Exam for the Lecture
Adversarial Machine Learning
Winter Term 2024/2025

Chair of Machine Learning and Security
BIFOLD & Technische Universität Berlin
Prof. Dr. Konrad Rieck

Test Exam

Winter Term 2024/2025

First name: .....

Last name: .....

Student number: .....

Important Notes

1. The exam has a total duration of 90 minutes.
2. Write your answers in the designated boxes provided. If additional space is required, use the extra pages at the end of the exam.
3. The achievable points for each task are indicated in square brackets next to the task.
4. No additional aids such as calculators, dictionaries, or other resources are allowed.
5. Before submitting your exam, carefully verify its completeness (9 pages).

Task	Points	Score
General questions	9	
White-box Adversarial Examples	14	
Black-box Adversarial Examples	14	
Model Stealing Attacks	14	

Task	Points	Score
Membership Inference Attacks	13	
Poisoning Attacks	13	
Backdoor Attacks	13	
Total	90	

Task 1: General questions ..... (9 points)

The exam includes various multiple-choice questions. All questions follow the *k-prim* format and your task is to determine whether each statement is true or false. If your decision is correct for all four statements, you get 100% of the points. If your decision is correct for three of the statements, you receive 50% of the points. If two or fewer statements are correct, you receive no points.

(a) Consider the security cycle of a machine learning system. Which of the following statements are correct? [3]

- ■ **The reactive security process only develops defenses after detecting an attack.**
- □ The proactive security process does not model adversaries since it can simulate attacks.
- ■ **The proactive security process aims to anticipate and prevent attacks.**
- □ The reactive security process is always more effective than the proactive process.

(b) Problem-space attacks aim to generate adversarial examples that remain valid in real-world applications. Which of the following statements are correct? [3]

- □ Attacks in the problem space can be fully characterized by common L_p-norm constraints, just like feature-space attacks.
- □ Feature-space and problem-space attacks are equivalent if the attacker knows the model parameters *and* the training data.
- □ Problem-space attacks require knowledge of a model's feature extractor and decision-making function.
- ■ **Problem-space attacks often require structured transformations that preserve syntactic and semantic constraints of the input domain.**

(c) System-level attacks aim to attack machine learning systems by focusing on the components around the model. Which of the following components can be attacked? [3]

- ■ **Pre-processing steps, such as feature extraction and normalization.**
- ■ **Inference platform, such as differences in the employed linear algebra backends.**
- ■ **Data filtering mechanisms, such as data deduplication components.**
- ■ **Data encoding steps, such as text or image encoding.**

Task 2: White-box Adversarial Examples ..... (14 points)

(a) A key step in many white-box adversarial examples is computing gradients. Which of the following statements are correct? [3]

- ☐ The gradient is computed with respect to the model's parameters.
- ■ **The gradient is computed with respect to the input.**
- ☐ White-box adversarial examples require access to the model's loss function.
- ☐ Adversarial examples are generated by updating the model parameters based on the computed gradient.

(b) Consider Projected Gradient Descent (PGD) used to compute untargeted adversarial examples. Given a sample (x, y), step size α, projection operator Π, and loss function $\ell(\hat{y}, y)$, provide the formal iterative definition of the attack at step k. [4]

Solution:

$$\delta_k = \Pi(\delta_{k-1} + \alpha \cdot \mathsf{sign}(\nabla_{\delta_{k-1}} \ell(f(x + \delta_{k-1}), y)))$$

(c) Perform two iterations of PGD using the L_∞-norm with $\epsilon = 0.75$. Assume the initial point is given by $x = \binom{1}{1}$ and the step size is set to $\alpha = 0.5$. The perturbation is initialized with $\delta = \binom{0}{0}$ and the gradients for the first two iterations are $g_1 = \binom{1}{1}$ and $g_2 = \binom{1.5}{-0.5}$. What is the adversarial example after two iterations of the method? [4]

Solution:

$$\begin{aligned}
\delta_1 &= \Pi(\tbinom{0}{0} + 0.5 \cdot \mathsf{sign}(g_1)) \\
&= \Pi(\tbinom{0}{0} + 0.5 \cdot \mathsf{sign}(\tbinom{1}{1})) \\
&= \tbinom{0.5}{0.5} \\
\delta_2 &= \Pi(\tbinom{0.5}{0.5} + 0.5 \cdot \mathsf{sign}(g_2)) \\
&= \Pi(\tbinom{0.5}{0.5} + 0.5 \cdot \mathsf{sign}(\tbinom{1.5}{-0.5})) \\
&= \Pi(\tbinom{1}{0}) \\
&= \tbinom{0.75}{0} \\
\overline{x} &= \tbinom{1.75}{1}
\end{aligned}$$

(d) Modify the definition of PGD to a targeted attack for a class $t \neq y$. [3]

Solution: The gradient should be computed using the target and the loss needs to be minimized instead of maximized:

$$\delta_k = \Pi(\delta_{k-1} - \alpha \cdot \mathsf{sign}(\nabla_{\delta_{k-1}} \ell(f(x + \delta_{k-1}), t)))$$

Task 3: Black-box Adversarial Examples . . . . . (14 points)

