7 Supplementary

7.1 Database Summary

We empirically evaluate our theoretical study on three data sets collected from real-world multi-label applications. They include cyber security practices (*Creepware*), object recognition (*VOC2012*) [4] and environment research (*Planet*) [12]. *Creepware* data include different stalkware app instances and each instance has 16 labels indicating different types of surveillance on the victim's mobile device. Besides, each app is profiled by the introductory texts of the app available in the app stores and signatures of its mobile service access. *VOC2012* is a well-known image data set and it is widely used in multi-label learning research. *Planet* data collects daily satellite imagery of the entire land surface of the earth. Each image is equipped with labels denoting different atmospheric conditions and various classes of land cover/land use.

7.2 Input Normalization and Parameter settings

When imposing attacks, we project the perturbed data in *VOC2012* and *Planet* to $[-1, 1]$, while we don't limit the value range of data in *Creepware*. The α in Eq.(15) is empirically set to 0.01 in all experiments. The regularization parameters λ in Eq.15 and other baselines are chosen empirically from the range $\left\{10^{-8}, 10^{-7}, \cdots, 10^{7}, 10^{8}\right\}$

Our codes were written in Python and all the models were built by Keras package [1]. The needed targeted evasion attack and adversarial training are implemented by adversarial-robustness-toolbox [14]. Our experiments were conducted on GPU rtx2080ti.

7.3 Proofs

We supple the proof of the theorem in our paper, especially the Eq.(3) and Eq.(5), and the proof from Eq.(11) to Eq.(12)

Lemma 2. *(Thomas 2020 [17], Corollary 5) Let E, E' and Z be independent random variables where E and E' have identical distributions. Let A be a random function whose randomness is independent from E, E' and Z. Let g be a fixed function. Then*

$$\begin{aligned} &\mathop{\mathbb{E}}_{A,E,Z}\left[g(A(E,Z),E,Z)\right] \\ &\leq \inf_{t>0} \frac{I(A(E,Z);E|Z)+\mathop{\mathbb{E}}_{Z}\left[\log \mathop{\mathbb{E}}_{A,E,E'Z}\left[e^{t\cdot g(A(E,Z),E',Z)}\right]\right]}{t} \end{aligned} \tag{17}$$

Lemma 3. *(Hoeffding 1963 [11].) Let $\mathbf{X} \in [a, b]$ be a random variable with mean μ. Then for all $t \in \mathbb{R}$,*

$$\mathbb{E}(e^{t\mathbf{X}}) \leq e^{t\mu + t^2(b-a)^2/8} \tag{18}$$

Proof from Eq.11 to Eq.12: We can rewrite Eq.9 as

$$A_{h(x),\tilde{r}} = \tilde{r'} \sum_{k=1}^{m} \frac{-\nabla h_k(x) * \max\{\text{sgn}(-\tilde{r'} y_k \nabla h_k(x)), 0\}}{h_k(x)}. \tag{19}$$

If there is no sgn function and max function in Eq.19, Eq.11 is actually the definition of dual norm. To eliminate the sgn and max function, we can break the domain of $\tilde{r}$ into a group of subsets according the output of those sgn functions. Denote the domain of $\tilde{r}$ as I and I_S is a subset of I which is defined by Eq.(20). S is an element from the power set of $\{1, \cdots, m\}$.

$$I_S = \left\{ \tilde{r} \,\middle|\, \begin{matrix} \tilde{r} y_k \nabla h_k < 0, k \in S \\ \tilde{r} y_k \nabla h_k \geq 0, k \notin T \end{matrix}, \tilde{r} \in \mathbb{R}^n \right\} \tag{20}$$

Based on Eq.(20), we redefine Eq.(9) and Eq.(11) over the sub-domain I_S of $\mathbf{r}$ as:

$$A_{h(x),\tilde{r}s} = \sum_{k \in S} \frac{-\tilde{r}' \nabla h_k(x)}{h_k(x)} \tag{21}$$

$$\begin{aligned} &\phi_s = \max_{\tilde{r} \in I_S} A_{h(x),\tilde{r}s}, \\ &s.t. \ \left\|\tilde{r}\right\|_p = 1 \end{aligned} \tag{22}$$

