Appendix of Dataset condensation with latent quantile matching

Wei Wei[1], Tom De Schepper[1], Kevin Mets[2]
University of Antwerp - imec, IDLab,
[1]Department of Computer Science, [2]Faculty of Applied Engineering,
Sint-Pietersvliet 7, 2000 Antwerp, Belgium
{wei.wei, tom.deschepper, kevin.mets}@uantwerpen.be

A. Ablation study

Choices of goodness of fit tests. In Barbiero and Hitaj [1], the optimal quantiles of the CvM and AD test statistics are reported. The optimal quantiles for AD test statistics can be obtained by an iterative process. We report the process proposed by Barbiero and Hitaj [1] in Appendix B. The quantiles obtained will minimize the AD test statistic, which is a variant of the CvM test statistic that gives more weight to the tails of the distribution. We have experimented with the optimal quantiles that minimize the AD test statistic with graph structured data. The result is shown in Appendix A.

Compared to the performance of optimal CvM quantiles, the optimal AD quantiles does not provide noticeable improvements across the four experimented dataset. In the one task setting with graph structured data, the variant that uses optimal CvM quantiles consistently outperform the variant that uses optimal AD quantiles. Next, the computation of CvM quantiles does not require an iterative process. Thus, the use of the CvM test is more suitable for DM based DC.

SETTING	METHODS	CORAFULL		ARXIV		REDDIT		PRODUCT	
		AA(%)↑	BWT(%)↑	AA(%)↑	BWT(%)↑	AA(%)↑	BWT(%)↑	AA(%)↑	BWT(%)↑
ONE TASK	CAT+LQM (AD)	57.0±0.2	-	67.2±0.2	-	**92.8±0.1**	-	84.3±0.1	-
	CAT+LQM (CVM)	**57.5±0.2**	-	**67.6±0.4**	-	**92.8±0.1**	-	**84.4±0.1**	-
CGL	CAT+LQM (AD)	**68.3±0.4**	**-8.7±0.2**	67.1±0.2	-11.8±0.6	**97.2±0.1**	**-0.5±0.1**	**71.0±0.4**	-5.0±0.3
	CAT+LQM (CVM)	68.1±0.2	**-8.7±0.3**	**68.0±0.3**	**-10.7±0.2**	97.1±0.0	**-0.5±0.0**	**71.0±0.2**	**-4.9±0.2**

Table 1. Comparison of AA of LQM using two different goodness of fit test statistic. The bold results are the best performance. ↑ denotes the greater value represents greater performance.

B. Algorithm of optimal quantiles for Anderson-Darling test statistic

Algorithm 1: Optimal quantile computation for Anderson-Darling test statistic

Input : Budget of discrete points k, epsilon ϵ_{max} for convergence checking

Params : For $i = \{1, 2, ..., k\}$: quantile of the target distribution q_i, quantile of the discrete approximating distribution Q_i, probability of the discrete approximations p_i, loop counter t.

initialize $t = 1$

for $i = 1$ **to** k **do**

  initialize $p_i^0 = \frac{1}{k}, Q_i^0 = \frac{1}{k}$.

while $\epsilon_t > \epsilon_{max}$ **do**

  $t \leftarrow t + 1$

  for $i = 1$ **to** k **do**

    $q_i^t = \frac{Q_{i-1}^{t-1} + Q_i^{t-1}}{2}$

    $Q_i^t = log(\frac{1-q_i^t}{1-q_{i+1}^t})/log(\frac{q_{i+1}^t(1-q_i^t)}{q_i^t(1-q_{i+1}^t)})$

    $p_i^t = Q_i^t - Q_{i-1}^t$

  $\epsilon_t = max_{i=1}^k |p_i^t - p_i^{t-1}|$

Output : Quantiles q_i that minimizes the Anderson-Darling test statistic.

C. Visualizations of synthetic image datasets learned by IDM+LQM

We visualize the synthetic image dataset learned by IDM+LQM from Fig. 1 to Fig. 4. We observe some repetitive dot patterns in the synthetic datasets learned in 1 image per setting, i.e., in Fig. 2 and Fig. 3. In the corresponding 10 image per class setting for TinyImageNet demonstrated in Fig. 4, this is less severe. This may indicate that in 1 image per class setting, the quantiles can't be matched perfectly if we want to maintain the image details.

Figure 1. Synthetic image dataset learned by IDM+LQM on CIFAR10 with 10 image per class, each row corresponds to a class.

Figure 2. Synthetic image dataset learned by IDM+LQM on CIFAR100 with 1 image per class.

Figure 3. Synthetic image dataset learned by IDM+LQM on TinyImageNet with 1 image per class. Only the first 100 classes are shown.

Figure 4. Synthetic image dataset learned by IDM+LQM on TinyImageNet with 10 image per class. Only the first 10 classes are shown, each row corresponds to a class.

References

[1] Alessandro Barbiero and Asmerilda Hitaj. Discrete approximations of continuous probability distributions obtained by minimizing cramér-von mises-type distances. *Statistical Papers*, 64(5):1669–1697, 2023. 2