7. Supplementary material

In what follows, we aim at proving a universal approximation theorem for the class of permutation invariant neural networks we have defined. To ease readings, products, sums and real function applications are assumed to be broadcasted when need be. Throughout the paper the batch dimension n is constant and ommited from set indices.

Definition 1. *A function* $f\colon \mathbb{R}^{n\times k} \mapsto \mathbb{R}^l$ *is* symmetric *if for any permutation of indexes* σ *and for all* $x \in \mathbb{R}^{n\times k}$, $f(x_{\sigma(1)}, \dots, x_{\sigma(n)}) = f(x_1, \dots, x_n)$. *The set of continuous symmetric functions from* $\mathbb{R}^{n\times k}$ *to* $\mathbb{R}^l$ *is denoted by* $\mathcal{I}_k^l$

Definition 2. *A function* $f\colon \mathbb{R}^{n\times k} \mapsto \mathbb{R}^{n\times l}$ *is* permutation equivariant *if for any permutation of indexes* σ *and for al* $x \in \mathbb{R}^n$, $f(x_{\sigma(1)}, \dots, x_{\sigma(n)}) = f(x)_{\sigma(1)}, \dots, f(x)_{\sigma(n)}$.

When symmetric functions and permutation equivariant functions are restricted to a compact, we assume that the compact itself is symmetric.

In what follows, we use ρ as a reducing operator on vectors defined for $x \in \mathbb{R}^{n\times k}$ by

$$\rho(x)_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}. \tag{15}$$

Definition 3. *Let the sets* E_k^l *be sets that contain* permutation equivariant neural networks *from* $\mathbb{R}^{n\times k}$ *to* $\mathbb{R}^{n\times l}$, *recursively defined thus:*

- *For all* $k \in \mathbb{N}$, *the identity function on* $\mathbb{R}^{n\times k}$ *belongs to* E_k^k.
- *For all* $f \in E_r^k$, $\Gamma \in \mathbb{R}^{l\times k}$, $\Lambda \in \mathbb{R}^{l\times k}$ *and* $\beta \in \mathbb{R}^l$, *and for act, a sigmoid activation function,* g *defined as*

$$g(x)_{i,j} = \sum_{p=1}^{k} \Gamma_{j,p} act(f(x))_{i,p} + \sum_{p=1}^{k} \Lambda_{j,p}\rho(act \circ f(x))_p + \beta_j) \tag{16}$$

  is in E_r^l.

The number of layers of the network is defined as the induction depth of the previous construction. The set of thus constructed permutation equivariant neural networks *with number of layers* L *is denoted by* $E(L)_k^l$. *Note that this class of function is trivially stable by composition, i.e. if* $g_1 \in E_{l_1}^{l_2}$ *and* $g_2 \in E_{l_2}^{l_3}$, *the* $g_2 \circ g_1 \in E_{l_1}^{l_3}$.

Definition 4. *Let* I_k^l *be a set containing* symmetric neural networks *from* $\mathbb{R}^{n\times k}$ *to* $\mathbb{R}^l$ *defined as*

$$I_k^l = \rho(E_k^l). \tag{17}$$

We have constructed sets I_k^l, containing permutation invarient networks. We now show that the way we they are constructed is not too restictive, i.e. that any analytical symmetric function can be approximated with arbitrary precision by a sufficiently expressive network of our construct. In other words we aim at proving Theorem 1 2.

Theorem 2. *For all* n, k, l *and for all compact* K, $I_k^l\big|_K$ *is dense in* $\mathcal{I}_k^l\big|_K$.

The first step of the proof is to show that the closure of $I_k^l\big|_K$ is a ring, i.e. that it is stable by sum, product and that each element has an inverse for $+$, as well as a vectorial space, making it an algebra. The second step is to prove that this closure contains a generative familly of the set of all polynomials that operate symmetrically on the batch dimension and because symmetric polynomials are dense in the set of all symmetric functions, this proves the theorem.

