

f -GANs in an Information Geometric Nutshell — Supplementary Material —

Abstract

This is the Supplementary Material to Paper ” f -GANs in an Information Geometric Nutshell” by R. Nock, Z. Cranko, A-K. Menon, L. Qu and and R.-C. Williamson. Theorems and Lemmata are numbered with letters (A, B, ...) to make a clear difference with the main file numbering.

Table of contents

Supplementary material on proofs and formal results	Pg 2
Proof of Theorem 5	Pg 2
Proof of Theorem 6	Pg 3
Proof of Theorem 7	Pg 4
Proof of Theorem 9	Pg 5
Proof of Lemma 11	Pg 11
Proof of Theorem 12	Pg 13
 Supplementary material on experiments	Pg 14
Architectures	Pg 14
Experimental setup for varying the activation function in the generator	Pg 14
Visual results	Pg 15
↔ MNIST results for GAN_DCGAN at varying μ ($\mu = 1$ is ReLU)	Pg 16
↔ MNIST results for WGAN_DCGAN at varying μ ($\mu = 1$ is ReLU)	Pg 17
↔ MNIST results for WGAN_MLP at varying μ ($\mu = 1$ is ReLU)	Pg 18
↔ MNIST results for GAN_MLP at varying μ ($\mu = 1$ is ReLU)	Pg 19
↔ LSUN results for GAN_DCGAN at varying μ ($\mu = 1$ is ReLU)	Pg 20

Supplementary material on proofs and formal results

I Proof of Theorem 5

Our basis for the proof of the Theorem is the following Lemma.

Lemma A [10, Proposition 1.6.1] *Let $f : I \rightarrow \mathbb{R}$ be continuous convex and let $\xi : I \rightarrow \mathbb{R}$ such that $\xi(z) \in \partial f(z), \forall z \in \text{int} I$. Then for any $a < b$ in I , it holds that:*

$$f(b) = f(a) + \int_a^b \xi(t) dt . \quad (1)$$

Suppose that $b < a$. Then Lemma A says that we have $f(a) = f(b) + \int_b^a \xi(t) dt$, that is, after re-ordering, $f(b) = f(a) - \int_b^a \xi(t) dt = f(a) + \int_a^b \xi(t) dt$, so in fact the requested ordering between the integral's bounds can be removed. Also, we can suppose that the integral may not be proper, in which case we compute it as a limit of a proper integral for which Lemma A therefore holds.

We now prove Theorem 5. Suppose there exists $M \in \mathbb{R}$ such that $\sup \xi(\mathbb{I}_{P,Q}) \leq M$, for some $\partial f \ni \xi : \text{int dom}(f) \rightarrow \mathbb{R}$. For any constants k , letting $f_k(z) \doteq f(z) - k(z - 1)$, which is convex since f is, we note that

$$\begin{aligned} \mathbb{E}_{X \sim Q} \left[f_k \left(\frac{P(X)}{Q(X)} \right) \right] &= \mathbb{E}_{X \sim Q} \left[f \left(\frac{P(X)}{Q(X)} \right) \right] - k \cdot \mathbb{E}_{X \sim Q} \left[\frac{P(X)}{Q(X)} - 1 \right] \\ &= \mathbb{E}_{X \sim Q} \left[f \left(\frac{P(X)}{Q(X)} \right) \right] - k \cdot \left(\int P(X) d\mu(X) - \int Q(X) d\mu(X) \right) \\ &= \mathbb{E}_{X \sim Q} \left[f \left(\frac{P(X)}{Q(X)} \right) \right] . \end{aligned} \quad (2)$$

Let $\xi_k \doteq \xi - k \in \partial f_k$. Since f_k is convex continuous, it follows from [10, Proposition 1.6.1] (Lemma A) that:

$$\begin{aligned} f_k \left(\frac{P(\mathbf{x})}{Q(\mathbf{x})} \right) &= f_k(1) + \lim_{\rho \rightarrow \frac{P(\mathbf{x})}{Q(\mathbf{x})}} \int_1^\rho \xi_k(t) dt \\ &= - \lim_{\rho \rightarrow \frac{P(\mathbf{x})}{Q(\mathbf{x})}} \int_1^\rho (-\xi(t) + k) dt . \end{aligned} \quad (3)$$

The second identity comes from the assumption that $f(1) = 0 = f_k(1)$. The limit appears to cope with a subdifferential that would diverge around a density ratio. Fix some constant $\epsilon > 0$ and let

$$\chi(t) = \begin{cases} \frac{1}{-\xi(t) + M + \epsilon} & \text{if } t < \sup \mathbb{I}_{P,Q} \\ \frac{1}{\epsilon} & \text{if } t \geq \sup \mathbb{I}_{P,Q} \end{cases} , \quad (4)$$

which, since $\sup \xi(\mathbb{I}_{P,Q}) \leq M$, guarantees $\chi \geq 0$ and χ is also increasing since ξ is increasing (f is convex). We then check, using eqs. (2) and (4) that:

$$\begin{aligned}
KL_\chi(\mathbb{Q} \parallel \mathbb{P}) &= \mathbb{E}_{\mathbf{X} \sim \mathbb{Q}} \left[-\log_\chi \left(\frac{P(\mathbf{X})}{Q(\mathbf{X})} \right) \right] \\
&= \mathbb{E}_{\mathbf{X} \sim \mathbb{Q}} \left[-\lim_{\rho \rightarrow \frac{P(\mathbf{X})}{Q(\mathbf{X})}} \int_1^\rho \frac{1}{\chi(t)} dt \right] \\
&= \mathbb{E}_{\mathbf{X} \sim \mathbb{Q}} \left[-\lim_{\rho \rightarrow \frac{P(\mathbf{X})}{Q(\mathbf{X})}} \int_1^\rho (-\xi(t) + M + \epsilon) dt \right] \\
&= \mathbb{E}_{\mathbf{X} \sim \mathbb{Q}} \left[f_{M+\epsilon} \left(\frac{P(\mathbf{X})}{Q(\mathbf{X})} \right) \right] \\
&= \mathbb{E}_{\mathbf{X} \sim \mathbb{Q}} \left[f \left(\frac{P(\mathbf{X})}{Q(\mathbf{X})} \right) \right] = I_f(\mathbb{P} \parallel \mathbb{Q}) .
\end{aligned} \tag{5}$$

This ends the proof of Theorem 5.

II Proof of Theorem 6

We have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_\chi(P(\mathbf{X})) - \log_\chi(Q(\mathbf{X})))] \\
&= \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_\chi(P(\mathbf{X})) - \log_\chi(\tilde{Q}(\mathbf{X})))] + \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_\chi(\tilde{Q}(\mathbf{X})) - \log_\chi(Q(\mathbf{X})))] \\
&= \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_\chi(P(\mathbf{X})) - \log_\chi(\tilde{Q}(\mathbf{X})))] - \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_\chi(Q(\mathbf{X})) - \log_\chi(\tilde{Q}(\mathbf{X})))] .
\end{aligned}$$

Consider some fixed $\mathbf{x} \in \mathcal{X}$. We have

$$\begin{aligned}
\log_\chi(P(\mathbf{x})) - \log_\chi(\tilde{Q}(\mathbf{x})) &= \int_1^{P(\mathbf{x})} \frac{1}{\chi(t)} \cdot dt - \int_1^{\tilde{Q}(\mathbf{x})} \frac{1}{\chi(t)} \cdot dt \\
&= \int_{\tilde{Q}(\mathbf{x})}^{P(\mathbf{x})} \frac{1}{\chi(t)} \cdot dt \\
&= \int_1^{\frac{P(\mathbf{x})}{\tilde{Q}(\mathbf{x})}} \frac{\tilde{Q}(\mathbf{x})}{\chi(t\tilde{Q}(\mathbf{x}))} \cdot dt \\
&= \int_1^{\frac{P(\mathbf{x})}{\tilde{Q}(\mathbf{x})}} \frac{1}{\chi_{\tilde{Q}(\mathbf{x})}(t)} \cdot dt \\
&= \log_{\chi_{\tilde{Q}(\mathbf{x})}} \left(\frac{P(\mathbf{x})}{\tilde{Q}(\mathbf{x})} \right) ,
\end{aligned} \tag{6}$$

with

$$\chi_{\tilde{Q}(\mathbf{x})}(t) \doteq \frac{1}{\tilde{Q}(\mathbf{x})} \cdot \chi(t\tilde{Q}(\mathbf{x})) . \tag{7}$$

To cope with the case where any of the integrals is improper, we derive the limit expression:

$$(\log_{\chi}(P(\mathbf{x})) - \log_{\chi}(\tilde{Q}(\mathbf{x}))) = \lim_{(p,q) \rightarrow (P(\mathbf{x}), \tilde{Q}(\mathbf{x}))} \log_{\chi_q} \left(\frac{p}{q} \right) , \quad (8)$$

so we get in all cases,

$$\mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_{\chi}(P(\mathbf{X})) - \log_{\chi}(\tilde{Q}(\mathbf{X})))] = KL_{\chi_{\tilde{\mathbb{Q}}}}(\tilde{Q} \| P) . \quad (9)$$

