Appendix

A. Proofs

We prove propositions by order of appearance in the main text.

A.1. Asymptotics of the Sinkhorn negentropy—Proof of Prop. 1

Proof. We start by showing the Shannon entropy limit of the Sinkhorn entropy, in the discrete case. In this case, we use the standard Kantorovich dual (Cuturi, 2013). Let $\varepsilon > 0$, $\alpha \in \triangle^d$, and

$$\Omega(\alpha) \triangleq \Omega_{C/\varepsilon}(\alpha) = -\max_{f \in \mathbb{R}^d} \langle \alpha, f \rangle - \langle \alpha \otimes \alpha, \exp(\frac{f \oplus f - C}{2}) \rangle + 1. \tag{9}$$

For all $f \in \mathbb{R}^d$

$$\Psi_\alpha(f) \triangleq \langle \alpha, f \rangle - \langle \alpha \otimes \alpha, \exp(\frac{f \oplus f - C/\varepsilon}{2}) + 1 \rangle = \sum_{i=1}^d f_i \alpha_i - \sum_{i,j=1}^d \alpha_i \alpha_j \exp(\frac{f_i + f_j - c_{i,j}/\varepsilon}{2}) + 1.$$

For f optimal in (9), letting $\varepsilon \to 0$, we have, using element-wise multiplication $*$,

$$\nabla \Psi_\alpha(f) = \alpha - \alpha^2 * e^f = 0 \quad \text{i.e.} \quad e^{f_i} = \frac{1}{\alpha_i} \quad \text{for all } i \in [d].$$

Replacing in (9), we obtain

$$\Omega(\alpha) = \langle \alpha, \log \alpha \rangle + \sum_{i=1}^d \alpha_i - 1 = \langle \alpha, \log \alpha \rangle.$$

Let us now consider the limit for $\varepsilon \to \infty$ of $\Omega_{C/\varepsilon}(\alpha)$, for an arbitrary symmetric cost matrix C. We rewrite

$$\Omega_{C/\varepsilon}(\alpha) = \max_{f \in \mathcal{C}(\mathcal{Y})} 2\langle \alpha, \frac{f}{2} \rangle - \varepsilon \langle \alpha \otimes \alpha, e^{\frac{\frac{f \oplus f}{2} - C}{\varepsilon}} \rangle = \mathrm{OT}_\varepsilon(\alpha, \alpha).$$

The asymptotic behavior of $\varepsilon \Omega_{C/\varepsilon}(\alpha)$, namely

$$\varepsilon \Omega_{C/\varepsilon}(\alpha) \overset{\varepsilon \to +\infty}{\longrightarrow} \frac{1}{2} \langle \alpha \otimes \alpha, -C \rangle,$$

is then a simple consequence of the asymptotics of Sinkhorn OT distances (Genevay et al., 2018), that we apply in the symmetric case. In the discrete setting, the result for $\varepsilon \to \infty$ becomes, if $C = 1 - I_{d \times d}$,

$$\frac{1}{2} \langle \alpha \otimes \alpha, I_{d \times d} - 1 \rangle = \frac{1}{2} \sum_{i=1}^d \alpha_i^2 - 1,$$

as $\langle \alpha \otimes \alpha, 1 \rangle = 1$, which concludes the proof.

□

A.2. Construction of the geometric softmax—Proof of Prop. 2

Proof. We can rewrite the self transport with the change of variable $\mu = \alpha e^{\frac{f}{2}} \in \mathcal{M}^+(\mathcal{Y})$, due to Feydy & Trouvé (2019). We then have $\frac{f}{2} = -\log \frac{\mathrm{d}\alpha}{\mathrm{d}\mu}$, and

$$\begin{aligned}\Omega(\alpha) \triangleq -\frac{1}{2}\mathrm{OT}_2(\alpha,\alpha) &= -\max_{f\in\mathcal{C}(\mathcal{Y})} \langle \alpha, f\rangle - \log\langle \alpha\otimes\alpha, \frac{\exp(f\oplus f - C)}{2}\rangle \\ &= -\max_{\mu\in\mathcal{M}^+(\mathcal{Y})} -2\langle \alpha, \log\frac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle - \log\|\mu\|_{k_2}^2, \\ \text{where}\quad \|\mu\|_{k_2} &\triangleq \int_{\mathcal{X}}\int_{\mathcal{X}} \exp(\frac{-C(x,y)}{2})\mathrm{d}\mu(x)\mathrm{d}\mu(y)\end{aligned}$$