Lemma 1. *If* $f_1 \in \overline{E_{l_1}^{l_2}\Big|_K}$ *and* $f_2 \in \overline{E_{l_2}^{l_3}\Big|_{f_1(K)}}$ *then* $f_2 \circ f_1 \in \overline{E_{l_1}^{l_3}\Big|_K}$.

Proof. Let $\varepsilon > 0$, f_2 is continuous on a compact set, thus uniformly continuous, and there exists an $\eta > 0$ such that $\|x - x'\| < \eta$ implies $\|f_2(x) - f_2(x')\| < \frac{\varepsilon}{2}$. Now let $g_1 \in E_{l_1}^{l_2}\Big|_K$ be such that $\|g_1 - f_1\|_\infty \leq \eta$ and $g_2 \in E_{l_2}^{l_3}\Big|_K$ such that $\|g_2 - f_2\|_\infty \leq \frac{\varepsilon}{2}$, then, for x in K

$$\begin{aligned}\|f_2 \circ f_1(x) - g_2 \circ g_1(x)\| &\leq \|f_2 \circ f_1(x) - g_2 \circ f_1(x)\| + \|g_2 \circ f_1(x) - g_2 \circ g_1(x)\| \\ &\leq \varepsilon\end{aligned}$$

□

Lemma 2. *For any continuous functions $g\colon \mathbb{R}^k \mapsto \mathbb{R}^l$, the restriction of the function $G\colon \mathbb{R}^{n\times k} \mapsto \mathbb{R}^{n\times k}$, defined as $G(x) = (g(x_1), \dots, g(x_n))$, to a compact K is in $\overline{E_k^l\big|_K}$. More precisely, for all $L \geq 2$, the restriction of G to K is in $\overline{E(L)_k^l\big|_K}$.*

Proof. This is a consequence of the neural network universal approximation theorem, as stated e.g. in (Cybenko, 1989). □

Lemma 3. *If $f_1 \in E_k^{l_1}\Big|_K$, $f_2 \in E_k^{l_2}\Big|_K$ and f_1 and f_2 have the same number of layers (i.e. they have the same induction depth), then $concat_1(f_1, f_2) \in E_k^{l_1,l_2}\Big|_K$, with*

$$concat_1(x,y)_{i,j} = \begin{cases} x_{i,j} \text{ if } j \leq l_1 \\ y_{i,j-l_1} \text{ otherwise} \end{cases} \tag{18}$$

Proof. By induction on the number of layers L,

- if $L = 0$, the result is clear.
- if $L > 0$, let g_1, Γ_1, Λ_1 and β_1 as well as g_2, Γ_2, Λ_2 and β_2 be the parameters associated to f_1 and f_2, then, by induction, $\text{concat}_1(g_1, g_2)$ is a permutation equivariant network, and $\text{concat}_1(f_1, f_2)$ is obtained by setting Γ to be the block diagonal matrix obtained with Γ_1 and Γ_2, Λ, the block diagonal matrix obtained with Λ_1 and Λ_2, and β the concatenation of both β's.

□

Lemma 4. *If $f_1 \in \overline{E_k^{l_1}\Big|_K}$, $f_2 \in \overline{E_k^{l_2}\Big|_K}$, then $concat_1(f_1, f_2) \in \overline{E_k^{l_1+l_2}\Big|_K}$.*

Proof. Let $\varepsilon > 0$, let $g_1 \in E_k^{l_1}\Big|_K$ and $g_2 \in E_k^{l_2}\Big|_K$ be such that $\|g_1 - f_1\|_\infty \leq \frac{\varepsilon}{4}$ and $\|g_2 - f_2\|_\infty \leq \frac{\varepsilon}{4}$. Denote by L_1 and L_2 the numbers of layers of g_1 and g_2. We assume $L_1 \geq L_2$ without loss of generality. By lemma 2, there exist $h_1 \in E_{l_1}^{l_1}\Big|_K$ and $h_2 \in E_{l_2}^{l_2}\Big|_K$ with h_1 of depth 2 and h_2 of depth $L_1 - L_2 + 2$ such that $\|h_1 - Id\|_\infty \leq \frac{\varepsilon}{4}$ on $g_1(K)$ and $\|h_2 - Id\|_\infty \leq \frac{\varepsilon}{4}$ on $g_2(K)$. The networks $h_1 \circ g_1$ and $h_2 \circ g_2$ have the same number of layers, consequently, $\text{concat}_1(h_1 \circ g_1, h_2 \circ g_2) \in E_k^{l_1,l_2}\Big|_K$. Besides,