We also note that

$$\begin{aligned} \log_{\chi}(Q(\mathbf{X})) - \log_{\chi}(\tilde{Q}(\mathbf{X})) &= \lim_{(q,q') \rightarrow (Q(\mathbf{x}), \tilde{Q}(\mathbf{x}))} \log_{\chi_{q'}} \left(\frac{q}{q'} \right) \\ &= \log_{\chi_{\tilde{Q}(\mathbf{x})}} \left(\frac{Q(\mathbf{x})}{\tilde{Q}(\mathbf{x})} \right) \end{aligned} \quad (10)$$

(if the limit exists) so we get

$$\mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_{\chi}(P(\mathbf{X})) - \log_{\chi}(\tilde{Q}(\mathbf{X})))] = KL_{\chi_{\tilde{\mathbb{Q}}}}(\tilde{Q} \| P) , \quad (11)$$

$$\mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [-(\log_{\chi}(Q(\mathbf{X})) - \log_{\chi}(\tilde{Q}(\mathbf{X})))] = KL_{\chi_{\tilde{\mathbb{Q}}}}(\tilde{Q} \| Q) , \quad (12)$$

and

$$\mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [\log_{\chi}(Q(\mathbf{X})) - \log_{\chi}(P(\mathbf{X}))] = KL_{\chi_{\tilde{\mathbb{Q}}}}(\tilde{\mathbb{Q}} \| \mathbb{P}) - KL_{\chi_{\tilde{\mathbb{Q}}}}(\tilde{\mathbb{Q}} \| \mathbb{Q}) , \quad (13)$$

as claimed.

III Proof of Theorem 7

Let us denote $\mathcal{F}_{\tilde{\mathbb{Q}}} \subseteq \mathbb{R}^{\mathcal{X}}$ denote the subset of functions : $\mathcal{X} \rightarrow \mathbb{R}$ whose values are constrained as follows:

$$\mathcal{F}_{\tilde{\mathbb{Q}}} \doteq \left\{ T \in \mathbb{R}^{\mathcal{X}} : T(\mathbf{x}) \in \text{dom} \left(-\log_{\chi_{\tilde{Q}(\mathbf{x})}} \right)^* \right\} . \quad (14)$$

Since $-\log_{\chi_{\tilde{Q}(\mathbf{x})}}$ is convex for any \mathbf{x} , it follows from Legendre duality,

$$\begin{aligned} KL_{\chi_{\tilde{\mathbb{Q}}}}(\tilde{\mathbb{Q}} \| \mathbb{P}) &= \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} \left[-\log_{\chi_{\tilde{Q}(\mathbf{X})}} \left(\frac{P(\mathbf{X})}{\tilde{Q}(\mathbf{X})} \right) \right] \\ &= \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} \left[\sup_{T(\mathbf{X}) \in \text{dom} \left(\log_{\chi_{\tilde{Q}(\mathbf{X})}} \right)^*} \left\{ T(\mathbf{X}) \cdot \frac{P(\mathbf{X})}{\tilde{Q}(\mathbf{X})} - (-\log_{\chi_{\tilde{Q}(\mathbf{X})}})^*(T(\mathbf{X})) \right\} \right] \\ &= \sup_{T \in \mathcal{F}_{\tilde{\mathbb{Q}}}} \left\{ \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} \left[T(\mathbf{X}) \cdot \frac{P(\mathbf{X})}{\tilde{Q}(\mathbf{X})} - (-\log_{\chi_{\tilde{Q}(\mathbf{X})}})^*(T(\mathbf{X})) \right] \right\} \\ &= \sup_{T \in \mathcal{F}_{\tilde{\mathbb{Q}}}} \left\{ \mathbb{E}_{\mathbf{X} \sim \mathbb{P}} [T(\mathbf{X})] - \mathbb{E}_{\mathbf{X} \sim \tilde{\mathbb{Q}}} [(-\log_{\chi_{\tilde{Q}(\mathbf{X})}})^*(T(\mathbf{X}))] \right\} . \end{aligned} \quad (15)$$

Now, we know that $-\log_{\chi_{\tilde{Q}(\mathbf{x})}}(z)$ is proper lower-semicontinuous and therefore $(-\log_{\chi_{\tilde{Q}(\mathbf{x})}})^{\star\star} = -\log_{\chi_{\tilde{Q}(\mathbf{x})}}$. Being closed, the domain of the derivative of $(-\log_{\chi_{\tilde{Q}(\mathbf{x})}})^{\star}$ is the image of the derivative of $-\log_{\chi_{\tilde{Q}(\mathbf{x})}}$, given by $-\tilde{Q}(\mathbf{x})/\chi(\tilde{Q}(\mathbf{x})t)$. If $\chi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, then $-\tilde{Q}(\mathbf{x})/\chi(\tilde{Q}(\mathbf{x})t) \in \overline{\mathbb{R}_{++}}, \forall \tilde{Q}(\mathbf{x})$ and so $\mathcal{F}_{\tilde{Q}} = \left\{T \in \overline{\mathbb{R}_{++}}^{\chi}\right\}$.

A pointwise differentiation of eq. (15) yields that at the optimum, we have

$$\begin{aligned} P(\mathbf{x}) - \tilde{Q}(\mathbf{x}) \cdot (-\log_{\chi_{\tilde{Q}(\mathbf{x})}})^{\star'}(T(\mathbf{x})) &= P(\mathbf{x}) - \tilde{Q}(\mathbf{x}) \cdot (-\log_{\chi_{\tilde{Q}(\mathbf{x})}})^{\prime-1}(T(\mathbf{x})) \\ &= 0, \end{aligned} \quad (16)$$

that is, exploiting the fact that $(-\log_{\chi_{\tilde{Q}(\mathbf{x})}})^{\prime} = -\tilde{Q}(\mathbf{x})/\chi(\tilde{Q}(\mathbf{x})t)$,

$$\begin{aligned} T^*(\mathbf{x}) &= (-\log_{\chi_{\tilde{Q}}})' \left(\frac{P(\mathbf{x})}{\tilde{Q}(\mathbf{x})} \right) \\ &= -\frac{\tilde{Q}(\mathbf{x})}{\chi \left(\frac{P(\mathbf{x})}{\tilde{Q}(\mathbf{x})} \cdot \tilde{Q}(\mathbf{x}) \right)} \end{aligned} \quad (17)$$

$$\begin{aligned} &= -\frac{\tilde{Q}(\mathbf{x})}{\chi(P(\mathbf{x}))} \\ &= -\frac{1}{Z} \cdot \frac{\chi(Q(\mathbf{x}))}{\chi(P(\mathbf{x}))}. \end{aligned} \quad (18)$$

IV Proof of Theorem 9

In the context of the proof, we simplify notations and replace signature χ_{net} by χ and output activation v_{out} by v_2 . Let us call $\mathbf{z} \in \mathbb{R}^d$ the output of g . We revert the transformation and check:

$$\phi_{l-1}(\mathbf{z}) \doteq \mathbf{W}_l^{-1}(\mathbf{v}^{-1}(\phi_l(\mathbf{z})) - \mathbf{b}_l), \forall l \in \{1, 2, \dots, L\}, \quad (19)$$

$$\phi_L(\mathbf{z}) = \Gamma^{-1}(\mathbf{v}_2^{-1}(\mathbf{z}) - \beta). \quad (20)$$

For the sake of readability, we shall sometimes remove the dependence in \mathbf{z} . Letting a_i denote coordinate i in vector \mathbf{a} , $(A)_{ij}$ the coordinate in row i and column j of matrix A , for any $i, j \in [d]$, and $a_{l,i}$ coordinate i in vector \mathbf{a}_l , we have

$$\frac{\partial \phi_{l-1,i}}{\partial \phi_{l,j}} = (\mathbf{W}_l^{-1})_{ij} \cdot \frac{1}{v'_i(\mathbf{v}^{-1}(\phi_l))}, \quad (21)$$

and furthermore

$$\frac{\partial \phi_{L,i}}{\partial z_j} = (\Gamma^{-1})_{ij} \cdot \frac{1}{v'_{2i}(\mathbf{v}_2^{-1}(\mathbf{z}))}. \quad (22)$$

