Over-the-Air Adversarial Attacks on Deep Learning Based Modulation Classifier over Wireless Channels

Kyle McClintick
Dept. of Electrical & Computer Engineering, Worcester Polytechnic Institute,
Worcester, MA 01609 USA

Wireless Innovation Laboratory

WPI

April 29th, 2020

Modulation

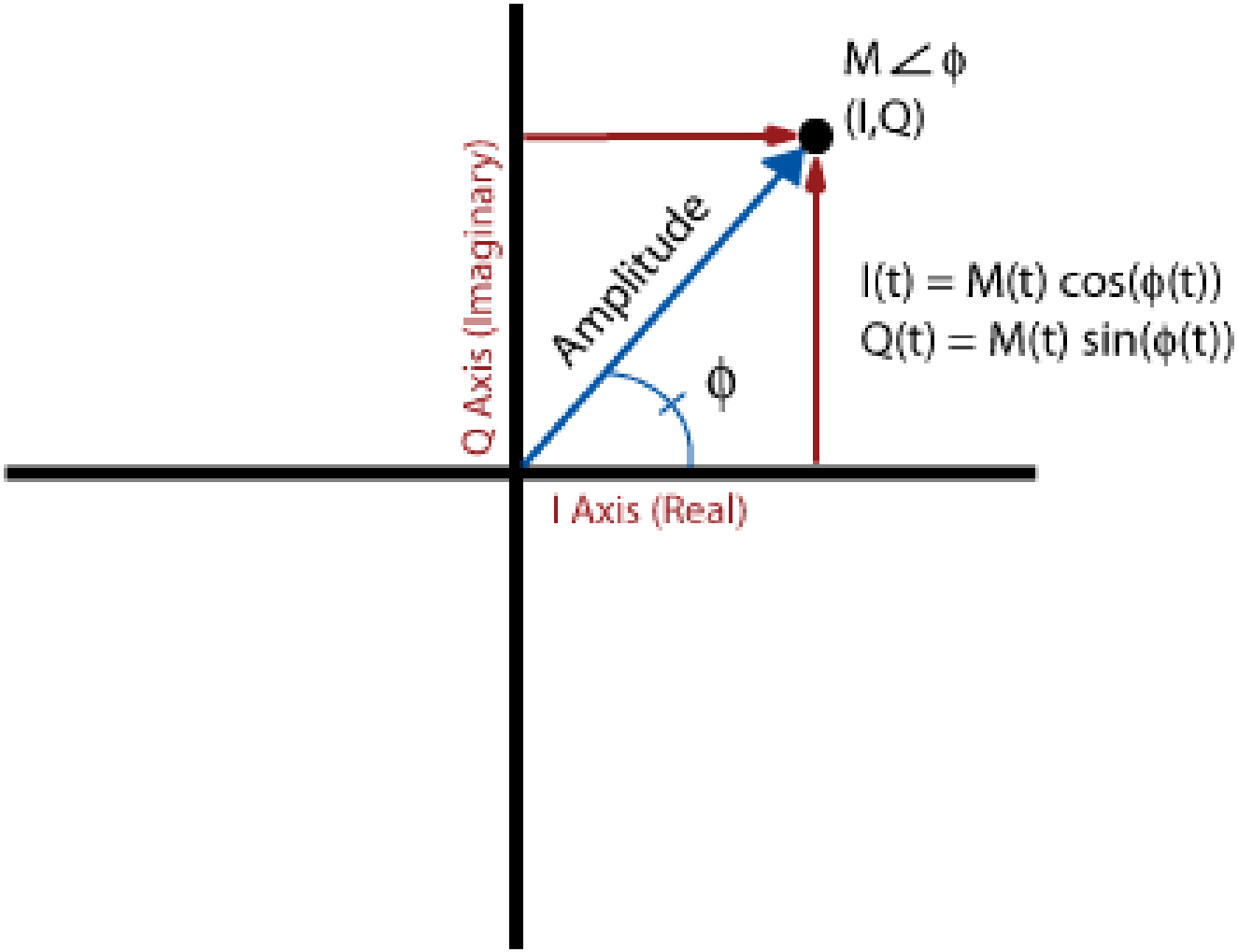


Figure: In-phase and Quadrature (IQ) plots are a polar coordinate mapping to a set of basis functions

