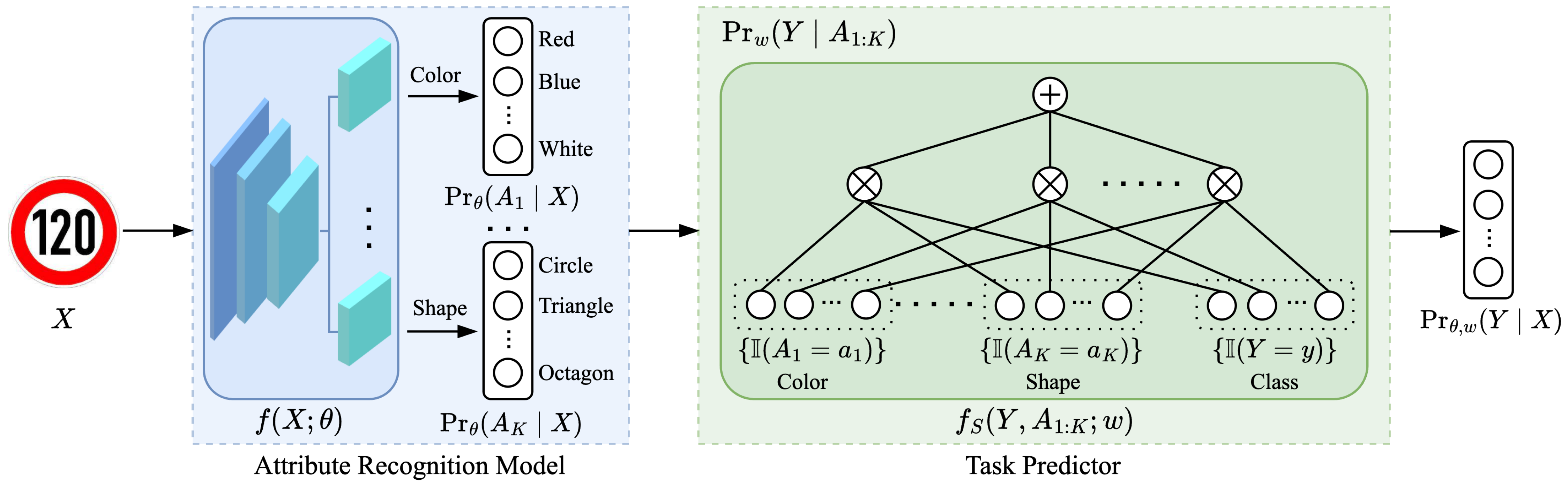
Understanding and Improving Adversarial Robustness of Neural Probabilistic Circuits

Weixin Chen, Han Zhao
University of Illinois Urbana-Champaign

Nov 2, 2025

Research Questions

- Neural Probabilistic Circuits (NPCs), a recent class of **concept bottleneck models**, combine a neural-network-based attribute prediction model and a **probabilistic-circuit**-based task predictor, enhancing interpretability and downstream performance.