Let us denote vector $\tilde{\mathbf{a}}$ as the vector whose coordinates are the inverses of those of \mathbf{a} , namely $\tilde{a}_i \doteq 1/a_i$. From eqs. (21) and (22), the layerwise Jacobians are:

$$\frac{\partial \phi_{l-1}}{\partial \phi_l^\top} = \mathbf{W}_l^{-1} \odot \tilde{\mathbf{v}}'(\mathbf{v}^{-1}(\phi_l)) \mathbf{1}^\top, \forall l \in \{1, 2, \dots, L\}, \quad (23)$$

$$\frac{\partial \phi_L}{\partial \mathbf{z}^\top} = \Gamma^{-1} \odot \tilde{\mathbf{v}}'_2(\mathbf{v}_2^{-1}(\mathbf{z})) \mathbf{1}^\top, \quad (24)$$

where \odot is Hadamard (coordinate-wise) product. These Jacobians have a very convenient form, since:

$$\begin{aligned} \det \left(\frac{\partial \phi_{l-1}}{\partial \phi_l^\top} \right) &= \sum_{\boldsymbol{\sigma} \in S_d} \text{sign}(\boldsymbol{\sigma}) \cdot \prod_{i=1}^d (\mathbf{W}_l^{-1} \odot \tilde{\mathbf{v}}'(\mathbf{v}^{-1}(\phi_l)) \mathbf{1}^\top)_{i, \sigma_i} \\ &= \sum_{\boldsymbol{\sigma} \in S_d} \text{sign}(\boldsymbol{\sigma}) \cdot \prod_{i=1}^d (\mathbf{W}^{-1})_{l, i, \sigma_i} (\tilde{\mathbf{v}}'(\mathbf{v}^{-1}(\phi_l)) \mathbf{1}^\top)_{i, \sigma_i} \\ &= \sum_{\boldsymbol{\sigma} \in S_d} \left(\prod_{i=1}^d \tilde{v}'_i(\mathbf{v}^{-1}(\phi_l)) \right) \cdot \text{sign}(\boldsymbol{\sigma}) \cdot \prod_{i=1}^d (\mathbf{W}^{-1})_{l, i, \sigma_i} \\ &= \left(\prod_{i=1}^d \tilde{v}'_i(\mathbf{v}^{-1}(\phi_l)) \right) \cdot \sum_{\boldsymbol{\sigma} \in S_d} \text{sign}(\boldsymbol{\sigma}) \cdot \prod_{i=1}^d (\mathbf{W}^{-1})_{l, i, \sigma_i} \\ &= \left(\prod_{i=1}^d \tilde{v}'_i(\mathbf{v}^{-1}(\phi_l)) \right) \cdot \det(\mathbf{W}_l^{-1}) \\ &= \left(\prod_{i=1}^d \tilde{v}'_i(\mathbf{v}^{-1}(\phi_l)) \right) \cdot (\det(\mathbf{W}_l))^{-1}, \forall l \in \{1, 2, \dots, L\}, \end{aligned}$$

and, using the same derivations,

$$\det \left(\frac{\partial \phi_L}{\partial \mathbf{z}^\top} \right) = \left(\prod_{i=1}^d \tilde{v}'_{2i}(\mathbf{v}_2^{-1}(\mathbf{z})) \right) \cdot (\det(\Gamma))^{-1}. \quad (25)$$

The change of variable formula [4] yields:

$$\begin{aligned}
Q_g(\mathbf{z}) &= Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z})) \cdot \left| \det \left(\frac{\partial \mathbf{g}^{-1}}{\partial \mathbf{z}^\top} \right) \right| \\
&= Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z})) \cdot \left| \det \left(\frac{\partial \phi_0}{\partial \mathbf{z}^\top} \right) \right| \\
&= Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z})) \cdot \left| \det \left(\prod_{l=1}^L \frac{\partial \phi_{l-1}}{\partial \phi_l^\top} \cdot \frac{\partial \phi_L}{\partial \mathbf{z}^\top} \right) \right| \\
&= Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z})) \cdot \left| \prod_{l=1}^L \det \left(\frac{\partial \phi_{l-1}}{\partial \phi_l^\top} \right) \cdot \det \left(\frac{\partial \phi_L}{\partial \mathbf{z}^\top} \right) \right| \\
&= Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z})) \cdot \prod_{l=1}^L \prod_{i=1}^d |\tilde{v}'_i(\mathbf{v}^{-1}(\phi_l))| \cdot \prod_{i=1}^d |\tilde{v}'_{2i}(\mathbf{v}_2^{-1}(\mathbf{z}))| \cdot \left| \det \left(\Gamma \cdot \prod_{l=1}^L \mathbf{W}_l \right) \right|^{-1} \\
&= Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z})) \cdot \frac{1}{\prod_{l=1}^L \prod_{i=1}^d |v'_i(\mathbf{v}^{-1}(\phi_l))| \cdot \prod_{i=1}^d |v'_{2i}(\mathbf{v}_2^{-1}(\mathbf{z}))|} \cdot \left| \det \left(\Gamma \cdot \prod_{l=1}^L \mathbf{W}_l \right) \right|^{-1} \\
&= Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z})) \cdot \frac{1}{\prod_{l=1}^L \prod_{i=1}^d |v'(v^{-1}(\phi_{l,i}))| \cdot \prod_{i=1}^d |v'_2(v_2^{-1}(z_i))|} \cdot \left| \det \left(\Gamma \cdot \prod_{l=1}^L \mathbf{W}_l \right) \right|^{-1} \\
&= \frac{Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z}))}{\prod_{l=1}^L \prod_{i=1}^d |v'(v^{-1}(\phi_{l,i}))|} \cdot \frac{1}{\prod_{i=1}^d |v'_2(v_2^{-1}(z_i))| \cdot |\det(\mathbf{N})|} ,
\end{aligned}$$

because v and v_2 are coordinatewise. We have let

$$\mathbf{N} = \Gamma \cdot \prod_{l=1}^L \mathbf{W}_l , \tag{26}$$

and also $\phi_{l,i} \doteq v(\mathbf{w}_{l,i}^\top \phi_{l-1} + b_{l,i})$, where $\mathbf{w}_{l,i} \doteq \mathbf{W}_l^\top \mathbf{1}_i$ is the (column) vector built from row i in \mathbf{W}_l and similarly $z_i \doteq v_2(\gamma_i^\top \phi_L + \beta_i)$ with $\gamma_i \doteq \Gamma^\top \mathbf{1}_i$. Notice that we can also write

$$\prod_{l=1}^L \prod_{i=1}^d |v'(v^{-1}(\phi_{l,i}))| = \prod_{l=1}^L \prod_{i=1}^d |v'(\mathbf{w}_{l,i}^\top \phi_{l-1} + b_{l,i})| . \tag{27}$$

So, letting $\tilde{Q}_{\text{deep}}^* \doteq \prod_{l=1}^L \prod_{i=1}^d |v'(\mathbf{w}_{l,i}^\top \phi_{l-1} + b_{l,i})|$, $H_{\text{out}} \doteq \prod_{i=1}^d |v'_{\text{out}}(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_i)|$ (with $\mathbf{x} \doteq \mathbf{g}^{-1}(\mathbf{z})$), and dropping the determinant which does not depend on \mathbf{z} , we get:

$$Q_g(\mathbf{z}) \propto \frac{Q_{\text{in}}(\mathbf{g}^{-1}(\mathbf{z}))}{\tilde{Q}_{\text{deep}}^*} \cdot \frac{1}{H_{\text{out}}} . \tag{28}$$

To finish up the proof, we are going to identify $\tilde{Q}_{\text{deep}}^*$ to (a constant times) the product of escorts in eq. (14). To do so, we are first going to design the general activation function v as a function of χ , and

choose:

$$v(z) \doteq k + k' \cdot \exp_\chi(z) , \quad (29)$$

for $k \in \mathbb{R}, k' > 0$ constants, which can be chosen *e.g.* to ensure that zero signal implies zero activation ($v(0) = 0$). Our choice for v has the following key properties.

Lemma B *v is C^1 , invertible and we have $v'(z) = k' \cdot \chi(\exp_\chi(z))$.*

Proof The derivative comes from [2, Eq. 84]. Notice that \exp_χ is continuous as an integral, χ is continuous by assumption and so v' is continuous, implying v is C^1 . We prove the invertibility. Because of the expression of v' , v is increasing, and in fact strictly increasing with the sole exception when $\exp_\chi(z) \in \chi^{-1}(0)$. However, note that $\chi^{-1}(0) \not\subset \text{dom}(\log_\chi)$ because of the definition of \log_χ . Since \exp_χ is the inverse of \log_χ [9, Section 10.1], it follows that $\chi^{-1}(0) \not\subset \text{im}(\exp_\chi)$ and so $\exp_\chi(z) \notin \chi^{-1}(0), \forall z \in \text{dom}(\exp_\chi)$, which implies v invertible. ■

What the Lemma shows is that we can plug v as in eq. (29) directly in $\tilde{Q}_{\text{deep}}^*$. To do so, let us now define strictly positive constants Z_{li} that shall be fixed later. We directly get from eq. (27)

$$\begin{aligned} \tilde{Q}_{\text{deep}}^* &= \left(\prod_{l=1}^L \prod_{i=1}^d Z_{li} \right) \cdot \prod_{l=1}^L \prod_{i=1}^d \frac{1}{Z_{li}} \cdot |v'(\mathbf{w}_{l,i}^\top \boldsymbol{\phi}_{l-1} + b_{l,i})| \\ &= \left(k'^{Ld} \cdot \prod_{l=1}^L \prod_{i=1}^d Z_{li} \right) \cdot \underbrace{\prod_{l=1}^L \prod_{i=1}^d \frac{1}{Z_{li}} \cdot \chi(\exp_\chi(\mathbf{w}_{l,i}^\top \boldsymbol{\phi}_{l-1} + b_{l,i}))}_{\doteq \tilde{Q}_{\text{deep}}} \end{aligned} \quad (30)$$