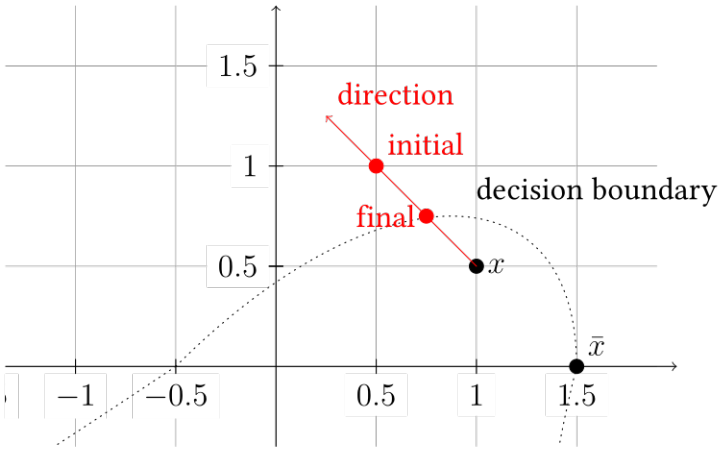
(a) In black-box attacks, the adversary has limited access to the model. Which of the following statements are correct? [3]

- ☐ In a black-box attack, the adversary has only access to the model's parameters but not the training data and model architecture.
- ■ **Black-box attacks operate by querying the model's prediction function and analyzing its outputs.**
- ☐ Black-box attacks cannot be applied if the prediction function is non-linear.
- ☐ Hard-label attacks are restricted to attacking only a specific label of the learning model.

(b) Many black-box attacks rely on the random walk framework, differing primarily in the generation procedure. Given the base point (x, y) and the current best adversarial example $\bar{x}$, the RayS attack defines the generation as: [8]

1: Choose new direction $w \in \{-1, +1\}^d$
2: **if** $f(x + r \cdot \bar{w}) = y$ **then**
3: **return** skip direction
4: $s = 0, e = r$
5: **while** $e - s > \varepsilon$ **do**
6: $m = (s + e)/2$
7: **if** $f(x + m \cdot \bar{w}) = y$ **then**
8: $e \leftarrow m$
9: **else**
10: $s \leftarrow m$
11: **return** $x + (s + e)/2 \cdot \bar{w}$

Perform the generation algorithm graphically using the diagram below. Assume that $w = \binom{-1}{1}$ is sampled in line 1. Draw the direction from the base point, the initial point, and the final point. Will the framework accept the resulting point? Explain your decision.



Solution: The final point will be accepted, as it is closer to the base point.

(c) How can the attack be adjusted to minimize the number of bad queries? Propose and explain one possible modification. [3]

Solution: The number of bad queries can be reduced by exchanging the binary search for the radius. For example, a line search can be used starting from the previous best radius and stopping when the boundary is crossed for the first time. This would only ever incur one bad query per candidate.

Task 4: Model Stealing Attacks ..... (14 points)

(a) The Jacobian-based dataset augmentation (JBDA) is a technique that generates queries to approximate a target model. Which of the following statements are correct? [3]

- ■ **JBDA estimates the decision boundary by computing the Jacobian of a locally trained model and generating new queries from it.**
- □ JBDA requires access to the target model's gradients to generate informative queries close to the decision boundary.
- ■ **JBDA expands the adversary's dataset iteratively, improving the stolen model's accuracy over time.**
- ■ **JBDA has limitations for model stealing because it explores the decision boundary locally, making it less effective for complex decision boundaries.**

(b) Consider a single-layer neural network, defined as [7]

$$f_\theta(x) = \log_2(\theta_1 \cdot x_1 + 1) + \log_2(\theta_2 \cdot x_2 + 1) + \theta_3.$$

Examine the following set of input-output pairs and perform an equation-solving attack to determine model parameters θ:

$$f_\theta(\binom{0}{0}) = 0.3, \quad f_\theta(\binom{1}{0}) = 2.3, \quad f_\theta(\binom{0}{1}) = 1.3\,.$$

Solution:

$$\theta_1 = 3, \theta_2 = 1, \theta_3 = 0.3$$

(c) Name the two different types of stealing. Briefly explain the difference. [4]

Solution: The attacker might focus on either fidelity or performance. A focus on performance is given when the attacker wants to steal the model functionality. Fidelity is needed to use the stolen model as a surrogate for further adversarial uses.

Task 5: Membership Inference Attacks ..... (13 points)

(a) The Likelihood Ratio Attack (LiRA) by Carlini et al. is a membership inference attack that improves upon previous methods. Which of the following statements are correct? [3]

- ☐ LiRA requires access to the target model's parameters and gradients to estimate the membership likelihood.
- ■ **LiRA performs likelihood ratio testing to compare the probability of a sample being in the training set versus being out of it.**
- ☐ LiRA assumes that the distributions of in-training and out-training samples are identical for most learning models.
- ☐ LiRA builds on a machine-learning classifier for detecting members, similar to the attack by Shokri & Shmatikov.