$$\begin{aligned} &\|\text{concat}_1(f_1, f_2) - \text{concat}_1(h_1 \circ g_1, h_2 \circ g_2)\|_\infty && (19)\\ \leq &\|f_1 - g_1\|_\infty + \|h_1 \circ g_1 - g_1\|_\infty + \|f_2 - g_2\|_\infty + \|h_2 \circ g_2 - g_2\|_\infty && (20)\\ \leq &\varepsilon && (21) \end{aligned}$$

yielding the result. □

Lemma 5. *If f_1 and f_2 are in $\overline{E_k^l\big|_K}$, then $f_1 + f_2$ is too.*

Proof. By lemma 3, $\text{concat}_1(f_1, f_2)$ is in $\overline{E_k^{2l}\big|_K}$. Consider the layer g, with kernels $\Gamma_{i,j} = \begin{cases} 1 \text{ if } j = i \text{ or } j = k + i \\ 0 \text{ otherwise} \end{cases}$, $1 \leq i \leq l, 1 \leq j \leq 2l$, $\Lambda = 0$, $\beta = 0$. By lemma 1, as both $\text{concat}_1(f_1, f_2)$ and g are in closures of permutation equivariant networks, their composition is too. This composition is $\text{act}(f_1 + f_2)$. By the universal approximation theorem act^{-1} is also in the closure so $f_1 + f_2$ is in the closure. □

More generally, following similar reasonings, closures of permutation equivariant networks are vectorial spaces. It follows that closures of permutation invariant networks are vectorial spaces too.

Lemma 6. *If $f \in \overline{I_k^l\big|_K}$, then F defined by*

$$F(x)_{i,j} = f(x)_j \tag{22}$$

for all i, j, is in $\overline{E_k^l\big|_K}$

Proof. By definition, for any $\varepsilon > 0$, there exists a G in $E_k^l\big|_K$ such that f and $\rho(G)$ are at distance at most $\frac{\varepsilon}{2}$. Let α be a non zero real number such that $\text{act}^{-1}(\alpha G(x))$ is well defined for any $x \in K$. Consider the equivariant layer

$$m(x)_{i,j} = \alpha^{-1}\rho(\text{act}(x))_j. \tag{23}$$

Let η_1 be a positive real number, and L_{η_1} be a compact set that contains both $\text{act}^{-1}(\alpha G(K))$ and any ball of radius η_1 contained in this set. m is uniformly continuous on L_{η_1}, and consequently there exists an η_2 such that if x and y are at distance at most η_2, $m(x)$ and $m(y)$ are at distance at most $\frac{\varepsilon}{2}$. Now, by composition and the universal approximation theorem, let $h \in E_k^l$ be such that h and $\text{act}^{-1}(\alpha G)$ are at distance at most $\min(\eta_1, \eta_2)$. Then $m \circ \text{act}^{-1}(\alpha G)$ and $m \circ h$ are at distance at most $\frac{\varepsilon}{2}$, and by triangular inequality, F and $m \circ h$ are at distance at most ε. □

Lemma 7. *If f_1 and f_2 are in $\overline{I_k^l\big|_K}$, then $f_1 f_2$ is too.*

Proof. Let F_1 and F_2 be the extensions of f_1, f_2 as defined in lemma 6. There exists a $C \in \mathbb{R}$ such that for all $i, j, x \in K$, $F_1(x)_{i,j} + C > 0$, and similarily for F_2. Consequently, by lemma 1, lemma 2 and lemma 5, $\exp(\log(F_1 + C) + \log(F_2 + C)) = F_1F_2 + F_1C + F_2C + C^2 \in \overline{E_k^l\big|_K}$. As this closure is a vectorial space, $F_1F_2 \in \overline{E_k^l\big|_K}$. Consequently, $f_1 f_2 = \rho(F_1F_2) \in \overline{I_k^l\big|_K}$. □

We have now shown that $\overline{I_k^l\big|_K}$ is a ring. We are left to prove that it contains a generative familly of the continuous symmetric functions. Let us first exhibit a familly of continuous symmetric functions that is contained in the set of interest, and that we will later show generate all continuous symmetric function.