(we can remove the absolute values since χ is non-negative). We now ensure that \tilde{Q}_{deep} is indeed a product of escorts: to do so, we just need to ensure that (i) $b_{l,i}$ normalizes the deformed exponential family, *i.e.* defines (negative) its cumulant (Definition 2), and (ii) Z_{li} normalizes its escort as in eq. (6). To be more explicit, we pick $b_{l,i}$ the solution of

$$\int_{\boldsymbol{\phi}} \exp_\chi(\mathbf{w}_{l,i}^\top \boldsymbol{\phi} - b_{l,i}) d\nu_{l-1}(\boldsymbol{\phi}) = 1 , \quad (31)$$

where $d\nu_{l-1}(\boldsymbol{\phi}) \doteq \int_{\boldsymbol{\phi}_{l-1}(\mathbf{x})=\boldsymbol{\phi}} d\mu(\mathbf{x})$ is the pushforward measure, and

$$Z_{li} = \int_{\mathbf{x}} \chi(P_{\chi, b_{l,i}}(\mathbf{x} | \mathbf{w}_{l,i}, \boldsymbol{\phi}_{l-1})) d\mu(\mathbf{x}) . \quad (32)$$

We get

$$\tilde{Q}_{\text{deep}} = \prod_{l=1}^L \prod_{i=1}^d \tilde{P}_{\chi, b_{l,i}}(\mathbf{x} | \mathbf{w}_{l,i}, \boldsymbol{\phi}_{l-1}) , \quad (33)$$

and finally,

$$\begin{aligned} Q_g(z) &= \frac{Q_{\text{in}}(\mathbf{x})}{\tilde{Q}_{\text{deep}}^*(\mathbf{x})} \cdot \frac{1}{H_{\text{out}}(\mathbf{x}) \cdot |\det(\mathbf{N})|} \\ &= \frac{Q_{\text{in}}(\mathbf{x})}{\tilde{Q}_{\text{deep}}(\mathbf{x})} \cdot \frac{1}{H_{\text{out}}(\mathbf{x}) \cdot Z_{\text{net}}} , \end{aligned} \quad (34)$$

with

$$Z_{\text{net}} \doteq \left(k'^{Ld} \cdot \prod_{l=1}^L \prod_{i=1}^d Z_{li} \right) \cdot |\det(\mathbf{N})| \quad (35)$$

a constant. We get the statement of Theorem 9.

Remark. (unnormalized densities) since in practice all b_l s are learned, we in fact work with deformed exponential families with unspecified normalization. We may also consider that the normalization of escorts is unspecified and therefore drop all Z_{li} s, which simplifies Z_{net} to $Z_{\text{net}} = |\det(\mathbf{N})|$. ■

Remark. (completely factoring Q_g as an escort) Denote for short $\mathbf{z}_p \doteq \phi_L(\mathbf{x})$ the penultimate layer of \mathbf{g} , and \mathbf{g}_p the net obtain from eliminating the last layer of \mathbf{g} , which allows us to drop $H_{\text{out}}(\cdot)$ from $Q_{g_p}(\mathbf{z})$ and we have $Q_g(\mathbf{z}) \propto Q_r \doteq Q_{\text{in}}(\mathbf{g}_p^{-1}(\mathbf{z}_p)) / \tilde{Q}_{\text{deep}}(\mathbf{g}_p^{-1}(\mathbf{z}_p))$. One can factor Q_r as a proper likelihood over escorts of χ_{net} -exponential families: for this, replace all Ld inner nodes of \mathbf{g}_p by random variables, say $\Phi_{l,i}$ (for $l \in \{0, 1, \dots, L-1\}, i \in \{1, 2, \dots, d\}$), treat the deep net \mathbf{g}_p as a directed graphical model whose connections are the dashed arcs. Now, if we let, say, $Q_{\text{in}}(\mathbf{g}_p^{-1}(\mathbf{z}_p)) \doteq \tilde{Q}_a(\cap_{l,i} \Phi_{l,i})$ and $\tilde{Q}_{\text{deep}}(\mathbf{g}_p^{-1}(\mathbf{z}_p)) \doteq \tilde{Q}_b(\cap_{l>0,i} \Phi_{l,i})$, and if we use as Q_{in} an uninformed escort (*i.e.* with constant coordinate, say for example $\theta = 1$, Definition 2), then assuming correct factorization one may obtain $Q_r = \tilde{Q}_c(\mathbf{g}_p^{-1}(\mathbf{z}_p) | \cap_{l>0,i} \Phi_{l,i})$ for some escort \tilde{Q}_c that we can plug directly in eq. (11). To properly understand the relationships between χ , Q_a , Q_b and how the escorts factor in Q_c requires a push of the state of the art: conjugacy in deformed exponential families is less understood than for exponential families; it is also unknown how product of deformed exponential families factor within the same deformed exponential families [1]; some factorizations are known but only on subsets of deformed exponential families and rely on particular notions of independence [8]; ■

Remark. (twist introduced by the last layer) We return to the twist introduced by the last layer of \mathbf{g} :

$$H_{\text{out}}(\mathbf{x}) = \prod_{i=1}^d |v'_2(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_L)| \quad (36)$$

It is clear that when v_2 is the identity, $H_{\text{out}}(\mathbf{x})$ is constant; so deep architectures, as experimentally carried out *e.g.* in Wasserstein GANs [3] or analyzed theoretically *e.g.* in [7] exactly fit to the escort factoring — notice that one can choose as input density one from some particular deformed exponential family, as *e.g.* done experimentally for [11, Section 2.5] (standard Gaussian), so that in this case $Q_g(\mathbf{z})$ factors completely as escorts.

Suppose now that v_2 is not the identity but chosen so that, for some couple (χ, g) where χ is differentiable and $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ is invertible,

$$(v'_2 \circ g)(z) = \frac{d}{dz} (\log_\chi \circ \chi)(z) = \frac{\chi'(z)}{\chi(z)} \quad (37)$$

which is equivalent, after a variable change, to having v_2 satisfy

$$v'_2(t) = \frac{\chi' \circ g^{-1}}{\chi \circ g^{-1}}(t) \quad (38)$$

In addition, suppose that g is chosen so that $\sum_i g^{-1}(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_i) = 1$. Call $D \doteq \{p_1, p_2, \dots, p_d\}$ this discrete distribution, removing reference to \mathbf{x} . We then have:

$$\begin{aligned}
H_{\text{out}}(\mathbf{x}) &= \prod_{i=1}^d \frac{\chi'(g^{-1}(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_i))}{\chi(g^{-1}(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_i))} \\
&= \prod_{i=1}^d \frac{\chi'(p_i)}{\chi(p_i)} \\
&= \prod_{i=1}^d \chi(p_i) \cdot \frac{\chi'(p_i)}{\chi(p_i)} \\
&= \left| \prod_{i=1}^d ((\exp_\chi)' \circ \log_\chi)(p_i) \cdot (\log_\chi)''(p_i) \right| \\
&\propto |\det(H)| .
\end{aligned} \tag{39}$$

Here, H is the χ -Fisher information metric of D [2, Theorem 12, eqs 119, 120]. In other words, $H_{\text{out}}(\mathbf{x})$ can be absorbed in the volume element in eq. (34).

As an example, pick a prop- τ activation (Table 1), for which $\log_\chi = (\tau^*)^{-1}(\tau^*(0)z)$ and

$$\chi(t) = \frac{(\tau^*)' \circ (\tau^*)^{-1}(\tau^*(0)z)}{\tau^*(0)} . \tag{40}$$

Now, pick $g(z) = \log_\chi(K \cdot z)$, where $K \doteq \sum_i \exp_\chi(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_i)$ guarantees:

$$\sum_i g^{-1}(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_i) = \frac{1}{K} \cdot \sum_i \exp_\chi(\gamma_i^\top \phi_L(\mathbf{x}) + \beta_i) = 1 . \tag{41}$$

Condition in eq. (37) becomes

$$\begin{aligned}
(v_2' \circ (\tau^*)^{-1})(\tau^*(0)Kz) &= \frac{\chi' \circ g^{-1}(t)}{\chi \circ g^{-1}(t)} \\
&= \tau^*(0) \cdot \frac{(\tau^*)'' \circ (\tau^*)^{-1}(\tau^*(0)Kz)}{((\tau^*)' \circ (\tau^*)^{-1}(\tau^*(0)Kz))^2} ,
\end{aligned} \tag{42}$$

and we obtain after a variable change,

$$v_2 = \tau^*(0) \cdot \int_t \frac{(\tau^*)''(t)}{((\tau^*)')^2(t)} dt , \tag{43}$$

which does not depend on K and, if τ^* is *strictly* convex, is strictly increasing. Notice that we can carry out the integration, $v_2(z) = K' - (\tau^*(0)/(\tau^*)'(z))$ for some constant K' . To make a parallel with a popular activation for the last layer, consider the sigmoid, $v_2 \doteq v_s(z) \doteq 1/(1 + \exp(-z))$, for which

$$v_s'(z) = \frac{\exp(z)}{(1 + \exp(z))^2} . \tag{44}$$

Fitting it to eq. (43),

$$\frac{\exp(z)}{(1 + \exp(z))^2} = \tau_s^*(0) \cdot \frac{(\tau_s^*)''(t)}{((\tau_s^*)')^2(t)} \quad (45)$$

reveals that we can pick $\tau_s^*(z) = z + \exp(z)$ (we control that $\tau_s^*(0) = 1$). Such a τ^* analytically fits to the prop- τ definition and in fact corresponds to a χ -exponential family, but it does not correspond to an entropy τ . This would be also true for affine scalings (argument and function) of the sigmoid of the type $v_2 = a + bv_s(c + dz)$. ■

