Supplementary Material Discriminative Adversarial Privacy: Balancing Accuracy and Membership Privacy in Neural Networks

Eugenio Lomurno
eugenio.lomurno@polimi.it
Alberto Archetti
alberto.archetti@polito.it
Francesca Ausonio
francesca.ausonio@mail.polimi.it
Matteo Matteucci
matteo.matteucci@polimi.it

Department of Electronics, Information and Bioengineering
Politecnico di Milano
Via Ponzio 34/5
20133 Milan, Italy

AOP λ Analysis

This section presents the Accuracy Over Privacy (AOP) results for different values of lambda (λ), distinct from those utilized in the primary paper. Specifically, Tables 1, 2, 3, 4, 5, 6 display the AOP metric outcomes for lambda values of 1, 2, 5, 10, 20 and 50, respectively. The experiments are referred to Cifar-10 [4], Cifar-100 [4], FMNIST [7], EuroSAT [3], TinyImagenet [5], OxfordFlowers [6], STL-10 [1] and Cinic-10 [2] datasets. Table 1 stands out as the only one displaying accuracy and Area Under the Curve (AUC) of the Membership Inference Attack (MIA) at equivalent magnitudes. In this particular scenario, it is evident that the trade-off between these two metrics exclusively favors the Reg model, followed by the DAP models, compared to the Baseline model. As the λ value increases, the AOP scores decrease, indicating a greater penalty in proportion to the AUC. Both Table 3 and Table 4 demonstrate how the average AOP of the Baseline diminishes at a much faster rate than that of the other models. Particularly in Table 4, the DAP models indisputably emerge as the optimal choice for managing the trade-off, trailed by the DP models, and as a last resort, the Reg model, which exhibits performance very similar to the Baseline model. It is important to note that DP results tend to deteriorate less than Baseline and Reg results, in many cases surpassing them, as λ increases. They do, however, suffer from problems with many classes, an indication of the poor accuracy achieved despite the guaranteed privacy.

Table 1: The AOP metric on the test sets ($\lambda = 1$). Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10	0.605	**0.643**	0.310	0.355	0.402	0.415	0.615	0.607
Cifar-100	0.399	**0.428**	0.039	0.080	0.089	0.071	0.305	0.273
FMNIST	0.844	0.824	0.603	0.698	0.730	0.766	0.854	**0.861**
EuroSAT	0.880	**0.900**	0.305	0.586	0.681	0.643	0.898	0.891
TinyImagenet	0.303	**0.319**	0.030	0.032	0.031	0.025	0.252	0.213
OxfordFlowers	0.372	**0.431**	0.028	0.047	0.083	0.131	0.269	0.247
STL-10	0.542	**0.577**	0.084	0.135	0.247	0.288	0.472	0.379
Cinic-10	0.588	0.577	0.279	0.339	0.389	0.402	0.565	**0.578**
Average	0.567	**0.587**	0.210	0.284	0.331	0.343	0.529	0.506

Table 2: The AOP metric on the test sets ($\lambda = 2$). Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10	0.467	0.509	0.307	0.338	0.387	0.413	**0.607**	0.601
Cifar-100	0.331	**0.345**	0.039	0.078	0.087	0.070	0.296	0.269
FMNIST	0.765	0.733	0.600	0.695	0.724	0.759	0.842	**0.850**
EuroSAT	0.809	0.852	0.302	0.583	0.681	0.641	**0.896**	0.889
TinyImagenet	0.251	**0.270**	0.029	0.032	0.029	0.025	0.244	0.209
OxfordFlowers	0.244	**0.281**	0.026	0.044	0.079	0.123	0.250	0.237
STL-10	0.449	**0.513**	0.083	0.129	0.245	0.288	0.465	0.375
Cinic-10	0.514	0.470	0.279	0.337	0.386	0.399	0.552	**0.570**
Average	0.479	0.497	0.208	0.280	0.327	0.340	**0.519**	0.500

Table 3: The AOP metric on the test sets ($\lambda = 5$). Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10	0.214	0.253	0.298	0.290	0.346	0.406	0.582	**0.583**
Cifar-100	0.188	0.180	0.039	0.072	0.084	0.068	**0.269**	0.260
FMNIST	0.568	0.516	0.593	0.687	0.707	0.736	0.808	**0.821**
EuroSAT	0.629	0.723	0.293	0.276	0.681	0.633	**0.891**	0.884
TinyImagenet	0.143	0.162	0.027	0.032	0.026	0.024	**0.222**	0.198
OxfordFlowers	0.069	0.079	0.020	0.036	0.067	0.102	0.201	**0.209**
STL-10	0.255	0.359	0.082	0.112	0.238	0.286	**0.443**	0.362
Cinic-10	0.343	0.254	0.277	0.331	0.379	0.389	0.517	**0.547**
Average	0.301	0.316	0.204	0.267	0.316	0.331	**0.492**	0.483

Table 4: The AOP metric on the test sets ($\lambda = 10$). Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10	0.059	0.079	0.283	0.225	0.287	0.394	0.543	**0.555**
Cifar-100	0.074	0.061	0.039	0.062	0.078	0.064	0.230	**0.245**
FMNIST	0.346	0.288	0.581	0.674	0.680	0.701	0.754	**0.773**
EuroSAT	0.412	0.551	0.279	0.565	0.681	0.621	0.882	**0.875**
TinyImagenet	0.056	0.069	0.023	0.031	0.021	0.023	**0.190**	0.181
OxfordFlowers	0.008	0.009	0.014	0.025	0.052	0.075	0.269	**0.247**
STL-10	0.099	0.198	0.081	0.089	0.226	0.283	**0.409**	0.341
Cinic-10	0.175	0.091	0.274	0.321	0.368	0.374	0.464	**0.510**
Average	0.154	0.168	0.197	0.249	0.299	0.317	0.451	**0.456**

