Supplementary Material

1 Description of Datasets

We provide details in preprocessing for datasets used in the experiments in the main text. In total, we tested AntiVAE on seven datasets: static MNIST, dynamic MNIST, FashionMNIST, OMNIGLOT, Caltech 101 Silhouettes, Frey Faces, and Histopathology patches. As in previous literature, static MNIST uses a fixed binarization of images whereas dynamic MNIST resamples images from the training dataset at each minibatch. In dynamic MNIST, the validation and test sets have fixed binarization. We do an identical dynamic resampling procedure for OMNIGLOT. Caltech101 is given as binary data, so we cannot resample at training time, which we find to cause overfitting on the test set. For grayscale images that cannot be binarized, we would like to parameterize the generative model as a Gaussian distribution. In practice, we found this choice to cause over-prioritization of the reconstruction term, essentially causing the VAE to behave like a regular autoencoder. Instead, we find that a logistic distribution over the 256 grayscale domain avoids this failure mode. We use the default variance constraints as in (Tomczak and Welling, 2017). We now provide a brief description to introduce each dataset.

MNIST is a dataset of hand-written digits from 0 to 9 split into 60,000 examples for training and 10,000 for testing. We use 10,000 randomly chosen images from the training set as a validation group.

FashionMNIST Similar to MNIST, this more difficult dataset contains 28x28 *grayscale* images of 10 different articles of clothing e.g. skirts, shoes, shirts, etc. The sizing and splits are identical to MNIST.

OMNIGLOT is a dataset with 1,623 hand-wrriten characters from 50 different alphabets. Unlike the MNIST family of datasets, each character is only represented by 20 images, making this dataset more difficult. The training set is 24,345 examples with 8,070 test images. We again take 10% of the training data as validation.

Caltech 101 Silhouettes contains silhouettes of 101 different object classes in black and white: each image has a filled polygon on a white background. There are 4,100 training images, 2,264 validation datapoints and 2,307 test examples. Like OMNIGLOT, this task is difficult due to the limited data set.

FreyFaces is a collection of portrait photos of one individual with varying emotional expressions for a total of 2,000 images. We use 1,565 for training, 200 validation, and 200 test examples.

Histopathology Patches is a dataset from ten different biopsies of patients with cancer (e.g. lymphona, leukemia) or anemia. The dataset is originally in color with 336 x 448 pixel images. The data was processed to be 28 x 28 grayscale. The images are split in 6,800 training, 2,000 validation, and 2,000 test images. We refer to (Tomczak and Welling, 2016) for exact details.

All splitting was either given by the dataset owners or decided by a random seed of 1.

1.1 Evaluation Details

To compute the test log-likelihood (in any of the experiments), we use $k = 100$ samples to estimate the following:

$$\log \mathbb{E}_{z\sim q_\phi(z|x)}[\frac{p_\theta(x,z)}{q_\phi(z|x)}] \approx \log \frac{1}{k}\sum_{i=1}^{k}[\frac{p_\theta(x,z)}{q_\phi(z|x)}] \tag{1}$$

where $q_\phi(z|x)$ is an amortized inference network, $p_\theta(x|z)$ is a generative model, and $p(z)$ is a simple prior. Notably, we use (unbiased) i.i.d. samples to estimate Eqn. 1. The final number reported is the average test log likelihood across the test split of the dataset.

1.2 Approximate Antithetic Sampling Algorithm

We explicitly write out the approximate algorithm for antithetic sampling. Note the similiarity to Alg. 2 in the main text; the only distinction is that we use a derived approximation to the inverse CDF tranform for a Chi-squared random variable.. Here, we refer to this as ApproxAntitheticSample. In the main text, this algorithm is often referred to as AntitheticSample.

Algorithm 1: ApproxAntitheticSample

Data: i.i.d. samples $(x_1, ..., x_k) \sim \mathcal{N}(\mu, \sigma^2)$; i.i.d. samples $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_{k-1}) \sim \mathcal{N}(0, 1)$; Population mean μ and variance σ^2; Number of samples $k \in \mathbb{N}$.

Result: A set of k samples $x_{k+1}, x_{k+2}, ..., x_{2k}$ marginally distributed as $\mathcal{N}(\mu, \sigma^2)$ with sample mean η' and sample standard deviation δ'.

$v = k - 1$;
$\eta = \frac{1}{k} \sum_{i=1}^{k} x_i$;
$\delta^2 = \frac{1}{k} \sum_{i=1}^{k} (x_i - \eta)^2$;
$\eta' = 2\mu - \eta$;
$\lambda = v\delta^2/\sigma^2$;
$\lambda' = v(2(1 - \frac{3}{16v} - \frac{7}{512v^2} + \frac{231}{8192v^3}) - (\frac{\lambda}{v})^{1/4})^4$;
$(\delta')^2 = \lambda'\sigma^2/v$;
$(x_{k+1}, ..., x_{2k}) = \text{MarsagliaSample}(\boldsymbol{\epsilon}, \eta', (\delta')^2, k)$;
Return $(x_{k+1}, ..., x_{2k})$;

2 Proofs of Propositions

In this section, we provide more rigor in proving (1) properties of antithetic samplers using Marsaglia's method and (2) properties of Marsaglia's method itself. In particular, we provide the proof of Proposition 1 (from the main text).

