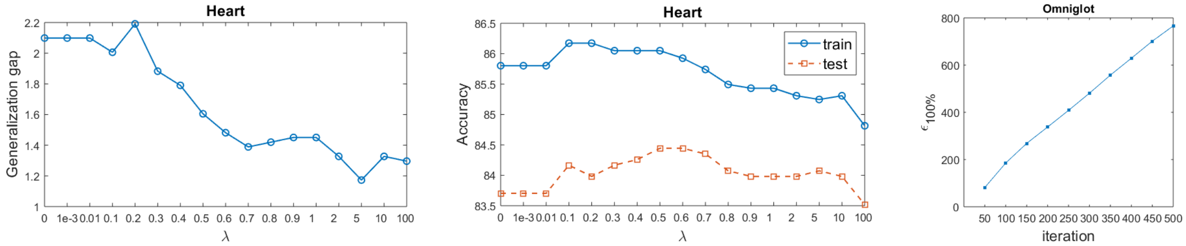
We thank all the reviewers (**R1**-**R4**) for their insightful comments and invaluable feedback on our manuscript. As the reviewers mentioned, our work shows the following strengths. (1) The proposed decomposition theorem and the resulting bound are interesting (R1,R2) and new (R3,R4); (2) The theoretical study of non-convex metric learning is important (R4), and the technique is general and of interest in establishing similar time-generalization tradeoff (R1,R3); (3) Our algorithm is novel (R4), intuitive, and adequately uses the theoretical insights provided by the bound (R1); (4) The theoretical analysis is complete (R2,R3), and the paper is easy to read (R1,R2,R3). We respond to the key comments below but will address all feedback in the final version.

[R3] Motivation. Our motivation is to provide an approach to deriving a new generalization bound that considers the parameters related to optimization, such as the number of iterations, which cannot be achieved by current uniform convergence bounds. Then as summarized by **R1**,"the theoretical insights provided by the bound are adequately used (smooth loss and classifiers, early stopping scheme, enforced small Lipschitz constants)" in the proposed algorithm.

[R3] Significance of the proved results. As pointed out by **R1**, "the proposed bound seems to be new from a metric learning point of view. Furthermore, it is general enough to be of potential interest for a broader range of researcher". As suggested by **R2**, "the theoretical techniques may be of interest in establishing similar time generalization trade-off bounds for non-convex problems, where the traditional route of uniform convergence does not make much sense" and "the paper overcomes a common hurdle of requiring convex formulation or even optimization accuracy". In Sec.3.3.4, we also provide discussion on the implications of Theorem 3. Greater stress on the significance of the proved results and their wider applications will be added to Sec.3.3.4, Abstract and Introduction to make the significance more clear.

[R3] Compared with generalization bounds of Cao et al. [6]. The bounds obtained in [6] is based on the framework of uniform convergence. However, as pointed out by **R2**, when considering "time-generalization trade-off bounds", i.e., the generalization bounds considering the iteration number, "the traditional route of uniform convergence does not make much sense". The approach proposed by our paper is fundamentally different from uniform convergence and it takes optimization parameters into account. We will, on top of discussing the shortcomings of uniform convergence, highlight the difference between our approach and their approaches including [6] in Introduction to make this more clear.

[R1] Study the influence of the parameter λ; [R3] Performance of the model determined by loss term or regularization term? Figs.(a) and (b) show the influence of λ on the generalization gap, training accuracy and test accuracy. The generalization gap decreases with λ, which is consistent with our theoretical result that constraining the norms of $\boldsymbol{L}, \boldsymbol{x}, \boldsymbol{r}$ gives smaller Lipschitz constants, thereby tightening the bound. As the training accuracy generally decreases with λ as well, the test accuracy is highest when $\lambda = 0.5$. More detailed analysis will be added into the final version.



(a) Effect of λ on the generalization gap. (b) Effect of λ on training and test accuracy. (c) Concentration of deep ML.

[R2, R4] Explore with deep metric learning (ML) algorithms; [R1] Extension to batch gradient descent. We have investigated the impact of training iterations on the concentration of parameters of deep ML and present the result in Fig.(c). The network consists of three convolutional blocks and one fully connected (FC) layer and is trained with mini-batch gradient descent (batch size=100); $\epsilon_{100\%}$ is calculated from the parameters of the FC layer over 10 rounds. Similar to Fig.2 in the main paper, we see again that $\epsilon_{100\%}$ increases along training, indicating that the variance of learned parameters becomes larger. While the empirical result is encouraging, we acknowledge that the proved theorem cannot guarantee the learnability of mini-batch/stochastic gradient descent algorithms due to the randomness of training instances introduced in each batch. This is indeed important and will be added into the Broader Impact Section.

[R4] The method is only applicable to binary classification problems. Our theoretical framework can be readily generalized to multi-class classification. Theorem 2 and Lemma 2 are proved regardless of the number of classes. Lemma 3 needs to be revised by considering the uniform convergence property of the hypothesis class defined over a multi-class risk R. This has been discussed in some references. For example, based on the risk defined on p185 of [Ref 1], uniform convergence is guaranteed by Theorem 8.1 on p187. More details will be added into the final version.
[Ref 1] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning.

[R2] Advantage gained by learning the representative instances. We will add experiments by awarding other methods, such as LMNN, the luxury of choosing representative points, as suggested by **R2**.