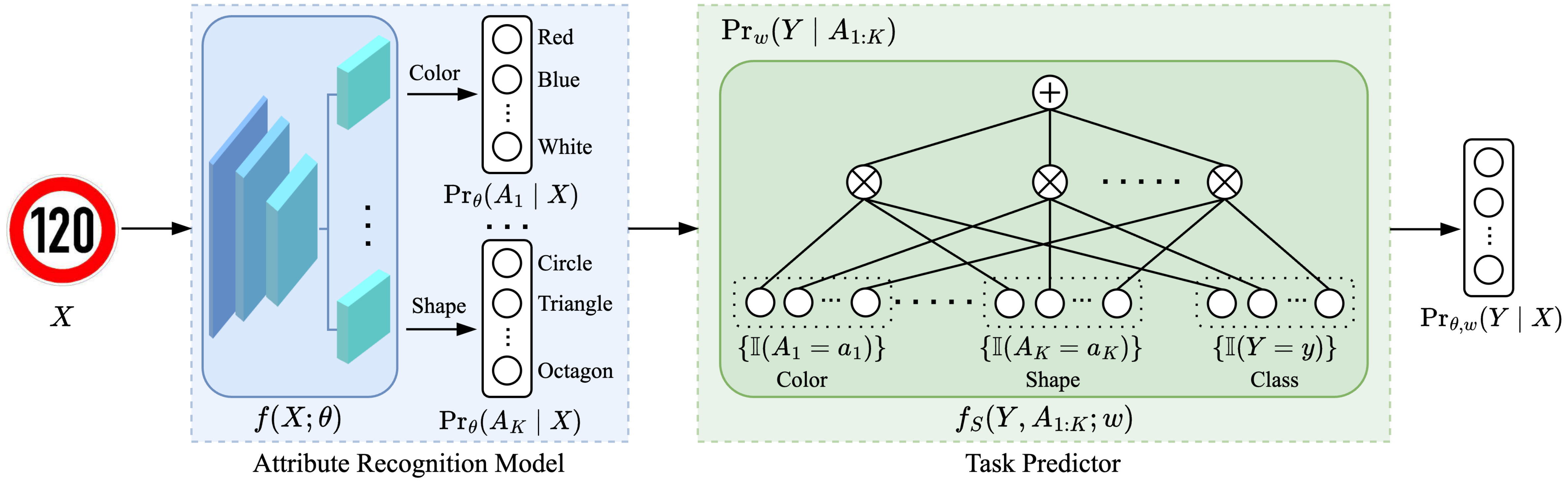
- However, the neural component remains a black box, leaving NPCs vulnerable to **adversarial attacks**.

- Question 1: *How robust are NPCs to adversarial attacks?*

- Question 2: *How can we improve their adversarial robustness?*

Preliminaries

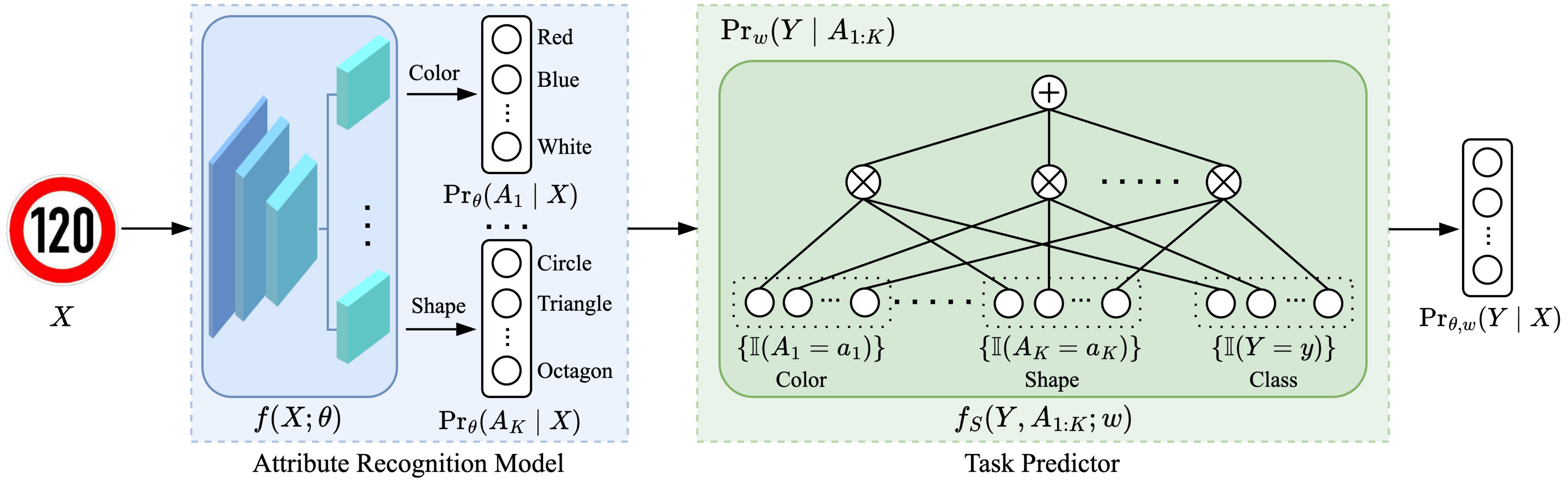
Neural Probabilistic Circuits



- A probabilistic circuit is a computational graph used to represent the joint distribution over a set of random variables. A smooth and decomposable probabilistic circuit supports **tractable** probabilistic reasoning tasks.

