}) - C(\hat{f}, \mathbf{x}) = \underbrace{\mathbb{E}_y\left[\ell(\bar{f}(\mathbf{x}), y) - \ell(\hat{f}(\mathbf{x}), y)\Big|\mathbf{x}\right]}_{T_1} + \lambda \underbrace{\left(\ell_{\text{adv}}(\bar{f}, g, \mathbf{x}) - \ell_{\text{adv}}(\hat{f}, g, \mathbf{x})\right)}_{T_2}.$$

We show that both T_1, T_2 are non-positive. Using the monotonicity property of logistic loss in Lemma 1, it is easy to verify that $T_1 < 0$. We now bound T_2. First, observe that based on our construction of $\bar{f}(\mathbf{x})$ and our definition of set X, we have

$$\inf_{\mathbf{x} \notin X} \bar{f}(\mathbf{x}) g(\mathbf{x}) \geq \sup_{\mathbf{x} \in X} \bar{f}(\mathbf{x}) g(\mathbf{x}), \quad \inf_{\mathbf{x} \notin X} \hat{f}(\mathbf{x}) g(\mathbf{x}) \geq \sup_{\mathbf{x} \in X} \hat{f}(\mathbf{x}) g(\mathbf{x}).$$

Since the logistic loss $\ell(z) = \log(1+e^{-z})$ is monotonically decreasing in z, this shows that both the inner maxima in T_2 are achieved at $\boldsymbol{\delta}$'s such that $\mathbf{x}+\boldsymbol{\delta} \in X$. Using this observation, T_2 can be rewritten as

$$\lambda\left(\max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon, \mathbf{x}+\boldsymbol{\delta} \in X \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \ell\left(\bar{f}(\mathbf{x}), g(\mathbf{x})\right)\right) - \lambda\left(\max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon, \mathbf{x}+\boldsymbol{\delta} \in X \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell(\hat{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})) - \ell(\hat{f}(\mathbf{x}), g(\mathbf{x}))\right).$$

The above expression can be rewritten as

$$\lambda\left(\max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon, \mathbf{x}+\boldsymbol{\delta} \in X \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon, \mathbf{x}+\boldsymbol{\delta} \in X \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell(\hat{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x}))\right) - \lambda\left(\ell\left(\bar{f}(\mathbf{x}), g(\mathbf{x})\right) - \ell(\hat{f}(\mathbf{x}), g(\mathbf{x}))\right).$$

Note that $\ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right)$ in the above expression can equivalently be written as $\ell\Big((1-\tau)\hat{f}(\mathbf{x}+\boldsymbol{\delta}) + \tau\xi g(\mathbf{x}), g(\mathbf{x})\Big)$. This shows that both $\ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right)$ and $\ell(\hat{f}(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x}))$ in the above expression are monotonically decreasing in $g(\mathbf{x})\hat{f}(\mathbf{x}+\boldsymbol{\delta})$ and as a result the maximum of both the inner

objectives is achieved at a $\boldsymbol{\delta}$ which minimizes $g(\mathbf{x})\hat{f}(\mathbf{x}+\boldsymbol{\delta})$. Let $\boldsymbol{\delta}_{\mathbf{x}}$ be the point at which the maxima is achieved. Then the above expression can be written as

$$T_2 = \lambda\left(\ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{x}}), g(\mathbf{x})\right) - \ell(\hat{f}(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{x}}), g(\mathbf{x}))\right) - \lambda\left(\ell\left(\bar{f}(\mathbf{x}), g(\mathbf{x})\right) - \ell(\hat{f}(\mathbf{x}), g(\mathbf{x}))\right).$$

From Lemma 2 we know that $\ell\left(\bar{f}(\mathbf{x}), g(\mathbf{x})\right) - \ell(\hat{f}(\mathbf{x}), g(\mathbf{x}))$ is an increasing function in $\hat{f}(\mathbf{x})g(\mathbf{x})$. Since $\hat{f}(\mathbf{x})g(\mathbf{x}) \geq \hat{f}(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{x}}), g(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{x}})$, we have

$$T_2 \leq 0.$$

Combining the bounds for T_1 and T_2 we obtain $C(\bar{f}, \mathbf{x}) - C(\hat{f}, \mathbf{x}) < 0$, for any $\mathbf{x} \in X$. This shows that $\bar{f}(\mathbf{x})$ has a strictly lower joint risk than $\hat{f}$. So $\hat{f}$ can't be a minimizer of the joint risk. This finishes the proof of Theorem 1.

B Proof of Theorem 2

The proof follows from the proof of Theorem 3, because under the margin condition $H_{\text{adv},0-1}(f)$ is equivalent to $G_{\text{adv},0-1}(f)$ when the label y is a deterministic function of $\mathbf{x}$.

C Proof of Theorem 3

0/1 loss. We first prove the Theorem for 0/1 loss. We use a similar proof strategy as Theorem 1 and prove the result by contradiction. Let $\eta(\mathbf{x})$ be a Bayes decision rule which satisfies the margin condition. Let f^* be a Bayes optimal classifier such that $\text{sign}(f^*(\mathbf{x})) = \eta(\mathbf{x})$ a.e. Suppose $\hat{f}$ is a minimizer of the joint objective. Let $\text{sign}(\hat{f}(\mathbf{x}))$ disagree with $\text{sign}(f^*(\mathbf{x}))$ over a set X of non-zero measure. From the proof of Theorem 1 we know that $R_{0-1}(\hat{f}) - R_{0-1}(f^*) > 0$.