Modulation

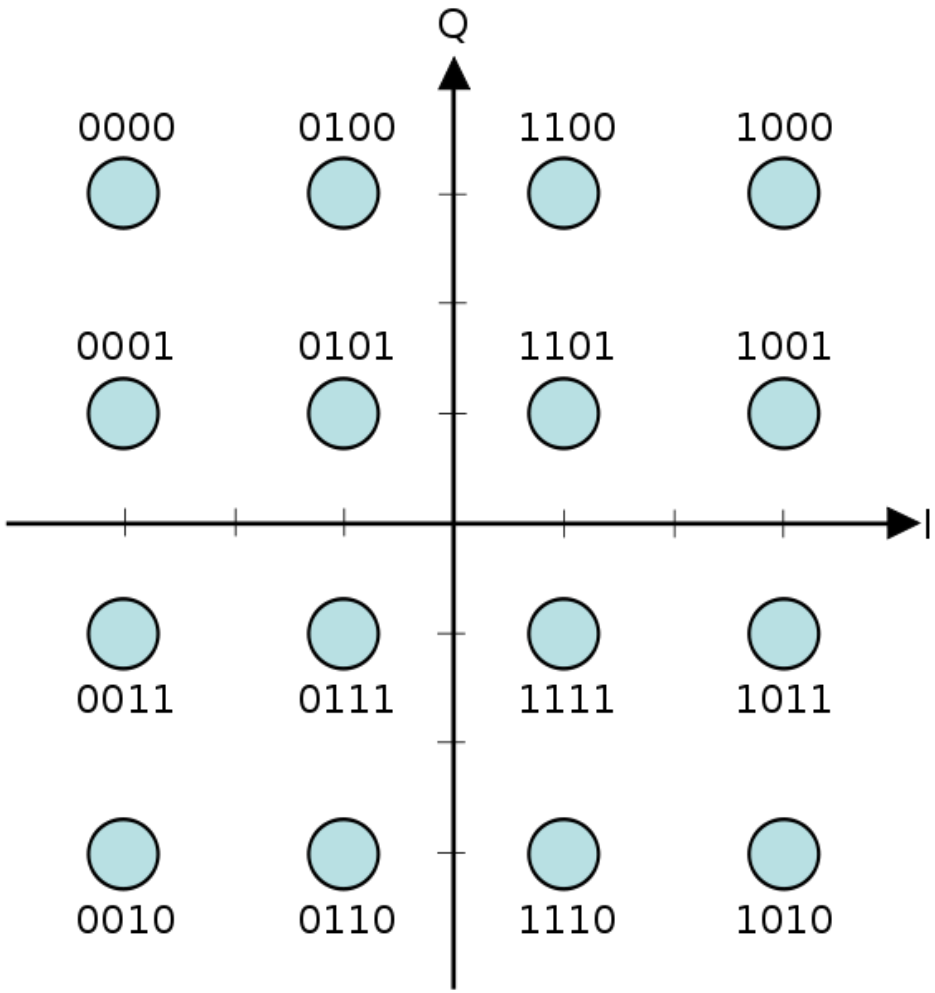


Figure: They are used to map bits to electronic signals (and back), judged by the metrics of energy efficiency and margin