V Proof of Lemma 11

We first show point (i). Define function

$$h(z) \doteq \frac{v(z) - \inf v(z)}{v(0) - \inf v(z)} , \quad (46)$$

and let $g(z) \doteq h^{-1}(z)$. Since $\text{dom}(v) \cap \overline{\mathbb{R}_+} \neq \emptyset$, $v(0) - \inf v(z) > 0$, so $h(z)$ bears the same properties as v . We first show that g is a valid χ -logarithm. Since v is convex increasing, $g(z)$ is concave increasing and $-g$ is convex decreasing. Therefore, since g is C^1 as well, letting $\xi \doteq g'$, we get:

$$g(z) = \int_1^z \frac{1}{\left(\frac{1}{\xi(t)}\right)} dt . \quad (47)$$

We also check that $g(1) = 0$ since $h(0) = 1$. If we let $\chi \doteq 1/\xi$, then because $\xi(z) \geq 0$, $\chi(z) \geq 0$ and also because ξ is decreasing, χ is increasing. Finally, $\chi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$. Summarizing, we have shown that χ defines a valid signature and $g(z) = \log_\chi(z)$. Therefore, $h(z) = \exp_\chi(z)$ and it comes that

$$v(z) = k + k' \cdot \exp_\chi(z) , \quad (48)$$

for $k \doteq \inf v(z) \in \mathbb{R}$ and $k' \doteq v(0) - \inf v(z) > 0$, so v matches the analytic expression in eq. (29), which allows to complete the proof of point (i) in the Lemma.

The strong admissibility results are easy to check in point (ii), so we concentrate on showing the weak admissibility of ReLU. We use a scaled perspective transform of the Softplus function and let:

$$v_\mu(z) \doteq (1 - \mu) \cdot \log \left(1 + \exp \left(\frac{z}{1 - \mu} \right) \right) , \quad (49)$$

with $\mu \in [0, 1]$. It is clear that v_μ is strongly admissible for any $\mu \in [0, 1]$.

Lemma C For any $z \geq 0, \mu \in [0, 1]$,

$$(1 - \mu) \cdot \log \left(\frac{1 + \exp \left(\frac{z}{1 - \mu} \right)}{1 + \exp(z)} \right) \leq \mu z . \quad (50)$$

Proof Equivalently, we want

$$\frac{1 + \exp\left(\frac{z}{1-\mu}\right)}{1 + \exp(z)} \leq \exp\left(\frac{\mu z}{1-\mu}\right), \quad (51)$$

or, equivalently,

$$\begin{aligned} 1 + \exp\left(\frac{z}{1-\mu}\right) &\leq \exp\left(\frac{\mu z}{1-\mu}\right) + \exp(z) \cdot \exp\left(\frac{\mu z}{1-\mu}\right) \\ &= \exp\left(\frac{\mu z}{1-\mu}\right) + \exp\left(\frac{z}{1-\mu}\right), \end{aligned} \quad (52)$$

which, after simplification, is equivalent to $\mu z/(1-\mu) \geq 0$, which indeed holds when $z \geq 0, \mu \in [0, 1]$. ■

We now have $v_\mu(z) \geq \max\{0, z\}, \forall \mu \in [0, 1]$, and we can also check that Lemma C implies

$$\begin{aligned} (1-\mu) \cdot \log\left(1 + \exp\left(\frac{z}{1-\mu}\right)\right) - z \\ \leq (1-\mu) \cdot (\log(1 + \exp(z)) - z), \forall z \geq 0, \mu \in [0, 1]. \end{aligned} \quad (53)$$

Let us denote, for any $z \geq 0$,

$$\begin{aligned} I_\mu(z) &\doteq \int_0^z |v_\mu(t) - \max\{0, t\}| dt \\ &= \int_0^z |v_\mu(t) - t| dt \\ &= \int_0^z (v_\mu(t) - t) dt. \end{aligned} \quad (54)$$

Since $\max\{0, -t\} = \max\{0, t\} - t$ and $v_\mu(-t) = v_\mu(t) - t$, we have $\|v_\mu - \text{ReLU}\|_{L1} = 2 \lim_{z \rightarrow +\infty} I_\mu(z)$. It also comes from ineq. (53) that

$$I_\mu(z) \leq (1-\mu)I_0(z), \forall z \leq 0, \quad (55)$$

furthermore, it can be shown by numerical integration that $\lim_{+\infty} I_0(z) = \pi^2/6$, so we get

$$\|v_\mu - \text{ReLU}\|_{L1} \leq \frac{(1-\mu)\pi^2}{3}, \forall \mu \in [0, 1], \quad (56)$$

and to have the right hand side smaller than $\epsilon > 0$, it suffices to take

$$\mu > 1 - \frac{3\epsilon}{\pi^2}, \quad (57)$$

which completes the proof of point (ii) and so the statement of the Lemma.

VI Proof of Theorem 12

The proof of the Theorem mainly follows from identifying the parameters of eq. (17) with the variational part of eq. (11). Recall that

$$(\chi^\bullet)_{\frac{1}{q}}(t) = \frac{1}{(\chi_q)^{-1}\left(\frac{1}{t}\right)}, \quad (58)$$

so, exploiting the fact that $K(\mathbb{Q})$ does not depend on T , we get:

$$\begin{aligned} \ell'_x(-1, z) &= \frac{d}{dz}(-\log_{\chi_{\tilde{Q}(x)}})^*(-z) \\ &= \frac{d}{dz} -\log_{(\chi^\bullet)_{\frac{1}{\tilde{Q}(x)}}}(-z) \\ &= \left(\chi_{\tilde{Q}(x)}\right)^{-1}\left(-\frac{1}{z}\right). \end{aligned} \quad (59)$$

Since $\ell'(+1, z) = -1$, we deduce that the loss is proper composite with inverse link function [13, Corollary 12] given by:

$$\begin{aligned} \Psi_x^{-1}(z) &= \frac{\ell'(-1, z)}{\ell'(-1, z) - \ell'(+1, z)} \\ &= \frac{\left(\chi_{\tilde{Q}(x)}\right)^{-1}\left(-\frac{1}{z}\right)}{\left(\chi_{\tilde{Q}(x)}\right)^{-1}\left(-\frac{1}{z}\right) + 1}, \end{aligned} \quad (60)$$

so that the link is

$$\Psi_x(z) = -\frac{1}{\chi_{\tilde{Q}(x)}\left(\frac{z}{1-z}\right)}. \quad (61)$$

Remark. We easily retrieve the optimal discriminator (Theorem 7) but this time from the proper composite loss, since (the first line is a general property of Ψ_x , see Section 5):

$$\begin{aligned} T^*(x) &= \Psi_x\left(\frac{P(x)}{P(x) + \tilde{Q}(x)}\right) \\ &= -\frac{1}{\chi_{\tilde{Q}(x)}\left(\frac{\frac{P(x)}{P(x) + \tilde{Q}(x)}}{1 - \frac{P(x)}{P(x) + \tilde{Q}(x)}}\right)} \\ &= -\frac{1}{\chi_{\tilde{Q}(x)}\left(\frac{P(x)}{\tilde{Q}(x)}\right)} \\ &= -\frac{1}{Z} \cdot \frac{\chi(Q(x))}{\chi(P(x))}. \end{aligned}$$

The last identity follows from eqs. (17) — (18). ■

Supplementary material on experiments

VII Architectures

We consider two architectures in our experiments: DCGAN [12] and the multilayer feedforward network (MLP) used in [11]. Suppose the size of input images is isize-by-isize , the details of architectures are given as follows:

Generator of DCGAN :

$\text{ConvTranspose}(\text{input}=100, \text{output}=8 \times \text{isize}, \text{stride}=1) \rightarrow \text{BatchNorm} \rightarrow \text{Activation} \rightarrow \text{Conv}(\text{input}=8 \times \text{isize}, \text{output}=4 \times \text{isize}, \text{stride}=2, \text{padding}=1) \rightarrow \text{BatchNorm} \rightarrow \text{Activation} \rightarrow \text{ConvTranspose}(\text{input}=4 \times \text{isize}, \text{output}=2 \times \text{isize}, \text{stride}=2, \text{padding}=2) \rightarrow \text{BatchNorm} \rightarrow \text{Activation} \rightarrow \text{ConvTranspose}(\text{input}=2 \times \text{isize}, \text{output}=\text{isize}, \text{stride}=2, \text{padding}=1) \rightarrow \text{BatchNorm} \rightarrow \text{Activation} \rightarrow \text{Conv}(\text{isize}, \text{number of channel}, \text{stride}=2, \text{padding}=1) \rightarrow \text{Last Activation}$

Discriminator of DCGAN :