We now show that $\hat{f}$ has a larger adversarial risk than f^*. From the definition of $G_{\text{adv},0-1}(f^*)$ we have

$$G_{\text{adv},0-1}(f^*) \quad = \quad \mathbb{E}_{(\mathbf{x},y)}\left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell_{0-1}\left(f^*(\mathbf{x}+\boldsymbol{\delta}), y\right) - \ell_{0-1}\left(f^*(\mathbf{x}), y\right)\right].$$

From margin condition in Equation (5) we know that $\forall \mathbf{x}, \|\boldsymbol{\delta}\| \leq \epsilon, \text{sign}(f^*(\mathbf{x}+\boldsymbol{\delta})) = \eta(\mathbf{x}+\boldsymbol{\delta}) = \eta(\mathbf{x})$. So $G_{\text{adv},0-1}(f^*) = 0$.

Since $G_{\text{adv},0-1}$ of any classifier is always non-negative, this shows that $G_{\text{adv},0-1}(\hat{f}) \geq G_{\text{adv},0-1}(f^*)$. Combining this with the above result on classification risk we get

$$R_{0-1}(\hat{f}) + \lambda G_{\text{adv},0-1}(\hat{f}) > R_{0-1}(f^*) + \lambda G_{\text{adv},0-1}(f^*).$$

This shows that $\hat{f}$ can't be a minimizer of the joint objective. This shows that any minimizer of Equation (7) is a Bayes optimal classifier.

Logistic loss. To prove the Theorem for logistic loss, we heavily rely on some of the intermediate results we proved for Theorem 1. Let $\xi = \log\frac{1+2\gamma}{1-2\gamma}$. Suppose $\hat{f}$ is a minimizer of the joint objective and is not Bayes optimal. Define set X as

$$X = \{\mathbf{x} : \hat{f}(\mathbf{x})\eta(\mathbf{x}) < \xi\}.$$

Note that X is a set with non-zero measure. Construct $\bar{f}$ as follows

$$\bar{f}(\mathbf{x}) = \begin{cases} \hat{f}(\mathbf{x}), & \text{if } \mathbf{x} \notin X \\ \hat{f}(\mathbf{x}) + \left(\xi - \hat{f}(\mathbf{x})\eta(\mathbf{x})\right)\tau\eta(\mathbf{x}), & \text{otherwise} \end{cases},$$

where $\tau \in (0,1)$ is a constant. We now show that $\bar{f}$ has a strictly lower joint risk than $\hat{f}$.

Define the conditional risk of f at $\mathbf{x}$ as

$$C(f, \mathbf{x}) = \mathbb{E}_y\left[\ell(f(\mathbf{x}), y)\middle|\mathbf{x}\right] + \lambda\mathbb{E}_y\left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(f(\mathbf{x}+\boldsymbol{\delta}), y\right) - \ell\left(f(\mathbf{x}), y\right)\middle|\mathbf{x}\right].$$

We consider two cases, $\mathbf{x} \in X$ and $\mathbf{x} \notin X$, and show that in both the cases $\bar{f}$ has a lower conditional risk than $\hat{f}$.

Case 1. Let $\mathbf{x} \notin X$. Then $\hat{f}(\mathbf{x}) = \bar{f}(\mathbf{x})$. So we have

$$\begin{aligned} C(\bar{f},\mathbf{x}) - C(\hat{f},\mathbf{x}) &= \lambda \mathbb{E}\left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}),y\right) - \max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(\hat{f}(\mathbf{x}+\boldsymbol{\delta}),y\right)\Big|\mathbf{x}\right] \\ &= \lambda P(y=\eta(\mathbf{x})|\mathbf{x}) \underbrace{\left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}),\eta(\mathbf{x})\right) - \max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(\hat{f}(\mathbf{x}+\boldsymbol{\delta}),\eta(\mathbf{x})\right)\right]}_{T_1} \\ &\quad + \lambda P(y=-\eta(\mathbf{x})|\mathbf{x}) \underbrace{\left[\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(\bar{f}(\mathbf{x}+\boldsymbol{\delta}),-\eta(\mathbf{x})\right) - \max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(\hat{f}(\mathbf{x}+\boldsymbol{\delta}),-\eta(\mathbf{x})\right)\right]}_{T_2} \end{aligned}$$

Using the margin condition on $\eta(\mathbf{x})$, and using the same technique as in proof of Case 1 of Theorem 1, we can show that $T_1 \leq 0$. Since both the inner maxima in T_2 are achieved at $\mathbf{x}+\boldsymbol{\delta} \notin X$, it is easy to verify that $T_2 = 0$. This shows that $\forall \mathbf{x} \notin X, C(\bar{f},\mathbf{x}) - C(\hat{f},\mathbf{x}) \leq 0$.

Case 2. Let $\mathbf{x} \in X$. Let $\ell_{\text{adv}}(f,\mathbf{x},y)$ be the adversarial risk at point $(\mathbf{x},y)$

$$\ell_{\text{adv}}(f,\mathbf{x},y) = \max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(f(\mathbf{x}+\boldsymbol{\delta}),y\right) - \ell\left(f(\mathbf{x}),y\right).$$

We have

$$C(\bar{f},\mathbf{x}) - C(\hat{f},\mathbf{x}) = \underbrace{\mathbb{E}\left[\ell(\bar{f}(\mathbf{x}),y) - \ell(\hat{f}(\mathbf{x}),y)\Big|\mathbf{x}\right]}_{T_3} + \lambda \underbrace{\mathbb{E}_y\left[\ell_{\text{adv}}(\bar{f},\mathbf{x},y) - \ell_{\text{adv}}(\hat{f},\mathbf{x},y)\right]}_{T_4}.$$

