A Population Annealing vs. Persistence Contrastive Divergence

In this section, we compare population annealing (PA) to persistence contrastive divergence (PCD) for sampling in the negative phase. In Table 2, we train DVAE# with the power-function smoothing on the binarized MNIST dataset using PA and PCD. As shown, PA results in a comparable generative model when there is one group of latent variables and better models in other cases.

Table 2: The performance of DVAE# with power-function smoothing for binarized MNIST when PCD or PA is used in the negative phase.

Struct.	K	PCD	PA
1 —	1	**89.25±0.04**	89.35±0.06
	5	**88.18±0.08**	**88.25±0.03**
	25	**87.66±0.09**	**87.67±0.07**
1 ∼	1	**84.95±0.05**	**84.93±0.02**
	5	**84.25±0.04**	**84.21±0.02**
	25	**83.91±0.05**	**83.93±0.06**
2 ∼	1	83.48±0.04	**83.37±0.02**
	5	83.12±0.04	**82.99±0.04**
	25	83.06±0.03	**82.85±0.03**
4 ∼	1	83.62±0.06	**83.18±0.05**
	5	83.34±0.06	**82.95±0.07**
	25	83.18±0.05	**82.82±0.02**

B On the Gradient Variance of the Power-function Smoothing

Our experiments show that power-function smoothing performs best because it provides a better approximation of the binary random variables. We demonstrate this qualitatively in Fig. 1 and quantitatively in Fig. 2(c) of the paper. This is also visualized in Fig. 3. Here, we generate 10^6 samples from $q(\zeta) = (1-q)r(\zeta|z=0) + qr(\zeta|z=1)$ for $q = 0.5$ using both the exponential and power smoothings with different values of β ($\beta \in \{8, 9, 10, \ldots, 15\}$ for exponential, and $\beta \in \{10, 20, 30, \ldots, 80\}$ for power smoothing). The value of β is increasing from left to right on each curve. The mean of $|\zeta_i - z_i|$ (for $z_i = \mathbb{1}_{[\zeta_i > 0.5]}$) vs. the variance of $\partial\zeta_i/\partial q$ is visualized in this figure. For a given gradient variance, power function smoothing provides a closer approximation to the binary variables.

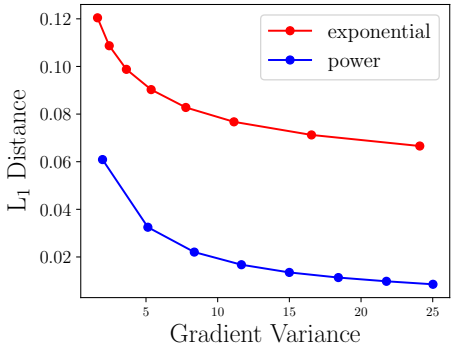

Figure 3: Average distance between ζ and its binarized z vs. variance of $\partial\zeta/\partial q$ measured on 10^6 samples from $q(\zeta)$. For a given gradient variance, power function smoothing provides a closer approximation to the binary variables.