A Maintaining log-submodularity in the generative model

Theorem 1. Let f be a strictly submodular function over subsets of a ground set $\mathcal{Y}$, and g be a function over the same space such that

$$\|f-g\|_\infty \leq \min_{\substack{S\neq T\\ S,T\notin\{\emptyset,\mathcal{Y}\}}} \frac{1}{4}\left[f(S)+f(T)-f(S\cup T)-f(S\cap T)\right]. \tag{4}$$

Then g is also submodular.

Proof. In all the following, we assume that S, T are subsets of a ground set $\mathcal{Y}$ such that $S \neq T$ and $S, T \notin \{\emptyset, \mathcal{Y}\}$ (the inequalities being immediate in these corner cases). Let

$$\epsilon := \min_{S,T} f(S)+f(T)-f(S\cup T)-f(S\cap T))$$

By the strict submodularity hypothesis, we know $\epsilon > 0$.

Let $S, T \subseteq \mathcal{Y}$ such that $S \neq T$ and $S, T \neq \emptyset, \mathcal{Y}$. To show the log-submodularity of g, it suffices to show that

$$g(S)+g(T) \geq g(S\cup T)+g(S\cap T).$$

By definition of ϵ,

$$f(S)+f(T)-f(S\cup T)-f(S\cap T)) \geq \epsilon$$

From equation 4, we know that

$$\max_{S\subseteq\mathcal{Y}} |f(S)-g(S)| \leq \epsilon/4.$$

It follows that

$$\begin{aligned}&g(S)+g(T)-g(S\cup T)+g(S\cap T)\\ &\quad\geq f(S)+f(T)-f(S\cup T)-f(S\cap T)-\epsilon\\ &\quad\geq 0\end{aligned}$$

which proves the submodularity of g. ☐

B Encoder details

For the MNIST encodings, the VAE encoder consists of a 2d-convolutional layer with 64 filters of height and width 4 and strides of 2, followed by a 2d convolution layer with 128 filters (same height, width and strides), then by a dense layer of 1024 neurons. The encodings are of length 32.

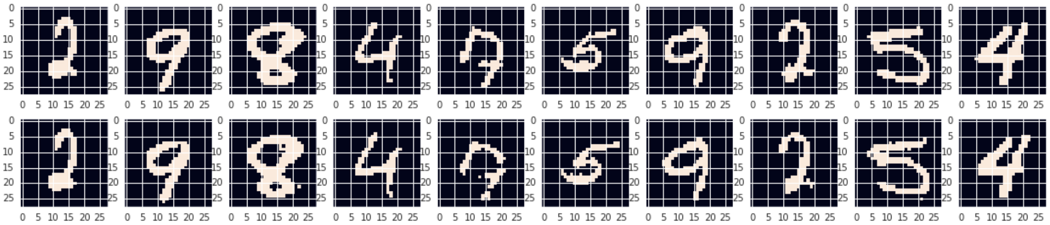

Figure 6: Digits and VAE reconstructions from the MNIST training set

CelebA encodings were generated by a VAE using a Wide Residual Network [47] encoder with 10 layers and filter-multiplier $k = 4$, a latent space of 32 full-covariance Gaussians, and a deconvolutional decoder trained end-to-end using an ELBO loss. In detail, the decoder architecture consists of a 16K dense layer followed by a sequence of 4×4 convolutions with $[512, 256, 128, 64]$ filters interleaved with $2\times$ upsampling layers and a final 6×6 convolution with 3 output channels for each of 5 components in a mixture of quantized logistic distributions representing the decoded image.