2.1 Properties of Antithetic Sampling

Theorem 2.1. *Let $p(x)$ be a distribution over $\mathbb{R}$ and let $p(\zeta) = \prod_i^k p(x_i)$ be the distribution of k i.i.d. samples. Let $T : \mathbb{R}^k \to \mathbb{R}^l$ be a (l-dimensional) statistic of ζ. Let $s(t)$ be the induced distribution of this statistic: $s(t) = \int_\zeta \delta_{T(\zeta)=t} p(\zeta) d\zeta$. Let $F : \mathbb{R}^l \to \mathbb{R}^l$ be a deterministic function such that $s(F(t)) = s(t)$. We now construct a sample $\bar{\zeta}$ by: sampling $\zeta \sim p(\zeta)$, computing $\bar{t} = F(T(\zeta))$, sampling $\bar{\zeta} \sim p(\zeta|\bar{t})$ from the conditional given $T(\zeta) = \bar{t}$. This $\bar{\zeta}$ is distributed according to $p(\zeta)$ and, in particular, it's elements are i.i.d. according to $p(x)$.*

Proof. We begin by noting that:

$$p(\zeta) = \int_t p(\zeta|t) s(t) \tag{2}$$

By assumption,

$$s(\bar{t}) = s(F(t)) \tag{3}$$
$$= s(t). \tag{4}$$

Thus,

$$p(\bar{\zeta}) = \int_{\bar{t}} p(\zeta|\bar{t}) s(\bar{t}) \tag{5}$$
$$= \int_t p(\zeta|t) s(t) \tag{6}$$
$$= p(\zeta) \tag{7}$$

Thus $\bar{\zeta} \sim p(\zeta)$. Since $p(\zeta)$ is the distribution over i.i.d. samples from $p(x)$, the resulting elements of $\bar{\zeta}$ are also i.i.d. from $p(x)$. □

We provide one example of a function F with the desired property.

Lemma 2.2. *Let $F(t) = \text{CDF}(1 - \text{CDF}^{-1}(t))$ where CDF is the cumulative distribution function for $s(t)$. Then $s(F(t)) = s(t)$.*

Proof. Let $X \sim \text{U}(0, 1)$. By definition, $\text{CDF}(X)$ will be distributed as $s(t)$. Trivially, $\text{CDF}^{-1}(t) \sim \text{U}(0, 1)$ when $t \sim s(t)$, and so too is $1 - \text{CDF}^{-1}(t)$. □

Corollary 2.2.1. *Let $\theta = \mathbb{E}_p[h(x)]$ be a function expectation of interest with respect to a distribution, $p(x), x \in \mathbb{R}$. Let $\hat{\theta}_1$ be an unbiased Monte Carlo estimate using i.i.d. samples $\zeta \sim p(\zeta)$. Let $\hat{\theta}_2$ be an "antithetic" estimate using samples $\bar{\zeta}$ generated as in Theorem 2.1. Then the following hold,*

- *$\hat{\theta}_2$ is unbiased estimate of θ*
- *$\hat{\theta}_3 = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$ is unbiased estimate of θ*
- *Let $F = \text{CDF}(1 - \text{CDF}^{-1}(T))$. Then the first and second moments of ζ are anti-correlated to those of $\bar{\zeta}$*

Proof. By Theorem 2.1, "antithetic" samples $\bar{\zeta} \sim p(\zeta)$ i.i.d., hence $\hat{\theta}_2$ is unbiased ($\hat{\theta}_2$ is equivalent to $\hat{\theta}_1$). $\hat{\theta}_3$ is also unbiased as a linear combination of two unbiased estimators is itself unbiased. Anti-correlation of moments falls trivially from our choice of F. □

Connection to Differentiable Antithetic Sampling In the paper, we proposed the following proposition,

Proposition 1. *For any $k > 2$, $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$, if $\eta \sim \mathcal{N}(\mu, \frac{\sigma^2}{k})$ and $\frac{(k-1)\delta^2}{\sigma^2} \sim \chi^2_{k-1}$, and $\bar{\eta} = f(\eta), \bar{\delta}^2 = g(\delta^2; \sigma^2)$ for some functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, and $\epsilon = (\epsilon_1, ..., \epsilon_k) \sim \mathcal{N}(0, 1)$, then the "antithetic" samples $\zeta = (x_1, ..., x_k) = \text{MarsagliaSample}(\epsilon, \bar{\eta}, \bar{\delta}^2, k)$ are independent normal variates sampled from $\mathcal{N}(\mu, \sigma^2)$ such that $\frac{1}{k}\sum_i^k x_i = \bar{\eta}$ and $\frac{1}{k}\sum_i^k (x_i - \bar{\eta})^2 = \bar{\delta}^2$.*

Define a statistic $T = [\bar{\eta}, \bar{\delta}^2]$, and function $F = [f, g]$. Marsaglia's algorithm (or Pullin's, Cheng's) can be seen as a method for sampling from $p(\zeta|t)$ for a fixed statistic t. In Proposition 1, we first sample $t \sim s(T(\zeta))$ where $\zeta \sim \mathcal{N}(\mu, \sigma^2)$. Then, we choose "antithetic" statistics using

$$f = \text{GaussianCDF}(1 - \text{GaussianCDF}^{-1}(\eta)) \tag{8}$$

$$g = \frac{\sigma^2}{(k-1)}\text{ChiSquaredCDF}(1 - \text{ChiSquaredCDF}^{-1}(\frac{(k-1)\delta^2}{\sigma^2})) \tag{9}$$

such that $s(F(t)) = s(t)$ by symmetry in $\text{U}(0, 1)$. By Theorem 2.1, antithetic samples $\bar{\zeta}$ are distributed as ζ is. In practice, we use both ζ and $\bar{\zeta}$ for stochastic estimation, as anti-correlated moments provide empirical benefits.