From the proof of Case 2 of Theorem 1 we know that $T_3 < 0$. We now show that $T_4 \leq 0$. Let $p_{\mathbf{x}} = P(y=\eta(\mathbf{x})|\mathbf{x})$. T_4 can be decomposed as follows

$$p_{\mathbf{x}} \underbrace{\left(\ell_{\text{adv}}(\bar{f},\mathbf{x},\eta(\mathbf{x})) - \ell_{\text{adv}}(\hat{f},\mathbf{x},\eta(\mathbf{x}))\right)}_{T_5} + (1-p_{\mathbf{x}}) \underbrace{\left(\ell_{\text{adv}}(\bar{f},\mathbf{x},-\eta(\mathbf{x})) - \ell_{\text{adv}}(\hat{f},\mathbf{x},-\eta(\mathbf{x}))\right)}_{T_6}.$$

Following the proof of Case 2 of Theorem 1 and using the margin condition we can show that $T_5 \leq 0$. We now show that $T_6 \leq 0$. First observe that both the suprema in T_6 either occur at the same point. Suppose both the suprema in T_6 occur outside X. Then T_6 is given by

$$T_6 = \ell(\hat{f}(\mathbf{x}),-\eta(\mathbf{x})) - \ell(\bar{f}(\mathbf{x}),-\eta(\mathbf{x})) \leq 0.$$

Suppose both the suprema occur inside X. Then using the observation that $\ell(\bar{f}(\mathbf{x}),-\eta(\mathbf{x})) - \ell(\hat{f}(\mathbf{x}),-\eta(\mathbf{x}))$ is a decreasing function of $\hat{f}(\mathbf{x})\eta(\mathbf{x})$ (see Lemma 2), we get $T_6 \leq 0$.

D Proof of Theorem 4

For any Bayes decision rule $\eta(\mathbf{x})$, let X_η be the set of points which violate the margin condition

$$X_\eta = \{\mathbf{x} : \exists \tilde{\mathbf{x}}, \|\tilde{\mathbf{x}}-\mathbf{x}\| \leq \epsilon \text{ and } \eta(\tilde{\mathbf{x}}) \neq \eta(\mathbf{x})\}.$$

Since no Bayes decision rule satisfies the margin condition, we have $\Pr(\mathbf{x} \in X_\eta) > 0, \forall \eta$. Let $p = \inf_\eta \Pr(\mathbf{x} \in X_\eta)$.

We first consider the joint risk $R_{0-1}(f) + \lambda G_{\text{adv},0-1}(f)$. To prove the Theorem we show that there exist classifiers which obtain strictly lower joint risk than any Bayes optimal classifier. Let $f^* : \mathbb{R}^d \to \mathbb{R}$ be any Bayes optimal classifier, with the corresponding Bayes decision rule $\eta(\mathbf{x}) = \text{sign}(f^*(\mathbf{x}))$. We first obtain a lower bound on

$G_{\text{adv},0-1}(f^*)$. Consider the following

$$\begin{aligned}
G_{\text{adv},0-1}(f^*) &\geq \Pr(\mathbf{x} \in X_\eta) \times \mathbb{E}_{(\mathbf{x},y)} \left[\sup_{\|\boldsymbol{\delta}\| \leq \epsilon} \ell_{0-1}\left(f^*(\mathbf{x}+\boldsymbol{\delta}), y\right) - \ell_{0-1}\left(f^*(\mathbf{x}), y\right) \Big| \mathbf{x} \in X_\eta \right] \\
&\geq \Pr(\mathbf{x} \in X_\eta) \times \mathbb{E}_{\mathbf{x}} \left[P(y = \eta(\mathbf{x})|\mathbf{x}) \left(\sup_{\|\boldsymbol{\delta}\| \leq \epsilon} \ell_{0-1}\left(f^*(\mathbf{x}+\boldsymbol{\delta}), \eta(\mathbf{x})\right) - \ell_{0-1}\left(f^*(\mathbf{x}), \eta(\mathbf{x})\right) \right) \Big| \mathbf{x} \in X_\eta \right] \\
&\geq \frac{\Pr(\mathbf{x} \in X_\eta)}{2} \mathbb{E}_{\mathbf{x}} \left[\sup_{\|\boldsymbol{\delta}\| \leq \epsilon} \ell_{0-1}\left(f^*(\mathbf{x}+\boldsymbol{\delta}), \eta(\mathbf{x})\right) - \ell_{0-1}\left(f^*(\mathbf{x}), \eta(\mathbf{x})\right) \Big| \mathbf{x} \in X_\eta \right] \\
&\geq \frac{\Pr(\mathbf{x} \in X_\eta)}{2} \\
&\geq \frac{p}{2}
\end{aligned}$$

where the third inequality follows from the fact that $P(y = \eta(\mathbf{x})|\mathbf{x}) \geq \frac{1}{2}$ and the fourth inequality follows from the observation that any $\mathbf{x} \in X_\eta$ violates the margin condition. This gives us the following lower bound on the joint risk of f^*

$$R_{0-1}(f^*) + \lambda G_{\text{adv},0-1}(f^*) \geq \frac{\lambda p}{2}. \tag{9}$$

Now consider the "constant" classifier f_{-1} which assigns all the points to the negative class. This classifier has 0 adversarial risk. So its joint risk can be upper bounded as follows

$$R_{0-1}(f_{-1}) + \lambda G_{\text{adv},0-1}(f_{-1}) \leq 1. \tag{10}$$

Equations (9), (10) show that $\forall \lambda > \frac{2}{p}$, there exist classifiers with strictly lower joint risk than any Bayes optimal classifier. Using the same argument we can show that similar results hold for the other joint risk $R_{0-1}(f) + \lambda H_{\text{adv},0-1}(f)$.

E Proofs of Section 6

Here we present the proofs of Section 6. To begin with, we first present a result which characterizes the standard and adversarial risk for the mixture model.