Preliminaries

Neural Probabilistic Circuits



Assumption 3.1 (Sufficient Attributes):

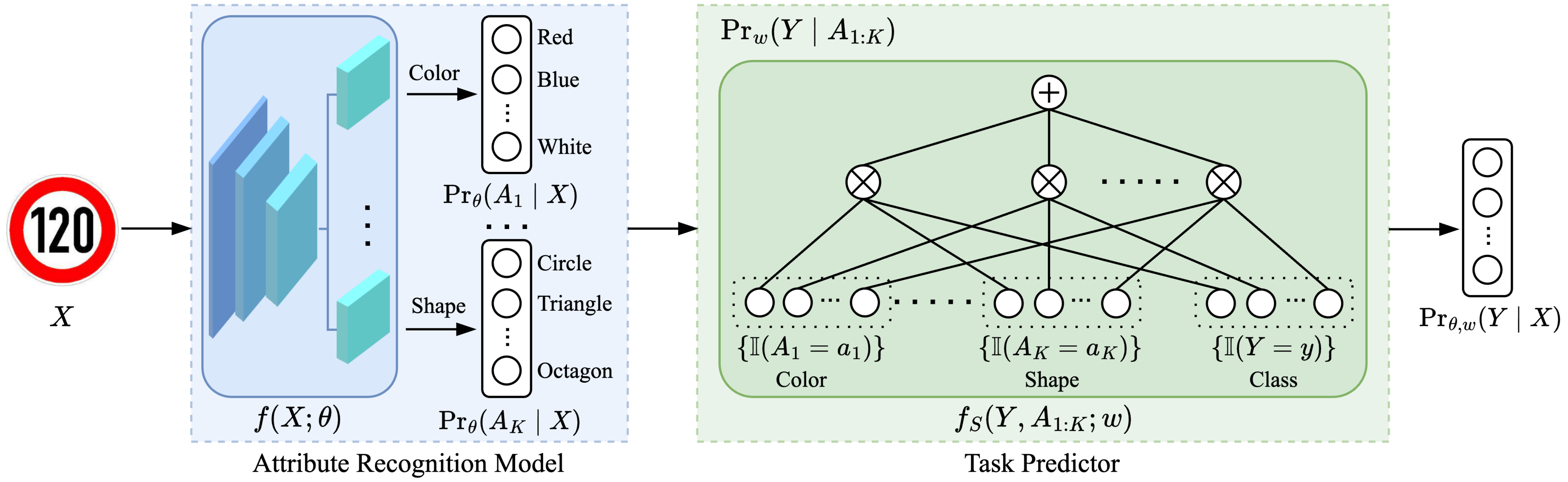
The class label Y and the input X are conditionally independent given the attributes $A_{1:K}$, i.e., $Y \perp X \mid A_{1:K}$.

Assumption 3.2 (Complete Information):

Given any input, all attributes are conditionally mutually independent, i.e., $A_1 \perp A_2 \perp \cdots \perp A_K \mid X$.

Preliminaries

Neural Probabilistic Circuits



- NPC Inference:

$$\mathbb{P}_{\theta,w}\left(Y = y \mid X = x\right) = \sum_{a_{1:K}} \mathbb{P}_\theta\left(A_{1:K} = a_{1:K} \mid X = x\right) \cdot \mathbb{P}_w\left(Y = y \mid A_{1:K} = a_{1:K}, X = x\right)$$

$$= \underbrace{\sum_{a_{1:K}} \prod_{k=1}^{K} \mathbb{P}_{\theta_k}\left(A_k = a_k \mid X = x\right)}_{\text{attribute recognition model}} \cdot \underbrace{\mathbb{P}_w\left(Y = y \mid A_{1:K} = a_{1:K}\right)}_{\text{task predictor}}$$

Preliminaries

Threat Model

- Consider a white-box, norm-bounded, untargeted adversarial attack against the **attribute recognition model**.
- Given an input $(x, a_{1:K})$, the attacker seeks to find a perturbed input $\tilde{x} \in \mathbb{B}_p(x, \ell)$, such that one or more attribute predictions become incorrect.