Table 5: The AOP metric on the test sets ($\lambda = 20$). Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10	0.004	0.008	0.256	0.136	0.198	0.371	0.472	**0.502**
Cifar-100	0.011	0.007	0.037	0.046	0.068	0.057	0.168	**0.217**
FMNIST	0.129	0.089	0.559	0.647	0.628	0.634	0.656	**0.686**
EuroSAT	0.177	0.319	0.252	0.543	0.681	0.596	**0.865**	0.858
TinyImagenet	0.009	0.013	0.018	0.031	0.014	0.021	0.138	**0.152**
OxfordFlowers	1e-4	1e-4	0.006	0.012	0.031	0.040	0.067	**0.113**
STL-10	0.015	0.060	0.077	0.056	0.205	0.278	**0.349**	0.302
Cinic-10	0.046	0.012	0.269	0.302	0.347	0.345	0.373	**0.444**
Average	0.049	0.064	0.184	0.222	0.271	0.292	0.386	**0.409**

Table 6: The AOP metric on the test sets ($\lambda = 50$). Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10	2e-6	7e-6	0.190	0.030	0.065	0.310	0.311	**0.373**
Cifar-100	4e-5	1e-5	0.035	0.019	0.045	0.040	0.065	**0.152**
FMNIST	0.007	0.003	0.495	**0.574**	0.494	0.471	0.432	0.480
EuroSAT	0.014	0.062	0.187	0.481	0.681	0.529	**0.814**	0.808
TinyImagenet	3e-5	8e-5	0.008	0.029	0.004	0.017	0.053	**0.089**
OxfordFlowers	4e-10	4e-10	5e-4	0.001	0.007	0.006	0.007	**0.039**
STL-10	5e-5	0.002	0.069	0.014	0.152	**0.261**	0.217	0.211
Cinic-10	8e-4	2e-5	0.253	0.253	0.290	0.272	0.194	**0.292**
Average	0.003	0.008	0.155	0.175	0.217	0.238	0.262	**0.305**

Table 7: The AUC metric of the MIAs against miss-classified samples. Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10	0.753	0.689	0.505	**0.503**	**0.503**	0.506	0.510	0.505
Cifar-100	0.610	0.627	**0.501**	0.506	0.538	0.505	0.518	0.509
FMNIST	0.659	0.768	0.502	0.506	0.502	**0.500**	0.523	0.527
EuroSAT	0.698	0.563	**0.500**	0.501	0.523	0.504	0.527	0.528
TinyImagenet	0.585	0.565	0.504	0.502	**0.501**	0.505	0.518	0.508
OxfordFlowers	0.785	0.837	0.529	0.555	**0.518**	0.532	0.547	0.537
STL-10	0.636	0.567	**0.500**	0.506	0.506	0.504	0.509	0.504
Cinic-10	0.560	0.649	**0.501**	0.503	0.503	0.502	0.508	0.510
Average	0.661	0.658	**0.505**	0.510	0.512	0.507	0.520	0.516

Table 8: The AUC metric of the MIAs against correctly classified samples. Results improving the baseline are coloured in green, while results worse than the baseline are red. The best results among them are in **bold**, while the second best are underlined.

Dataset	Baseline	Reg	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	DAP_t	DAP_v
Cifar-10 [4]	0.545	0.537	**0.500**	0.502	0.503	0.506	0.507	0.504
Cifar-100 [4]	0.516	0.517	0.512	0.521	0.508	0.511	**0.507**	0.509
FMNIST [7]	0.514	0.532	0.504	0.505	0.511	0.510	0.505	**0.501**
EuroSAT [3]	0.514	0.503	0.516	0.508	0.508	0.532	**0.502**	**0.502**
TinyImagenet [5]	0.509	0.506	0.504	0.530	0.544	0.520	0.502	**0.501**
OxfordFlowers [6]	0.556	0.595	0.877	0.677	0.589	0.590	**0.529**	0.535
STL-10 [1]	0.522	0.508	0.521	0.513	**0.501**	0.505	0.503	0.504
Cinic-10 [2]	0.508	0.534	0.505	**0.501**	**0.501**	**0.501**	**0.501**	**0.501**
Average	0.523	0.529	0.555	0.532	0.521	0.522	**0.507**	**0.507**

MIA Against Slices

The effectiveness of Membership Inference Attacks (MIAs) relies on whether they target correctly classified samples from the attacked model or not. This assertion is precisely demonstrated in Table 7 and Table 8. The experiments are referred to Cifar-10 [4], Cifar-100 [4], FMNIST [7], EuroSAT [3], TinyImagenet [5], OxfordFlowers [6], STL-10 [1] and Cinic-10 [2] datasets. In Table 7, showing the results of the MIA attack against the miss-classified slices, the Baseline and Reg models are particularly vulnerable, while the DP models exhibit high levels of security. The DAP technique achieves results close to random guessing, thus ensuring strong protection. On the other hand, Table 8 reveals the poor performance of MIAs on correctly classified samples on average. Moreover, both the Baseline and Reg models enable the attacking model to achieve an AUC close to 0.5, which, in many cases, surpasses the AUC obtained by the DP models. This suggests that correctly predicted test set samples possess similar characteristics to the training data, making them challenging to distinguish. However, the most noteworthy result is observed in the DAP models. They not only outperform the Baseline model but also emerge as the overall best, surpassing even the DP models.

References

[1] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[2] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[3] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226, 2019.

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto University press*, 2009.

[5] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[6] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[7] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.