Theorem 10. *Suppose the perturbations are measured w.r.t L_∞ norm. Let $\mathbf{w} \in \mathbb{R}^p$ be a linear separator and moreover suppose the base classifier $g(\mathbf{x})$ is the Bayes optimal decision rule. Then, for any linear classifier $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, we have that*

1. $R_{0-1}(f_{\mathbf{w}}) = \Phi\left(-\frac{\mathbf{w}^T\mathbf{w}^*}{\sigma\|\mathbf{w}\|_2}\right)$,
2. $G_{adv,0-1}(f_{\mathbf{w}}) = \Phi\left(\frac{\|\mathbf{w}\|_1\epsilon - \mathbf{w}^T\mathbf{w}^*}{\sigma\|\mathbf{w}\|_2}\right)$,
3. $R_{adv,0-1}(f_{\mathbf{w}}) \leq \Phi\left(\frac{\|\mathbf{w}-\mathbf{w}^*\|_1\epsilon - (\mathbf{w}-\mathbf{w}^*)^T\mathbf{w}^*}{\sigma\|\mathbf{w}-\mathbf{w}^*\|_2}\right)$,

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Proof. To see the first part, we begin by observing that $\mathbf{w}^T\mathbf{x}$ is a univariate normal random variable when conditioned on the label y, one can derive the 0-1 error for the classifier in closed form. In particular,

$$R_{0-1}(f_{\mathbf{w}}) = 1 - \frac{1}{2}\Phi\left(\frac{\mathbf{w}^T\mathbf{w}^*}{\sigma\|\mathbf{w}\|_2}\right) - \frac{1}{2}\Phi\left(\frac{\mathbf{w}^T\mathbf{w}^*}{\sigma\|\mathbf{w}\|_2}\right) = 1 - \Phi\left(\frac{\mathbf{w}^T\mathbf{w}^*}{\sigma\|\mathbf{w}\|_2}\right) = \Phi\left(\frac{-\mathbf{w}^T\mathbf{w}^*}{\sigma\|\mathbf{w}\|_2}\right)$$

Following the existing definition of adversarial risk, we see that

$$G_{adv,0-1}(f) = \mathbb{E}\left[\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \ell_{0-1}(f(\mathbf{x}+\boldsymbol{\delta}), y)\right]$$

We consider the case of $y = 1$. We know that $\mathbf{x}|y = 1 \sim \mathcal{N}(\mathbf{w}^*, \sigma^2\mathcal{I}_d)$. So, $\mathbf{x} = \mathbf{w}^* + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2\mathcal{I}_d)$. Now, for any $\mathbf{z}$, we incur a loss of 1, whenever there exists a $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$ and,

$$\mathbf{w}^T(\mathbf{x} + \boldsymbol{\delta}) = \mathbf{w}^T(\mathbf{w}^*) + \mathbf{w}^T(\mathbf{z}) + \mathbf{w}^T\boldsymbol{\delta} \leq 0,$$

As long as $\mathbf{z}$ is such that, $\mathbf{w}^T\mathbf{z} \leq \|\mathbf{w}\|_1 \epsilon - \mathbf{w}^T\mathbf{w}^*$, we will always incur a penalty. Now, $\mathbf{w}^T\mathbf{z} \sim \mathcal{N}(0, \sigma^2\|\mathbf{w}\|_2^2)$, therefore, $Pr(\mathbf{w}^T\mathbf{z} \leq \|\mathbf{w}\|_1 \epsilon - \mathbf{w}^T\mathbf{w}^*) = \Phi(\frac{\|\mathbf{w}\|_1 \epsilon - \mathbf{w}^T\mathbf{w}^*}{\|\mathbf{w}\|_2 \sigma})$. Symmetric argument holds for $y = -1$. Hence, we get that,

$$G_{adv,0-1}(f_{\mathbf{w}}) = \Phi\left(\frac{\|\mathbf{w}\|_1 \epsilon - \mathbf{w}^T\mathbf{w}^*}{\|\mathbf{w}\|_2 \sigma}\right)$$

Now to prove the third claim, we have that

- Suppose $y = 1$, then $\mathbf{x} = \mathbf{w}^* + z$ where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2\mathcal{I}_d)$. Suppose ${\mathbf{w}^*}^T\mathbf{x} > 0$.
- Then, for a given $\mathbf{z}$, we will incur a penalty if $\mathbf{z}$ satisfies the following constraints:
  - We have that $\mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{w}^* + {\mathbf{w}^*}^T\mathbf{z} > 0$.
  - There exists a $\boldsymbol{\delta}$ *s.t.* $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$, such that,

    $$ {\mathbf{w}^*}^T(\mathbf{x} + \boldsymbol{\delta}) > 0 \;\text{ and }\; \mathbf{w}^T(\mathbf{x} + \boldsymbol{\delta}) < 0$$

  - Note that whenever the above event happens, the following also happens:

    $$(\mathbf{w} - \mathbf{w}^*)^T(\mathbf{x} + \boldsymbol{\delta}) = (\mathbf{w} - \mathbf{w}^*)^T(\mathbf{z}) + (\mathbf{w} - \mathbf{w}^*)^T(\mathbf{w}^*) + (\mathbf{w} - \mathbf{w}^*)^T\boldsymbol{\delta} < 0$$

    Now, for a given $\mathbf{z}$, $(\mathbf{w} - \mathbf{w}^*)^T(\mathbf{z}) \sim \mathcal{N}(0, \|\mathbf{w} - \mathbf{w}^*\|_2^2)\sigma^2$. Also, as long as $\mathbf{z}$ is such that $(\mathbf{w} - \mathbf{w}^*)^T(\mathbf{z}) \leq \|\mathbf{w} - \mathbf{w}^*\|_1 \epsilon - (\mathbf{w} - \mathbf{w}^*)^T\mathbf{w}^*$, we will incur a penalty. This event happens with probability,

    $$\Phi\left(\frac{\|\mathbf{w} - \mathbf{w}^*\|_1 \epsilon - (\mathbf{w} - \mathbf{w}^*)^T\mathbf{w}^*}{\sigma\|\mathbf{w} - \mathbf{w}^*\|_2}\right)$$

    This establishes the upper bound.