2.2 Properties of Marsaglia's Method

Theorem 2.3. *Let $\epsilon = (\epsilon_1, ..., \epsilon_{k-1}) \sim \mathcal{N}(0, 1)$ auxiliary variables. Let η, δ be known variables. Then $\zeta = (x_1, ..., x_k) = \text{MarsagliaSample}(\epsilon, \eta, \delta^2, k)$ are uniform samples from the sphere*

$$S = \{(x_1, ..., x_k) | \sum_i^k x_i = k\eta, \sum_i^k (x_i - \eta)^2 = k\delta^2\}$$

Proof. S is the intersection of a hyperplane and the surface of a k-sphere: the surface of a $(k-1)$-sphere. Marsaglia uses the following to sample from S:

Let $z = (z_1, ..., z_{k-1})$ be a sample drawn uniformly from the unit $(k-1)$-sphere centered at the origin. (In practice, set $z_i = \epsilon_i / \sqrt{\sum_j^k \epsilon_j^2}$.) Let

$$\zeta = rzB + \eta v \tag{10}$$

where $v = (1, 1, ..., 1)$ and choose B to be a $(k-1)$ by k matrix whose rows form an orthonormal basis with the null space of v. By definition, $BB^t = I$ and $Bv^t = 0$ where I is the identity matrix. We note the following consequence:

$$\begin{aligned}
\zeta v^t &= (rzB + \eta v)v^t && (11)\\
&= rzBv^t + \eta v v^t && (12)\\
&= 0 + \eta v v^t && (13)\\
&= k\eta && (14)
\end{aligned}$$

$$\begin{aligned}
(\zeta - \eta v)(\zeta - \eta v)^t &= (rzB + \eta v - \eta v)(rzB + \eta v - \eta v)^t && (15)\\
&= (rzB)(rzB)^t && (16)\\
&= r^2 zBB^t z^t && (17)\\
&= r^2 zz^t && (18)\\
&= r^2 && (19)
\end{aligned}$$

Eqn. 14, 19 exactly match the constraints defined in S. So $\zeta \in S$. Further ζ is uniformly distributed in S as z is uniform over the $(k-1)$-sphere. □

Theorem 2.4. *Let $\zeta = (x_1, ..., x_k) \sim p(\zeta)$ be a random vector of i.i.d. Gaussians $\mathcal{N}(\mu, \sigma^2)$. Let $\eta = \frac{1}{k}\sum_i^k x_i$ and $\delta^2 = \frac{1}{k}\sum_i^k (x_i - \eta)^2$. Then $\eta \sim \mathcal{N}(\mu, \frac{\sigma^2}{k})$ and $\frac{(k-1)\delta^2}{\sigma^2} \sim \chi^2_{k-1}$ and η, δ^2 are independent random variables.*

Proof. This is a known property of Gaussian distributions. Reference *Statistics: An introductory analysis* or any introductory statistics textbook. □

Theorem 2.5. *Let $\zeta = (x_1, ..., x_k)$ be a random vector of i.i.d. Gaussians $\mathcal{N}(\mu, \sigma^2)$. Let $\eta = \frac{1}{k}\sum_i^k x_i =$ and $\delta^2 = \frac{1}{k}\sum_i^k (x_i - \eta)^2$ and $T = [\eta, \delta^2]$. Let $p(\zeta, T(\zeta)) = p(\zeta, \eta, \delta^2)$ denote their joint distribution.*

Then, the conditional density is of the form

$$p(\zeta | \eta = \eta, \delta^2 = \delta^2) = \begin{cases} a \text{ if } \zeta \in S \\ 0 \text{ if } \zeta \notin S. \end{cases} \quad (20)$$

where $S = \{(x_1, ..., x_k) | \sum_i x_i = k\eta, \sum_i (x_i - \eta)^2 = k\delta^2\}$, $0 < a < 1$ is a constant.

Proof.
Intuition: Level sets of a multivariate isotropic Gaussian density function are spheres. The event we are conditioning on is a sphere.

Formal Proof: Let $f(x_1, ..., x_k) = (2\pi\sigma^2)^{-k/2} e^{(-\sum_i (x_i - \mu)^2/(2\sigma^2))}$ denote a Gaussian density. Note the following derivation:

$$\begin{aligned}
\sum_{i=1}^k (x_i - \mu)^2 &= \sum_i (x_i - \eta)^2 + 2(\eta - \mu)\sum_i (x_i - \eta) + k(\eta - \mu)^2 && (21)\\
&= \sum_i (x_i - \eta)^2 + k(\eta - \mu)^2 && (22)\\
&= r^2 + k(\eta - \mu)^2 && (23)
\end{aligned}$$

This implies $f(x_1, ..., x_k)$ is equal for any $(x_1, ..., x_k) \in S$. Thus, the conditional distribution $p(\zeta | \zeta \in S)$ is the uniform distribution over S for any μ, σ. □

Finally, proof of Proposition 1 from the paper (denoted as Proposition 2 here):

Proposition 2. *For any $k > 2$, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, if $\eta \sim \mathcal{N}(\mu, \frac{\sigma^2}{k})$ and $\frac{(k-1)\delta^2}{\sigma^2} \sim \chi^2_{k-1}$ and $\boldsymbol{\epsilon} = \epsilon_1, ..., \epsilon_{k-1} \sim \mathcal{N}(0,1)$ i.i.d., then the generated samples $x_1, ..., x_k = \text{MARSAGLIASAMPLE}(\boldsymbol{\epsilon}, \eta, \delta^2, k)$ are independent normal variates sampled from $\mathcal{N}(\mu, \sigma^2)$ such that $\frac{1}{k}\sum_i x_i = \eta$ and $\frac{1}{k}\sum_i (x_i - \eta)^2 = \delta^2$.*