How robust are NPCs to adversarial attacks?

Definition 3.3 (Prediction Perturbation of NPCs):
It is defined as the worst-case TV distance between the class distributions conditioned on the vanilla and perturbed inputs, i.e.,

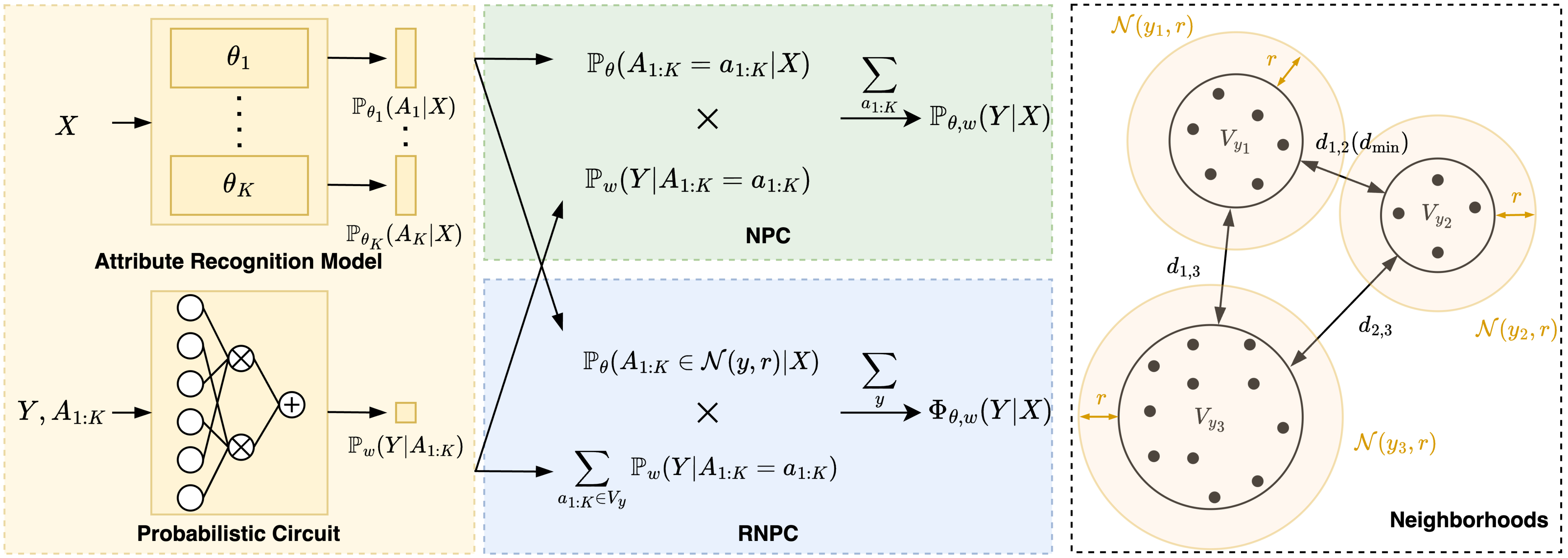
$$\Delta_{\theta,w}^{NPC} := \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} d_{\mathrm{TV}} \left(\mathbb{P}_{\theta,w}(Y \mid X), \mathbb{P}_{\theta,w}(Y \mid \tilde{X}) \right) \right].$$