□

E.1 Proof of Theorem 5

We use the same notation as in the proof of Theorem 10. Let $R^* = R_{0-1}(f_{\mathbf{w}^*})$. Using Theorem 10, we can write the excess 0-1 risk of $\mathbf{w}$ as:

$$R_{0-1}(f_{\mathbf{w}}) - R^* = \Phi\left(-\frac{\|\mathbf{w}^*\|_2^2}{\sigma\left(\sqrt{\|\mathbf{w}^*\|_2^2 + 1}\right)}\right) - \Phi\left(-\frac{\|\mathbf{w}^*\|_2}{\sigma}\right)$$

$$R_{0-1}(f_{\mathbf{w}}) - R^* = \Phi\left(-\frac{\|\mathbf{w}^*\|_2}{\sigma\left(\sqrt{1 + \frac{1}{\|\mathbf{w}^*\|_2^2}}\right)}\right) - \Phi\left(-\frac{\|\mathbf{w}^*\|_2}{\sigma}\right)$$

Next, we lower bound the adversarial risk. Suppose that $y = 1$, then we have that $\mathbf{x} = \mathbf{w}^* + \mathbf{z}_S + \mathbf{z}_{S^c}$. Similarly, let $\mathbf{w} = \mathbf{w}_S + \mathbf{w}_{S^c}$. In our case, $\mathbf{w}_S = \mathbf{w}^*$ and $\mathbf{w}_{S^c} = \boldsymbol{\alpha} = [\frac{\pm 1}{\sqrt{d-k}}, \frac{\pm 1}{\sqrt{d-k}}, \ldots, \frac{\pm 1}{\sqrt{d-k}}]^T$. Then, we have that $\mathbf{w}^T\mathbf{x} = {\mathbf{w}^*}^T\mathbf{w}^* + {\mathbf{w}^*}^T\mathbf{z}_S + \boldsymbol{\alpha}^T\mathbf{z}_{S^c}$.

- Consider the Event $\boldsymbol{\alpha}^T\mathbf{z}_{S^c} > -{\mathbf{w}^*}^T\mathbf{w}^* - {\mathbf{w}^*}^T\mathbf{z}_S$ This is the event that $\mathbf{w}, \mathbf{w}^*$ agree before perturbation.

- Consider the Event B,
$$\boldsymbol{\alpha}^T \mathbf{z}_{S^c} < \|\boldsymbol{\alpha}\|_1 \epsilon - \mathbf{w}^{*T}\mathbf{w}^* - \mathbf{w}^{*T}\mathbf{z}_S$$
This is the event that there exists a perturbation restricted to the subspace S^c such that, $\mathbf{w}^T(\mathbf{x}+\boldsymbol{\delta}) < 0$. Note that since the perturbation is restricted to S^c, $\mathbf{w}^*$'s prediction doesn't change.

- Now for the probability that both events happen:
  - Observe that $A = (\boldsymbol{\alpha}^T \mathbf{z}_{S^c} + \mathbf{w}^{*T}\mathbf{z}) \sim \mathcal{N}(0, \sigma^2(\|\boldsymbol{\alpha}\|_2^2 + \|\mathbf{w}^*\|_2^2))$.
  - So, the probability of both events happening is that the random variable $-\mathbf{w}^{*T}\mathbf{w}^* \leq A \leq \|\boldsymbol{\alpha}\|_1 \epsilon - \mathbf{w}^{*T}\mathbf{w}^*$
$$\Phi\left(\frac{\|\boldsymbol{\alpha}\|_1 \epsilon - \mathbf{w}^{*T}\mathbf{w}^*}{\sigma\sqrt{(\|\boldsymbol{\alpha}\|_2^2 + \|\mathbf{w}^*\|_2^2)}}\right) - \Phi\left(\frac{-\mathbf{w}^{*T}\mathbf{w}^*}{\sigma\sqrt{(\|\boldsymbol{\alpha}\|_2^2 + \|\mathbf{w}^*\|_2^2)}}\right)$$
  - Now, for $\epsilon = 2\|\mathbf{w}^*\|_2^2/\sqrt{d-k}$, we get that the probability that both events happens is:
$$\Phi\left(\frac{\mathbf{w}^{*T}\mathbf{w}^*}{\sigma\sqrt{(\|\boldsymbol{\alpha}\|_2^2 + \|\mathbf{w}^*\|_2^2)}}\right) - \Phi\left(\frac{-\mathbf{w}^{*T}\mathbf{w}^*}{\sigma\sqrt{(\|\boldsymbol{\alpha}\|_2^2 + \|\mathbf{w}^*\|_2^2)}}\right) = 2\Phi\left(\frac{\mathbf{w}^{*T}\mathbf{w}^*}{\sigma\sqrt{(\|\boldsymbol{\alpha}\|_2^2 + \|\mathbf{w}^*\|_2^2)}}\right) - 1$$
  - Now, for $\frac{\mathbf{w}^{*T}\mathbf{w}^*}{\sigma\sqrt{(\|\boldsymbol{\alpha}\|_2^2+\|\mathbf{w}^*\|_2^2)}} = 2$, $R_{adv,0-1}(f) > 0.95$
  - Note that $\|\boldsymbol{\alpha}\|_2^2 = 1$. Therefore for $\sigma = 1$, we get that $\|\mathbf{w}^*\|_2^2 = 2 + 2 * \sqrt{2}$.
  - At this value, we have that excess 0-1 risk < 0.02, which completes the proof.

