The authors would like to thank the reviewers for their constructive feedback. Our response (for each reviewer) follows.

Reviewer #1 points out that "the proposed theoretical results do not imply that the consistency of ensemble methods are better than any individual predictor". We agree and we will refine the explanation further and state that the method is appropriate when average performance of all the predictors in the ensemble is considered. However, it should be noted that this is a well-known issue when using ensemble methods to improve other performance metrics as well (e.g. accuracy) and has been discussed in previous research. In 'R.Polikar,*Ensem. Based Syst. in Decis. Makin.*,IEEE Circ. and Syst. Magaz.,6(3):21-45,2006', 'R. Polikar, *Ensem. Learn.*, Scholarpedia,4(1):2776,2009', authors state that *"...there is no guarantee that the combination of multiple classifiers will always perform better than the best individual classifier in the ensemble. Nor an improvement on the ensemble's average performance can be guaranteed except for certain special cases. ...it certainly reduces the overall risk of making a particularly poor selection."* Our theoretical results are consistent with the above statements and show ensemble methods can but not necessarily be better than the best component. In practice for online systems during model re-training we cannot know the best possible model in advance, hence for practical reasons we should compare with average performance of all predictors in the ensemble. Further, we provide a theoretical support (Corollary 3.1) to show how an improvement on the ensemble's performance (correct-consistency) can be guaranteed with a quantifiable probability. We will clarify this in the revised version and also add the statement 'combining classifiers may not necessarily beat the consistency performance metric of the best classifier in the ensemble'. We hope the introduction of an important problem, originality and strong empirical feasibility of our work will spark more interest in consistency estimation.

Reviewer #2 points out that "the scope of our metrics and theorems seem limited to single-label classification problems". We agree and stated in lines 145-147 (page 4), multi-label classification is a challenging problem where the statement *"A smaller Euclidean distance corresponds to a higher consistency."* is not always true. The problem will be explored in the future. Reviewer suggested to show "how the accuracy and percentage of correct->incorrect and incorrect->correct predictions change for D1->D2 and D2->D3". We show these additional results in Table 1. In general, we want more incorrect->correct (ItoC) and correct->correct (CtoC) but less correct-incorrect (CtoI), so we compute Com=CtoC+ItoC-CtoI as an additional metric. The observations are consistent with findings in our paper. We will add these additional results to the paper. Reviewer mentioned that "the effect of the random initialization and shuffling (RIS) in the algorithm is not clear". We discussed this in lines 284-293 (page 8) where we show that given the same ensemble size $m = 20$ ExtBagging, DynSnap-cyc and DynSnap-step that use RIS outperform Snapshot which does not use RIS. In sensitivity analysis for snapshot number N (results and settings presented in Figure 1c and Appendix J, respectively), where using DynSnap-cyc with $\beta = 1$ i.e. ensemble of N best models without RIS, we show that as N doubles, the performance improves but the improvement is minor after $N = 20$. However, performance of DynSnap-cyc with $\beta = \beta^*$ is further improved partially due to RIS.

Reviewer #3 points out "the proposed method is not significantly more powerful than the ExtBagging". We discussed this in lines 284-287 (page 8). Our theoretical and empirical results demonstrate that ExtBagging performs very well in both accuracy and consistency as it selects components with the best estimated accuracy and utilizes RIS. Our proposed method DynSnap-cyc combines techniques from ExtBagging and Snapshot to achieve a slightly better performance than ExtBagging but *with substantially reduced training cost*. The reviewer also points that "ACC and CON seem to be correlated". In our experiments, we observe examples like ExtBagging, that have better ACC but lower CON than DynSnap-cyc on CIFAR10+ResNet20 and YAHOO!Answer+fastText. Another example is DynSnap-cyc (ACC 75.64%, CON 85.70%), that has improved accuracy by 3.7% and consistency by 11.3% (significant improvement) compared with Snapshot (ACC 72.96%, CON 76.98%) on CIFAR100+ResNet56. The reviewer also asks "whether the theoretical findings work correctly when any other distance measure is used". Yes, indeed, our theoretical findings are true for Hamming, Euclidean, Manhattan, and Minkowski distances. The proofs can be generalized by using Minkowski inequality 'M.Voitsekhovskii,*Minkowski inequality*,Encyclopedia of Math.,2001' with corresponding order p (replace order 2 in Equation 8 of Appendix B with order p). We will provide the generalization proof in supplementary materials. We thank the reviewer for the suggestion "to boldface the highest hits" (we will do this) and "provide more details about replication results" (5 replicates per method).

