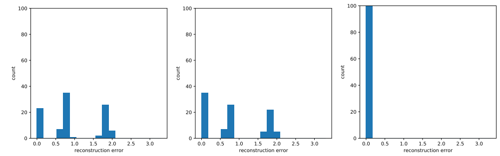


(a) GD (b) NGD (c) AIS-HMC

Figure 1: Error bar plots for GAN after 15k iterations on synthetic data (following Fig. 2 in main paper). AIS-HMC performs best.

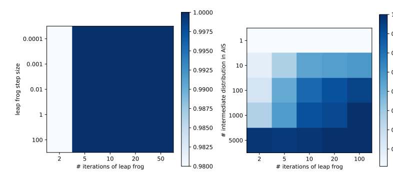
We thank all reviewers for useful feedback.

To Reviewer 1:

Re. noisy gradient descent (NGD) baseline: Thanks for the suggestion. We ran this baseline on synthetic (Fig. 1) and image (Fig. 4) data, using sigmoid annealing schedule. We observe NGD to improve over GD on synthetic data. This is intuitive as NGD is better able to escape some local optima. However, even on synthetic data it performs nowhere close to the AIS-HMC method.

Re. exploration of complex distributions: Sampling from a multi-modal distribution is challenging. Particularly if the modes are well separated it is important to adequately explore the domain in order not to get stuck in a single mode. To observe this we study the ability to sample from a multi-modal distribution on our synthetic data. We use observation $x_o = x_1 = -1$ which retains an ambiguous $x_2 = 0.5$ or $x_2 = -0.5$. Results are shown in Fig. 3.

Re. other applications: While other applications are also insightful, we think that synthetic, inpainting and super resolution already display the major challenges: the loss is ragged when optimizing w.r.t. the latent variable. AIS-HMC addresses this ill-posed challenge well, particularly also ambiguity (see Fig. 3 where there exists more than one correct prediction given $x_o = x_1 = -1$).



(a) (b)

Figure 2: Ratio of trials reaching a reconstruction error less than 0.2: (a) leap frog step size over the number leap frog iterations and (b) the number of leap frog iterations over the number of intermediate distributions in AIS.

Re. ablation study: We perform two studies and show results in Fig. 2: (1) leap frog step size over leap frog iterations; and (2) leap frog iterations over number of intermediate AIS distributions, i.e., the number of HMC iterations. From (1) we see the method is stable. This is due to HMC adjusting leapfrog step size and acceptance probability. From (2) we note that performance is suboptimal with few intermediate AIS distributions. With one AIS distribution, we run vanilla HMC. This shows that AIS-based HMC has a big advantage over just HMC.

Re. FID: FID is not an appropriate metric for co-generation. We are interested in retrieving a reconstruction which best fits the given data, hence accuracy matters. Note, diversity does not necessarily exist. In contrast FID assesses diversity (among others). This being said, we do agree that research on how to better assess this task is necessary. This is however beyond the scope.

Re. AIS and necessity: It is AIS-based HMC because we use an annealing process to move samples from a tractable distribution to the target distribution via a sequence of intermediate steps (line 4 in the algorithm). HMC would directly sample from the target distribution. For complex distributions obtained from GANs we found plain HMC to be challenging (see Fig. 2b when the number of intermediate distributions in AIS=1).

Re. Fig. 4 (paper): It is first referenced five lines after discussing the experiment (L254). We'll present better. Error bars and additional baselines are shown in Fig. 4 (rebuttal).

Re. SGD: Thanks for pointing out, it should be GD, we'll revise.

Re. train/test data: We follow and use prior work, e.g., Progressive GAN. They do not split train/test data. We use 100 images for the metric evaluation for all real image tasks. We added error bars to Fig. 4 (rebuttal).

To Reviewer 2:

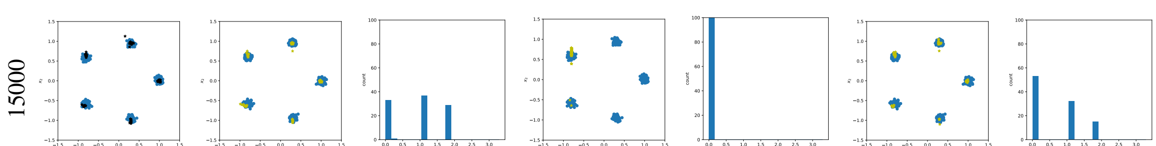
Re. optimizing N points: Thanks for suggesting. This improves slightly compared to gradient descent, but is computationally more expensive. The AIS-HMC method still has a significant edge. See 'MultiOpt+GD' in Fig. 4(c, d).

Re. run-time: For CelebA data, AIS-HMC takes approximately 13min (0.1min for one HMC step). GD and NGD both take approximately 15min for the 30,000 GD update iterations that we use.

Re. other techniques & clarity: We'll discuss, add references, e.g., to Dinh et al., and clarify.

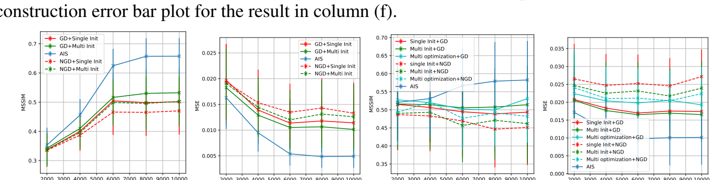
To Reviewer 3:

Re. inpainting & writing: Please note that this isn't an inpainting paper. We are interested in studying the co-generation task, i.e., how to optimize w.r.t. latent samples. This is more general than inpainting. Agreed, specific methods can be trained for each task, however, this isn't the point. Thanks for pointing out writing, we'll clarify.



(a) (b) (c) (d) (e) (f) (g)

Figure 3: Columns illustrate: (a) Samples generated with a vanilla GAN (black); (b) GD reconstructions from 100 random initializations; (c) Reconstruction error bar plot for the result in column (b); (d) Reconstructions recovered with Alg. 1; (e) Reconstruction error bar plot for the results in column (d). (f) NGD reconstructions from 100 random initializations; (g) Reconstruction error bar plot for the result in column (f).



(a) higher is better (b) lower is better (c) higher is better (d) lower is better

Figure 4: Reconstruction error over Progressive GAN training iterations: (a) MSSIM on CelebA; (b) MSE on CelebA; (c) MSSIM on LSUN; (d) MSE on LSUN.