E.2 Proof of Theorem 6

We know that $\mathbf{w}^* = [1/\sqrt{d/2}, 1/\sqrt{d/2}, \ldots, 1/\sqrt{d/2}] \in \mathbb{R}^p$. When restricted to only top half co-ordinates, it is easy to see that $\mathbf{w} = [\underbrace{1/\sqrt{d/2}, \ldots, 1/\sqrt{d/2}}_{d/2}, 0, \ldots, 0]$ is the optimizer of the standard risk. For this setting, from Theorem 10, we get that,

$$R_{0-1}(f_{\mathbf{w}^*}) = \Phi(-\sqrt{2}) = 0.07, \quad R_{0-1(f_{\mathbf{w}})} = \Phi(-1) = 0.16$$

Hence, we have that $R_{0-1}(f_{\mathbf{w}}) - R_{0-1}(f_{\mathbf{w}^*}) < 0.1$. Now, to get a lower bound on the adversarial risk of $\mathbf{w}$, consider the perturbations of the form $\boldsymbol{\gamma} = [\underbrace{-\epsilon, -\epsilon, \ldots, -\epsilon}_{d/2}, \epsilon, \epsilon, \ldots, \epsilon]$. Note that for such a perturbation γ, we have that,

$$\mathbf{w}^{*T}\mathbf{x} = \mathbf{w}^{*T}(\mathbf{x}+\boldsymbol{\gamma}) \;\text{ and }\; \mathbf{w}^T(\mathbf{x}+\boldsymbol{\gamma}) = \mathbf{w}^T\mathbf{x} - \epsilon\sqrt{\frac{d}{2}}$$

Now, suppose $y = 1$. Then, $\mathbf{x} = \mathbf{w}^* + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \mathcal{I}_d)$. For this, we have that,

$$\mathbf{w}^{*T}(\mathbf{x}) = \mathbf{w}^{*T}\mathbf{w}^* + \mathbf{w}^{*T}\mathbf{z} = 2 + \underbrace{\mathbf{w}_{1:d/2}^T \mathbf{z}_{1:d/2}}_{A} + \underbrace{\mathbf{w}_{1:d/2}^T \mathbf{z}_{d/2:d}}_{B},$$

On the other hand, $\mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{w}^* + \mathbf{w}^T\mathbf{z} = 1 + \underbrace{\mathbf{w}_{1:d/2}^T \mathbf{z}_{1:d/2}}_{A}$. Consider the event such that

$$\mathbf{w}^{*T}\mathbf{x} > 0 \;\&\; \mathbf{w}^T\mathbf{x} > 0 \;\&\; \mathbf{w}^T(\mathbf{x}+\boldsymbol{\gamma}) < 0.$$

This is the event that x is such that both of $\mathbf{w}$ and $\mathbf{w}^*$ agree before, but after adding the perturbation γ the prediction of $\mathbf{w}$ changes. Following the form of $\boldsymbol{\gamma}$, this event can be rewritten as:

$$\mathbf{w}^{*T}\mathbf{x} > 0 \;\&\; \mathbf{w}^T\mathbf{x} > 0 \;\&\; \mathbf{w}^T(\mathbf{x}) < \epsilon\sqrt{d/2}$$

Rewriting this event in terms of the random variables A and B, we get the equivalent event,

$$2 + A + B > 0 \;\;\&\;\; A + 1 > 0 \;\;\&\;\; A + 1 < \epsilon\sqrt{d/2},$$

where A and B are independent and zero-mean unit variance gaussians, *i.e.* $A, B \sim \mathcal{N}(0, 1)$. We just need to lower bound the probability of this event. Consider the distribution of A conditioned on $A + B > -2$, suppose its CDF is F, then the probability of the event above is $F(\epsilon\sqrt{d/2} - 1) - F(-1)$. Now, to derive an expression for F,

$$F(a) = P(A \le a | (A + B) > -2) = \frac{P(A \le a \;\&\; ((A + B) > -2))}{P((A + B) > -2)},$$

Using that $A + B \sim \mathcal{N}(0, 2)$, we get

$$F(a) = (1 - \Phi(-\sqrt{2}))) \int_{-\infty}^{a} P(B > -2 - u)\phi(u)du$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of standard normal. Hence, we get that for a suitable constant $\epsilon \ge C/\sqrt{d}$ the probability of this event is lower bounded by 0.95.

E.3 Proof of Theorem 7

Suppose gradient descent is initialized at $\mathbf{w}^0$. Let $\mathbf{w}^t$ be the t^{th} iterate of GD. Note that the gradients of the loss function are always in the span of the covariates $\mathbf{x}_i$. Hence, any iterate of gradient descent lies in $\mathbf{w}^0 + \text{span}(\{\mathbf{x}_i\}_{i=1}^n)$. Let S be the indices corresponding to the non-zero entries in $\mathbf{w}^*$. Since the covariates lie in a low dimensional subspace and are 0 outside the subspace, the co-ordinates of $\mathbf{w}^t$ satisfy the invariant,

$$\mathbf{w}^t_{S^c} = \mathbf{w}^0_{S^c}.$$

Moreover, since we initialized $\mathbf{w}^0$ using a random gaussian initialization with covariance $\frac{1}{\sqrt{d-k}}\mathcal{I}_d$, we know that with high probability,

$$\left\|\mathbf{w}^0_{S^c}\right\|_1 = \sqrt{d-k} \quad \text{and} \quad \left\|\mathbf{w}^0_{S^c}\right\|_2 = O(1)$$

Next, we lower bound the adversarial risk. Suppose we fix $y = 1$, then we have that for any $\mathbf{x} = \mathbf{w}^* + \mathbf{z}_S$. Note that $\mathbf{z}^c_S = 0$.

