Evasion Attacks Against Bayesian Predictive Models

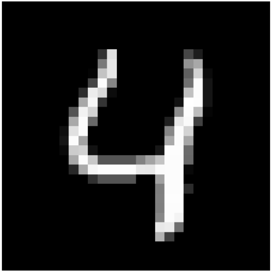
Targeting Posterior Predictive Distributions and Uncertainty

Pablo G. Arce[1,2] Roi Naveiro[3] David Ríos Insua[1]

[1]Institute of Mathematical Sciences, Spanish National Research Council
[2]Universidad Autónoma de Madrid
[3]CUNEF Universidad

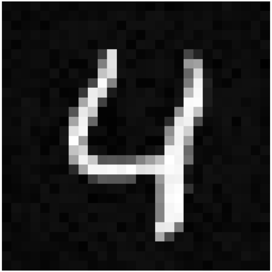
July 24, 2025

Adversarial Machine Learning



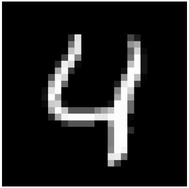
Prediction: 4

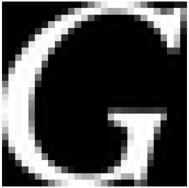
→



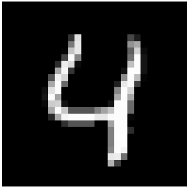
Prediction: 9

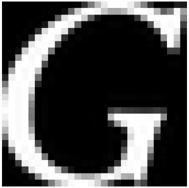
Out of Distribution

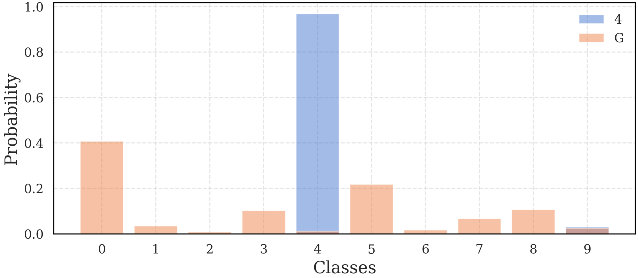




Out of Distribution

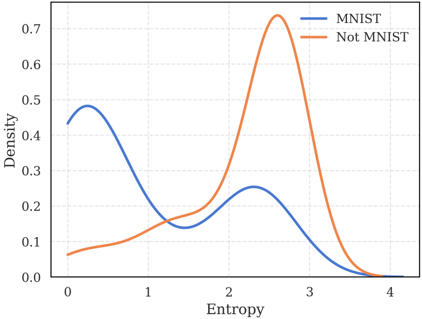






Out-of-Distribution Detection

$$\pi(y|x, D) \Rightarrow \mathbb{H}(Y) = -\int \pi(y|x, D) \log \pi(y|x, D) dy$$



Our Goal

Key Question

Are Bayesian predictive models really more robust to adversarial attacks?

Adversarial Machine Learning:

Modify (minimally) the input of the model to achieve some specific goal.

Our Goal

Key Question

Are Bayesian predictive models really more robust to adversarial attacks?

Adversarial Machine Learning:

Modify (minimally) the input of the model to achieve some specific goal.

Research Gap

Most AML research focuses on **frequentist models** and **point predictions**. Vulnerabilities of **Bayesian models** and their **uncertainty estimates** remain largely unexplored.

Problem Setup

Predictor: Bayesian model with posterior predictive distribution (PPD)

$$\pi(y|x, D) = \int \pi(y|f_\beta(x), \phi)\pi(\gamma|D)d\gamma \ , \qquad \gamma \equiv (\beta, \phi)$$

Attacker: Seeks to manipulate inputs $x \to x'$ to achieve some objective.