Personal Interest

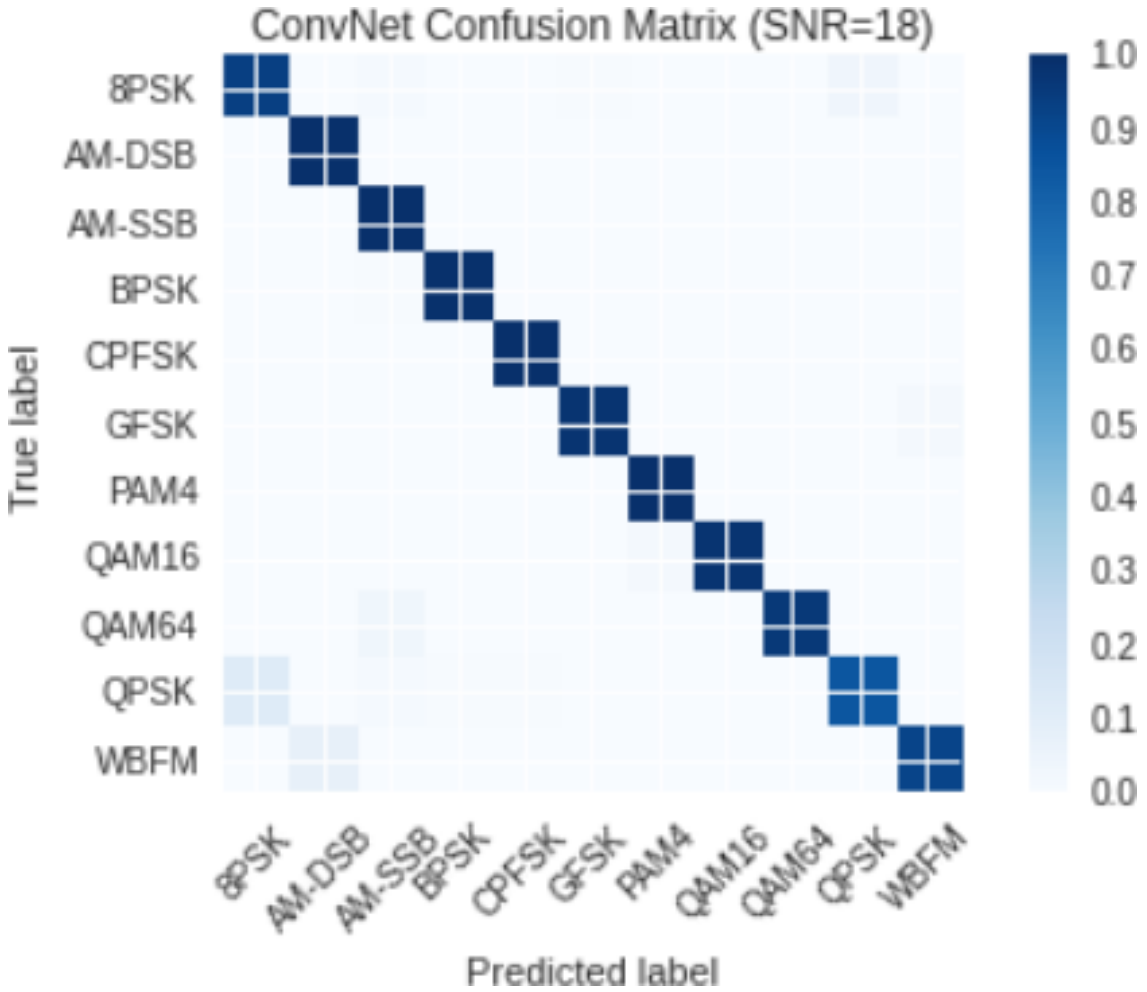
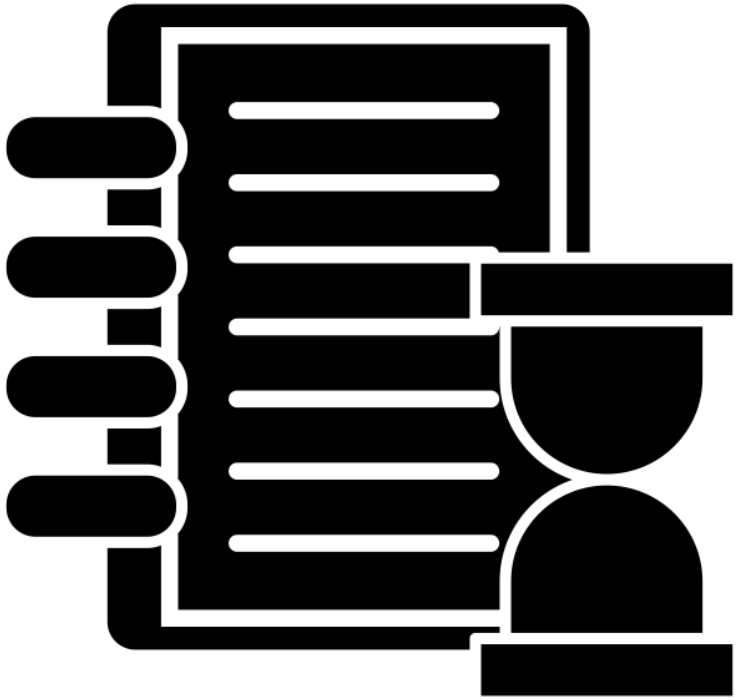


Figure: Deepsig, founded by Virgina Tech's Tim O'Shea, is a leading company in digital signal processing using machine learning

