UNCOVERING DIRECTIONS OF INSTABILITY VIA QUADRATIC APPROXIMATION OF DEEP NEURAL LOSS IN REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

1 TRUE POSITIVE RATES (TPRS)

In this section we provide true positive rates (TPRs) for SO-INRD, FO-INRD and Roth et al. (2019) for various additional FPR values. Table 1 shows the TPR values when false positive rate (FPR) is equal to 0.001 with FGSM, MI-FGSM and Nesterov Momentum computed adversarial directions for Riverraid, RoadRunner, Alien, Seaquest, Boxing, Pong, and Robotank. Table 2 shows the TPR values when false positive rate (FPR) is equal to 0.025 with Carlini & Wagner (2017), Elastic Net and DeepFool adversarial directions for Riverraid, RoadRunner, Alien, Seaquest, Boxing, Pong, and Robotank.

Table 1: True Positive Rates (TPR) for FGSM, MI-FGSM, and Nesterov Momentum when False Positive Rate (FPR) is equal to 0.001. The proposed methods SO-INRD and FO-INRD are evaluated in Riverraid, RoadRunner, Alien, Seaquest, Boxing, Pong, and Robotank.

Identification Method-Attack Method	RiverRaid	RoadRunner	Alien	Seaquest	Boxing	Pong	Robotank
SO-INRD FGSM	0.988	1.0	1.0	0.989	0.989	1.0	0.999
FO-INRD FGSM	0.953	0.75	0.765	0.509	0.783	0.581	0.206
Roth et al. (2019) FGSM	0.160	0.166	0.635	0.054	0.032	0.174	0.599
SO-INRD MI-FGSM	0.996	1.0	1.0	0.954	0.882	1.0	0.983
FO-INRD MI-FGSM	0.793	0.778	0.981	0.776	0.805	0.577	0.173
Roth et al. (2019) MI-FGSM	0.391	0.291	0.786	0.156	0.170	0.419	0.505
SO-INRD Nesterov Momentum	0.975	0.982	0.994	0.923	0.865	1.0	0.947
FO-INRD Nesterov Momentum	0.881	0.450	0.990	0.759	0.723	0.628	0.304
Roth et al. (2019) Nesterov Momentum	0.430	0.314	0.817	0.276	0.228	0.461	0.545

Table 2: True Positive Rates (TPR) for Carlini & Wagner (2017), Elastic-Net and DeepFool when False Positive Rate (FPR) is equal to 0.025. The proposed methods SO-INRD and FO-INRD are evaluated in Riverraid, RoadRunner, Alien, Seaquest, Boxing, Pong, and Robotank.

Identification Method-Attack Method	RiverRaid	RoadRunner	Alien	Seaquest	Boxing	Pong	Robotank
SO-INRD Carlini&Wagner	0.940	0.991	0.960	0.778	0.891	0.865	0.733
FO-INRD Carlini&Wagner	0.759	0.598	0.749	0.595	0.883	0.583	0.191
Roth et al. (2019) Carlini&Wagner	0.032	0.166	0.040	0.026	0.202	0.038	0.112
SO-INRD Elastic Net	0.839	0.951	0.912	0.753	0.829	0.753	0.832
FO-INRD Elastic Net	0.765	0.516	0.676	0.577	0.818	0.433	0.317
Roth et al. (2019) Elastic Net	0.118	0.280	0.158	0.046	0.316	0.102	0.106
SO-INRD DeepFool	0.955	0.997	0.994	0.908	0.964	0.903	0.910
FO-INRD DeepFool	0.922	0.872	0.961	0.826	0.958	0.851	0.416
Roth et al. (2019) DeepFool	0.468	0.555	0.266	0.376	0.567	0.369	0.703

2 SO-INRD CODE

Figure 1 shows the code for SO-INRD algorithm. Our algorithm is less than 15 lines of code, quite simple and fast. SO-INRD requires only one gradient evaluation and two function evaluations.

```python
#SO-INRD

def so_inrd(normal_obs, count_frame):
    sgrad_dir = sgrad(normal_obs, eps_dir)
    obs_grad = gradient(normal_obs[None], stochastic=stochastic)[0]
    l2norm_grad = l2_norm(obs_grad)

    normal_obs_sgrad = normal_obs + sgrad_dir
    dot_dir = np.tensordot(obs_grad, sgrad_dir, axes = ((0,1,2),(0,1,2)))

    Q_dir = act(normal_obs_sgrad[None], stochastic=stochastic)[1][0]
    Q_normal = act(normal_obs[None], stochastic=stochastic)[1][0]
    act_normal = act(normal_obs[None], stochastic=stochastic)[0][0]
    cost_normal = -np.log(Q_normal[act_normal])
    cost_dir = -np.log(Q_dir[act_normal])

    dot_array[count_frame] = -(dot_dir + cost_normal - cost_dir)
    ls_metric = -(dot_dir + cost_normal - cost_dir)
return ls_metric, dot_array
```

Figure 1: Second Order Identification of Non-Robust Directions (SO-INRD) code.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *In 2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5498–5507. PMLR, 2019.