Table 1. Performance of geometric loss as a drop-in replacement in linear models for ordinal regression. Our method performs better w.r.t. its natural metric, the Hausdorff divergence.

	LR	LR(AT)	LR(IT)	g-logistic
Haus. div.	.46±.12	.47±.14	.59±.16	**.44±.08**
MAE	.44±.06	**.42±.03**	.44±.08	.45±.09
Acc.	**.66±.07**	.65±.06	.65±.06	.65±.07

slightly better in term of mean absolute error (MAE, the reference metric in ordinal regression). It thus provides a viable alternative to thresholding techniques, that performs worse in accuracy but better in MAE. It has the further advantage of naturally providing a distribution of output given an input x. We simply have, for all $y \in [d]$, $p(Y = y|\, X = x) = (\text{g-softmax}(g_{W,b}(x)))_y$.

Calibration of the geometric loss. We validate Prop. 5 experimentally on the ordinal regression dataset *car*. During training, we measure the geometric cross-entropy loss and the Hausdorff divergence on the train and validation set. Figure 3 shows that ℓ_Ω is indeed an upper bound of D_Ω, and that the difference between both terms reduces to almost 0 on the train set. Prop. 5 ensures this finding provided that the set of scoring function is large enough, which appears to be approximately the case here.

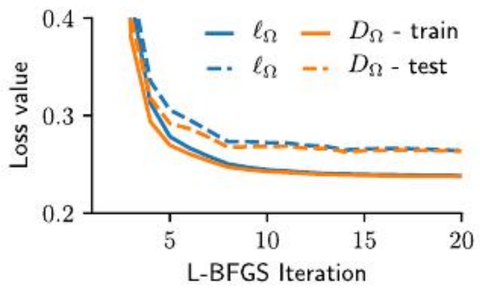

Figure 3. Training curves for ordinal regression on dataset *car*. The difference between the g-logistic loss and the Hausdorff divergence vanishes on the train set.

5.2. Drawing generation with variational auto-encoders

The proposed geometric loss and softmax are suitable to estimate distributions from inputs. As a proof-of-concept experiment, we therefore focus on a setting in which distributional output is natural: generation of hand-drawn doodles and digits, using the Google QuickDraw (Ha & Eck, 2018) and MNIST dataset. We train variational autoencoders on these datasets using, as output layers, (1) the KL divergence with normalized output and (2) our geometric loss with normalized output. These approaches output an image prediction using a softmax/g-softmax over all pixels, which is justified when we seek to output a concentrated distributional output. This is the case for doodles and digits, which can be seen as 1D distributions in a 2D space. It differs from the more common approach that uses a binary cross-entropy loss for every pixel and enables to capture interactions between pixels at the feature extraction level. We use standard KL penalty on the latent space distribution.

Using the g-softmax takes into account a cost between pixels (i, j) and (k, l), that we set to be the Euclidean cost C/σ, where C is the ℓ_2^2 cost and σ is the typical distance of interaction—we choose $\sigma = 2$ in our experiments. We therefore made the hypothesis that it would help in reconstructing the input distributions, forming a non-linear layer that captures interaction between inputs in a non-parametric way.

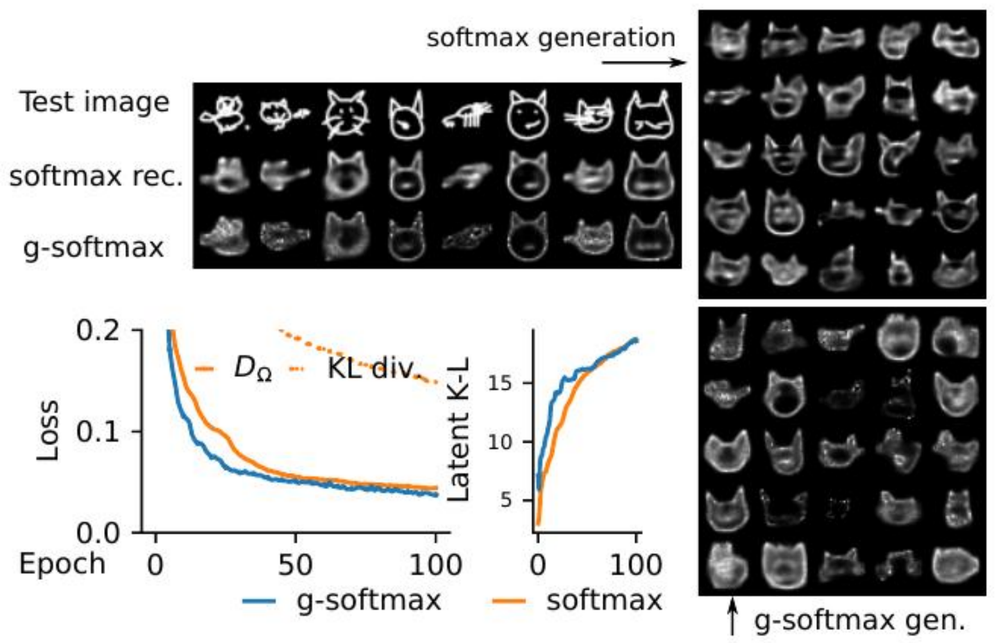

Figure 4. The g-softmax layer permits to generate and reconstruct drawing in a more concentrated manner. For a same level of variational penalty, the g-softmax better and faster minimizes the asymmetric Hausdorff divergence. See also Figure 6.

Results. We fit a simple MLP VAE on 28x28 images from the QuickDraw Cat dataset. Experimental details are reported in Appendix B (see Figure 6). We also present an experiment with 64x64 images and a DCGAN architecture, as well as visualization of a VAE fitted on MNIST. In Figure 4, we compare the reconstruction and the samples after training our model with the g-softmax and simple softmax loss. Using the g-softmax, which has a deconvolutional effect, yields images that are concentrated near the edges we want to reconstruct. We compare the training curves for both the softmax and g-softmax version: using the g-softmax link function and its associated loss better minimizes the asymmetric Hausdorff divergence. The cost of computation is again increased by a factor 10.

6. Conclusion

We introduced a principled way of learning distributional predictors in potentially continuous output spaces, taking into account a cost function in between inputs. We constructed a geometric softmax layer, that we derived from Fenchel conjugation theory in Banach spaces. The key to our construction is an entropy function derived from regularized optimal transport, convex and weak* continuous on probability measures. Beyond the experiments in discrete measure spaces that we presented, our framework opens the doors for new applications that are intrinsically off-the-grid, such as super-resolution.