Lemma 8. *For all f, restriction of a function from R^l to R^k to a compact set K, the symmetric function F, defined on $K^{n\times l}$ by*

$$F(x) = \sum_{i=1}^{n} f(x_i) \tag{24}$$

is in I_k^l.

Proof. By the universal approximation theorem, f is in $\overline{I_k^l\big|_K}$. By lemma 6, there exists a G in $\overline{E_k^l\big|_K}$ that replicates f along the batch axis of an equivariant network. Consequently, $\rho(G) = F$ is in $\overline{I_k^l\big|_K}$. □

We are going to prove that this familly of functions generates the set of all symmetric polynomials. Deriving a generalization of Stone Weierstrass theorem to symmetric functions, we obtain the final result.

To keep things general, in what follows, X denotes an arbitrary set, F an algebra of functions on X, and S is the symmetrization operator on functions of X^n, i.e. for all $(x_1, \dots, x_n) \in X^n$,

$$(Sf)(x_1, \dots, x_n) = \sum_{\sigma} f(x_{\sigma(1)}, \dots, x_{\sigma(n)}) \tag{25}$$

where the sum is over all permutations of $[1, n]$.

Let P be the algebra of functions of X^n generated by the functions $f(x_k)\colon x \to f(x_k)$ for f in F, with a slight abuse of notations. P is linearly generated by the monomials $f_1(x_1) \dots f_n(x_n)$ for f_k arbitrary functions of F. We are interested in the symmetrization of P, SP. By linearity of S, SP is generated by the symmetrized monomials,

$$Sf_1(x_1) \dots f_n(x_n) = \sum_{\sigma} \prod_{k=1}^{n} f_k(x_{\sigma(k)}). \tag{26}$$

Lemma 9. *SP is generated as an algebra by $Sf(x_1)$ for $f \in F$. Notably, $Sf(x_1)$ takes the special form*

$$Sf(x_1) = \sum_{\sigma} f(x_{\sigma(1)}) = (n-1)! \sum_{k=1}^{n} f(x_k). \tag{27}$$

Typically, for our case, $X = \mathbb{R}^l$ for l the number of input features, F is an algebra of functions containing the multivariate polynomials on $\mathbb{R}^l$, and SP thus contains the set of all polynomials which are symmetric along the batch dimension.

Proof. Call *rank* of a monomial $f_1(x_1) \dots f_n(x_n)$, the number of functions f_k such that $f_k \neq 1$. Let $k_1, \dots, k_r$ be these indices. Up to renaming f_{k_1} to f_1, etc., the monomial can be written as $f_1(x_{k_1}) \dots f_r(x_{k_r})$.

We will work by induction on r. For $r = 1$ the claim is trivial.

Since S does not care about permuting the variables, we have

$$Sf_1(x_{k_1}) \dots f_r(x_{k_r}) = Sf_1(x_1) \dots f_r(x_r) = \sum_{\sigma \in \mathcal{S}_{\mathrm{K}}} \prod_{i=1}^{r} f_i(x_{\sigma(i)}) \tag{28}$$

and now the values $\sigma(r+1), \dots, \sigma(n)$ have no influence so that

$$Sf_1(x_1) \dots f_n(x_n) = (n-r)! \sum_{\sigma \in \mathrm{Inj}_r^n} \prod_{i=1}^{r} f_i(x_{\sigma(i)}) \tag{29}$$

where Inj_r^n is the set of injective functions from r to n.