Proof. Let $\zeta = (x_1, ..., x_k)$ be a random vector of i.i.d. Gaussians $\mathcal{N}(\mu, \sigma^2)$. Compute $\eta = \frac{1}{k}\sum_i^k x_i$ and $\delta^2 = \frac{1}{k}\sum_i^k (x_i - \eta)^2$ and $T = [\eta, \delta^2]$. Let $p(\zeta, T(\zeta)) = p(\zeta, \eta, \delta^2)$ denote their joint distribution. Factoring

$$p(\zeta, \eta, \delta^2) = p(\eta, \delta^2) p(\zeta \mid \eta, \delta^2)$$

, it is clear that we can sample from the joint by first sampling $\eta, \delta^2 \sim p(\eta, \delta^2)$ and then $\zeta' \sim p(\zeta \mid \eta = \eta, \delta^2 = \delta^2)$. From Theorem 2.4, we know $p(\eta, \delta^2)$ analytically and from Theorem 2.5 we know $p(\zeta \mid \eta, \delta^2)$ is a uniform distribution over the sphere. By assumption, η, δ^2 are sampled independently from the correct marginal distributions from Theorem 2.4. Then, from Theorem 2.3, we know $\text{MARSAGLIASAMPLE}(\epsilon, \eta, \delta^2, k)$ samples from the correct conditional density (i.e. from S). Thus, samples ζ' from MARSAGLIASAMPLE will have the same distribution as ζ, namely i.i.d. Gaussian. □

3 Additional Experiments

3.1 Convolutional Architectures

In the main text, we present results where $q_\phi(z|x)$ and $p_\theta(x|z)$ are parameterized by feedforward neural networks (multilayer perceptrons). While that architecture choice was made for simplicitly, we recognize that modern encoder/decoders have evolved beyond linear layers. Thus, we ran a subset of the experiments using DCGAN architectures (Radford et al., 2015). Specifically, we design $q_\phi(z|x)$ using 3 convolutional layers and $p_\theta(x|z)$ with 3 deconvolutional layers and 1 convolutional layer.

Model	stat. MNIST	dyn. MNIST	FashionMNIST	Omniglot	Caltech	Hist.
VAE	-90.58	-90.02	-2767.97	-108.97	-116.15	-3218.16
AntiVAE	-90.25	-89.53	-2762.02	-108.40	-115.14	-3213.83
VAE+IWAE	-89.19	-88.61	-2758.72	-107.52	-116.25	-3213.05
AntiVAE+IWAE	-89.01	-88.13	-2751.11	-107.44	-115.04	-3209.98

Table 1: Test log likelihoods between the VAE and AntiVAE using (de)convolutional architectures for encoders and decoders. All images were reshaped to 32 by 32 to match standard DCGAN input sizes.

Table 1 shows log-likelihoods on a test set for a variety of image datasets. Like experiments presented in the main text, we find improvements in density estimation when using antithetics. This agrees with our intuition that more representative samples benefit learning regardless of architecture choice.

3.2 Variance over Independent Runs

In the main text, we report the average test log likelihoods over 5 runs, each with a different random seed. Here, we report in Table. 2 the variance as well (which we could not fit in the main table).

Dataset	VAE	AntiVAE	VAE+IWAE	AntiVAE+IWAE	VAE+10-NF	AntiVAE+10-NF
StaticMNIST	-90.44 ± 0.031	-89.74 ± 0.066	-89.78 ± 0.080	-89.71 ± 0.059	-90.07 ± 0.033	-89.77 ± 0.042
DynamicMNIST	-86.96 ± 1.398	-86.94 ± 1.412	-86.71 ± 1.778	-86.62 ± 1.426	-86.93 ± 1.132	-86.57 ± 1.173
FashionMNIST	-2819.13 ± 1.769	-2807.06 ± 1.591	-2797.02 ± 1.714	-2793.01 ± 1.174	-2803.98 ± 1.487	-2801.90 ± 1.459
Omniglot	-110.65 ± 0.141	-110.13 ± 0.063	-109.32 ± 0.134	-109.48 ± 0.104	-110.03 ± 0.178	-109.43 ± 0.057
Caltech101	-127.26 ± 0.254	-124.87 ± 0.213	-123.99 ± 0.262	-123.35 ± 0.195	128.62 ± 0.278	-126.72 ± 0.247
FreyFaces	-1778.78 ± 4.649	-1758.66 ± 7.581	-1772.06 ± 7.275	-1771.47 ± 5.783	-1780.61 ± 4.595	-1777.26 ± 6.467
Histopathology	-3320.37 ± 6.136	-3294.23 ± 1.543	-3311.23 ± 2.859	-3305.91 ± 1.972	-3328.68 ± 5.426	-3303.00 ± 1.517

Table 2: Identical to Table 1 in the main text but we include an errorbar over 5 runs. We find the differences induced by antithetics to be significant.

3.3 Runtime Experiments

To measure runtime, we compute the average wall-time of the forward and backward pass over a single epoch with fixed hyperparameters for VAE and AntiVAE. Namely, we use a minibatch size of 128 and vary the number of samples $k = 8, 16$. The measurements are in seconds using a Titan X GPU with CUDA 9.0. The implementation of the forward pass in PyTorch is vectorized across samples for both VAE and AntiVAE. Thus the comparison of runtime should be fair. We report the results in the Table. 3.

