Supplementary Materials
Differentially Private Bagging: Improved utility and cheaper privacy than subsample-and-aggregate

James Jordon
University of Oxford
james.jordon@wolfson.ox.ac.uk

Jinsung Yoon
University of California, Los Angeles
jsyoon0823@g.ucla.edu

Mihaela van der Schaar
University of Cambridge
University of California, Los Angeles
Alan Turing Institute
mv472@cam.ac.uk, mihaela@ee.ucla.edu

1 Discussion continued

As alluded to in the main paper, we believe that our work is a fundamental improvement over the subsample-and-aggregate model and as such can be used to improve (most) methods that build on the subsample-and-aggregate framework. To illustrate how we think this might look, we take Private Aggregation of Teacher Ensembles (PATE) [1] as an example (note that we do not prove the following result - it is a conjecture).

PATE builds on the standard subsample-and-aggregate model by deriving a data-dependent bound on the (global) moments accountant at each step of the form

$$\alpha(l) \leq \log((1-q)\left(\frac{1-q}{1-e^{2\lambda}q}\right)^{l} + qe^{2\lambda l}) \tag{1}$$

where

$$q = \sum_{c \neq c^*} \frac{2+\lambda|n_c^* - n_c|}{4\exp(\lambda|n_c^* - n_c|)} \tag{2}$$

with n_c^* denoting the number of teachers voting for the majority class.

What we believe our algorithm offers here is to improve on this data-dependent bound in a similar fashion to the improvement we make in the paper, which amounts to replacing λ with $\lambda \times m$ and instead bounding the *personalised* moments accountants, so that

$$\check{\alpha}_{\check{c}_\lambda(x_{new})}(l;\mathcal{D},u) \leq \log((1-q)\left(\frac{1-q}{1-e^{2\lambda m(x_{new};u)}q}\right)^{l} + qe^{2\lambda m(x_{new};u)l}). \tag{3}$$

While we are confident in this bound, it is unclear whether we can also replace the λ term in the definition of q, which we leave as an open question for future research.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2 Full tables for the results on the Heart Failure dataset

2.1 Logistic Regression

Table 1: Prediction performance (Accuracy, AUROC, AUPRC) of DPBag and SAA with $\delta = 10^{-5}$ on the Heart Failure dataset using Logistic Regression. **Bold** indicates the best performance achieved for the given metric and fixed ϵ. NPB is a non-private baseline model, included to indicate an upper bound on our performance.

Model	n	k	Accuracy			AUROC			AUPRC		
			$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	50	10	.5538	.5933	.6073	.5478	.6015	.6310	.4738	.5187	.5459
		50	.5628	.5941	.6123	.5516	.6134	.6385	.4760	.5317	.5560
		100	.5478	.5886	.6083	.5475	.6037	.6333	.4739	.5255	.5520
	100	10	.5635	.6051	.6137	.5581	.6285	.6431	.4828	.5467	.5625
		50	.5639	**.6085**	**.6154**	.5547	.6326	**.6453**	.4793	.5530	**.5656**
		100	.5667	.6050	.6142	.5626	.6295	.6448	.4895	.5496	.5652
	250	10	.5919	.6032	.6066	.6008	.6330	.6399	.5234	**.5542**	.5623
		50	.5888	.6061	.6099	.5954	.6320	.6391	.5161	.5526	.5607
		100	**.5986**	.6077	.6091	**.6096**	**.6373**	.6398	**.5289**	**.5542**	.5644
DPBag-	100	50	.5614	.6019	.6128	.5544	.6288	.6429	.4792	.5411	.5612
		100	.5596	.6007	.6108	.5609	.6174	.6354	.4767	.5338	.5525
	250	50	.5855	.6051	.6086	.5896	.6295	.6366	.5093	.5498	.5565
		100	.5875	.6061	.6110	.5884	.6321	.6407	.5103	.5518	.5615
SAA	50	-	.5468	.5842	.6002	.5367	.5927	.6159	.4596	.5086	.5310
	100		.5613	.6006	.6077	.5567	.6144	.6280	.4837	.5342	.5471
	250		.5798	.6019	.6024	.5778	.6284	.6356	.5023	.5496	.5559
NPB	1	-		.6527			.6992			.6281	

Table 2: Number of labels provided by each method using Logistic Regression before the privacy budget, ϵ, is used up on the Heart Failure dataset with $\delta = 10^{-5}$. Note that DPBAG- and SAA have the same, data-independent privacy analysis and so provide the same number of labels as each other.