$\text{Conv}(1, 2 \times \text{isize}, \text{stride}=2) \rightarrow \text{BatchNorm} \rightarrow \text{LeakyReLU} \rightarrow \text{Conv}(\text{input}=2 \times \text{isize}, \text{output}=4 \times \text{isize}, \text{stride}=2, \text{padding}=1) \rightarrow \text{BatchNorm} \rightarrow \text{LeakyReLU} \rightarrow \text{Conv}(\text{input}=4 \times \text{isize}, \text{output}=8 \times \text{isize}, \text{stride}=2, \text{padding}=2) \rightarrow \text{BatchNorm} \rightarrow \text{LeakyReLU} \rightarrow \text{Conv}(\text{input}=8 \times \text{isize}, \text{output}=1, \text{stride}=2, \text{padding}=1) \rightarrow \text{Link function}$

Generator of MLP :

$z \rightarrow \text{Linear}(100, 1024) \rightarrow \text{BatchNorm} \rightarrow \text{Activation} \rightarrow \text{Linear}(1024, 1024) \rightarrow \text{BatchNorm} \rightarrow \text{Activation} \rightarrow \text{Linear}(1024, \text{isize} \times \text{isize}) \rightarrow \text{last Activation}$

Discriminator of MLP :

$x \rightarrow \text{Linear}(\text{isize} \times \text{isize}, 1024) \rightarrow \text{ELU} \rightarrow \text{Linear}(1024, 1024) \rightarrow \text{ELU} \rightarrow \text{Linear}(1024, 1) \rightarrow \text{Link function}$

VIII Experimental setup for varying the activation function in the generator

Setup. We train adversarial networks with varying activation functions for the generators on the MNIST [6] and LSUN [15] datasets. In particular, we compare ReLU, Softplus, Least Square loss as an example of $\text{prop-}\tau$, and μ -ReLU with varying μ in $[0, 0.1, \dots, 1]$ by using them as the activation functions in all hidden layers of the generators. For all models, we fix the learning rate to 0.0002 and batch size to 64 throughout all experiments after tuning on a hold-out set.

MNIST. We evaluate the activation functions by using both DCGAN and the MLP used in [11] as the architectures. As training divergence, we adopt both GAN and Wasserstein distance (WGAN) because GAN belongs to variational f -divergence formulation while WGAN does not. The link function of the discriminators is specific to the respective divergence, which is sigmoid for GAN and linear for WGAN. We sample random noise $z \in \text{Uniform}_{100}(0, 1)$ for MLP and $z \in \text{Gaussian}(0, 1)$ for DCGAN, which is found slightly better than sampling from $\text{Uniform}_{100}(-1, 1)$. As the best practice, we apply Adam [5] to optimize models with GAN and RMSprop [14] to optimize WGAN based models. For GAN, we train one batch for discriminator and one batch for generator iteratively during training. For WGAN, we apply weight clipping with 0.01 and train five batches for discriminator and one batch for generator interchangeably during training.

We train all models on the full MNIST training data set and evaluate the performance on the test set by using the kernel density estimation (KDE). Since the size of images accepted by DCGAN should be n -fold of 16, all images are rescaled to 32-by-32 for all models. Following [11], we apply three-fold cross validation to find optimal bandwidth for the isotropic Gaussian kernel of KDE on a hold-out set. To estimate the log probability of the test set, we sample 16k images from the models in the same way as [11]. We observe that the initialization of model parameters has significant influence on performance. Therefore, we conduct three runs with different random seeds for each experimental setting and report the mean and standard deviation of the results.

LSUN. We also evaluate all activation functions in consideration for the generator on LSUN natural scene images. We train DCGAN with GAN as the divergence on the *tower* category of images, which are rescaled and center-cropped to 64-by-64 pixels, as in [12]. Due to the center-cropped images, we apply tanh as last activation of generators instead of sigmoid for GAN based models.

IX Visual results on MNIST

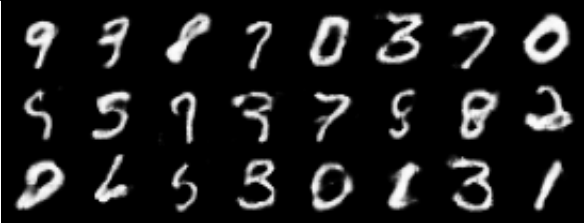
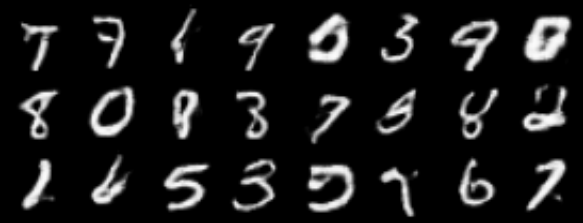
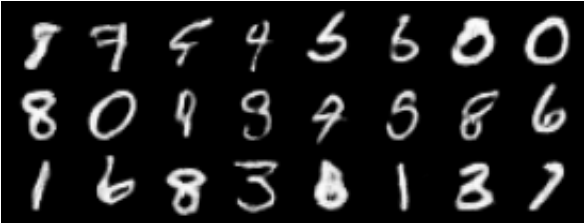
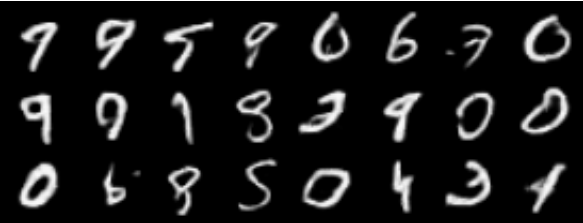
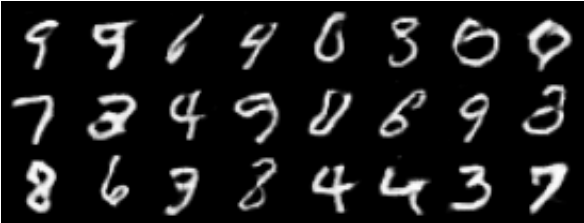
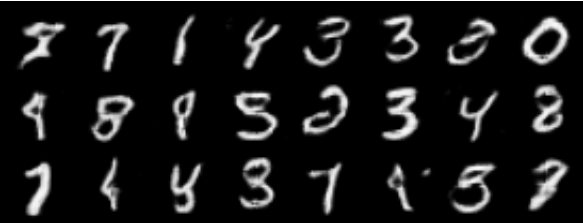
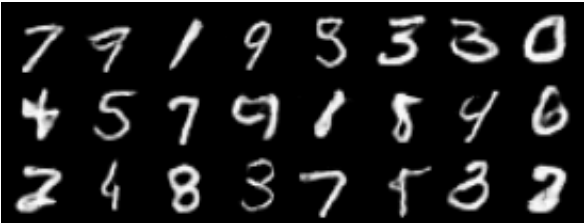
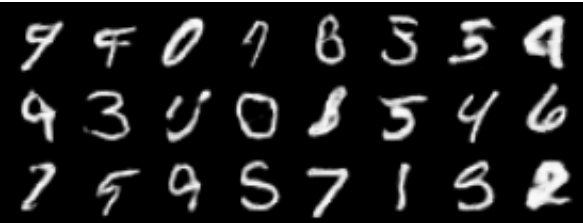
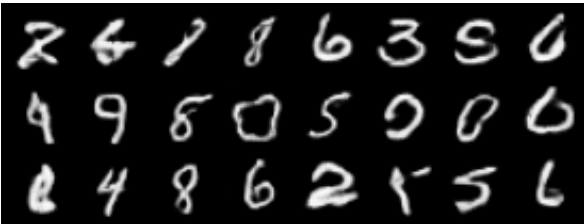
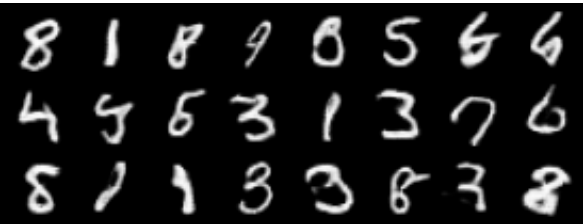

μ		μ	
0		0.1	
0.2		0.3	
0.4		0.5	
0.6		0.7	
0.8		0.9	
1			

Table A1: MNIST results for GAN_DCGAN at varying μ ($\mu = 1$ is ReLU).