Theorem 3.4 (Adversarial Robustness of NPCs):

$$\Delta_{\theta,w}^{NPC} \leqslant \underbrace{\mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} d_{\mathrm{TV}} \left(\mathbb{P}_\theta \left(A_{1:K} \mid X \right), \mathbb{P}_\theta \left(A_{1:K} \mid \tilde{X} \right) \right) \right]}_{\Lambda_{NPC}} \leqslant \sum_{k=1}^{K} \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} d_{\mathrm{TV}} \left(\mathbb{P}_{\theta_k} \left(A_k \mid X \right), \mathbb{P}_{\theta_k} \left(A_k \mid \tilde{X} \right) \right) \right].$$

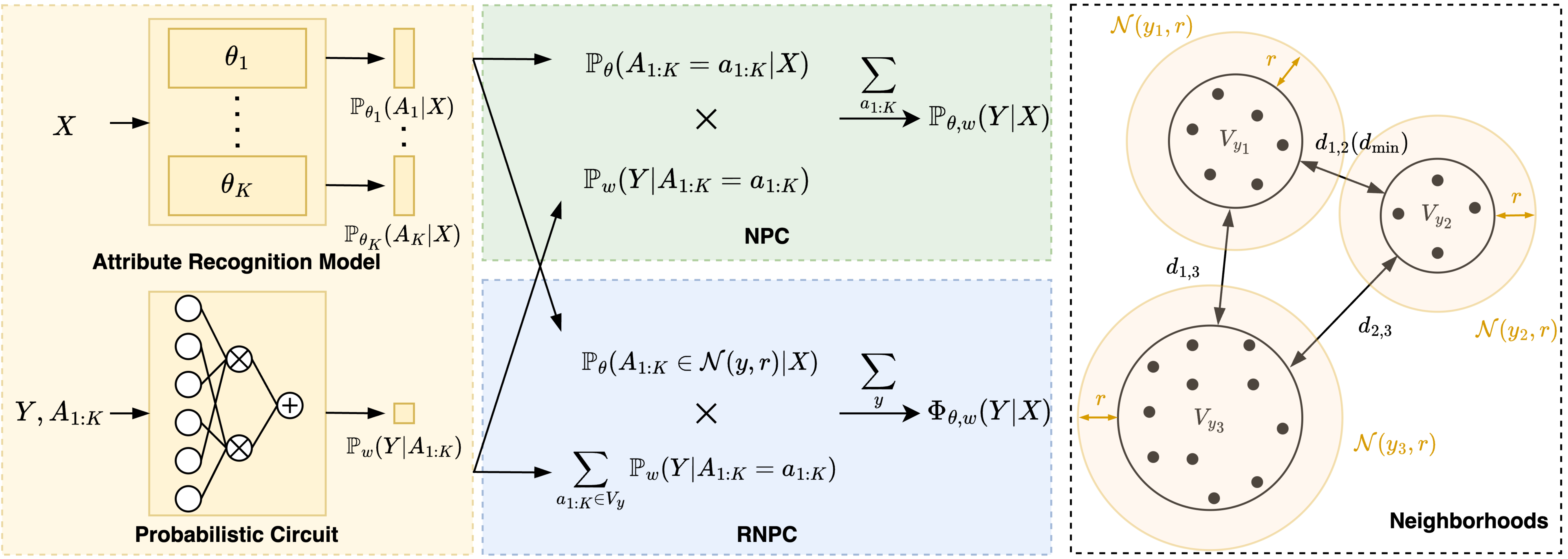
The robustness of NPC depends solely on that of the attribute recognition model. Adding a probabilistic circuit on top does not affect the robustness of NPC.

How to improve adversarial robustness of NPCs?