Models	n	k	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	50	10	17	120	301
		50	20	137	345
		100	20	139	351
	100	10	62	507	1277
		50	74	593	1487
		100	76	609	1538
	250	10	372	2978	7452
		50	468	3785	6380
		100	507	4044	6805
SAA	50	-	11	85	211
	100		43	338	843
	250		264	2108	5269

2.2 GBM

Table 3: Prediction performance (Accuracy, AUROC, AUPRC) of DPBag and SAA with $\delta = 10^{-5}$ on the Heart Failure dataset using GBM. **Bold** indicates the best performance achieved for the given metric and fixed ϵ. DPBAG- is our method without the improved privacy analysis. NPB is a non-private baseline model, included to indicate an upper bound on our performance.

Model	n	k	Accuracy			AUROC			AUPRC		
			$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	100	50	.5740	.6002	.6121	.5827	.6239	.6307	.5073	.5473	.5641
		100	.5906	**.6165**	**.6239**	.5908	.6265	.6444	.5115	.5502	.5613
	150	50	**.5912**	.6060	.6164	**.5987**	.6281	.6322	**.5182**	.5423	.5656
		100	.5835	.6101	.6218	.5936	**.6289**	**.6451**	.5170	**.5504**	**.5691**
DPBag-	100	50	.5633	.5978	.6054	.5723	.6189	.6231	.4887	.5387	.5424
		100	.5665	.6061	.6136	.5714	.6203	.6334	.4894	.5393	.5521
	150	50	.5786	.6022	.6113	.5911	.6195	.6292	.5066	.5377	.5556
		100	.5751	.6014	.6186	.5864	.6193	.6355	.5158	.5433	.5523
SAA	100	-	.5763	.5959	.6068	.5796	.6132	.6261	.4948	.5284	.5457
	150		.5731	.5977	.6111	.5839	.6137	.6276	.5005	.5353	.5511
NPB	1	-		.6482			.6945			.6215	

Table 4: Number of labels provided by each method using GBM before the privacy budget, ϵ, is used up on the Heart Failure dataset with $\delta = 10^{-5}$. Note that DPBAG- and SAA have the same, data-independent privacy analysis and so provide the same number of labels as each other.

Models	n	k	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	100	50	73	615	1527
		100	79	635	1591
	150	50	161	1332	3329
		100	174	1386	3473
SAA	100	-	43	338	843
	150		95	759	1897

3 Results on UCI Adult Dataset

UCI Adult dataset: UCI Adult dataset `https://archive.ics.uci.edu/ml/datasets/adult` is a public dataset for binary classification. The total number of features is 108 (after one-hot encoding) and the number of samples is 48841. Among 48841 samples, 11687 samples (23.9%) have class 1.

3.1 Logistic Regression

Table 5: Prediction performance (Accuracy, AUROC, AUPRC) of DPBag and SAA with $\delta = 10^{-5}$ on the UCI Adult Dataset using Logistic Regression. **Bold** indicates the best performance achieved for the given metric and fixed ϵ. NPB is a non-private baseline model, included to indicate an upper bound on our performance.

Model	n	k	Accuracy			AUROC			AUPRC		
			$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	50	10	.7532	.7934	.8042	.7005	.8069	.8176	.4031	.5505	.5751
		50	.7384	.7908	.8056	.6816	.7714	.8187	.3837	.4988	.5690
		100	.7467	.7872	.8018	.6609	.7811	.8160	.3990	.5159	.5717
	100	10	.7721	.8061	.8221	.7386	.8005	.8237	.4514	.5503	.6068
		50	.7736	.8089	.8201	.7382	.8053	.8299	.4558	.5625	.6031
		100	.7707	.8092	.8233	.7379	.8051	.8388	.4497	.5648	.6216
	250	10	.7950	.8170	.8213	**.8062**	.8339	.8445	.5384	.6091	.6225
		50	.7938	.8199	**.8249**	.7831	**.8448**	**.8524**	.5216	**.6191**	**.6389**
		100	**.8035**	**.8207**	.8236	.7920	.8393	.8505	**.5433**	.6137	.6350
DPBag-	100	50	.7644	.7922	.8117	.7426	.7963	.8253	.4548	.5406	.5977
		100	.7697	.7989	.8130	.7490	.8002	.8291	.4631	.5503	.5985
	250	50	.7918	.8155	.8206	.7907	.8355	.8482	.5287	.6012	.6274
		100	.7919	.8156	.8220	.7922	.8301	.8466	.5259	.5993	.6291
SAA	50	-	.7107	.7766	.7939	.6553	.7701	.7934	.3755	.5072	.5485
	100		.7579	.7986	.8149	.7652	.8057	.8291	.4795	.5590	.6147
	250		.7883	.8103	.8137	.7825	.8261	.8378	.5193	.5976	.6163
NPB	1	-	.8472			.9029			.7526		

Table 6: Number of labels provided by each method using Logistic Regression before the privacy budget, ϵ, is used up on the UCI Adult Dataset with $\delta = 10^{-5}$. Note that DPBAG- and SAA have the same, data-independent privacy analysis and so provide the same number of labels as each other.