Two Attack Types

1. **Point Attacks**: Target specific predictions (mean, quantiles, utilities,...)

$$\min_{x' \in \mathcal{X}} \|\mathbb{E}[g(x', y)] - G^*\|_2$$

2. **Distribution Attacks**: Reshape the entire PPD

$$\min_{x' \in \mathcal{X}} \mathrm{KL}(\pi_A(y) \| \pi(y|x', D))$$

Type 2: Targeting Full Distribution

Objective: Steer PPD towards adversarial distribution $\pi_A(y)$

$$\min_{x' \in \mathcal{X}} \mathrm{KL}(\pi_A(y) \| \pi(y|x', D))$$

Proposition (2)

Under some regularity conditions, the gradient can be expressed as:

$$\nabla_{x'} KL = -E_y \left[\frac{E_{\gamma|D}[\nabla_{x'} \pi(y|x', \gamma)]}{E_{\gamma|D}[\pi(y|x', \gamma)]} \right]$$

Challenge: Gradient involves ratio of expectations.

Solution: Multi-level Monte Carlo for unbiased estimation.

Gradient Estimation for PPD Attacks

Define

$$g_{x',M}(y) \equiv -\frac{\frac{1}{M}\sum_{m=1}^{M} \nabla_{x'} \pi(y|x', \gamma_m)}{\frac{1}{M}\sum_{m=1}^{M} \pi(y|x', \gamma_m)}.$$

We have:

$$\nabla_{x'} \mathrm{KL} = \sum_{\ell=0}^{\infty} \mathbb{E}[g_{x',M_\ell}(y) - g_{x',M_{\ell-1}}(y)],$$

where we take $g_{x',M_{-1}}(y) \equiv 0$.

Unbiased MLMC gradient estimator

Sample $\ell^{(1)}, \ldots, \ell^{(R)}$ with probabilities $\omega_\ell \propto 2^{-\tau\ell}$, and estimate:

$$\hat{\nabla}_{x'} \mathrm{KL} = \frac{1}{R} \sum_{r=1}^{R} \frac{g_{x',M_{\ell^{(r)}}}(y) - g_{x',M_{\ell^{(r)}-1}}(y)}{\omega_{\ell^{(r)}}}$$

Practical Implementation: $\Delta g_{x',\ell}(y)$ computed using antithetic coupling to reduce variance.

Algorithm (simplified): PPD Attacks

Input: x, $\pi_A(y)$, $\pi(\gamma|\mathcal{D})$, $\mathcal{X}$, η, steps T, samples R, sequence $\{M_\ell\}$ and weights $\{\omega_\ell\}$

for $t = 1$ to T **do**

    Sample $y \sim \pi_A(y)$ and R levels $\ell^{(r)} \sim \omega_\ell$

    Compute $\Delta g_{x',\ell^{(r)}}(y)$ for each r

    Estimate gradient: $\widehat{\nabla}_{x'} J(x') = \frac{1}{R} \sum \frac{\Delta g}{\omega_{\ell^{(r)}}}$

    Update $x' \leftarrow \text{Proj}_{\mathcal{X}} \left(x' - \eta \widehat{\nabla}_{x'} J(x') \right)$

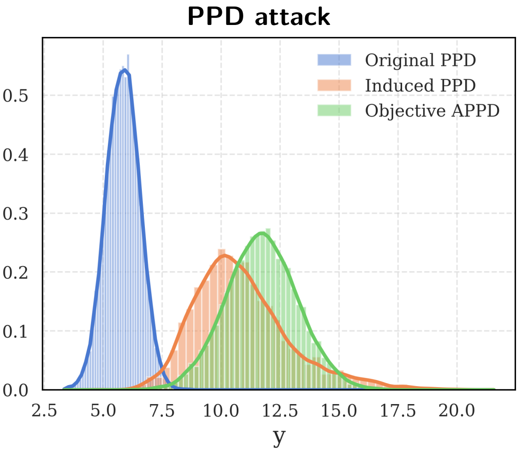
end for

Return: x'

