Supplementary Material

Karim Tit[1,2] **Teddy Furon**[2]

[1]University of Luxembourg, Luxembourg, LU[†]
[2]Inria, CNRS, IRISA, University of Rennes, Rennes, FR

A ADDITIONAL SIMULATION RESULTS FOR FORM AND SORM

We report below additional results for the FORM and SORM methods.

Table 4: FORM/SORM estimations of $P_{\text{F}} \approx 1.69 \cdot 10^{-8}$ for the model M_2 and input $\mathbf{x}_{0,2}$, with uniform noise ($\varepsilon = 0.18$).

Attack	$P_{\text{F}}^{\text{FORM}}$	$P_{\text{F}}^{\text{SORM}}$	$\cos(\tilde{u}^*, \nabla G(\tilde{u}^*))$
CW	$1.74 \cdot 10^{-5}$	$6.88 \cdot 10^{-9}$	-0.95
FMNA	$3.17 \cdot 10^{-5}$	$6.12 \cdot 10^{-9}$	-0.996
HLRF	$1.89 \cdot 10^{-5}$	$7.56 \cdot 10^{-9}$	-0.97
	$\|\tilde{u}^*\|_2$	$G(\tilde{u}^*)$	Time (in sec.)
CW	4.14	$-2.1 \cdot 10^{-5}$	1.05
FMNA	4.0	$-1.9 \cdot 10^{-5}$	0.25
HLRF	4.12	$-2.3 \cdot 10^{-2}$	0.01

Table 5: FORM/SORM estimations of $P_{\text{F}} \approx 8.1 \cdot 10^{-3}$ for the model M_2 and input $\mathbf{x}_{0,3}$, with uniform noise ($\varepsilon = 0.18$)

Attack	$P_{\text{F}}^{\text{FORM}}$	$P_{\text{F}}^{\text{SORM}}$	$\cos(\tilde{u}^*, \nabla G(\tilde{u}^*))$
CW	$3.22 \cdot 10^{-2}$	$5.23 \cdot 10^{-3}$	-0.95
FMNA	$3.84 \cdot 10^{-2}$	$5.37 \cdot 10^{-3}$	-0.988
HLRF	$3.29 \cdot 10^{-2}$	$5.35 \cdot 10^{-3}$	-0.957
	$\|\tilde{u}^*\|_2$	$G(\tilde{u}^*)$	Time (in sec.)
CW	1.85	$-1.0 \cdot 10^{-5}$	0.9
FMNA	1.77	$-1.1 \cdot 10^{-5}$	0.01
HLRF	1.84	$-1.8 \cdot 10^{-2}$	0.16

Table 6: FORM/SORM estimations of $P_{\text{F}} \approx 9.92 \cdot 10^{-6}$ for the model M_2 and input $\mathbf{x}_{0,1}$, with uniform noise ($\varepsilon = 0.18$).

Attack	$P_{\text{F}}^{\text{FORM}}$	$P_{\text{F}}^{\text{SORM}}$	$\cos(\tilde{u}^*, \nabla G(\tilde{u}^*))$
CW	$6.64 \cdot 10^{-4}$	$4.66 \cdot 10^{-6}$	-0.96
FMNA	$8.45 \cdot 10^{-4}$	$3.83 \cdot 10^{-6}$	-0.993
HLRF	$7.36 \cdot 10^{-4}$	$6.53 \cdot 10^{-6}$	-0.96
	$\|\tilde{u}^*\|_2$	$G(\tilde{u}^*)$	Time (in sec.)
CW	3.21	$-1.9 \cdot 10^{-5}$	1.07
FMNA	3.14	$-2.5 \cdot 10^{-5}$	0.30
HLRF	3.18	$-2.1 \cdot 10^{-2}$	0.05

Accepted for the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024).

Table 7: FORM/SORM estimations of $P_{\text{F}} \approx 5.7 \cdot 10^{-6}$ for the custom CNN model, with gaussian noise ($\sigma = 0.02$).

Attack	$P_{\text{F}}^{\text{FORM}}$	$P_{\text{F}}^{\text{SORM}}$	$\cos(\tilde{u}^*, \nabla G(\tilde{u}^*))$
CW	$3.91 \cdot 10^{-5}$	NA	-0.96
FMNA	$5.22 \cdot 10^{-5}$	NA	-0.985
HLRF	$2.16 \cdot 10^{-5}$	NA	-0.973
	$\|\tilde{u}^*\|_2$	$G(\tilde{u}^*)$	Time (in sec.)
CW	3.95	$-3.7 \cdot 10^{-5}$	1.49
FMNA	3.88	$-8.0 \cdot 10^{-4}$	0.23
HLRF	4.09	$-1.9 \cdot 10^{-2}$	0.03