We thank all the reviewers for their insightful comments.

Reviewer 2: **Answer 1** - We have conducted a further experiment in which we use the bagging classifier but the simple, non-data-dependent bound on the privacy (DPBAG-) to demonstrate how much is gained from the bagging classifier and how much is gained from the improved privacy bound we derive. The table below shows the results for each ϵ and each metric for the best choice of n and k (which are not the same across DPBAG and DPBAG-). The full table will be included in the revised manuscript. (Also see **Reviewer 3: Answer 3** for results with GBM as the base learner.)

Model	Accuracy			AUROC			AUPRC		
	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	.5986	.6085	.6154	.6096	.6373	.6453	.5289	.5542	.5656
DPBag-	.5875	.6061	.6128	.5896	.6321	.6429	.5103	.5518	.5615

Answer 2 - We will revise the related works section, removing any opinion language, in the revised manuscript.

Answer 3 - (3) and (4) will only be equal when there is some personalised moments accountant that dominates all other personalised moments accountants for *every* query. If we consider the bounds derived in 4.2 for our personalised moments accountant, we see that this would mean that $m(x_{new}; u^*) > m(x_{new}, u)$ for some u^*, for **all** u and for **all** x_{new}. This corresponds to there being some set of teachers (corresponding to u^*) which essentially always disagree on every queried x_{new}. We will clarify this in the revised manuscript.

Answer 4 - While the personalised moments accountant is introduced as an intermediate quantity in Abadi et. al, our defining it as an explicit quantity which we then go on to show can lead to meaningful analysis justifies its inclusion in a section dedicated to our contributions. In order to separate Theorem's 2 and 3 which are inherited from Abadi et al. we will create a new subsection within section 4 that makes this clear. Thank you for the suggestion.

Answer 5 - Thank you, we will replace l with ℓ and "the mechanism" with "any mechanism".

Answer 6 - Theorem 2 is being stated with respect to the personalised moments accountants, for which the downwards accountant does depend on u. There is however a typo on line 158, the RHS of the inequality should be $\check{\alpha}$ rather than α. We will correct this in the revised manuscript.

Answer 7 - Thank you, this is correct. $m = 1$ when there is any class for which no teachers vote, rather than there being unanimity. In the case of binary classification, these two are the same, which is where the confusing language originated. We will correct this in the main manuscript.

Answer 8 - We agree that a theoretical result for accuracy would be nice, however, we have yet to derive one. We have since conducted a further experiment using GBM as the base learner (in place of logistic regression) to further verify that DPBag outperforms standard subsample-and-aggregate empirically, see Answer 3 to Reviewer 3.

Reviewer 3: **Answer 1** - While it is true that the main contribution of the paper is the bagging variation of subsample-and-aggregate, the explicit definition and demonstration of utility of the personalised moments accountants amount to a contribution that may have consequences for other techniques. In addition, the PATE framework was recently shown to be a useful and practical tool for building a differentially private GAN in (Jordon et al. 2019).

Answer 2 - As stated in the paper, our work in this paper is in parallel to the work of PATE, with improvements made in this paper being applicable to PATE as well. The improvements are over the underlying subsample-and-aggregate framework rather than a different privacy analysis (which is what PATE provides). We conjecture that the privacy bounds in PATE can be translated over to our work in much the same way the naive bounds can be (see the Supplementary Materials, section 1 for details). Moreover, we have ran experiments with PATE and found that their data-dependent bound was very rarely, or never, smaller than the naive bound, and as such PATE simply reduced to standard subsample-and-aggregate. But we stress that the key reason for comparing with standard subsample-and-aggregate is because we believe both PATE and Scalable PATE can be applied **on top** of our method.

Answer 3 - We have conducted additional experiments using gradient boosting method (GBM) as the base classifier for the teachers. The following table shows the results for DPBAG, DPBAG- (see **Reviewer 2: Answer 1**) and SAA (with the best setting of n and k for each metric and ϵ). The full table will be included in the revised supplementary materials.

Model	Accuracy			AUROC			AUPRC		
	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$
DPBag	.5912	.6165	.6239	.5987	.6289	.6451	.5182	.5504	.5691
DPBAG-	.5786	.6061	.6186	.5911	.6203	.6355	.5158	.5433	.5556
SAA	.5763	.5977	.6111	.5839	.6137	.6276	.5005	.5353	.5511