

## References

- [1] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.



Figure 1: **WAE-MMD**: Architecture A2,  $\lambda = 100$  (our hyper-param settings)



Figure 2: **WAE-MMD**: Architecture A2, our hyper-param settings except  $\lambda = 500$

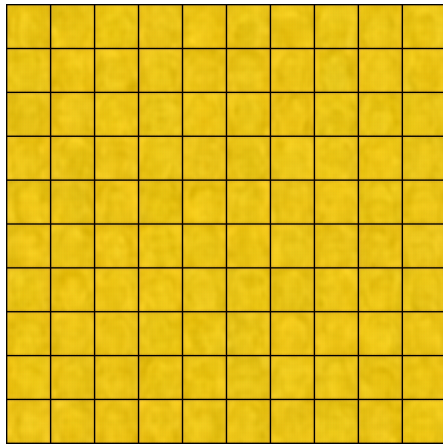


Figure 3: **WAE-MMD**: Architecture A1, our hyper-param settings except  $\lambda = 1$



Figure 4: **WAE-MMD**: Architecture A1,  $\lambda = 100$  (our hyper-param settings)



Figure 5: **WAE-MMD**: Architecture A1,  $\lambda = 100$  (hyper-params from [1])



Figure 6: **MoCA**: Architecture A2,  $\lambda = 100$  (our hyper-param settings)

Figure 7: Randomly generated samples using WAE-MMD using various hyper-parameter configurations, and sample from MoCA model.

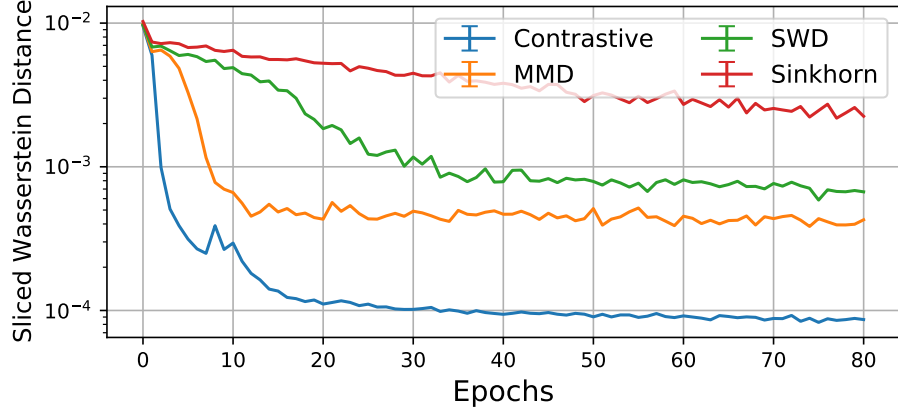


Figure 8: Systematic analysis of convergence speed for various objectives for matching the marginal distribution of the latent space to a prior, without the reconstruction term. These experiments are on the **synthetic dataset**, identical to that used in Fig 1 in the main paper. Each model is run using 5 different seeds and mean and standard deviation values are reported (variance between runs was typically extremely small and unnoticeable). Contrastive loss is faster than existing methods at this task and achieves lower SWD values.