Assume we can generate all symmetric monomials up to rank r. By definition we can generate $Sf_{r+1}(x_1)$ for any $f_{r+1} \in F$. Then we can generate the product

$$\begin{aligned}\frac{1}{(n-r-1)!}(Sf_{r+1}(x_1)) \left(\sum_{\sigma \in \mathrm{Inj}_r^n} \prod_{i=1}^{r} f_i(x_{\sigma(i)}) \right) &= (\sum_{k \in n} f_{r+1}(x_k)) \left(\sum_{\sigma \in \mathrm{Inj}_r^n} \prod_{i=1}^{r} f_i(x_{\sigma(i)}) \right) \\ &= \sum_{\sigma \in \mathrm{Inj}_r^n} \sum_{k \in n} f_{r+1}(x_k) \prod_{i=1}^{r} f_i(x_{\sigma(i)})\end{aligned}$$

Now, for each σ, we can decompose according to whether $k \in \operatorname{Im} \sigma$ or $k \in n \setminus \operatorname{Im} \sigma$, where $\operatorname{Im} \sigma = \{\sigma(1), \dots, \sigma(r)\}$ is the image of σ. We obtain two terms

$$\dots = \sum_{\sigma \in \mathrm{Inj}_r^n} \sum_{k \in \operatorname{Im} \sigma} f_{r+1}(x_k) \prod_{i=1}^{r} f_i(x_{\sigma(i)}) + \sum_{\sigma \in \mathrm{Inj}_r^n} \sum_{k \in n \setminus \operatorname{Im} \sigma} f_{r+1}(x_k) \prod_{i=1}^{r} f_i(x_{\sigma(i)})$$

But if k is not in $\operatorname{Im} \sigma$, then $(\sigma(1), \dots, \sigma(r), k)$ is an injective function from $r+1$ to n. So summing over σ then on $k \in n \setminus \operatorname{Im} \sigma$ is exactly equivalent to summing over $\sigma \in \mathrm{Inj}_{r+1}^n$. So the second term above is

$$\sum_{\sigma \in \mathrm{Inj}_{r+1}^n} \left(\prod_{i=1}^{r} f_i(x_{\sigma(i)}) \right) f_{r+1}(\sigma(r+1)) = \sum_{\sigma \in \mathrm{Inj}_{r+1}^n} \prod_{i=1}^{r+1} f_i(x_{\sigma(i)}) = Sf_1(x_{k_1}) \dots f_{r+1}(x_{k_{r+1}})$$

which is the one we are interested in.

So if we prove that we can generate the first term, we are done.

Let us consider the first term, with $k \in \operatorname{Im} \sigma$. Now, since $k \in \operatorname{Im} \sigma$, we can decompose over the cases $k = \sigma(1), \dots, k = \sigma(r)$, namely,

$$\sum_{\sigma \in \mathrm{Inj}_r^n} \sum_{k \in \operatorname{Im} \sigma} f_{r+1}(x_k) \prod_{i=1}^{r} f_i(x_{\sigma(i)}) = \sum_{\sigma \in \mathrm{Inj}_r^n} \sum_{j=1}^{r} f_{r+1}(x_{\sigma(j)}) \prod_{i=1}^{r} f_i(x_{\sigma(i)}) \tag{30}$$

$$= \sum_{j=1}^{r} \sum_{\sigma \in \mathrm{Inj}_r^n} \prod_{i=1}^{r} \tilde{f}_{ij}(x_{\sigma(i)}) \tag{31}$$

where

$$\tilde{f}_{ij} := \begin{cases} f_i & i \neq j \\ f_i f_{r+1} & i = j \end{cases} \tag{32}$$

Now since F is a ring, $f_i f_{r+1} \in F$. For each j the term

$$\sum_{\sigma \in \mathrm{Inj}_r^n} \prod_{i=1}^{r} \tilde{f}_{ij}(x_{\sigma(i)}) \tag{33}$$

is equal to $S\tilde{f}_{1j} \dots \tilde{f}_{rj}$ up to a factor $(n-(r+1))!$. By our induction hypothesis, each term can be generated. This ends the proof. □