We can rewrite the $\hat{\mathbf{w}}_{GD} = \mathbf{w} = \mathbf{w}_S + \underbrace{\mathbf{w}_{S^c}}_{=\boldsymbol{\alpha}}$ where $\mathbf{w}_S$ is the component in the low dimensional mixture subspace, and $\boldsymbol{\alpha}$ is the component in the complementary subspace S^c. As stated above, since the covariates lie in a low dimensional subspace, hence, the component in the complementary subspace doesn't get updated. Therefore, $\alpha = \mathbf{w}^0_{S^c}$.

Now, for any x, we have that

$$\begin{aligned}
\hat{\mathbf{w}}^T_{GD} x &= \mathbf{w}^T \mathbf{x} \\
&= \mathbf{w}^T (\mathbf{w}^* + \mathbf{z}_S + \underbrace{\mathbf{z}_{S^c}}_{=0}) \\
&= \mathbf{w}^T_S \mathbf{w}^* + \mathbf{w}^T_S \mathbf{z}_S
\end{aligned}$$

- Consider the event $\mathbf{w}^T_S \mathbf{z}_S > -\mathbf{w}^T_S \mathbf{w}^*$. This is the event that $\mathbf{w}, \mathbf{w}^*$ agree before perturbation.
- Consider the event B such that,
$$\mathbf{w}^T_S \mathbf{z}_S < \|\boldsymbol{\alpha}\|_1 \epsilon - \mathbf{w}^T_S \mathbf{w}^*$$
This is the event that there exists a perturbation restricted to the subspace S^c such that, the prediction of $\hat{w}_{GD}$ changes, *i.e.* $\mathbf{w}^T(\mathbf{x} + \boldsymbol{\delta}) < 0$. Note that since the perturbation is restricted to S^c, the prediction of $\mathbf{w}^*$ doesn't change.
- Hence, both events happen if
$$-\mathbf{w}^T_S \mathbf{w}^* \le \mathbf{w}^T_S \mathbf{z}_S \le \|\boldsymbol{\alpha}\|_1 \epsilon - \mathbf{w}^T_S \mathbf{w}^*$$

- To bound this probability, observe that $\mathbf{w}_S^T \mathbf{z}_S \sim \mathcal{N}(0, \sigma^2 \|\mathbf{w}_S\|_2^2)$. Hence,

$$\Pr\left(-\mathbf{w}_S^T \mathbf{w}^* \leq \mathbf{w}_S^T \mathbf{z}_S \leq \|\boldsymbol{\alpha}\|_1 \epsilon - \mathbf{w}_S^T \mathbf{w}^*\right) = \Phi\left(\frac{\|\boldsymbol{\alpha}\|_1 \epsilon - \mathbf{w}_S^T \mathbf{w}^*}{\sigma \|\mathbf{w}_S\|_2}\right) - \Phi\left(\frac{-\mathbf{w}_S^T \mathbf{w}^*}{\sigma \|\mathbf{w}_S\|_2}\right)$$

- We know that from our initialization, $\|\boldsymbol{\alpha}\|_1 = \left\|w^0_{S^c}\right\|_1 = \sqrt{d-k}$. Hence, for $\epsilon = 2\mathbf{w}_S^T \mathbf{w}^*/(\sqrt{d-k})$, we get that both the events happen with probability,

$$\Phi\left(\frac{\mathbf{w}_S^T \mathbf{w}^*}{\sigma \|\mathbf{w}_S\|_2}\right) - \Phi\left(\frac{-\mathbf{w}_S^T \mathbf{w}^*}{\sigma \|\mathbf{w}_S\|_2}\right) = 2\Phi\left(\frac{\mathbf{w}_S^T \mathbf{w}^*}{\sigma \|\mathbf{w}_S\|_2}\right) - 1$$

- Since as gradient descent progresses, $\mathbf{w}_S \to \mathbf{w}^*$, this implies that for $\sigma = 1$, and $\|w_S\|_2 = 2$, we have that $R_{adv,0-1}(\mathbf{w}_S) > 0.95$ for a very small ϵ such that $\epsilon = \frac{C}{\sqrt{d-k}}$, where $C > 0$ is a small constant.

Plugging this into Theorem 10, we recover the result.

F Proof of Theorem 8

1. First note that $f(\mathbf{x}+\boldsymbol{\delta})$ can be written as

$$f(\mathbf{x}+\boldsymbol{\delta}) = f(\mathbf{x}) + \int_{t=0}^{1} \nabla f(\mathbf{x}+t\boldsymbol{\delta})^T \boldsymbol{\delta} dt.$$

Rearranging the terms gives us:

$$|f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})| \leq \left|\int_{t=0}^{1} \nabla f(\mathbf{x}+t\boldsymbol{\delta})^T \boldsymbol{\delta} dt\right| \leq \epsilon \sup_{\|\boldsymbol{\delta}\| \leq \epsilon} \|\nabla f(\mathbf{x}+\boldsymbol{\delta})\|_*.$$

Let $u(\mathbf{x}) = \epsilon \sup_{\|\boldsymbol{\delta}\| \leq \epsilon} \|\nabla f(\mathbf{x}+\boldsymbol{\delta})\|_*$. Since the loss ℓ is 1-Lipschitz, we can upper bound $\ell(f(\mathbf{x}+\boldsymbol{\delta}), y)$ as

$$\ell(f(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})) - \ell(f(\mathbf{x}), g(\mathbf{x})) \leq |f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})| \leq \epsilon \sup_{\|\boldsymbol{\delta}\| \leq \epsilon} \|\nabla f(\mathbf{x}+\boldsymbol{\delta})\|_*.$$

So we have the following upper bound for the objective in Equation (4)

$$R(f) + \lambda R_{\text{adv}}(f) \leq R(f) + \epsilon\lambda \mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\| \leq \epsilon} \|\nabla f(\mathbf{x}+\boldsymbol{\delta})\|_*\right]. \tag{11}$$