Agenda

- Abstract
- Key Contributions
- State of the Art
- Novel Method
- Results
- Conclusion



Abstract

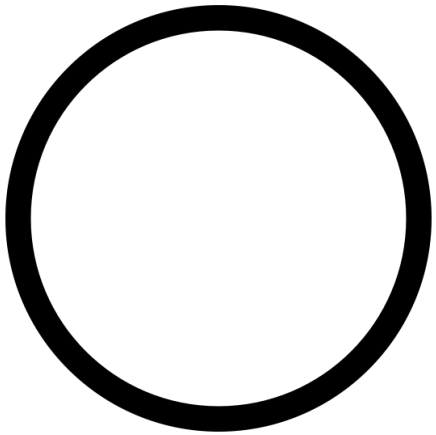


Figure: Given a transmitter, a receiver, and an adversary, modulation classification by the receiver is vulnerable over a Rayleigh fading channel with path loss and shadowing. Targeted attack (with minimum power) and non-targeted attacks are conducted by the adversary using white-box attacks that are transmitter input-specific and use channel information.

Abstract

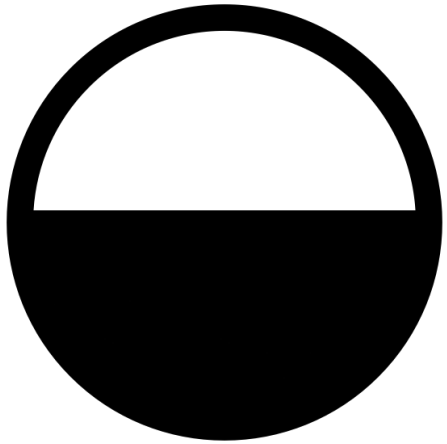


Figure: Additionally, adversarial attacks are generated using where the adversary only knows the transmitted data.

Abstract

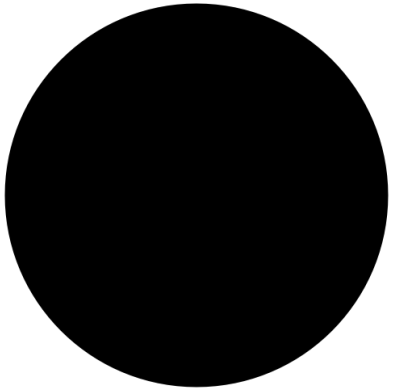
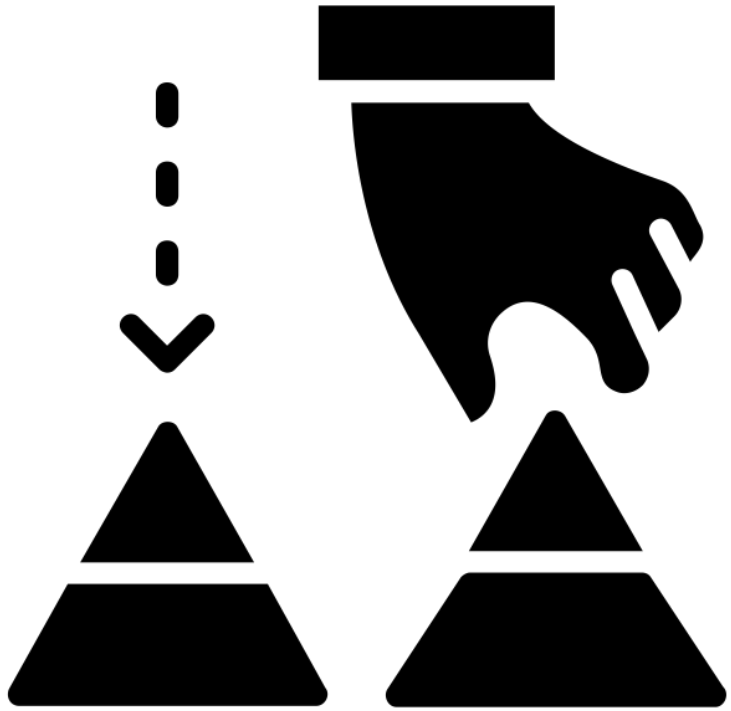


Figure: A generative model for black-box universal adversarial perturbation (UAP) attacks where the adversary has limited knowledge about both channel and transmitter input.