Now, we get $\phi_{h,x} = \max\limits_{S \in P(S)} \phi_s$. It's easy to know that:

$$\begin{aligned} &\phi_s = \max_{\tilde{r} \in I_S} A_{h(x),\tilde{r}s}, \\ &s.t. \ \left\|\tilde{r}\right\|_p = 1 \end{aligned} \leq \begin{aligned} &\phi_s = \max_{e \in \mathbb{R}^n} A_{h(x),es}, \\ &s.t. \ \left\|e\right\|_p = 1 \end{aligned} = \left\| \sum_{k \in S} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q \tag{23}$$

The equality holds when the optimal e^* exactly locates in I_S. Now, if we want to prove that Eq.(11) = Eq.(12), we just need to prove that $\phi_{S^*} = \left\| \sum\limits_{k \in S^*} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q$, that is we need to prove that the optimal e^* for S^* locates in I_{S^*}. We can prove that by contradiction. That is we assume $e^*_{S^*} \in I_{S'} (S' \neq S^*)$, then it is proved by Eq.(24).

$$\begin{aligned} \left\| \sum_{k \in S^*} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q &= \sum_{k \in S^*} \frac{-e^*_{S^*} \nabla h_k(x)}{h_k(x)} \\ &< \sum_{k \in S^* \cap S'} \frac{-e^*_{S^*} \nabla h_k(x)}{h_k(x)} \\ &\leq \left\| \sum_{k \in S^* \cap S'} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q \\ &< \left\| \sum_{k \in S^*} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q \end{aligned} \tag{24}$$

Proof of Eq.(3): We define the worst-case loss $l(h, z, \varepsilon)$ as:

$$\begin{aligned} &l(h, z, \varepsilon) = \max_{z' \in N(z)} l(h, z'), \\ &\text{where } N(z) = \left\{ (x', y') \,\middle|\, \left\|x' - x\right\|_p \leq \varepsilon, \, y' = y \right\}. \end{aligned} \tag{25}$$

We first upperly bound $l(h, z, \varepsilon)$ defined in Eq.(25) with the setting of linear classifier and hinge loss:

$$\begin{aligned} & l(h,z,\varepsilon) \le l(h,z) + \\ \max_{\|r\|_2 \le \varepsilon} & \left\| \sum_{k=1}^{m} y_k r' \cdot \mathbf{w}_k * \max\{\mathrm{sgn}(y_k r' \cdot \mathbf{w}_k), 0\} \right\|_2 \\ & \le l(h,z) + C_{\mathbf{W},z}\varepsilon. \end{aligned} \tag{26}$$