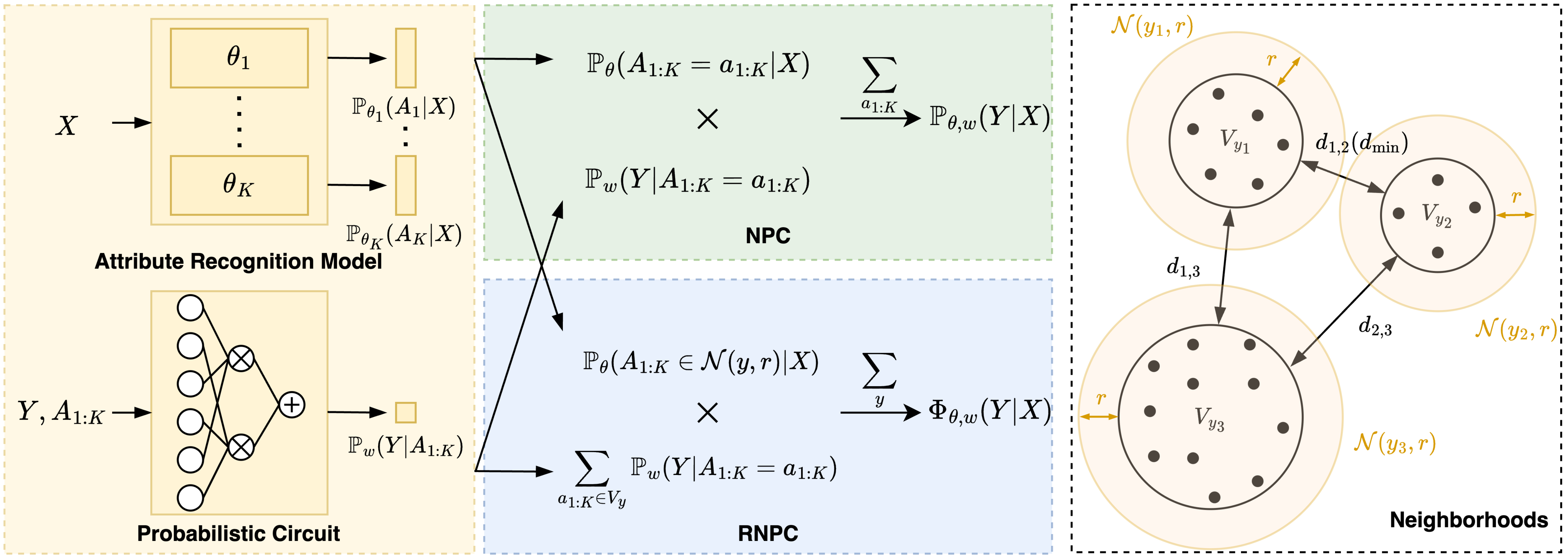
- Dataset: $D = \{(x, a_{1:K}, y)\}$
- Attribute set with a high prob mass: $V = \{a_{1:K} : \mathbb{P}_D(A_{1:K} = a_{1:K}) \geqslant \gamma\}$
- Partition V according to the most probable class of $a_{1:K}$
- Inter-class distance: $d_{i,j} = \min_{v_i \in V_i, v_j \in V_j} \{\text{Hamming}(v_i, v_j)\}$
- Radius: $r = \lfloor (d_{\min} - 1)/2 \rfloor$, where $d_{\min} = \min_{i \neq j} \{d_{i,j}\}$
- Neighborhood: $\mathcal{N}(y, r) := V_y \bigcup \{a^c_{1:K} \in V^c : \min_{a_{1:K} \in V_y} \text{Hamming}\left(a^c_{1:K}, a_{1:K}\right) \leqslant r\}$

How to improve adversarial robustness of NPCs?



- Intuition
  - If an attacker perturbs $m \leqslant r$ attributes, then the probabilities $\mathbb{P}_\theta(A_{1:K}|X)$ originally assigned to V_{y*} will now be assigned to $\mathcal{N}(y*, r) \setminus V_{y*}$.
  - We can **aggregate** these perturbed probabilities to alleviate the impact of attacks.

How to improve adversarial robustness of NPCs?



- RNPC Inference
  - Introduce the class-wise integration, instead of the node-wise integration.

$$\Phi_{\theta,w}(Y \mid X) = \sum_{\tilde{y} \in \mathcal{Y}} \left(\mathbb{P}_\theta \left(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X \right) \cdot \sum_{a_{1:K} \in V_{\tilde{y}}} \mathbb{P}_w \left(Y \mid A_{1:K} = a_{1:K} \right) \right)$$

How robust are RNPCs to adversarial attacks?