Experimental Results: Regression

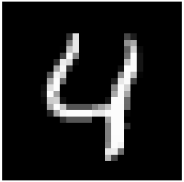
With $\|x' - x\|_2 \leq 0.5$

- Dataset: Wine quality. 11 features

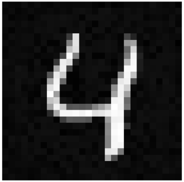


Experimental Results: Classification

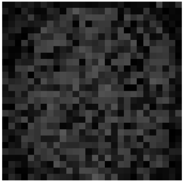
1) Unattacked x



2) $x' : \|x' - x\|_2 \leq 0.5$

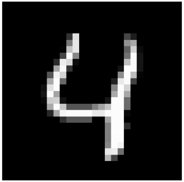


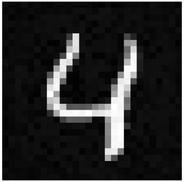
3) $10 \cdot |x' - x|$

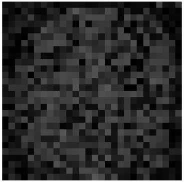


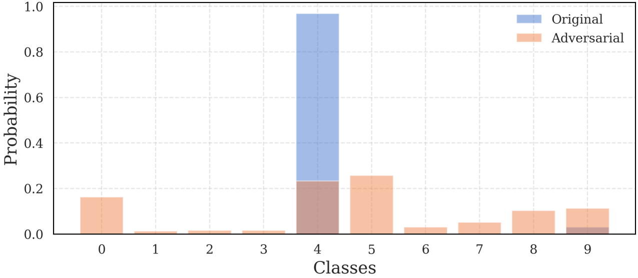
Experimental Results: Classification

1) Unattacked x **2)** $x' : \|x' - x\|_2 \leq 0.5$ **3)** $10 \cdot |x' - x|$



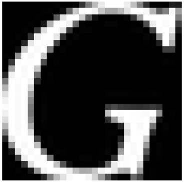




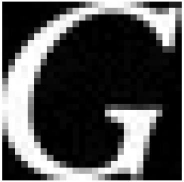


Experimental Results: Classification

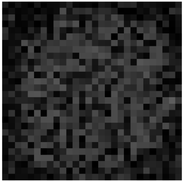
1) Unattacked x



2) $x' : \|x' - x\|_2 \leq 0.5$

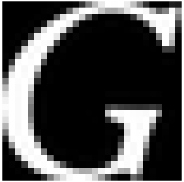


3) $10 \cdot |x' - x|$

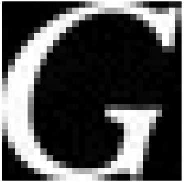


Experimental Results: Classification

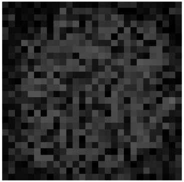
1) Unattacked x

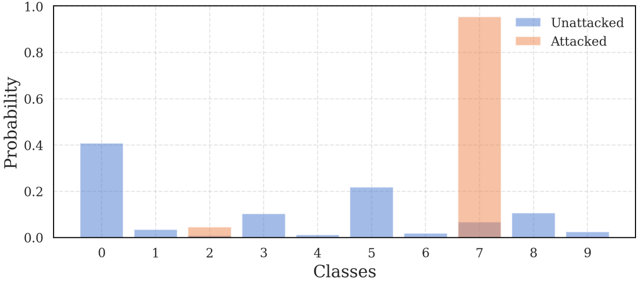


2) $x' : \|x' - x\|_2 \leq 0.5$



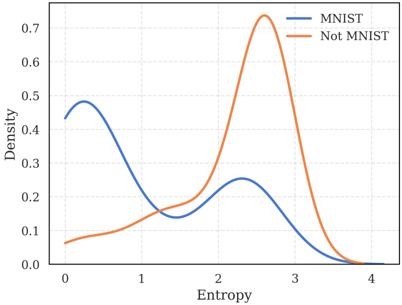
3) $10 \cdot |x' - x|$



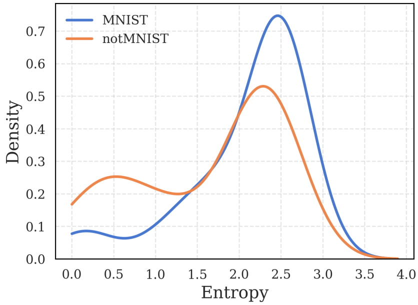


Experimental Results: Classification

1) Unattacked



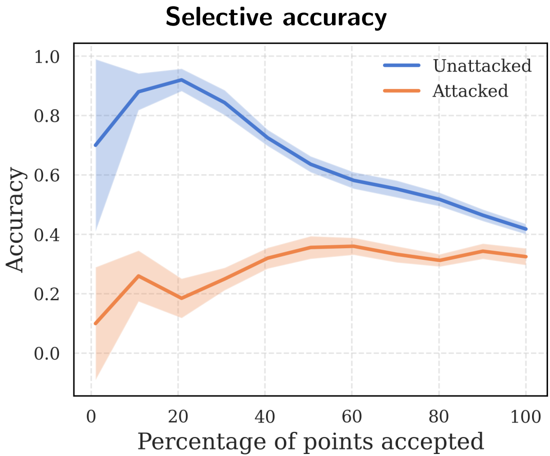
2) Attack with $\|x' - x\|_2 \leq \mathbf{0.5}$



Experimental Results: Classification

With $\|x' - x\|_2 \leq 0.5$

- Dataset: 50% of samples from MNIST and 50% from notMNIST
- Setup: Keep the % with lowest predictive entropy



Key Takeaways

Contributions

- **Novel Attack Framework** to attack Bayesian predictive models
- Can be applied to **any** inference paradigm that allows sampling
- Evidence across **white-box and gray-box** settings

Key Takeaways

Contributions

- **Novel Attack Framework** to attack Bayesian predictive models
- Can be applied to **any** inference paradigm that allows sampling