k	Model	StaticMNIST	DynamicMNIST	FashionMNIST	OMNIGLOT	Caltech101	FreyFaces	Hist. Patches
8	VAE	0.0132 ± 0.011	0.0122 ± 0.010	0.0142 ± 0.009	0.0144 ± 0.015	0.0188 ± 0.034	0.0283 ± 0.052	0.0173 ± 0.028
8	AntiVAE	0.0179 ± 0.011	0.0156 ± 0.009	0.0173 ± 0.010	0.0164 ± 0.017	0.0220 ± 0.036	0.0334 ± 0.054	0.0196 ± 0.029
8	AntiVAE (Cheng)	0.0242 ± 0.014	0.0210 ± 0.010	0.0231 ± 0.009	0.0221 ± 0.015	0.0353 ± 0.040	0.040 ± 0.062	0.0303 ± 0.026
16	VAE	0.0228 ± 0.009	0.0182 ± 0.011	0.0207 ± 0.010	0.0181 ± 0.015	0.0275 ± 0.035	0.0351 ± 0.049	0.0245 ± 0.027
16	AntiVAE	0.0252 ± 0.009	0.0240 ± 0.011	0.0288 ± 0.010	0.0256 ± 0.015	0.0308 ± 0.035	0.0384 ± 0.049	0.0315 ± 0.027
16	AntiVAE (Cheng)	0.0388 ± 0.011	0.0396 ± 0.010	0.0452 ± 0.011	0.0399 ± 0.015	0.0461 ± 0.038	0.0550 ± 0.054	0.0505 ± 0.033

Table 3: A comparison of runtime estimates between VAE and AntiVAE over different datasets. The number reported is the number of seconds for 1 forward and backward pass of a minibatch of size 128.

To compute the additional cost of antithetic sampling, we divided the average runtimes of AntiVAE by the average runtimes of VAE and took the mean, resulting in 22.8% increase in running time (about 0.004 seconds). We note that AntiVAE (Cheng) is much more expensive as it is difficult to vectorize Helmert's transformation.

3.4 Importance of Differentiability

We report the numbers plotted in Fig.4e, which showed that differentiability in antithetic sampling is the driving force behind sample diversity. The numbers reported are averaged over 5 runs on Histopathology.

Epoch	VAE	AntiVAE (no backprop)	AntiVAE (with backprop)
1	0.302 ± 0.031	0.301 ± 0.026	0.479 ± 0.021
10	0.102 ± 0.008	0.103 ± 0.022	0.348 ± 0.024
20	0.068 ± 0.006	0.065 ± 0.010	0.143 ± 0.016
50	0.040 ± 0.005	0.033 ± 0.006	0.063 ± 0.004
100	0.030 ± 0.002	0.028 ± 0.008	0.042 ± 0.009

Table 4: Variance of the first $k/2$ samples (non-antithetics) as measured over five independent runs on Histopathology. Without backprop, the variance is roughly equivalent to regular VAE.

As an aside, we provide the following remark: it is important to check that by adding differentiability, we do not introduce any unintended effects. For example, one might ask if differentiability leads to collapse of the VAE to a deterministic autoencoder (AE), thereby learning to "sample" only the mean. To confirm that this is not the case, we measure the average variance (across dimensions and examples in the test set) of the variational posterior $q(z|x)$ when trained as a VAE versus as a AntiVAE.

Dataset	VAE	AntiVAE
StaticMNIST	0.253	0.290
DynamicMNIST	0.269	0.290
FashionMNIST	0.049	0.049
OMNIGLOT	0.208	0.285
Caltech101	0.179	0.182
FreyFaces	0.048	0.061
Histopathology	0.029	0.028

Table 5: Learned variance of the approximate Gaussian posterior with and without antithetics. We measure variance on a variety of datasets.

If differentiating through antithetic sampling led to ignoring noise, we would expect q(z—x) to be deterministic i.e. near 0 variance. This does not appear to be the case, as shown in Table. 5.

4 Deriving One-Liner Transformations

We provide a step-by-step derviation for $g(\cdot)$ in one-liner transformations, namely from Gaussian to Cauchy and Exponential. We skip Log Normal as its formulation from a Gaussian variate is trivial. Below, let X represent a normal variate and let Y be a random variable in the desired distribution family.

Exponential Let $F(X) = 1 - \exp^{-\lambda X}$. Parameters: λ.

We start with $F(F^{-1}(Y)) = Y$.

$$\begin{aligned} 1 - \exp^{-(\lambda F^{-1}(Y))} &= Y \\ \exp^{-(\lambda F^{-1}(Y))} &= 1 - Y \\ -(\lambda F^{-1}(Y)) &= \log(1 - Y) \\ \lambda F^{-1}(Y) &= -\log(1 - Y) \\ F^{-1}(Y) &= -\frac{1}{\lambda}\log(1 - Y) \end{aligned}$$

Since $1 - Y \in \mathrm{U}(0, 1)$ and $Y \in \mathrm{U}(0, 1)$, we can replace $1 - Y$ with Y.

$$F^{-1}(Y) = -\frac{1}{\lambda}\log Y$$

Cauchy Let $F(X) = \frac{1}{2} + \frac{1}{\pi}\arctan(\frac{X - x_0}{\gamma})$. Parameters: x_0, γ.

$$\begin{aligned} \frac{1}{2} + \frac{1}{\pi}\arctan(\frac{F^{-1}(Y) - x_0}{\gamma}) &= Y \\ \arctan(\frac{F^{-1}(Y) - x_0}{\gamma}) &= \pi(Y - \frac{1}{2}) \\ F^{-1}(Y) &= \gamma(\tan(\pi(Y - \frac{1}{2})) + x_0) \\ F^{-1}(Y) &= \gamma(\tan(\pi Y) + x_0) \end{aligned}$$

In practice, we only optimize over γ, fixing x_0 to be 0.

5 Deriving Antithetic Hawkins-Wixley

We provide the following derivation for computing an antithetic χ^2 variate using a normal approximation to the χ^2 distribution. We assume the reader is familiar with the inverse CDF transform (as reviewed in the main text).