Lemma 10. *For any compact K, any $l \in \mathbb{N}$, the intersection of $\mathcal{I}_l^1$ with the set of multivariate polynomials is dense in $\mathcal{I}_l^1$ for the infinity norm.*

Proof. Let $\varepsilon > 0$, and f be in I_l^1. There exists a multivariate polynomials P such that $\|P - f\|_\infty \leq \varepsilon$. Let us consider the symmetrized polynomial

$$\tilde{P}(x_1, \dots, x_n) = \frac{1}{n!} \sum_\sigma P(x_{\sigma(1)}, \dots, x_{\sigma(n)}). \tag{34}$$

Then $\tilde{P}$ is in the intersection, and, for $x \in K$,

$$\|\tilde{P}(x) - f(x)\| = \|\frac{1}{n!} \sum_\sigma (P(x_{\sigma(1)}, \dots, x_{\sigma(n)}) - f(x_{\sigma(1)}, \dots, x_{\sigma(n)}))\| \tag{35}$$

$$\leq \frac{1}{n!} \sum_\sigma \|P(x_{\sigma(1)}, \dots, x_{\sigma(n)}) - f(x_{\sigma(1)}, \dots, x_{\sigma(n)})\| \tag{36}$$

$$\leq \varepsilon. \tag{37}$$

□

We now have all the ingredients to end the proof. For a given compact K of $\mathbb{R}^l$, for any multivariate polynomial P of $\mathbb{R}^l$, any $\varepsilon > 0$, there trivially exists an element f of I_k^1 at distance at most ε of $x \to \sum_{i=1}^{n} P(x_i)$. This means that the closure of the considered set contains all such functions. As this closure is an algebra (it is both a ring and a vectorial space), by lemma 8, it contains the intersection of $\mathcal{I}_l^2$ with the set of multivariate polynomials. By lemma 10, it contains $\mathcal{I}_l^1$, which ends the proof.

8. Other details

8.1. $p_{balanced}$ and $p_{unbalanced}$ are well normalized:

We now show that $p_{unbalanced}$ is well defined. The computation for $p_{balanced}$ is almost identical and left to the reader.

$$\begin{aligned} \int_y p_{unbalanced}(y)dy &= \frac{2}{B+1} \sum_{\beta \in \{0,1\}^B} \frac{\#\beta}{BC_B^{\#\beta}} \int_y p_x(y)^\beta p_{\tilde{x}}(y)^{1-\beta} dy \\ &= \frac{2}{B+1} \sum_{\#\beta=1}^{B} C_B^{\#\beta} \frac{\#\beta}{BC_B^{\#\beta}} \\ &= \frac{2}{(B+1)B} \sum_{\#\beta=1}^{B} \#\beta \\ &= \frac{2}{(B+1)B} \frac{B(B+1)}{2} \\ &= 1 \end{aligned}$$

9. Optimal discriminator for general beta prior

We hereby give a derivation of the optimal discriminator expression, when mixing parameters, p's are drawn from Beta(a, b). This extends Eq. (7), as Beta$(1, 1) = \mathcal{U}([0, 1])$.

Beta prior on batch mixing proportion. Consider mixed batches of samples of size B. The i-th sample of the batch is a real sample if $\beta_i = 1$ and a false sample if $\beta_i = 0$. Given a certain mixing proportion p, assuming that sample origine are sampled independantly according to a Bernoulli of parameter p, the probability of a certain β is

$$\mathbb{P}(\beta \mid p) = \prod_i p^{\beta_i}(1-p)^{1-\beta_i}, \tag{38}$$

Considering a beta prior distribution Beta(a, b) on the mixing parameter $p \in [0, 1]$, the posterior distribution on the number of real sample in the batch $\#\beta = \sum_i \beta_i$ is given by the beta-binomial compound distribution