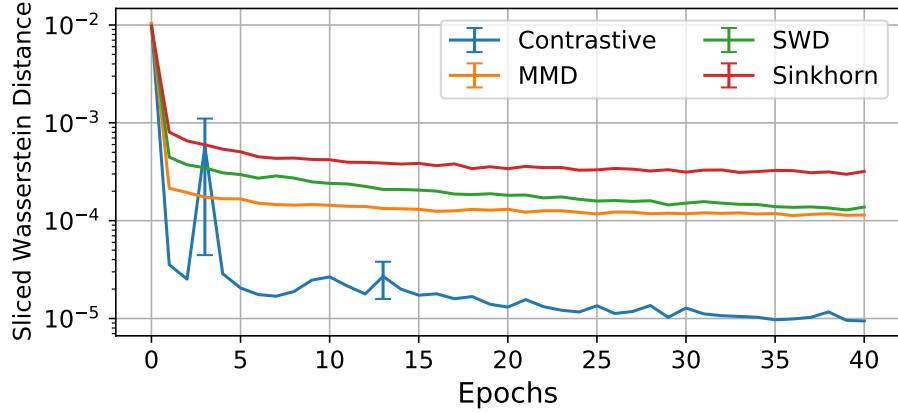


Figure 9: Systematic analysis of convergence speed for various objectives for matching the marginal distribution of the latent space to a prior, without the reconstruction term. These experiments are on the **CIAFR-10** dataset. Each model is run using 5 different seeds and mean and standard deviation values are reported (variance between runs was typically extremely small and unnoticeable). Contrastive loss is faster than existing methods at this task and achieves lower SWD values.

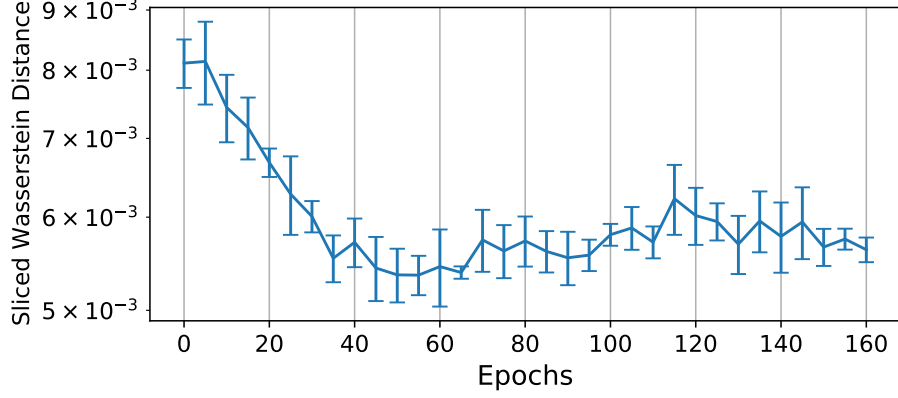


Figure 10: Systematic analysis of convergence for **GAN** objective for matching the marginal distribution of the latent space to a prior, without the reconstruction term. This experiment is on the **synthetic dataset**, identical to that used in Fig 1 in the main paper. Each model is run using 5 different seeds and mean and standard deviation values are reported.

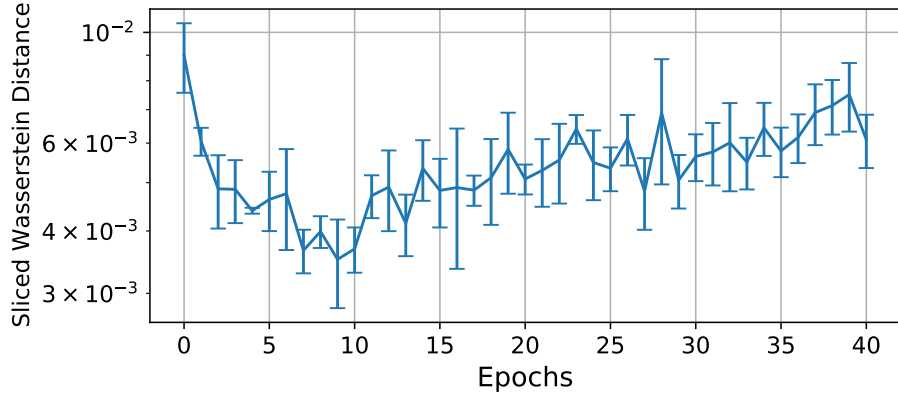


Figure 11: Systematic analysis of convergence for **GAN** objective for matching the marginal distribution of the latent space to a prior, without the reconstruction term. This experiment is on the **CIFAR-10 dataset**, identical to that used in Fig 1 in the main paper. Each model is run using 5 different seeds and mean and standard deviation values are reported.