Theoretical Results

Definition 4.5 (Prediction Perturbation of RNPCs):

It is defined as the worst-case TV distance between the class distributions conditioned on the vanilla and perturbed inputs,

$$\Delta^{RNPC}_{\theta,w} := \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} d_{\mathrm{TV}} \left(\hat{\Phi}_{\theta,w}(Y \mid X), \hat{\Phi}_{\theta,w}(Y \mid \tilde{X}) \right) \right].$$

Lemma 4.6 (Adversarial Robustness of RNPCs):

$$\Delta^{RNPC}_{\theta,w} \leqslant \underbrace{\mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} \left\{ \max_{\tilde{y} \in \mathcal{Y}} \left| 1 - \frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \right| \right\} \right]}_{\Lambda_{RNPC}}.$$

Theorem 4.7 (Comparison in Adversarial Robustness):

Assume the attribute recognition model is ϵ-DP. Under certain conditions,

$\Lambda_{NPC} \leqslant \frac{|\mathcal{A}_1| \ldots |\mathcal{A}_K|}{2} \alpha_\epsilon$ and $\Lambda_{RNPC} \leqslant \alpha_\epsilon$, where $\alpha_\epsilon := \max\{1 - e^{-K\epsilon}, e^{K\epsilon} - 1\}$

Compared to Λ_{RNPC}, Λ_{NPC} is bounded by an exponentially larger value that scales exponentially with the number of attributes.

How robust are RNPCs to adversarial attacks?

Empirical Results

- Adversarial Performance

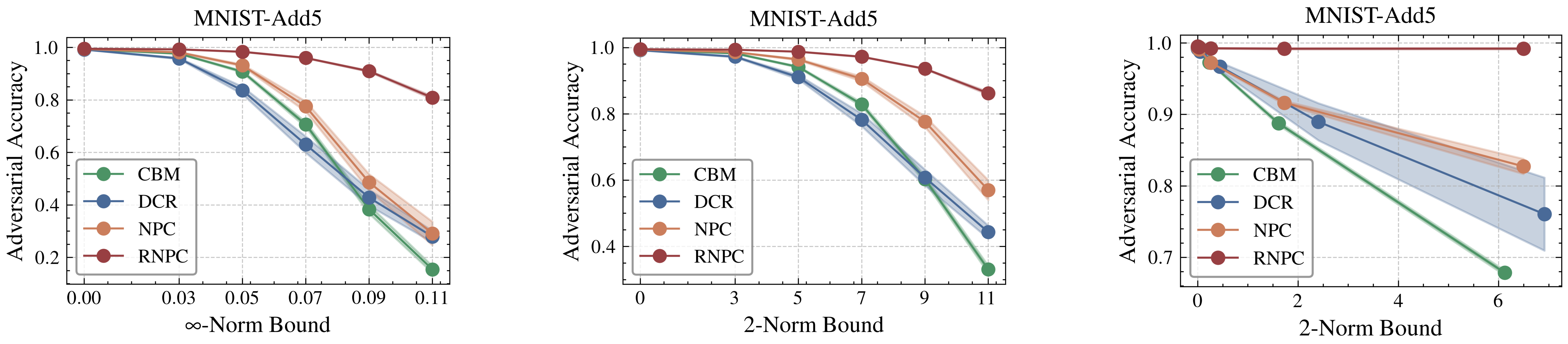


Figure 1: Adversarial accuracy under the ℓ_∞ and ℓ_2-bounded PGD attack. The attacker attacks a single attribute at a time.

RNPC achieves superior robustness against diverse adversarial attacks compared to various concept bottleneck models.

- Benign Performance

Table 1: Benign accuracy on four image classification datasets.

Dataset	**CBM**	**DCR**	**NPC**	**RNPC**
MNIST-Add3	99.02	98.54	99.32	**99.37**
MNIST-Add5	99.37	99.21	99.40	**99.51**
CelebA-Syn	99.83	99.45	**99.95**	**99.95**
GTSRB-Sub	99.42	99.42	**99.57**	99.49

RNPC maintains high accuracy on benign inputs.