$$\begin{aligned} \mathbb{P}(\#\beta) &= \int_p \text{Beta}(p \mid a, b)\mathbb{P}(\#\beta|p) && (39)\\ &= \binom{B}{\#\beta}\frac{\mathcal{B}(\#\beta + a, B - \#\beta + b)}{\mathcal{B}(a, b)} && (40)\end{aligned}$$

where $\mathcal{B}(\cdot, \cdot)$ is the beta function. For $a = 1, b = 1$, i.e. a uniform distribution on mixing parameters, the beta-binomial compound distribution reduces to a uniform distribution on $\#\beta$. From the expression of $\mathbb{P}(\#\beta)$ it follows that

$$\mathbb{P}(\beta) = \frac{\mathcal{B}(\#\beta + a, B - \#\beta + b)}{\mathcal{B}(a, b)}. \tag{41}$$

Optimal discriminator. Let $y = m_\beta(x, \tilde{x})$ denote a mixed batch of samples. The discriminator minimizes the KL divergence between $D(y)$ and β, averaged over batches and mixing vectors β, see Eq. (4) in the main paper. This reduces to minimizing the expected cross-entropy. For a given batch and mixing vector β,

$$L(D(y), \#\beta) = -\frac{\#\beta}{B}\ln D(y) - \frac{B - \#\beta}{B}\ln(1 - D(y)). \tag{42}$$

Averaging over batches and mixing vectors,

$$\begin{aligned} \mathbb{E}_{\beta,y}[L(D(y), \#\beta)] &= \int_y \mathbb{P}(y)\sum_\beta \mathbb{P}(\beta|y)L(D(y), \#\beta) && (43)\\ &= -\int_y \mathbb{P}(y)\left[\mathbb{E}_{\beta|y}\left[\frac{\#\beta}{B}\right]\ln D(y) + \mathbb{E}_{\beta|y}\left[\frac{B - \#\beta}{B}\right]\ln(1 - D(y))\right] && (44)\end{aligned}$$

From the latter it yields that for any y, the optimal discriminator value $D^*(y)$ is

$$D^*(y) = \mathbb{E}_{\beta|y}\left[\frac{\#\beta}{B}\right], \tag{45}$$

i.e. the posterior expectation of the fraction of training samples in the batch.

Posterior analysis. Through Bayes rule, the posterior expectation yields

$$D^*(y) = \mathbb{E}_{\beta|y}\left[\frac{\#\beta}{B}\right] = \frac{\sum_\beta \frac{\#\beta}{B}\mathbb{P}(y|\beta)\mathbb{P}(\beta)}{\mathbb{P}(y)}. \tag{46}$$

The marginal on the batch y is

$$\begin{aligned} \mathbb{P}(y) &= \sum_\beta \mathbb{P}(y \mid \beta)\mathbb{P}(\beta) && (47)\\ &= \sum_\beta \mathbb{P}(y|\beta)\frac{\mathcal{B}(\#\beta + a, B - \#\beta + b)}{\mathcal{B}(a, b)}. && (48)\end{aligned}$$

The numerator in Eq. (46) can be written as a distribution on y,

$$\mathbb{Q}(y) = \sum_{\beta} \mathbb{P}(y|\beta)\mathbb{Q}(\beta) \tag{49}$$

$$\mathbb{Q}(\beta) = \frac{a+b}{a}\mathbb{P}(\beta)\frac{\#\beta}{B}. \tag{50}$$

The distribution $\mathbb{Q}$ on β sums to 1, as $\mathbb{E}_{\mathbb{P}(\#\beta)}[\#\beta] = \frac{Ba}{a+b}$.

This finally yields

$$D^*(y) = \frac{a}{a+b}\frac{\mathbb{Q}(y)}{\mathbb{P}(y)}, \tag{51}$$

which for the uniform prior on p simplifies to

$$D^*(y) = \frac{1}{2}\frac{\mathbb{Q}(y)}{\mathbb{P}(y)}. \tag{52}$$

Expressing $\mathbb{P}(y \mid \beta)$. Notice that $m_\beta(x, \tilde{x}) = y$ is equivalent to for all i in $\{1, ..., B\}$, $x_i = y_i$ and $\beta_i = 1$ or $\tilde{x}_i = y_i$ and $\beta_i = 0$. Denote by p_1 (resp. p_2) the distribution of real samples (resp. generated samples).