- Evidence across **white-box and gray-box** settings

Bayesian models are NOT inherently robust

- **Uncertainty estimates can be manipulated** with small perturbations
- **Both point predictions and full distributions** are vulnerable
- **Attacks transfer** across models and limited information settings
- **Critical need** for robust Bayesian inference methods

Need for Security-by-Design

Partial solutions are insufficient. We need fundamental advances in robust Bayesian inference.

In the paper

More results on:

- Toy dataset with analytical solution
- Point attack derivation and experiments
- Regression tasks
- Transferability of attacks
- MCMC and VI based inference

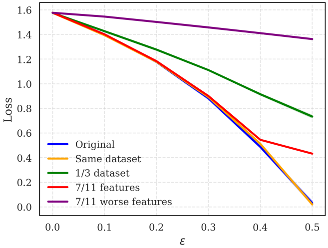
Questions?

pablo.garcia@icmat.es
`https://pablogarciarce.github.io`

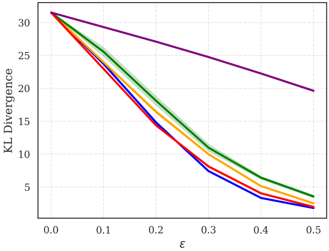
Gray-Box Attack Transferability

Limited information scenarios (Avoiding game-theoretic CK assumptions):

1. **Unknown architecture**: Different BNN arch
2. **Limited training data**: 1/3 of training dataset
3. **Partial features**: 7 best/worst predictive features (out of 11)



(a) Point attacks.



(b) Attacks to full PPD.

Figure: Security evaluation plot (SEP) of attacks.

Implication: Attacks remain effective even with partial information.

Backup: Mathematical Details

Proposition (2)

Under some regularity conditions, the gradient can be expressed as:

$$\nabla_{x'} KL = -E_y \left[\frac{E_{\gamma|D}[\nabla_{x'} \pi(y|x', \gamma)]}{E_{\gamma|D}[\pi(y|x', \gamma)]} \right]$$

Regularity Conditions for Proposition 2:

1. $y \mapsto \log \pi(y \mid x', D)\, \pi_A(y)$ is integrable for each x'
2. $x' \mapsto \log \pi(y \mid x', D)$ is differentiable for almost every y
3. There exists an integrable $H(y)$ with $\|\nabla_{x'} \log \pi(y \mid x', D)\| \leq H(y)$ for all x'
4. The map $\gamma \mapsto \pi(y \mid x', \gamma)\, \pi(\gamma \mid D)$ is integrable for each x', with $\nabla_{x'} \pi(y \mid x', \gamma)$ dominated by an integrable function