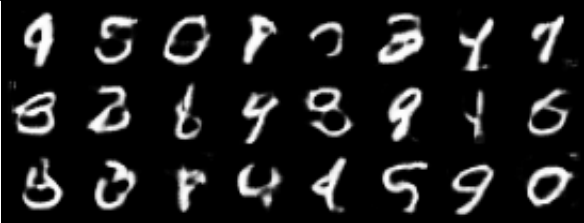
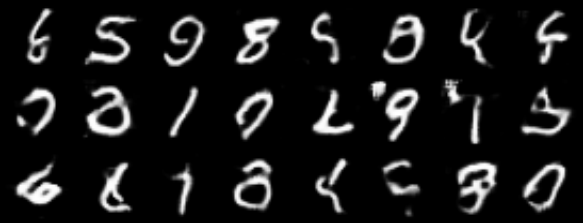
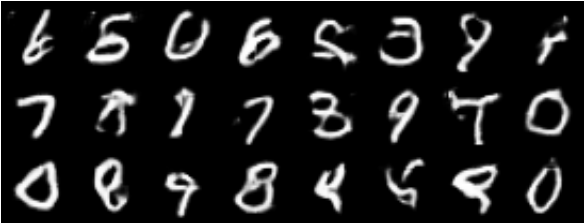
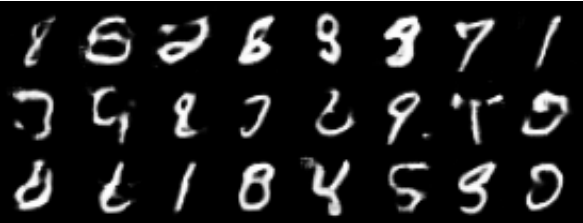
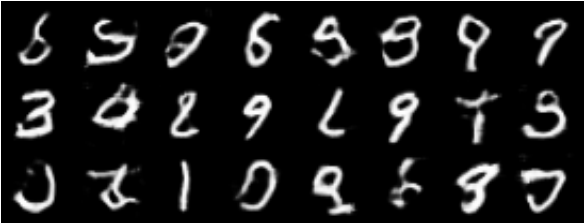
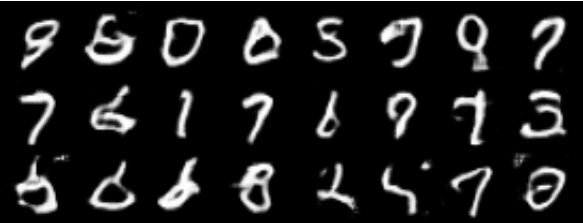
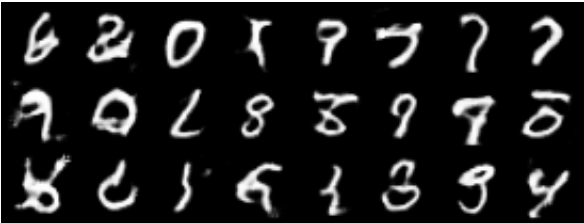
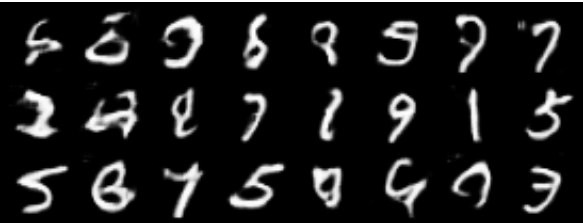
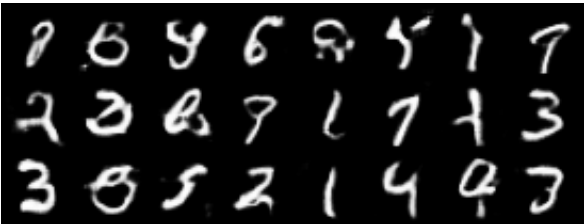
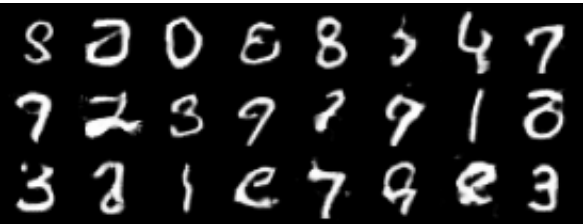
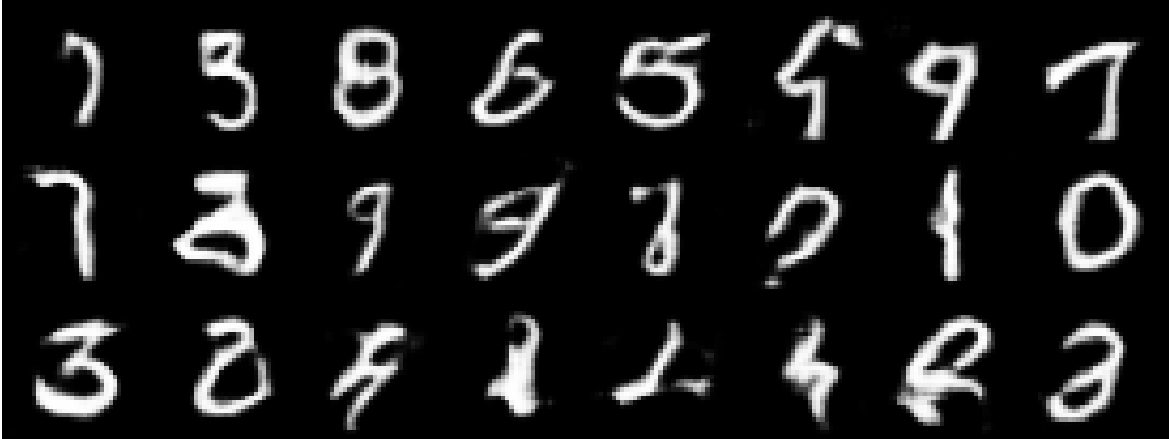
μ		μ	
0		0.1	
0.2		0.3	
0.4		0.5	
0.6		0.7	
0.8		0.9	
1			

Table A2: MNIST results for WGAN_DCGAN at varying μ ($\mu = 1$ is ReLU).

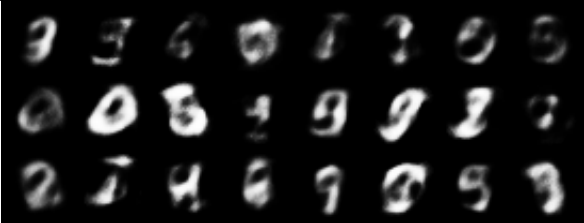
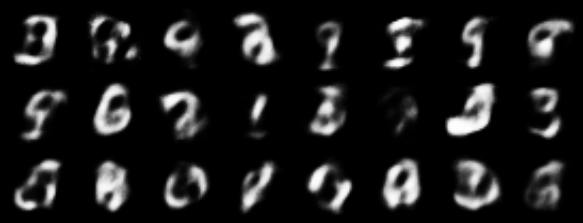
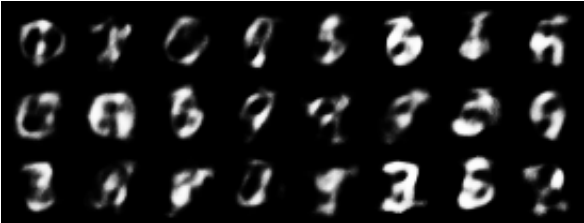

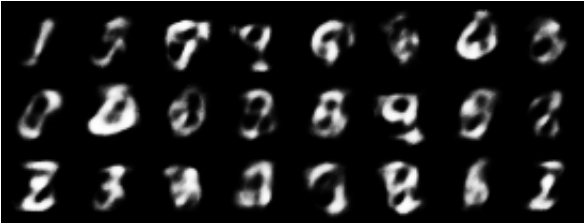
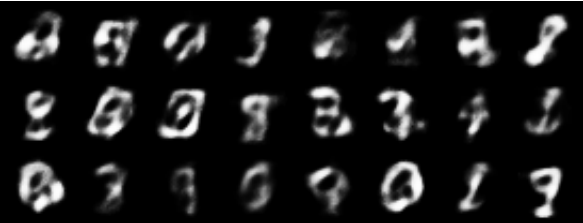
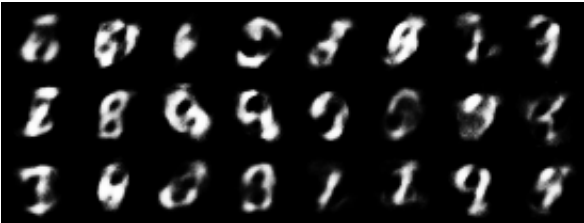
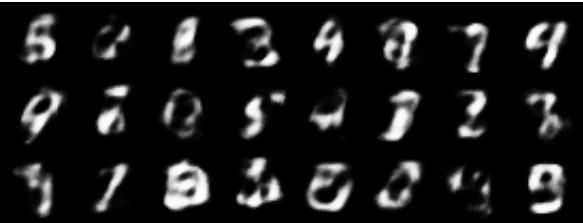
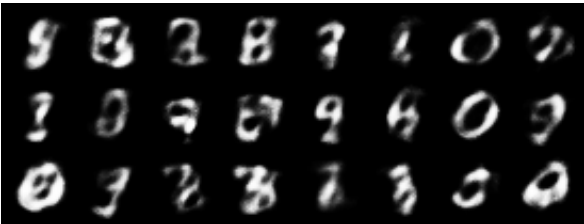
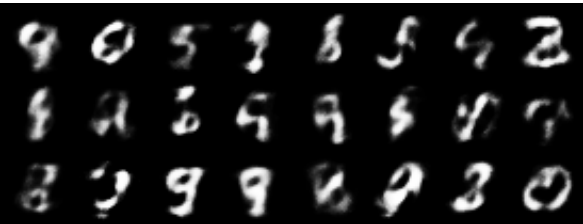
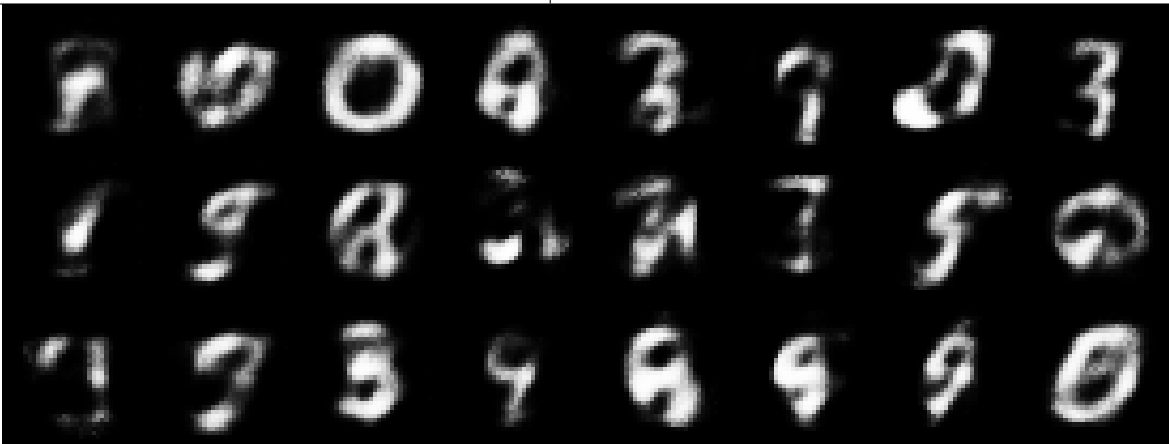
μ		μ	
0		0.1	
0.2		0.3	
0.4		0.5	
0.6		0.7	
0.8		0.9	
1			