The last step borrows the proof from Eq.11 to Eq.12. Then we have

$$\begin{aligned} & R_{\mathcal{D}}(A,\varepsilon) - R_{Z^n}(A,\varepsilon) \\ & = \mathop{\mathbb{E}}_{Z^n \leftarrow \mathcal{D}^n, A} l(A(Z^n), \mathcal{D}, \varepsilon) - \mathop{\mathbb{E}}_{Z^n \leftarrow \mathcal{D}^n, A} l(A(Z^n), Z^n, \varepsilon) \\ & = \mathop{\mathbb{E}}_{\bar{Z}, E, A} \left[l(A(\bar{Z}_E), \bar{Z}_{\bar{E}}, \varepsilon) - l(A(\bar{Z}_E), \bar{Z}_E, \varepsilon) \right] , \quad (\bar{Z} \leftarrow \mathcal{D}^{n\times 2}) \\ & = \mathop{\mathbb{E}}_{\bar{Z}, E, A} \left[f_{\bar{Z}}(A(\bar{Z}_E), E, \varepsilon) \right] \\ & \quad by\ LEMMA\,2 \\ & \le \inf_{t>0} \frac{I(A(\bar{Z}_E); E|\bar{Z}) + \mathop{\mathbb{E}}_{\bar{Z}} \left[\log \mathop{\mathbb{E}}_{\mathbf{W}, E'} \left[e^{t f_{\bar{Z}}(\mathbf{W}, E', \varepsilon)} \right] \right]}{t}, \\ & \quad by\ independence \\ & = \inf_{t>0} \frac{CMI_{\mathcal{D},A} + \mathop{\mathbb{E}}_{\bar{Z}} \left[\log \mathop{\mathbb{E}}_{\mathbf{W}} \left[\prod_{i=1}^{n} \mathop{\mathbb{E}}_{E'_i} \left[e^{\frac{t}{n}(l(W, (\bar{Z}_{E'})_i, \varepsilon) - l(\mathbf{W}, (\bar{Z}_{\bar{E}'})_i, \varepsilon))} \right] \right] \right]}{t}, \\ & = \inf_{t>0} \frac{CMI_{\mathcal{D},A} + \mathop{\mathbb{E}}_{\bar{Z}} \left[\log \mathop{\mathbb{E}}_{\mathbf{W}} \left[\prod_{i=1}^{n} \mathop{E}_{E'_i} \left[e^{\frac{t}{n}(1-2E'_i)(l(\mathbf{W}, \bar{Z}_{i,1}, \varepsilon) - l(\mathbf{W}, \bar{Z}_{i,2}, \varepsilon))} \right] \right] \right]}{t} \\ & \quad by\ LEMMA\ 3 \\ & \le \inf_{t>0} \frac{CMI_{\mathcal{D},A} + \mathop{\mathbb{E}}_{\bar{Z}} \left[\log \mathop{\mathbb{E}}_{\mathbf{W}} \left[\prod_{i=1}^{n} e^{\frac{t^2}{2n^2}(l(\mathbf{W}, \bar{Z}_{i,1}, \varepsilon) - l(\mathbf{W}, \bar{Z}_{i,2}, \varepsilon))^2} \right] \right]}{t}, \\ & \le \inf_{t>0} \frac{CMI_{\mathcal{D},A}}{t} + \frac{t}{2n} \mathop{\mathbb{E}}_{\bar{Z}} \left[\sup_{\mathbf{W} \in \mathcal{W}_A} \frac{1}{n} \sum_{i=1}^{n} \left(l(\mathbf{W}, \bar{Z}_{i,1}, \varepsilon) - l(\mathbf{W}, \bar{Z}_{i,2}, \varepsilon) \right)^2 \right] \\ & \le \inf_{t>0} \frac{CMI_{\mathcal{D},A}}{t} + \frac{t}{2n} \mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}} \left[\sup_{\mathbf{W} \in \mathcal{W}_A} l(\mathbf{W}, Z, \varepsilon)^2 \right] \\ & \le \inf_{t>0} \frac{CMI_{\mathcal{D},A}}{t} + \frac{t}{2n} \mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}} \left[\sup_{\mathbf{W} \in \mathcal{W}_A} \left(l(\mathbf{W}, Z) + C_{\mathbf{W},Z} \cdot \varepsilon \right)^2 \right] \\ & = \sqrt{\frac{2}{n} CMI_{\mathcal{D},A} \cdot \mathop{\mathbb{E}}_{Z \leftarrow \mathcal{D}} \left[\sup_{\mathbf{W} \in \mathcal{W}_A} \left(l(\mathbf{W}, Z) + C_{\mathbf{W},Z} \cdot \varepsilon \right)^2 \right]} \end{aligned} \tag{27}$$

Proof of Eq.(5): Here we use H to denote the entropy.

$$
\begin{aligned}
&CMI_{\mathcal{D},A} \\
&= I(A; S, \bar{Z}) - I(A; \bar{Z}) \\
&= H(A) + H(S, \bar{Z}) - H(A, S, \bar{Z}) - H(A) - H(\bar{Z}) + H(A, \bar{Z}) \\
&= H(A, \bar{Z}) + H(S|\bar{Z}) - H(S) - H(A, \bar{Z}|S) \quad : S \; is independent to \, Z \\
&= H(A, \bar{Z}) - H(A, \bar{Z}|S) \qquad (28) \\
&\leq H(A, \bar{Z}) \\
&\leq H(A) + H(\bar{Z}) \\
&= H(\mathbf{W}) + H(\bar{Z}) \\
&= ent(\mathbf{w}_1, \cdots, \mathbf{w}_m) + ent(\mathcal{D}_1, \cdots, \mathcal{D}_m)
\end{aligned}
$$