Reviewer #4 has asked "why having learning rate scheduling was important". In 'I.Loshchilov, et.al.,*Sgdr: Stochas. grad. desc. with warm restarts*, arXiv preprnt. arXiv:1608.03983,2016', authors suggest that cycling annealing schedule perturbs the parameters of a converged model, which allows the model to find a better local minimum. In 'G.Huang,et.al.,*Snapshot Ensembles: Train 1, get m for free*,arXiv preprnt. arXiv:1704.00109,2017', authors claim that there is a significant diversity in the local minima when visited during each cycle. Our proposed DynSnap-cyc is inspired by their findings. The results show that DynSnap-cyc (using cycling schedule) outperforms DynSnap-step (using step-wise decay). The reviewer also asks "how does the theoretical justification work in a setting where training data distribution changes over successive model generations". Our theoretical findings are invariant with respect to changes in training data distribution since representation (Equation 1) of consistency is invariant as long as r_t, s_{tj} and $\tilde{s}_{tj}$ are all represented in a p-dimensional space. If p changes, then it is a different problem where consistency loses its meaning. We will add the discussion in the paper.

Table 1: (WAVG) Percentage of correct→incorrect (CtoI), incorrect→correct (ItoC) and CtoC+ItoC-CtoI (Com) predictions for $D_1 \rightarrow D_2$, $D_2 \rightarrow D_3$ and $D_1 \rightarrow D_3$.

	CIFAR10+ResNet20									**CIFAR100+ResNet56**									**YAHOO!Answers+fastText**								
	$D_1 \rightarrow D_2$			$D_2 \rightarrow D_3$			$D_1 \rightarrow D_3$			$D_1 \rightarrow D_2$			$D_2 \rightarrow D_3$			$D_1 \rightarrow D_3$			$D_1 \rightarrow D_2$			$D_2 \rightarrow D_3$			$D_1 \rightarrow D_3$		
	ItoC	CtoI	Com	ItoC	CtoI	Com	ItoC	CtoI	Com	ItoC	CtoI	Com	ItoC	CtoI	Com	ItoC	CtoI	Com	ItoC	CtoI	Com	ItoC	CtoI	Com	ItoC	CtoI	Com
SingleBase	6.60	5.38	80.79	7.08	6.83	79.59	5.74	4.27	82.15	11.18	8.80	59.56	9.36	9.26	59.19	11.32	8.86	59.60	2.32	2.15	60.83	5.07	2.75	62.56	5.14	2.65	62.65
ExtBagging	3.11	2.13	**87.13**	4.40	3.80	**86.07**	3.13	1.56	**88.31**	6.57	4.42	70.65	4.87	4.71	70.53	6.81	4.51	70.73	1.72	1.82	61.54	4.43	2.29	63.21	4.65	2.61	62.88
MCDropout	6.69	5.20	80.67	7.29	7.16	78.84	5.78	4.16	81.84	10.02	9.35	58.08	10.00	9.21	59.01	10.69	9.23	58.99	2.84	1.65	62.00	7.88	5.25	61.04	7.70	3.88	62.41
Snapshot	4.89	3.11	84.98	5.89	5.38	83.22	3.91	1.62	86.98	8.46	5.92	67.66	6.32	5.64	68.63	8.91	5.68	68.59	2.05	1.78	62.07	3.35	1.56	64.07	4.48	2.42	63.21
DynSnap-cyc	2.84	2.11	86.36	4.47	3.71	85.51	2.78	1.29	87.93	5.74	3.35	**73.19**	3.58	3.90	**72.32**	5.60	3.54	**72.69**	1.38	1.05	**63.47**	3.46	1.67	**64.63**	4.08	1.97	**64.34**
DynSnap-step	3.09	2.47	86.38	4.36	3.64	85.91	3.47	2.13	87.42	6.44	5.92	68.14	5.62	4.93	69.82	5.98	4.77	69.98	1.77	1.28	63.23	3.52	1.71	64.62	4.29	1.99	**64.34**