Table A3: MNIST results for WGAN_MLP at varying μ ($\mu = 1$ is ReLU).


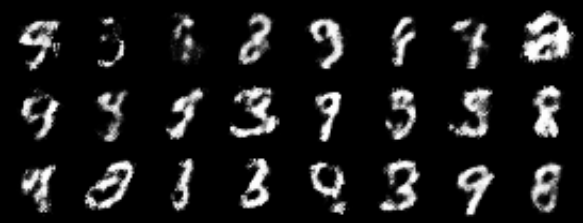

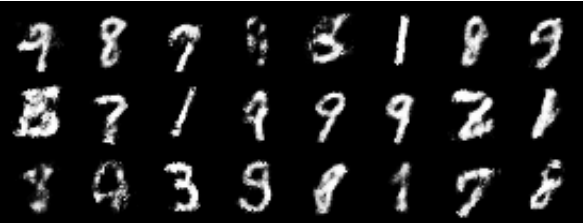
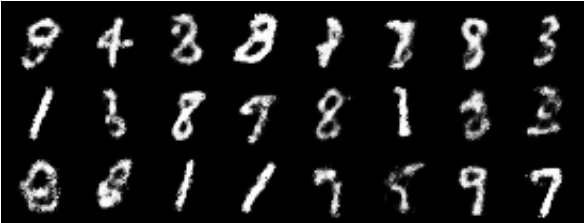

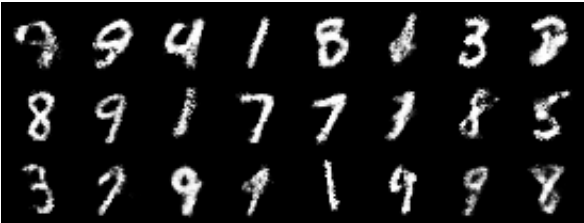
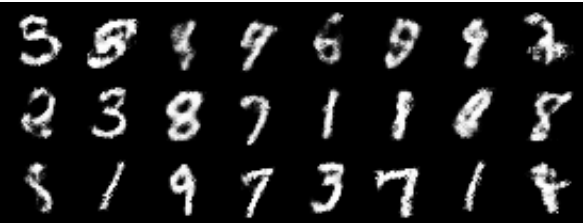
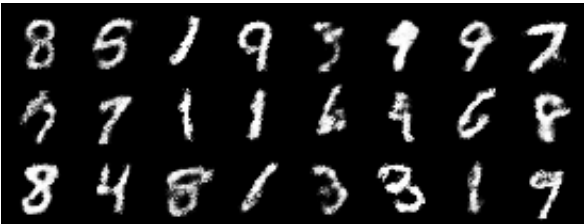
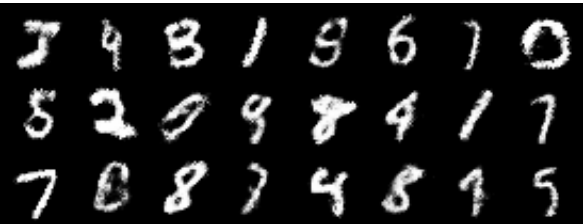
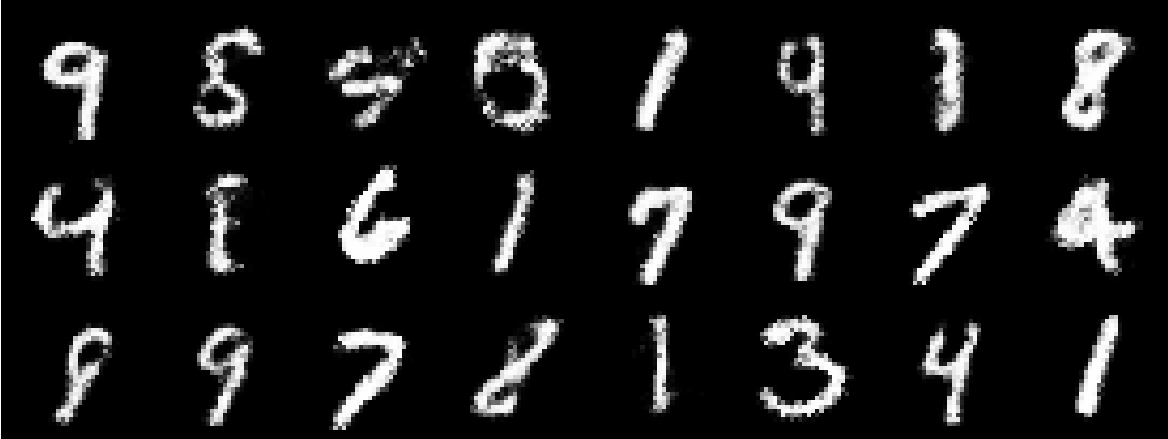
μ		μ	
0		0.1	
0.2		0.3	
0.4		0.5	
0.6		0.7	
0.8		0.9	
1			

Table A4: MNIST results for GAN_MLP at varying μ ($\mu = 1$ is ReLU).





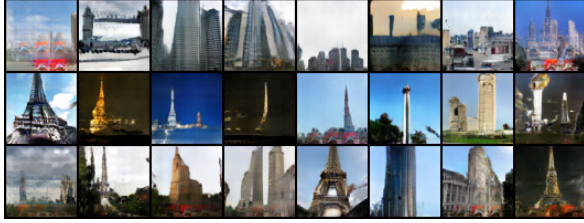
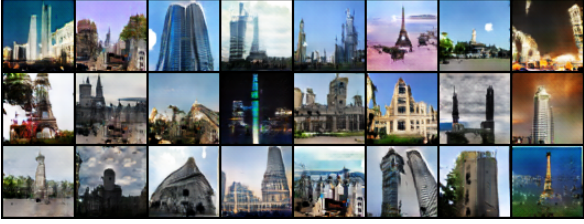
μ		μ	
0		0.1	
0.2		0.3	
0.4		0.5	
0.6		0.7	
0.8		0.9	
1			

Table A5: LSUN results for GAN_DCGAN at varying μ ($\mu = 1$ is ReLU).

References

- [1] S.-I. Amari. Personnal communication, 2017.
- [2] S.-I. Amari, A. Ohara, and H. Matsuzoe. Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries. *Physica A: Statistical Mechanics and its Applications*, 391:4308–4319, 2012.
- [3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.
- [4] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *5th ICLR*, 2017.
- [5] D.-P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [7] H. Lee, R. Ge, T. Ma, A. Risteski, and S. Arora. On the ability of neural nets to express distributions. *CoRR*, abs/1702.07028, 2017.
- [8] H. Matsuzoe and T. Wada. Deformed algebras and generalizations of independence on deformed exponential families. *Entropy*, 17:5729–5751, 2015.
- [9] J. Naudts. *Generalized thermostatics*. Springer, 2011.
- [10] C. Niculescu and L.-E. Persson. *Convex Functions and their Applications, A Contemporary Approach*. Springer, 2006.
- [11] S. Nowozin, B. Cseke, and R. Tomioka. f -GAN: training generative neural samplers using variational divergence minimization. In *NIPS*29*, pages 271–279, 2016.
- [12] A. Radford, L. Metz, and S. Chintala. unsupervised representation learning with deep convolutional generative adversarial networks. In *4th ICLR*, 2016.
- [13] M.-D. Reid and R.-C. Williamson. Composite binary losses. *JMLR*, 11, 2010.
- [14] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.
- [15] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.