(b) Suppose you want to perform a classification-based membership inference attack on a binary classifier. To do so, you prepared two shadow datasets $\tilde{D}_A$ and $\tilde{D}_B$ such that [7]

$$\{a, b, c, d, e\} \subset \tilde{D}_A \ \wedge \ \tilde{D}_A \cap \{f, g, h, i, j\} = \emptyset$$
$$\{f, g, h, i, j\} \subset \tilde{D}_B \ \wedge \ \tilde{D}_B \cap \{a, b, c, d, e\} = \emptyset\,,$$

indicating member and non-membership of data points a to j.

Next, you train two models θ_A and θ_B using $\tilde{D}_A$ and $\tilde{D}_B$, respectively. The table below contains confidence scores from these shadow models for data points a to j:

Data point	a	b	c	d	e	f	g	h	i	j
Model	A	B	A	B	A	B	A	B	A	B
$f_{\theta_i}(x)$	0.42	0.55	0.40	0.80	0.15	0.35	0.45	0.40	0.85	0.10

Determine a classification threshold τ that separates data points as either member or non-member based on the confidence score of the shadow models. Ensure that no sample is misclassified by the chosen threshold. Justify your answer.

Solution: Any $\tau \in [0.42, 0.45]$ is a correct solution.

(c) How can the attack model be targeted to each class? [3]

Solution: To make membership inference attack models targeted to each class, we would need to train a separate attack model for each class.

Task 6: Poisoning Attacks ..... (13 points)

(a) Label-flip poisoning attacks can be formulated as an optimization problem. Which of the following statements are correct? [3]

- ☐ The optimization is performed directly on discrete labels without any relaxation, providing advantages over other attacks.
- ■ **Label-flip attacks can be phrased as optimizing a bilevel problem, maximizing test loss and minimizing training loss.**
- ☐ Label-flip attacks can be used for clean-label attacks if during the optimization the label flips are gradually reversed.
- ☐ Label-flip attacks are limited to models with two classes, as for more classes the optimization problem becomes undecidable.

(b) You are preparing to perform a clean-label poisoning attack on a target t. During the execution of the attack algorithm, you compute the gradient at step i as $\nabla_{\hat{x}_i} d(\hat{x}_i)$. Provide the formal definition of $d(\hat{x}_i)$ as it appears in the forward step of the algorithm. [3]

Solution:

$$d(\hat{x}_i) = \|z(\hat{x}_i) - z(t)\|$$

(c) Suppose the gradient is given by $\nabla_{\hat{x}_i} d(\hat{x}_i) = \binom{0.02}{0.04}$ and parameter λ is set to $\lambda = 0.5$. Compute the forward step for $\hat{x}_i = \binom{0.1}{0.3}$. [4]

Solution:

$$\begin{aligned}\hat{x}_{i+1} &= \hat{x}_i - \lambda \nabla_{\hat{x}_i} d(\hat{x}_i) \\ &= \tbinom{0.1}{0.3} - \tbinom{0.01}{0.02} \\ &= \tbinom{0.09}{0.28}\end{aligned}$$

(d) Formally define the backward step as a function of the base point b, parameter β and λ. What constraint is optimized by the backward step? [3]

Solution:

$$\hat{x}_{i+1} = (\hat{x}_{i+1} + \lambda\beta b)/(1 + \lambda\beta)$$

The backward step optimizes the deviation from the base point to ensure that the perturbed point does not deviate too far.

Task 7: Backdoor Attacks . . . . . (13 points)

(a) Which of the following statements are correct? [3]

- ☐ Data-based backdoor attacks manipulate training samples by embedding a trigger in the input while keeping the label unchanged.
- ☐ Minimal backdoors as introduced in the lecture aim to reduce the size of the trigger while maintaining high attack effectiveness.
- ☐ Model-based backdoor attacks require access to the original training data for fine-tuning the learning model with triggers.
- ☐ A backdoor attack in practice optimizes only the attack success rate.

(b) Assume that we want to perform a model-based backdoor attack. During trigger generation, we optimize a trigger τ to induce target values t_k at the k-th layer. You are given the following values: [8]

$$x = \begin{bmatrix} 226 & 5 & 231 \\ 46 & 125 & 45 \\ 35 & 255 & 201 \end{bmatrix}, \nabla_x \|f_k(x) - t_k\| = \begin{bmatrix} 0.11 & 0.82 & 0.56 \\ -0.67 & 0.15 & 0.48 \\ 0.30 & -0.42 & -0.32 \end{bmatrix}, m = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Perform one step of the trigger generation with step size $\alpha = 1$. Calculate the updated point x and give the trigger τ extracted from x.

Solution:

$$x = \begin{bmatrix} 226 & 5 & 231 \\ 46 & 124.85 & 44.52 \\ 35 & 255.42 & 201.32 \end{bmatrix} \quad \tau = m \circ x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 124.85 & 44.52 \\ 0 & 255.42 & 201.32 \end{bmatrix}$$

(c) We can reduce the number of modified model parameters in a model-based backdoor by incorporating L_0 regularization into the learning objective. What practical consideration might arise when including such a regularization? [2]

Solution: L_0 regularization is non-differentiable, making it difficult to optimize using standard gradient-based methods.