Key Contributions

- First case study of attacks on supervised learning models in the area of digital signal processing of wireless signals
- Novel delivery methods of attacks



State of The Art

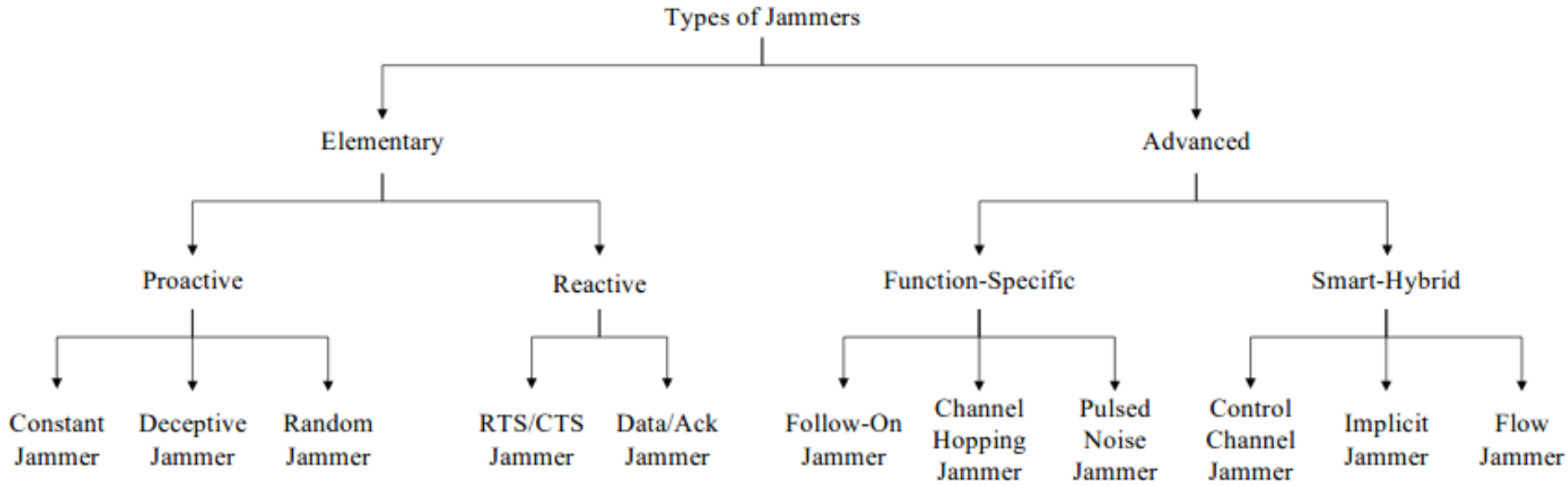


Figure 1 Types of jammers in wireless networks

White Box Attacks

Adversary has both channel and transmitter $h, x \in \mathbb{C}^p$ knowledge $r = hx + h\delta + n$, all attacks are minimum perturbation attack $||\delta||_\infty < \epsilon$ with following delivery methods:

- Channel Inversion Attack: divided by channel $\frac{\delta}{h}$
- Minimum Mean Squared Error (MMSE) Attack: Lagrangian optimization and KKT conditions give $\frac{\gamma h^* \delta}{h^* h + \lambda}$
- Maximum Received Perturbation Power (MRPP) Attack: channel is used to maximize power of attack $h^* \delta$

Grey/Black-Box Attack

Grey box attacks: adversary has only transmitter knowledge, channel is estimated $\hat{h}$ by PCA of samples
Black box: Adversary has limited channel knowledge and transmitter knowledge:

- Universal Adversarial Attack with Pre-Collected Input at the Receiver: estimate transmitted signals using PCA, $\hat{x}$
- Universal Adversarial Attack with Limited Channel Information
- Black-box Universal Adversarial Attack: standard FGSM method using stand-in CNN

