We thank all reviewers for the constructive comments! Itemized responses to each reviewer are appended below:
Abbreviations: imbalanced learning (IL), under-sampling (US), over-sampling (OS), cost-sensitive learning (CSL).

To all reviewers: Thanks for your careful reading! We will carefully resolve all writing, format, and notation issues.

Response to reviewer #1: **1)** MESA focuses on US, which is worrisome as it was compared with SMOTE-like baselines. **R:** We highlight that we conducted an extensive comparison including 19 different baselines (balanced/cleaning US, distance/ranking-based OS, their ensemble variants, etc.). SMOTE-*like algorithms are only a small part of them.* **2)** Results of random OS (ROS). **R:** We have tested ROS but didn't report the results due to the page limit. Generally, it yields poor performance compared with the other baselines. These results will be included in the camera-ready version. **3)** About the use of Gaussian function. **R:** Our main goal is to design an *efficient*, *concise*, and *practical* IL framework. It is nearly impossible to make instance-level decisions by using a complex meta-sampler (e.g., set a large output layer or use RNN), as the complexity of a single update will grow from $\mathcal{O}(1)$ to $\mathcal{O}(n)$, where the number of instances n is usually large in real-world datasets. Besides, complex model architecture also brings extra memory cost and hardship in optimization. For these reasons, we choose to use such a Gaussian function trick to simplify the meta-sampler, while maintains the ability to perform controlled sampling for generating accurate and diverse base learners. **4)** About weight normalization. **R:** For clarity, Eq. 3 shows the unnormalized sampling weights (noted in the paper). **5)** About SAC. **R:** SAC has better suitability for continuous state & action spaces compares to other methods like DQN. **6)** Why SMOTE-like methods perform poorly? **R:** We have discussed this in the paper, please see our discussions at lines 33-37, lines 224-227, lines 256-259. Also, please see the analysis and results in [1, 2] and references therein. **7)** About related works. **R:** We conducted an extensive review of related works including 3 papers of CSL (lines 78-82). As MESA is US-based, we mainly focused on more closely related works rather than reweighting ones, such as CSL. **8)** Reproducibility **R:** We have elaborated on all the implementation details in the appendix due to the page limit. The code of this work was released via GitHub, the link can be found in the paper (footnote #1).

Response to reviewer #2:
1) Theoretical analysis. **R:** Thanks for the advice! Most of the existing theoretical results of ensemble learning are limited to a specific model (e.g., perceptron/SVM) and ensemble schema (e.g., boosting/bagging). As MESA is a very general framework, it is not easy to derive *useful* theoretical results, so instead, we show our insights by providing solid experimental results and analysis. We will list the theoretical analysis of MESA as an important future direction. **2)** More datasets. **R:** Due to both the resource and space limitations, we have tried our best to fit all important results in the paper. We will include more datasets in the future, the results will possibly be released online (via Github).

Response to reviewer #3:
1) About algorithm 1. **R:** Sorry for the confusion! Please note that $\mathcal{P}_\tau$ and $\mathcal{N}_\tau$ stands for minority and majority set respectively (not the other way around). The algorithm 1 actually defines a function: $\texttt{Sample}(\mathcal{D}_\tau; F, \mu, \sigma)$, where F represents an ensemble classifier that is one of the function inputs. This function was called in algorithm 2, line 7, and you can see that the corresponding input is F_t. We will carefully revise this part to make it more readable.

Response to reviewer #4:
1) Algorithm design. **R:** Sorry for the confusion. Please refer to our response to reviewer #1, item 3. Due to the space limitation, we end up removed lots of discussions about out motivation. They will be included in the final version. **2)** About ggregation rule. **R:** Thanks for your advice! Our work mainly focuses on sampling strategy, which is the core application of meta-sampler. Adaptive aggregation, also known as "dynamic ensemble selection", is another interesting big topic, which may be a promising extension of this paper. We will study more on this topic in future work. **3)** About experiment 4.1. **R:** According to previous works (please see Fig. 2 in [2] and references therein) and our experiment, we found that class-overlapping is a more important property of an IL task compares to the imbalance ratio. Therefore, we choose to simulate different levels of overlapping for more illustrative results and visualization. **4)** About synthetic dataset. **R:** Number of instances: $|\mathcal{P}|$=200, $|\mathcal{N}|$=2,000. We will clarify this in the final version. **5)** About k in table 2. **R:** Baselines are *resampling methods* with a single classifier trained on the resampled data (k=1). **6)** About error bars. **R:** Due to page limit, we can only end up with removing the error bar from some results to fit the 8-page manuscript. We will adjust the layout to guarantee all the important results come with an error bar. **7)** Data split. **R:** Sorry for the confusion! In our implementation, we keep-out the validation set and report the mean score of 4-fold stratified CV (i.e., 60%/20%/20% train/valid/test split). We will clarify these in the final version. **8)** Origin of variance. **R:** The variance comes from the randomness in the data sampling and model training process. **9)** Comparison to other meta-ML methods. **R:** Thanks for the advice. However, due to the huge difference between them in the aspect of task (tabular/image), model (traditional classifier/neural network), and learning schema (ensemble/standalone), it is hard to conduct a fair comparative experiment. We will depict the difference more clearly.

References: **[1]** Krawczyk B. Learning from imbalanced data: open challenges and future directions[J]. Progress in Artificial Intelligence, 2016, 5(4): 221-232. **[2]** Liu Z, Cao W, Gao Z, et al. Self-paced Ensemble for Highly Imbalanced Massive Data Classification[C], 2020 IEEE 36th International Conference on Data Engineering. IEEE, 2020: 841-852.