(Hawkins and Wixley, 1986) presented the following fourth root approximation of a χ^2_n variate, denoted $X^{(1)}$ with n degrees of freedom as distributed according to the following Gaussian:

$$(X^{(1)}/n)^{1/4} \sim \mathcal{N}(1 - \frac{3}{16n} - \frac{7}{512n^2} + \frac{231}{8192n^3}, \frac{1}{8n} + \frac{3}{128n^2} - \frac{23}{1024n^3}) \tag{24}$$

We can separately define a unit Gaussian variate, $Z^{(1)} \sim \mathcal{N}(0, 1)$ such that

$$Z^{(1)} = ((X^{(1)}/n)^{1/4} - (1 - \frac{3}{16n} - \frac{7}{512n^2} + \frac{231}{8192n^3})) \cdot \frac{1}{\sqrt{\frac{1}{8n} + \frac{3}{128n^2} - \frac{23}{1024n^3}}} \tag{25}$$

Notice this is just the standard reparameterization trick reversed (Rezende et al., 2014).

Independently, we can define a second χ^2_n variate, $X^{(2)}$ and unit Gaussian variate $Z^{(2)}$ in the same manner.

$$Z^{(2)} = ((X^{(2)}/n)^{1/4} - (1 - \frac{3}{16n} - \frac{7}{512n^2} + \frac{231}{8192n^3})) \cdot \frac{1}{\sqrt{\frac{1}{8n} + \frac{3}{128n^2} - \frac{23}{1024n^3}}} \tag{26}$$

As each Z is distributed as $\mathcal{N}(0, 1)$, the inverse CDF transform amounts to:

$$Z^{(2)} = -Z^{(1)} \tag{27}$$

Expanding each Z, we can derive a closed form solution:

$$((X^{(2)}/n)^{1/4} - (1 - \frac{3}{16n} - \frac{7}{512n^2} + \frac{231}{8192n^3})) = ((X^{(1)}/n)^{1/4} - (1 - \frac{3}{16n} - \frac{7}{512n^2} + \frac{231}{8192n^3})) \tag{28}$$

$$(X^{(2)}/n)^{1/4} = 2(1 - \frac{3}{16n} - \frac{7}{512n^2} + \frac{231}{8192n^3})) - (X^{(1)}/n)^{1/4} \tag{29}$$

$$X^{(2)} = n[2(1 - \frac{3}{16n} - \frac{7}{512n^2} + \frac{231}{8192n^3})) - (X^{(1)}/n)^{1/4}]^4 \tag{30}$$

This is the approximation we use in the main text. Coincidentally, (Wilson and Hilferty, 1931) present a similar approximation but as a third root that is more popular. In the main text, we noted that we could not use this as it led negative antithetic variances. To see why, we first write their approximation:

$$(X^{(1)}/n)^{1/3} \sim \mathcal{N}(1 - \frac{2}{9n}, \frac{2}{9n}) \tag{31}$$

Following a similar derivation, we end with the following antithetic Wilson-Hilferty equation:

$$X^{(2)} = n[2(1 - \frac{2}{9n}) - (x^{(1)}/n)^{1/3}]^3 \tag{32}$$

The issue lies in the cube root. If $(x^{(1)}/n)^{1/3} \geq 2(1 - \frac{2}{9n})$, then inference is ill-posed as a Normal distribution with 0 or negative variance does not exist.

6 Cheng's Solution to the Constrained Sampling Problem

In the main text, we frequently reference a second algorithm, other than (Marsaglia and Good, 1980) to solve the constrained sampling problem. Here we walk through the derivation of (Cheng, 1984; Pullin, 1979) (which we present results for in the main text):

We first review a few useful characteristics of Gamma variables, then review an important transformation with desirable properties, and finally apply it to draw representative samples from a Gaussian distribution.

6.1 Invariance of Scaling Gamma Variates

We wish to show that Gamma random variables are closed under scaling by a constant and under normalization by independent Gamma variates.

Lemma 6.1. *If $x \sim \text{Gamma}(\mu, \alpha)$ where $\mu > 0$ represents shape and $\alpha > 0$ represents rate, and $y = cx$ for some constant $c \in \mathbb{R}^+$, $y \sim \text{Gamma}(\mu, \frac{\alpha}{c})$.*

Proof. In generality, let the chain rule be $f_y(y) = F_x(g^{-1}(y))|\frac{dx}{dy}|$ where f is the cumulative distribution function for a random variable. Applying this to a Gamma: $F_y(y) = \frac{\alpha^\mu (y/k)^{\mu-1} \exp^{-\alpha y/k}}{\Gamma(\mu)} = \frac{(\alpha/k)^\mu Y^{\mu-1} \exp^{-y \cdot \alpha/k}}{\Gamma(\mu)} = \text{Gamma}(\mu, \frac{\alpha}{k})$. □

Lemma 6.2. *Let $x_1, x_2, ..., x_k$ be* Gamma(μ, α) *variates and let x_{k+1} be a* Gamma$(k\mu, \alpha)$ *variate independent of x_i, for $i = 1, ..., k$. Then, $y_i = x_{k+1}(\frac{x_i}{\sum_{j=1}^{k} x_j})$ where $y_i \sim$* Gamma(μ, α).

Proof. See Aitchison (1963). □

Lemma 6.3. *If $x \sim \mathcal{N}(0, 1)$, then $x^2 \sim \chi_1^2$. Additionally, $x^2 \sim$* Gamma$(\frac{1}{2}, \frac{1}{2})$.