is the kernel norm defined with kernel $k_2 \triangleq e^{-\frac{C}{2}}$. Then, the conjugate of $\Omega(\alpha)$ reads, for all $f \in \mathcal{C}(\mathcal{Y})$,

$$\begin{aligned}\Omega^\star(f) &= \max_{\alpha\in\mathcal{M}_1^+(\mathcal{Y})} \langle \alpha, f\rangle - \Omega(\alpha) \\ &= \max_{\substack{\alpha\in\mathcal{M}_1^+(\mathcal{Y})\\ \mu\in\mathcal{M}^+(\mathcal{Y})}} \langle \alpha, f\rangle - 2\langle \alpha, \log\frac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle - \log\|\mu\|_{k_2}^2 \\ &= \max_{\mu\in\mathcal{M}^+(\mathcal{Y})} \log\frac{\iint_{\mathcal{X}^2}\exp\frac{(f(x)+f(y)}{2})\mathrm{d}\mu(x)\mathrm{d}\mu(y)}{\iint_{\mathcal{X}^2}\exp(-\frac{C(x,y)}{2})\mathrm{d}\mu(x)\mathrm{d}\mu(y)},\end{aligned}$$

where we have used the conjugation of the relative entropy over the space of probability measure $\mathcal{M}_1^+(\mathcal{Y})$:

$$\max_{\alpha\in\mathcal{M}_1^+(\mathcal{Y})} \langle \alpha, f\rangle - 2\langle \alpha, \log\frac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle = 2\log\int_{\mathcal{X}}\exp(\frac{f(x)}{2})\mathrm{d}\mu(x).$$

We now revert the first change of variable, setting $\beta = \mu e^{\frac{f}{2}} \in \mathcal{M}^+(\mathcal{Y})$, and $\alpha = \frac{\beta}{\int_{\mathcal{X}}\mathrm{d}\nu} \in \mathcal{M}_1^+(\mathcal{Y})$. We have

$$\Omega^\star(f) = \max_{\alpha\in\mathcal{M}_1^+(\mathcal{Y})} -\log\iint_{\mathcal{X}^2}\exp(-\frac{f(x)+f(y)+C(x,y)}{2})\mathrm{d}\alpha(x)\mathrm{d}\alpha(y),$$

and the first part of the proposition follows:

$$\text{g-LSE}(f) = \Omega^\star(f) = -\min_{\alpha\in\mathcal{M}_1^+(\mathcal{Y})} \langle \alpha\otimes\alpha, \exp(-\frac{f\oplus f + C}{2})\rangle.$$

We have assumed that $\exp(-\frac{C}{2})$ is positive definite, which ensures that the bivariate function

$$\Phi(f,\alpha) \triangleq \langle \alpha\otimes\alpha, \exp(-\frac{f\oplus f + C}{2})\rangle \tag{10}$$

is strictly convex in α and in f. Let $\alpha^\star \triangleq \mathrm{argmin}_{\alpha\in\mathcal{M}_1^+(\mathcal{Y})}\Phi(f,\alpha)$. The gradient of Φ with respect to f is a measure that reads

$$\begin{aligned}\nabla_f\Phi(f,\alpha) &= -\alpha\,\exp(-f - T_C(-f,\alpha)) \in \mathcal{M}(\mathcal{Y}), \quad \text{where we recall} \\ T_C(f,\alpha) &\triangleq -2\log\langle \alpha, \exp(\frac{f-C}{2})\rangle.\end{aligned}$$

From a generalized version of the Danskin theorem (Bernhard & Rapaport, 1995), the function

$$f \to \operatorname*{argmin}_{\alpha\in\mathcal{M}_1^+(\mathcal{Y})} \langle \alpha\otimes\alpha, \exp(-\frac{f\oplus f + C}{2})\rangle$$

is differentiable everywhere and has for gradient $\nabla_f \Phi(f, \alpha^\star)$. Composing with the $\log$, we obtain

$$\nabla\Omega^\star(f) \in \mathcal{M}_1^+(\mathcal{Y}), \quad \text{and} \quad \nabla\Omega^\star(f) \propto \alpha^\star \exp(-f - T_C(-f, \alpha^\star)),$$

where $\propto$ indicates proportionality. To conclude, we use Lemma 2, that describes the minimizers of (10), and that we prove in the next section. It ensures that $-f - T_C(-f, \alpha^\star) = 0$ on the support of $\alpha^\star$. Therefore

$$\nabla\Omega^\star(f) = \alpha^\star \in \mathcal{M}_1^+(\mathcal{Y}),$$

and the proposition follows. □

A.3. Geometry of the link function—Proofs of Lemma 1 and Prop. 3

We first state and proof Lemma 2 on optimality condition in the minimization of $\alpha \to \Phi(\alpha, f)$. We then prove Lemma 1, establish some basic properties of the extrapolation operator and prove Prop. 3.