2. We now get a different upper bound for $|\ell(f(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x}+\boldsymbol{\delta})) - \ell(f(\mathbf{x}), g(\mathbf{x}))|$ in terms of $\|f-g\|_\infty$. Since ℓ is 1-Lipschitz we have

$$|\ell(f(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x}+\boldsymbol{\delta})) - \ell(f(\mathbf{x}), g(\mathbf{x}))| \leq |f(\mathbf{x}+\boldsymbol{\delta})g(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})g(\mathbf{x})|.$$

Note that $|f(\mathbf{x})g(\mathbf{x})|$ can be upper bounded by $|f(\mathbf{x}) - g(\mathbf{x})|$. This gives us the following bound

$$|\ell(f(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x}+\boldsymbol{\delta})) - \ell(f(\mathbf{x}), g(\mathbf{x}))| \leq |f(\mathbf{x}) - g(\mathbf{x})| + |f(\mathbf{x}+\boldsymbol{\delta}) - g(\mathbf{x}+\boldsymbol{\delta})|$$

Substituting this in the definition of $R_{\text{adv}}(f)$ gives us the following upper bound for the objective in Equation (4)

$$R(f) + \lambda R_{\text{adv}}(f) \leq R(f) + 2\lambda \|f-g\|_\infty. \tag{12}$$

Combining Equations (11), (12) gives us the required result.

G Proof of Theorem 9

The proof of part (a) and upper bound of part (b) of the Theorem follow from the proof of Theorem 8. Here, we focus on proving the lower bound of part (b). The adversarial risk used in Equation (7) can be rewritten as

$$G_{\text{adv}}(f_{\mathbf{w}}) = \mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell\left(\mathbf{w}^T(\mathbf{x}+\boldsymbol{\delta}), y\right) - \ell\left(\mathbf{w}^T\mathbf{x}, y\right)\right].$$

Since $\ell\left(\mathbf{w}^T(\mathbf{x}+\boldsymbol{\delta}), y\right)$ is maximized at a point where $y\mathbf{w}^T(\mathbf{x}+\boldsymbol{\delta})$ is minimized, we get the following expression for $G_{\text{adv}}(f_{\mathbf{w}})$

$$G_{\text{adv}}(f_{\mathbf{w}}) = \mathbb{E}\left[\ell\left(\mathbf{w}^T\mathbf{x} - y\epsilon\|\mathbf{w}\|_*, y\right) - \ell\left(\mathbf{w}^T\mathbf{x}, y\right)\right].$$

We now obtain a lower bound for $G_{\text{adv}}(f_{\mathbf{w}})$

$$\begin{aligned} G_{\text{adv}}(f_{\mathbf{w}}) &= P(y\mathbf{w}^T\mathbf{x}\leq 0)\times\mathbb{E}\left[\ell\left(\mathbf{w}^T\mathbf{x} - y\epsilon\|\mathbf{w}\|_*, y\right) - \ell\left(\mathbf{w}^T\mathbf{x}, y\right)\Big| y\mathbf{w}^T\mathbf{x}\leq 0\right] \\ &\quad + P(y\mathbf{w}^T\mathbf{x}>0)\times\mathbb{E}\left[\ell\left(\mathbf{w}^T\mathbf{x} - y\epsilon\|\mathbf{w}\|_*, y\right) - \ell\left(\mathbf{w}^T\mathbf{x}, y\right)\Big| y\mathbf{w}^T\mathbf{x}>0\right] \qquad (13) \\ &\geq P(y\mathbf{w}^T\mathbf{x}\leq 0)\times\mathbb{E}\left[\ell\left(\mathbf{w}^T\mathbf{x} - y\epsilon\|\mathbf{w}\|_*, y\right) - \ell\left(\mathbf{w}^T\mathbf{x}, y\right)\Big| y\mathbf{w}^T\mathbf{x}\leq 0\right]. \end{aligned}$$

Consider the logistic loss $\ell(z) = \log 1 + e^{-z}$. For $z<0$, the absolute value of derivative of logistic loss is greater than $\frac{1}{2}$. This shows that for $(\mathbf{x}, y)$ such that $y\mathbf{w}^T\mathbf{x}\leq 0$, we have

$$\ell\left(\mathbf{w}^T\mathbf{x} - y\epsilon\|\mathbf{w}\|_*, y\right) - \ell\left(\mathbf{w}^T\mathbf{x}, y\right) \geq \frac{1}{2}\epsilon\|\mathbf{w}\|_*.$$

This completes the proof of the Theorem. Substituting this in the above lower bound for the adversarial risk $G_{\text{adv}}(f_{\mathbf{w}})$, we get

$$G_{\text{adv}}(f_{\mathbf{w}}) \geq \frac{1}{2}\epsilon R_{0-1}(f_{\mathbf{w}})\|\mathbf{w}\|_*.$$

H Experimental Settings

In all our experiments we use the following network architectures:

MNIST. For all our experiments on MNIST, we use 1 hidden layer neural network with ReLU activations. To control the capacity of the network we vary the number of hidden units.

CIFAR10. For all our experiments on CIFAR10, we use VGG11 network. To control the capacity of the network we scale the number of units in each layer. By a capacity scale of α, we mean that we use α times the number of units in each layer of original VGG network.

PGD Training. In all our experiments we measure adversarial perturbations w.r.t L_∞ norm and use projected gradient descent as our adversary. For PGD training on MNIST, we optimize the inner maximization problem for 50 iterations with step size 0.01. For PGD training on CIFAR10, we optimize the inner maximization problem for 15 iterations with step size 0.005. The outer minimization is run for 40 epochs for MNIST and 50 epochs for CIFAR10 and we use SGD+momentum with learning rate 0.01 and batch size 128.

Computation of adversarial risk. We use adversarial examples generated by PGD to compute the adversarial risk of a classifier. The hyper-parameter settings are the same as the one used for PGD training.