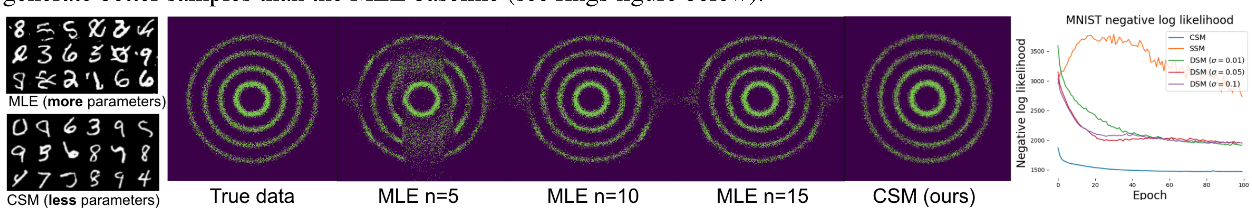
We thank all reviewers for providing constructive feedback. We are glad that [R1] believes our work is a **timely and relevant** contribution. We thank [R1, R2, R3, R4] for acknowledging the **theoretical contribution** of the paper (a new **divergence** for learning unnormalized autoregressive (AR) models), and [R2, R3, R4] for appreciating the **novelty** and **motivation** of our work. [R1, R4] raise questions about the configuration in the experiments. We believe this is due to a misunderstanding on the architectures used and number of parameters.

[R4] Q1: Model architectures for image experiments. The PixelCNN++ baseline model is a deep network with $>$100 layers. "ResNet" in appendix refers to a **group** of convolution blocks for each of the many gated ResNet layers we use. We apologize for the confusion and will clarify this as well as upload the code. For comparison, our AR-CSM model uses **exactly the same** AR model architecture (also with **convolutional architecture**) as the MLE baseline. However, unlike the MLE baseline which passes the output of the AR model to a pre-specified **normalized** density function (*e.g.* mixture of logistics), we pass to a score network (with $< 1\%$ parameters compared to the AR part) and learn an **unnormalized** density function via the proposed CSM divergence. We provide additional experiments showing that CSM can **outperform** MLE baselines even with **strictly less** parameters (see **[R1] Q1** MNIST and rings below).

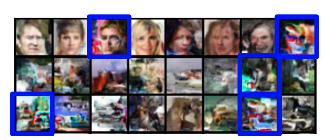
[R4] Q2: Clarification on experiments setups. We run all the experiments using exactly the same setting on a 12 GB TITAN Xp GPU. We briefly mention this at line 186. We will clarify this more in the revision. We use $\sigma = 0$ for both CSM and SSM in Figure 2 as they already worked well without noise perturbation in this setting.

[R1] Q1: Extra parameters introduced by score network. The extra number of parameters from the score network is almost **negligible** (*i.e.* $< 1\%$ compared to the autoregressive part). Empirically, we find that CSM is able to outperform an MLE baseline even with **strictly less** parameters (including the score network) by generating better MNIST digit samples (see MNIST samples below). We also provide a "rings" synthetic experiment where we use **strictly less** parameters for the AR-CSM model than the baseline MLE model. We use a MADE architecture for the AR model and n mixtures of logistic components for the MLE experiments. Even with **strictly less** parameters, CSM is still able to generate better samples than the MLE baseline (see rings figure below).



[R1] Q2: Figure 2 loss curves and advantage over SSM. We use Figure 2 to provide insights into the training challenges of DSM and SSM, and we do not intent to claim better density estimation from Figure 2. To compare density estimation performance, we train a MADE model with tractable likelihood using the three score matching methods on MNIST (a **more complicated** distribution than the one in Figure 2) and report the negative log likelihood (see the figure above). The loss curves in the above figure match our discussion in Section 5.1. For DSM, a smaller σ introduces less bias, but also makes training slower to converge. SSM can introduce a **high variance** when approximating the trace of the Hessian matrix. CSM, however, converges quickly. We believe this shows the efficacy of CSM over the other score matching methods for density estimation.

[R1] Q3: Whether expressivity gained by unnormalized density is helpful. In Section 6 Table 2, **all** the methods except for ELBO use unnormalized densities; CSM (unnormalized) outperforms ELBO (normalized) by a significant amount in all the settings. We believe the expressivity provided by an unnormalized density is helpful.



[R1] Q4: Less shifted color compared to baseline and denoising results. Although it is difficult to quantitatively measure "shifted color" in samples, we believe the samples marked in blue (from baseline) have inconsistent "shifted colors". CSM samples, in contrast, have more consistent colors according to human observers. We believe image denoising is not a simple task, and while we do not claim SOTA results, Figure 4 shows the capability of CSM to capture complex distributions. Our image results also show the effectiveness of CSM compared to the previous approach for training unnormalized AR models (see Figure 6).

[R2] Q1: Writing suggestions and ancestral sampling. We thank Reviewer #2 for pointing out the typos and writing suggestions. We will fix them in the revision. The quality of subsequent samples $x_{>d}$ does depend on earlier samples $x_{<d}$. We find our sampling algorithm able to generate $x_{<d}$ that work reasonably well in practice.

[R3] Q1: Comparison with normalizing flows. We thank Reviewer #3 for the advice. Due to time constraint, we only perform normalizing flow experiments on MNIST. We use flow models with comparable number of parameters as CSM. The flow models obtain AIS scores of 95.69 and 88.91 for VAE experiments with latent dimension 8 and 16 on MNIST. We notice that CSM **outperforms** the flow models with these settings.