A.3.1. Necessary and sufficient condition of optimality in $\nabla\Omega^\star(f)$

Finding the minimizer α of $\alpha \to \Phi(\alpha, f)$ amounts to finding the distribution for which $-f$ and its C-transform $T(-f, \alpha)$ are the less distant, as it appears in the following lemma.

Lemma 2 ($\nabla\Omega^\star$ from first order optimality condition)**.** *$\nabla\Omega^\star(f)$ is the only distribution $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$ such that there exists a constant $A \in \mathbb{R}$ such that*

$$\begin{aligned} \frac{f(y) + T(-f, \alpha)(y)}{2} &= A \quad \forall\, y \in \operatorname{supp} \alpha \\ \frac{f(y) + T(-f, \alpha)(y)}{2} &\leq A \quad \forall\, y \in \mathcal{Y} / \operatorname{supp} \alpha, \end{aligned} \tag{11}$$

We then have $A = 2\Omega^\star(f)$. (11) form sufficient optimality conditions for finding $\nabla\Omega^\star(f) = \alpha$.

Proof. We use an infinite version of the KKT condition (Luenberger, 1997, Section 9) to solve the optimality of ϕ, as defined in (10). We fix $f \in \mathcal{C}(\mathcal{Y})$. The Lagrangian associated to the minimization of $\alpha \to \phi(f, \alpha)$ over the space of probability measure $\mathcal{M}(\mathcal{X})$ reads

$$L(\alpha, \mu, \nu) \triangleq \Phi(f, \alpha) + \langle \alpha, \mu \rangle + \nu(\langle \alpha, 1 \rangle - 1).$$

A necessary and sufficient condition for $\alpha^\star$ to be optimal is the existence of a function $\mu \in \mathcal{C}(\mathcal{Y})$ and a real $\nu \in \mathbb{R}$ such that,

$$\begin{aligned} \alpha^\star \in \mathcal{M}_1^+(\mathcal{Y}) \quad &\text{(primal feasibility)}, \\ \forall\, y \in \mathcal{Y}, \quad -\nabla_\alpha \Phi(f, \alpha^\star)(y) = \mu(y) + \nu \quad &\text{(stationarity)}, \\ \forall\, y \in \mathcal{Y}, \quad \mu(y) \leq 0 \quad &\text{(dual feasibility)}, \\ \forall\, y \in \operatorname{supp}(\alpha^\star), \quad \mu(y) = 0 \quad &\text{(complementary slackness)}, \end{aligned}$$

where the derivative $\nabla_\alpha \Phi(f, \alpha^\star)$ is the displacement derivative (5), computed as

$$\nabla_\alpha \Phi(f, \alpha^\star)(y) = 2 \exp(-\frac{f + T(-f, \alpha)}{2}).$$

Therefore

$$\begin{aligned} \frac{f + T(-f, \alpha^\star)}{2} &= -\log(-\frac{\nu}{2}) \quad \text{on the support of } \alpha^\star, \text{ and} \\ \frac{f + T(-f, \alpha^\star)}{2} &= -\log(-\frac{\mu(y) + \nu}{2}) \leq -\log(-\frac{\nu}{2}) \quad \text{otherwise.} \end{aligned} \tag{12}$$

Replacing in the definition $\Omega^\star(f) = -\log \Phi(f, \alpha^\star)$, and using the equality

$$\Phi(f, \alpha) = \langle \alpha, \exp(-\frac{f + T(-f, \alpha)}{2}) \rangle$$

we obtain

$$-\log(-\frac{\nu}{2}) = \Omega^\star(f),$$

and the first part of the lemma follows. Then, note that $T(f + c, \alpha) = T(f, \alpha) - c$ for all $c \in \mathbb{R}$, $f \in \mathcal{C}(\mathcal{Y})$, $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$. Removing $\Omega^\star(f)$ from both side of inequality (12), we obtain

$$f - \Omega^\star(f) + T\big(-(f - \Omega^\star(f)), \nabla\Omega^\star(f)\big) \leq 0,$$

with equality on the support of $\nabla\Omega^\star(f)$, which brings the second part of the lemma. □