Proof. By definition. □

Corollary 6.3.1. *If $x \sim \mathcal{N}(0, \sigma^2)$, then $\frac{x^2}{\sigma^2} \sim \chi_1^2$. Furthermore, we can say $x^2 \sim \sigma^2\chi_1^2 = \sigma^2 \cdot$* Gamma$(\frac{1}{2}, \frac{1}{2}) =$ Gamma$(\frac{1}{2}, \frac{1}{2\sigma^2})$.

Proof. Direct application of Lemma 6.1, 6.3. □

6.2 Helmert's Transformation

Given a random sample of size k from any Gaussian distribution, Helmert's transformation (Helmert, 1875; Pegoraro, 2012) allows us to get $k - 1$ new i.i.d. samples normally distributed with zero mean and the same variance as the original distribution:

Let $x_1, ..., x_k \sim \mathcal{N}(\mu, \sigma^2)$ be k i.i.d. samples. We define the Helmert transformed variables, $y_2, ..., y_k$ as:

$$y_j = \frac{\sum_{i=j}^{k} x_i - (k + 1 - j)x_{j-1}}{[(k + 1 - j)(k + 2 - j)]^{1/2}} \tag{33}$$

for $j = 2, ..., k$. Helmert's transformation guarantees the following for new samples:

Proposition 3. *$y_2, ..., y_k$ are independently distributed according to $\mathcal{N}(0, \sigma^2)$ such that $\sum_{i=2}^{k} y_i^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2$ where $\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i$.*

Proof. See Helmert (1875) or Kruskal (1946). □

Critically, Prop. 3 also informs us that (1) the sample variance of $y_2, ..., y_k$ is equal to the sample variance of $x_1, ..., x_k$, and (2) $y_i, i = 2, ..., k$ can be chosen independently of $\bar{x}$. These properties will be important in the next subsection.

6.3 Choosing Representative Samples

We are tasked with the following problem: we wish to generate k i.i.d. samples $x_1, ..., x_k \sim \mathcal{N}(\mu, \sigma^2)$ subject to the following constraints:

$$\frac{1}{k}\sum_{i=1}^{k} x_i = \bar{x} = \eta \tag{34}$$

$$\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2 = s^2 = \delta^2 \tag{35}$$

where by definition $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{k})$ and $(k - 1)s^2/\sigma^2 \sim \chi_{k-1}^2$. We assume that η and $(k - 1)\delta^2/\sigma^2$ are particular values drawn from these respective sample distributions. In other words, given all the possible sets of k samples, we wish to choose a single set such that the sample moments match a particular value, $\bar{x} = \eta$ and $s^2 = \delta^2$. Note that this is *not* the same as choosing any $\eta \in \mathbb{R}$ and $\delta^2 \in \mathbb{R}$.

This problem is difficult as the number of sets of k samples that do not satisfy Constraints 34 and 35 is much larger than the number of sets that do. Thus, randomly choosing k samples will not work. Furthermore, preserving

that the samples are i.i.d. makes this much more difficult as we cannot rely on common methods like sampling without replacement, rejecting samples, etc.

To tackle this, Pullin (1979) used the two following observations: (1) we can handle Constraint 34 independently, and (2) as a linear transformation, Helmert is invertible. First, we investigate observation 1:

Helmert's transformations allows us to untie Constraint 34 from Constraint 35 as $y_2, ..., y_k$ are not dependent on μ or η. Suppose we instantiate a new variable, y_1 (Kendall et al., 1946) such that

$$\eta = \mu + y_1/\sqrt{k} \tag{36}$$

As $\eta \sim \mathcal{N}(\mu, \frac{\sigma^2}{k})$, y_1 is then distributed as $\mathcal{N}(0, \sigma^2)$ by reparameterization. This means that we can deterministically choose a value for y_1 given μ and η to satisfy Constraint 34.

Next, satisfying Constraint 35 amounts to sampling $y_2, ..., y_k$ according to Prop. 3. To do this, we follow (Cheng, 1984) and use the Gamma properties we introduced in Section 6.1:

First, we draw $k-1$ independent samples from $z_2, ..., z_k \sim \mathcal{N}(0, 1)$. Compute $c_2, ..., c_k$ where $c_i = (z_i * \sigma)^2$. Cheng (1984) defines $y_i, i = 2, ..., k$ such that

$$y_i^2 = \frac{(k-1)\delta^2 \cdot c_i}{\sum_{j=2}^{k} c_j} \tag{37}$$

By design, $\sum_i y_i^2 = (k-1)\delta^2$, as desired by Prop. 3. Furthermore, as $c_i \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2\sigma^2})$ and $(k-1)\delta^2 \sim \text{Gamma}(\frac{k-1}{2}, \frac{1}{2\sigma^2})$, Lemma 6.2 tells us that y_i^2 are also distributed as $\text{Gamma}(\frac{1}{2}, \frac{1}{2\sigma^2})$, which crucially guarantees $y_i \sim \mathcal{N}(0, \sigma^2)$ by Corollary 6.3.1. For $i = 2, ..., k$, we do the following:

$$y_i = b_i' \cdot \sqrt{y_i^2} \tag{38}$$

where $b_i = \text{Bern}(0.5)$ and $b_i' = 2b_i - 1$ i.e. we randomly attach a sign to y_i. Finally, now that we know how to generate $y_1, ..., y_k$, we use Pullin (1979)'s second observation to transform y_i back to x_i:

Precisely, the inverse of Eqn. 33 (Helmert) is the following:

$$x_1 = \frac{1}{k}(k\eta - \sqrt{k(k-1)}y_2) \tag{39}$$

$$x_j = x_{j-1} + \frac{(k+2-j)^{1/2}y_j - (k-j)^{1/2}y_{j+1}}{(k+1-j)^{1/2}} \tag{40}$$

for $j = 2, ..., k$. By the "inverse" of Prop. 3, Eqn. 40 will transform $y_1, ..., y_k$ to samples $x_1, ..., x_k \sim \mathcal{N}(0, \sigma^2)$ such that the sample mean is $\eta - \mu$ and the sample variance is δ^2. Lastly, adding $x_i = x_i + \mu$ for $i = 1, ..., k$ ensures samples from the correct marginal distribution along with the correct sample moments.