Results

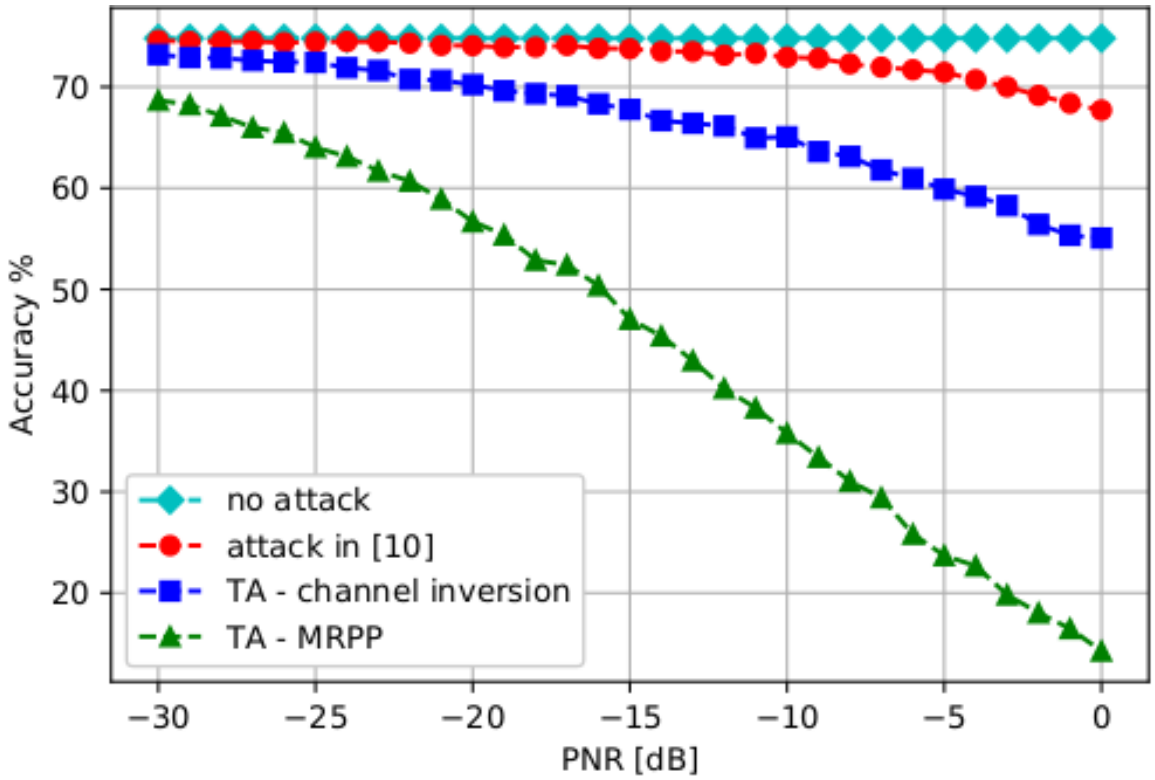


Fig. 1. Classifier accuracy with and without considering wireless channel when SNR = 10 dB.