A.3.2. Proof of Lemma 1

Proof. Let $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$ and $f \triangleq \nabla\Omega(\alpha)$. From the optimality condition of Sinkhorn dual minimization (4),

$$T(-f, \alpha) = -f,$$

hence, α meets the sufficient conditions for optimality in Lemma 2. Therefore $\nabla\Omega^\star(f) = \alpha$, $\Omega^\star(f) = 0$, and the first part of the lemma follows. To demonstrate the second part, we consider $f \in \mathcal{F}$. There exists $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$ such that $f = \nabla\Omega(\alpha)$, and thus

$$\nabla\Omega \circ \nabla\Omega^\star(f) = \nabla\Omega \circ \nabla\Omega^\star \circ \Omega(\alpha) = \nabla\Omega(\alpha) = f.$$

The lemma follows. □

A.3.3. Extrapolation effect of $\nabla\Omega^\star$—Proof of Prop. 3

We start by establishing some basic properties of the extrapolation operator.

Lemma 3 (Properties of f^E)**.** *The following properties hold, for all $f \in \mathcal{C}(\mathcal{Y})$,*

i. The extrapolated potential f^E verifies

$$f \leq f^E, \quad f_{|\operatorname{supp}\nabla\Omega^\star(f)} = f^E_{|\operatorname{supp}\nabla\Omega^\star(f)}.$$

ii. The extrapolation operator maintain the following values:

$$f^{EE} = f^E,\ \Omega^\star(f^E) = \Omega^\star(f),\ \nabla\Omega^\star(f^E) = \nabla\Omega^\star(f).$$

Proof. We demonstrate (i), then (ii).

i. Note that $T(f + c, \alpha) = T(f, \alpha) - c$ for all $c \in \mathbb{R}$, $f \in \mathcal{C}(\mathcal{Y})$, $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$. Removing $\Omega^\star(f)$ from both side of inequality (12), we obtain

$$f - \Omega^\star(f) + T\big(-(f - \Omega^\star(f)), \nabla\Omega^\star(f)\big) \leq 0,$$

with equality on the support of $\nabla\Omega^\star(f)$.

ii. We set $\alpha = \nabla\Omega^\star(f)$. According to Lemma 2, for all $y \in \operatorname{supp}\alpha$, $f^E(y) = f(y)$ and

$$\frac{f^E(y) + T(-f^E, \alpha)(y)}{2} = 2\Omega^\star(f).$$

Furthermore, for all $y \in \mathcal{Y}$, $-f^E(y) \leq -f(y)$, and therefore, as the soft C-transform operator is non-increasing with respect to f,

$$2\Omega^\star(f) - f^E = T(-f, \nabla\Omega^\star(f)) \leq T(-f^E, \nabla\Omega^\star(f)),$$

where the left equality stems from the definition of f^E. Therefore

$$\frac{f^E(y) + T(-f^E, \eta)(y)}{2} \leq 2\Omega^\star(f),$$

on all $\mathcal{Y}$, and we meet the sufficient condition of Lemma 2 for the optimality of η in

$$\min_{\alpha \in \mathcal{M}_1^+(\mathcal{Y})} \Phi(f^E, \alpha).$$

We thus have $\Omega^\star(f^E) = \Omega^\star(f)$, $\nabla\Omega^\star(f) = \nabla\Omega^\star(f^E)$. Therefore

$$\begin{aligned} f^{EE} &= -T(-f^E, \nabla\Omega^\star(f^E)) + 2\Omega^\star(f^E) \\ &= -T(-f^E, \nabla\Omega^\star(f)) + 2\Omega^\star(f) \\ &= -T(-f, \nabla\Omega^\star(f)) + 2\Omega^\star(f) = f^E, \end{aligned}$$

where we have used on the third line the fact that the value of $T(f, \alpha)$ depends only on the values of f on the support of α. In our case, we have $f^E_{|\operatorname{supp} \nabla\Omega^\star(f)} = f_{|\operatorname{supp} \nabla\Omega^\star(f)}$, from Lemma 2. The lemma follows. □

With Lemma 1 and Lemma 3 at hand, we are now ready to prove Prop. 3.