From the previous observation, it yields that

$$\mathbb{P}(y \mid \beta) = \prod_{i=1}^{B} p_1(y_i)^{\beta_i} p_2(y_i)^{1-\beta_i}. \tag{53}$$

From the latter and Eq. (52) we obtain the optimal discriminator expression.

10. Additional experiments

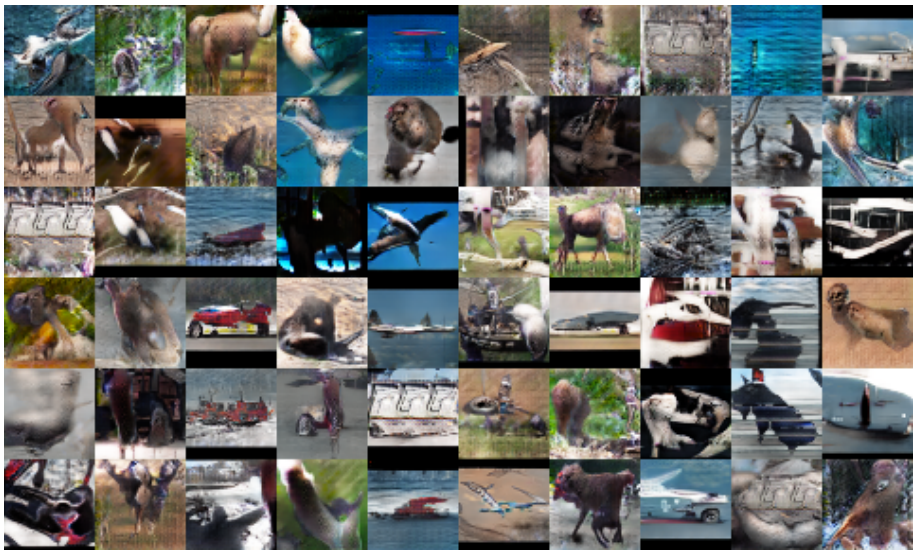

Figure 7. Sample images generated by our best model trained on STL10.

We additionally provide results on the STL-10 dataset, where M-BGAN yields numerical results slightly below Spectral Normalization. Except for the adaptation of the network to 48×48 images, as done in (Miyato et al., 2018), the experimental setup of the experimental section is left unchanged.

Table 2. Comparison to the state of the art in terms of inception score (IS) and Fréchet inception distance (FID) on the STL-10 dataset.

Model	IS	FID
WGP (Miyato et al., 2018)	8.4	55
M-BGAN	8.7	51
SN (Miyato et al., 2018)	8.7	47.5
SN (Hinge loss)(Miyato et al., 2018)	8.8	43.2

11. M-BGAN as an ensembling method

Intuitively, the M-BGAN loss performs a simple ensembling of many strongly dependant permutation invariant discriminators, at no additional cost.

In the general case, ensembling of N independant discriminators $D_1, \dots, D_N$ amounts to training each discriminator independently, and using the averaged gradient signal to train the generator. Ensembling is expected to alleviate some of the difficulties of GAN training: as long as one of the discriminators still provides a significant gradient signal, training of the generator is possible. With equation (14), M-BGAN is an ensemble of B permutation invariant discriminators, with respective outputs $1\text{-th}(o_1, \dots, o_B), \dots, B\text{-th}(o_1, \dots, o_B)$, where i-th is the function that returns the i-th greatest element of a B dimensional vector. Indeed,

$$\frac{1}{N}\sum_{i=1}^{N} l(i\text{-th}(o_1, \dots, o_B), y) = \frac{1}{N}\sum_{i=1}^{N} l(o_i, y). \tag{54}$$

which is the M-BGAN loss. The ensembled discriminators of the M-BGAN all share the same weights. We believe this ensembling effect at least partially explains the improved performance of M-BGAN.