Results

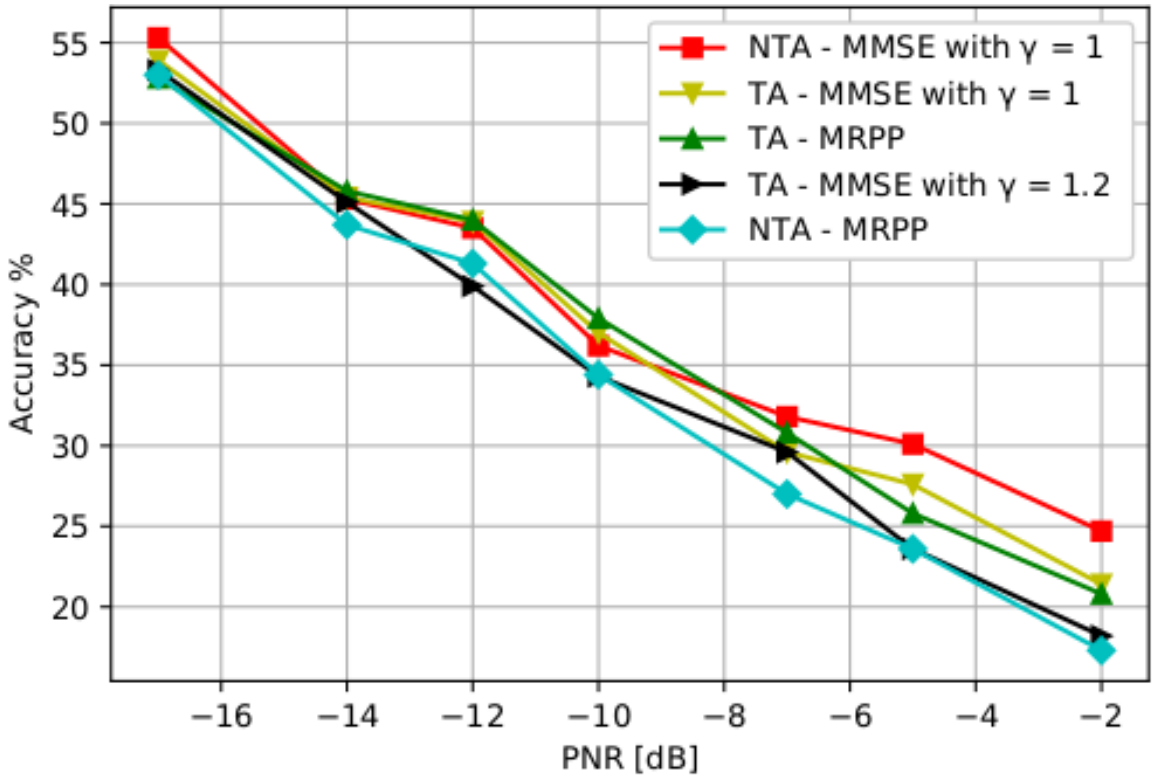


Fig. 2. Classifier accuracy under different white-box attack schemes when SNR = 10 dB.

Results

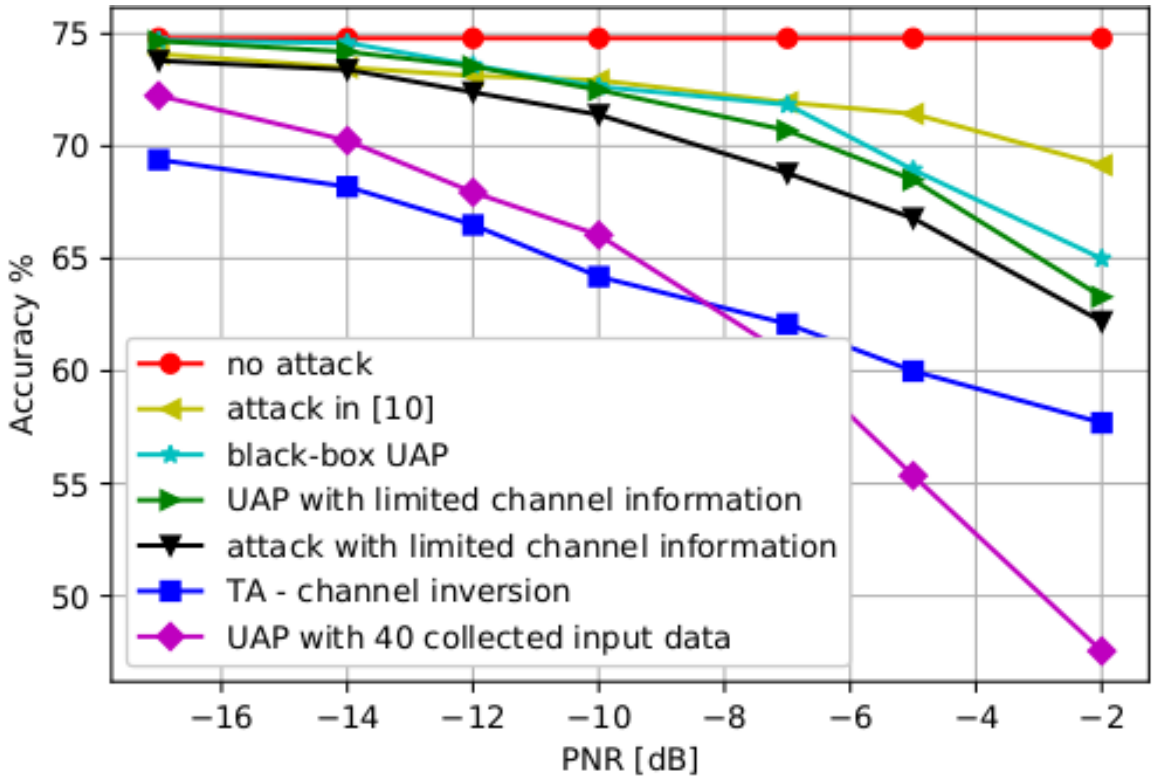


Fig. 3. Classifier accuracy using the UAP with different levels of information availability when SNR = 10 dB.

Conclusion

- Collecting and using channel and transmission information real time seems unfeasable
- Only FGSM seems feasable to me
- FGSM reduces accuracy from 75% to 65% at $-2dB$ PNR