Proof. We consider a function $f \in \mathcal{C}(\mathcal{Y})$. By construction of the extrapolation f^E,

$$g = f^E - \nabla\Omega^\star(f)$$

is a negative symmetric Sinkhorn potentials, as $T(-g, \nabla\Omega^\star(f)) = -g$. Therefore, from Lemma 1,

$$\begin{aligned} \nabla\Omega \circ \nabla\Omega^\star(g) &= g \\ \nabla\Omega \circ \nabla\Omega^\star(f^E) &= f^E - \nabla\Omega^\star(f) \\ \nabla\Omega \circ \nabla\Omega^\star(f) &= f^E - \nabla\Omega^\star(f), \end{aligned}$$

where the third equality stems from Lemma 3, property (ii), and the second from (8). □

A.4. Relation to Hausdorff divergence—Proofs of Prop. 4 and Prop. 5

We now turn to proving Prop. 4 and Prop. 5, that justifies the validity of the geometric logistic loss for a certain Bregman divergence, dubbed the asymmetric Hausdorff divergence.

A.4.1. Proof of Prop. 4

Proof. Let $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$ and $f \in \mathcal{C}(\mathcal{Y})$. By definition, the Hausdorff divergence $H = D_\Omega$ between α and $\nabla\Omega^\star(f)$ rewrites

$$\begin{aligned} D_\Omega(\alpha|\nabla\Omega^\star(f)) &= \Omega(\alpha) - \Omega(\nabla\Omega^\star(f)) - \langle \nabla\Omega \circ \nabla\Omega^\star(f), \alpha - \Omega^\star(f^E) \rangle \\ &= \Omega(\alpha) + \langle f, \nabla\Omega^\star(f) \rangle - \Omega(\nabla\Omega^\star(f)) - \langle f, \alpha \rangle + \langle f - \nabla\Omega \circ \nabla\Omega^\star(f), \alpha - \nabla\Omega^\star(f) \rangle \\ &= \ell_\Omega(\alpha, f) + \langle f - \nabla\Omega \circ \nabla\Omega^\star(f), \alpha - \nabla\Omega^\star(f) \rangle. \end{aligned}$$

This decomposition is a generic way of decomposing a Bregman divergence into a Fenchel-Young loss plus a perturbation term that depends on the "projection" $\nabla\Omega \circ \nabla\Omega^\star(f)$. In our case, thanks to Lemma 3, property (iv), this term rewrites

$$\langle f - \nabla\Omega \circ \nabla\Omega^\star(f), \alpha - \nabla\Omega^\star(f) \rangle = \langle f - f^E, \alpha \rangle + \langle f - f^E, \nabla\Omega^\star(f) \rangle + \Omega^\star(f) \langle 1, \alpha - \nabla\Omega^\star(f) \rangle.$$

The second term is null as a consequence of Lemma 3, while the third is null because α and $\nabla\Omega^\star(f)$ are both probability measures. The first one is null in case $\operatorname{supp} \nabla\Omega^\star(f) \in \operatorname{supp} \alpha$, in accordance to Lemma 3, property (i). The proposition follows from the fact that $f^E \geq f$ on the space $\mathcal{Y}$, according to the same property. □

A.4.2. Proof of Prop. 5

Proof. As a consequence of Prop. 4, for any true and estimated distribution $\alpha, \hat{\alpha} \in \mathcal{M}_1^+(\mathcal{Y})$, we have

$$D_\Omega(\alpha|\hat{\alpha}) = D_\Omega(\alpha|\nabla\Omega^\star(\nabla\Omega(\alpha))) = \ell_\Omega(\alpha, \nabla\Omega(\alpha)) - \langle \alpha, (\nabla\Omega(\alpha))^E - \nabla\Omega(\alpha)\rangle,$$

where the last term is null as $T(-\nabla\Omega(\alpha), \alpha) = -\nabla\Omega(\alpha)$ and $\Omega^\star(\nabla\Omega(\alpha)) = 0$ from Lemma 1. Therefore

$$D_\Omega(\alpha|\hat{\alpha}) = \ell_\Omega(\alpha, \nabla\Omega(\alpha)).$$

The equality of risks and the connection between minimizers immediately follows. To establish the Fisher consistency of the g-FY loss with respect to the Hausdorff divergence, note that, from Prop. 4, we have, for all $\hat{f} : \mathcal{X} \to \mathcal{C}(\mathcal{Y})$, for all $x \in \mathcal{X}$, $\alpha \in \mathcal{M}_1^+(\mathcal{Y})$,

$$D_\Omega(\alpha|\nabla\Omega^\star(\hat{f}(x)) \leq \ell_\Omega(\alpha, \hat{f}(x)).$$