Models	n	k	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	50	10	20	91	228
		50	21	94	235
		100	22	94	236
	100	10	52	372	938
		50	54	385	963
		100	55	390	973
	250	10	297	2371	5917
		50	310	2496	6240
		100	317	2533	6327
SAA	50	-	11	85	211
	100		43	338	843
	250		264	2108	5269

3.2 GBM

Table 7: Prediction performance (Accuracy, AUROC, AUPRC) of DPBag and SAA with $\delta = 10^{-5}$ on the UCI Adult Dataset using GBM. **Bold** indicates the best performance achieved for the given metric and fixed ϵ. DPBAG- is our method without the improved privacy analysis. NPB is a non-private baseline model, included to indicate an upper bound on our performance.

Model	n	k	Accuracy			AUROC			AUPRC		
			$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	100	50	.7743	**.7986**	.8069	.7655	.7861	.7997	.4908	.5548	.5746
		100	**.7763**	.7947	.8028	.7406	.7820	.7879	.4641	.5505	.5648
	150	50	.7697	.7974	.8157	.7548	**.7873**	.8033	.4984	**.5581**	.5873
		100	.7730	.7963	**.8183**	.7582	.7802	**.8039**	**.5084**	.5561	**.5883**
DPBag-	100	50	.7407	.7901	.8018	.6527	.7684	.7932	.4325	.5492	.5701
		100	.7673	.7945	.8004	.7149	.7812	.7956	.4048	.5472	.5693
	150	50	.7717	.7952	.8159	**.7688**	.7852	.7963	.5036	.5367	.5803
		100	.7720	.7942	.8101	.7613	.7729	.7919	.4956	.5414	.5723
SAA	100	-	.7364	.7747	.7896	.6864	.7631	.7821	.3861	.5268	.5517
	150		.7693	.7879	.8014	.7540	.7717	.7889	.4971	.5466	.5748
NPB	1	-		.8713			.9245			.8199	

Table 8: Number of labels provided by each method using GBM before the privacy budget, ϵ, is used up on the UCI Adult Dataset with $\delta = 10^{-5}$. Note that DPBAG- and SAA have the same, data-independent privacy analysis and so provide the same number of labels as each other.

Models	n	k	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	100	50	51	413	1037
		100	52	413	1042
	150	50	115	931	2323
		100	117	933	2336
SAA	100	-	44	338	843
	150		95	759	1897

4 Pseudo-code for Subsample-and-aggregate

Algorithm 1 Semi-supervised differentially private knowledge transfer using subsample-and-aggregate

Input: ϵ, δ, $\mathcal{D}$, batch size n_{mb}, number of teachers n, noise size λ, maximum order of moments to be explored, L, unlabelled public data $\mathcal{D}_{pub}$
Initialize: $\{\theta_T^i\}_{i=1}^n$, θ_S, $\hat{\epsilon} = 0$, $\alpha(l) = 0$ for $l = 1, ..., L$
Partition the dataset n disjoint subsets $\mathcal{D}_i$, $i = 1, ..., n$ such that $\bigcup_i \mathcal{D}_i = \mathcal{D}$ and $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$ for all i, j
while Teachers have not converged **do**
    for $i = 1, ..., n$ **do**
        Sample $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{n_{mb}}, y_{n_{mb}}) \overset{\text{i.i.d.}}{\sim} D_i$
        Update teacher, T_i, using SGD
        $\nabla_{\theta_T^i} - \left[\sum_{s=1}^{n_m b} \sum_{c \in \mathcal{C}} y_{s,c} \log(T_{i,j}^c(\mathbf{x}_s)) \right]$ (multi-task cross-entropy loss)
while $\hat{\epsilon} < \epsilon$ **do**
    Sample $\mathbf{x}_1, ..., \mathbf{x}_{n_{mb}} \sim \mathcal{D}_{pub}$
    for $s = 1, ..., n_{mb}$ **do**
        for $c \in \mathcal{C}$ **do**
            $n_c \leftarrow |\{(i, j) : T_{i,j}(\mathbf{x}_s) = c\}|$
        $r_s \leftarrow \arg\max\{n_c + Y_c : c \in \mathcal{C}\}$ where Y_c are i.i.d. $Lap(\frac{1}{\lambda})$
        Update the moments accountants
        for $l = 1, ..., L$ **do**
            $\alpha(l) \leftarrow \alpha(l) + 2\lambda^2 l(l+1)$
    Update the student, S, using SGD
    $\nabla_{\theta_S} - \sum_{s=1}^{n_{mb}} \sum_{c \in \mathcal{C}} r_{s,c} \log S^c(\mathbf{x}_s)$ (multi-task cross-entropy loss)
    $\hat{\epsilon} \leftarrow \min_l \frac{\alpha(l) + \log(\frac{1}{\delta})}{l}$
Output: S

References

[1] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.