We refer to this procedure as ChengSample, detailed in Alg. 2. We summarize the properties of ChengSample in the following proposition.

Proposition 4. *Given $k-1$ i.i.d samples $z_1, ..., z_{k-1} \sim \mathcal{N}(0, 1)$; $k-1$ i.i.d samples $b_1, ..., b_{k-1} \sim$ Bern(0.5); population moments from a Gaussian distribution $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}$; and desired sample moments η, δ^2 such that $\eta \sim \mathcal{N}(\mu, \frac{\sigma^2}{k})$ and $(k-1)\delta^2/\sigma^2 \sim \chi^2_{k-1}$, generated samples $x_1, ..., x_k$ from* ChengSample$([z_1, ..., z_{k-1}], [b_1, ..., b_{k-1}], \mu, \sigma, \eta, \delta, k)$ *are (1) i.i.d., (2) marginally distributed as $\mathcal{N}(\mu, \sigma^2)$, and (3) have a sample mean of η and a sample variance of δ^2.*

As a final note, we chose to use Marsaglia's solution instead of Cheng's as the former as a nice geometric interpretation and requires half as many random draws (no Bernoulli variables needed in Marsaglia's algorithm).

Algorithm 2: ChengSample

Data: i.i.d. samples $z_1, ..., z_{k-1} \sim \mathcal{N}(0,1)$; i.i.d. samples $b_1, ..., b_{k-1} \sim \text{Bern}(0.5)$; Population mean μ and variance σ^2; Desired sample mean η and variance δ^2; Number of samples $k \in \mathbb{N}$.

Result: A set of k samples $x_1, x_2, ..., x_k$ marginally distributed as $\mathcal{N}(\mu, \sigma^2)$ with sample mean η and sample variance δ^2.

$c_i = (z_{i-1} \cdot \sigma)^2$ for $i = 2, ..., k$;
$a = (k-1)\delta^2 / \sum_i c_i$;
$y_i^2 = a \cdot c_i$ for $i = 2, ..., k$;
$y_i = (2b_{i-1} - 1) \cdot \sqrt{y_i^2}$ for $i = 2, ..., k$;
$y_1 = \sqrt{k}(\eta - \mu)$;
$\alpha_k = k^{-\frac{1}{2}}$;
$\alpha_j = (j(j+1))^{-\frac{1}{2}}$ for $j = 2, ..., k$;
$s_k = \alpha_k^{-1} y_k$;
for $j \leftarrow k$ **to** 2 **do**
  $x_j = (s_j - \alpha_{j-1}^{-1} y_{j-1})/j$;
  $s_{j-1} = s_j - x_j$;
end
$x_1 = s_1$;
$x_i = x_i + \mu$ for $i = 1, ..., k$;
Return $x_1, ..., x_k$;

7 Miscellaneous

In the AntitheticSample proposition in the main text, we use the fact that the average of two unbiased estimators is an unbiased estimator. We provide the proof here.

Lemma 7.1. *A linear combination of two unbiased estimators is unbiased.*

Proof. Let e_1 and e_2 denote two unbiased estimators that $\mathbb{E}[e_1] = \mathbb{E}[e_2] = \theta$ for some underlying parameter θ. Define a third estimator $e_3 = k_1 e_1 + k_2 e_2$ where $k_1, k_2 \in \mathbb{R}$. We note that $\mathbb{E}[e_3] = k_1 \mathbb{E}[e_1] + k_2 \mathbb{E}[e_2] = (k_1 + k_2)\theta$. Thus e_3 is unbiased if $k_1 + k_2 = 1$. □

References

J. Aitchison. Inverse distributions and independent gamma-distributed products of random variables. *Biometrika*, 50(3/4):505–508, 1963.

R. C. Cheng. Generation of inverse gaussian variates with given sample mean and dispersion. *Applied statistics*, pages 309–316, 1984.

D. M. Hawkins and R. Wixley. A note on the transformation of chi-squared variables to normality. *The American Statistician*, 40(4):296–298, 1986.

F. Helmert. Über die berechnung des wahrscheinlichen fehlers aus einer endlichen anzahl wahrer beobachtungsfehler. *Z. Math. U. Physik*, 20(1875):300–303, 1875.

M. G. Kendall et al. The advanced theory of statistics. *The advanced theory of statistics.*, (2nd Ed), 1946.

W. Kruskal. Helmert's distribution. *The American Mathematical Monthly*, 53(8):435–438, 1946.

G. Marsaglia and I. Good. C69. generating a normal sample with given sample mean and variance. *Journal of Statistical Computation and Simulation*, 11(1):71–74, 1980.

I. Pegoraro. A transformation characterizing the normal distribution. *Communications in Statistics-Theory and Methods*, 41(16-17):3060–3067, 2012.

D. Pullin. Generation of normal variates with given sample mean and variance. *Journal of Statistical Computation and Simulation*, 9(4):303–309, 1979.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

J. M. Tomczak and M. Welling. Improving variational auto-encoders using householder flow. *arXiv preprint arXiv:1611.09630*, 2016.

J. M. Tomczak and M. Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

E. B. Wilson and M. M. Hilferty. The distribution of chi-square. *Proceedings of the National Academy of Sciences*, 17(12):684–688, 1931.