Taking the expectation with respect to the data distribution $\mathcal{D}$, we obtain

$$\mathcal{E}(\nabla\Omega^\star \circ \hat{f}) \leq \mathcal{R}(\hat{f}),$$

and the proposition follows. □

B. Further experiments and details

B.1. Variational auto-encoders

High definition experiment. As a complementary experiment, we generate a dataset of cat doodles from the Google QuickDraw dataset, with a line width of one pixel. We test the g-softmax link function and the geometric Fenchel-Young loss functions to train a VAE with a DC-GAN architecure (Radford et al., 2016). We reuse the architecture of the authors, using the discriminator as an encoder, with a final layer with a size of output twice the size of the latent dimension, to model the mean and variance of the latent encoding, and the generator as a decoder. Similarly to the experiment in the main text, we observe that the generated samples and the reconstructions are more concentrated on thin measures.

MNIST. We display a visualization of generates images and reconstruction of test image in Figure 5. The output distributions are well concentrated, despite the low resolution of the dataset.

Architecture Our multi-layer perceptron is simple: encoder and decoder are two layer MLP with 400 hidden units and ReLU activation.

Hyperparameters. We use a latent size of 100 in the experiment on QuickDraw 28x28, and 256 for the high resolution experiment. We set the KL weight to 1, and rescale the KL loss with a factor $h \times w$, to make its gradient of the same order as the one computed with separated binary cross entropy. We use $\sigma = 2$ as the scaling parameter of the Euclidean cost function.

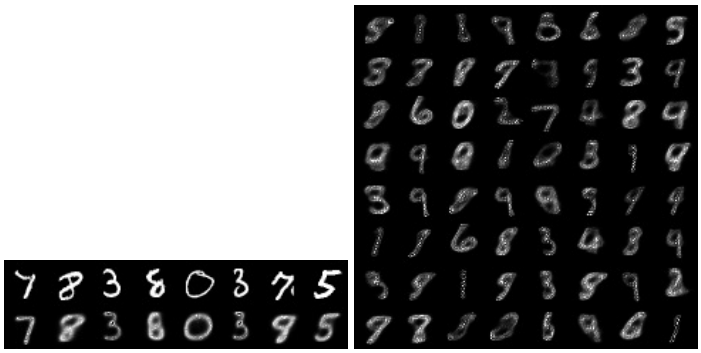

Figure 5. Examples of generated images and reconstruction of test images with an MLP VAE on MNIST dataset.

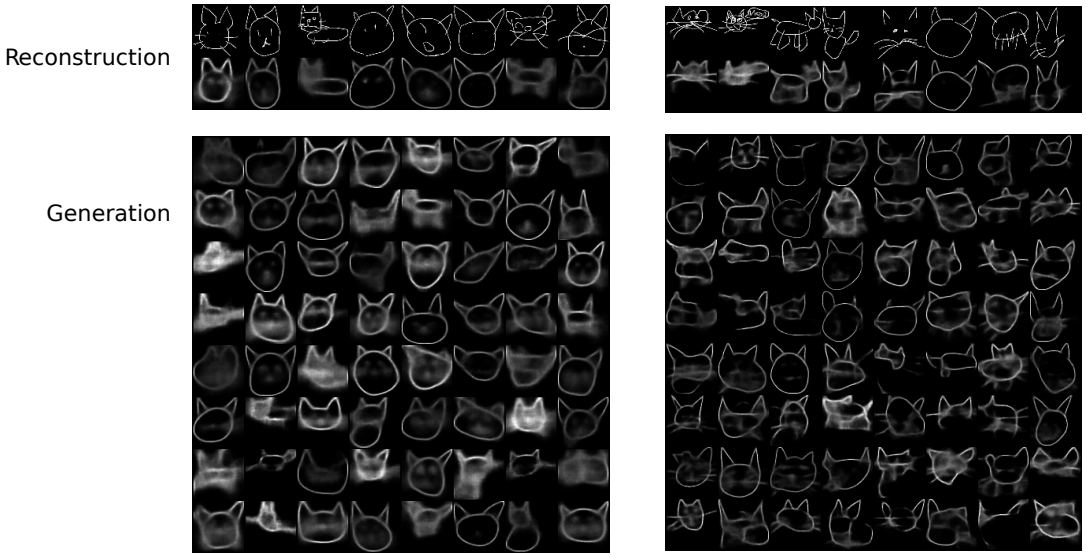

Figure 6. Examples of generated images and reconstruction of test images with a VAE-DC-GAN and a geometric softmax last layer. The generated images are sharper than when using a standard softmax layer and a KL divergence training.