Appendix

In Section A we give the proofs of all the Propositions and the Theorem. In Section B we give other theoretical results to validate statements made in the paper. Section C presents the algorithm from Maclaurin et al. (2015). In Section D we illustrate with codes that Momentum ResNets are a drop-in replacement for ResNets. Section E gives details for the experiments in the paper. We derive the formula for backpropagation in Momentum ResNets in Section F. Finally, we present additional figures in Section G.

A. Proofs

Notations

- $C_0^\infty([0,1], \mathbb{R}^d)$ is the set of infinitely differentiable functions from $[0,1]$ to $\mathbb{R}^d$ with value 0 in 0.
- If $f : U \times V \to W$ is a function, we denote by $\partial_u f$, when it exists, the partial derivative of f with respect to $u \in U$.
- For a matrix $A \in \mathbb{R}^{d\times d}$, we denote by $(\lambda - z)^a$ the Jordan block of size $a \in \mathbb{N}$ associated to the eigenvalue $z \in \mathbb{C}$.

A.0. Instability of fixed points – Proof of Proposition 1

Proof. Since (x^*, v^*) is a fixed point of the RevNet iteration, we have

$$\varphi(x^*) = 0$$

$$\psi(v^*) = 0$$

Then, a first order expansion, writing $x = x^* + \varepsilon$ and $v = v^* + \delta$ gives at order one

$$\Psi(v, x) = (v^* + \delta + A\varepsilon, x^* + \varepsilon + B(\delta + A\varepsilon)) \tag{9}$$

We therefore obtain at order one

$$\Psi(v, x) = \Psi(v^*, x^*) + J(A, B)\begin{pmatrix}\delta\\ \varepsilon\end{pmatrix}$$

which shows that $J(A, B)$ is indeed the Jacobian of Ψ at (v^*, x^*). We now turn to a study of the spectrum of $J(A, B)$. We let $\lambda \in \mathbb{C}$ an eigenvalue of $J(A, B)$, and vectors $u \in \mathbb{C}^d$, $w \in \mathbb{C}^d$ such that (u, w) is the corresponding eigenvector, and study the eigenvalue equation

$$J(A, B)\begin{pmatrix}u\\ w\end{pmatrix} = \lambda\begin{pmatrix}u\\ w\end{pmatrix}$$

which gives the two equations

$$u + Aw = \lambda u \tag{10}$$

$$w + Bu + BAw = \lambda w \tag{11}$$

We start by showing that $\lambda \neq 1$ by contradiction. Indeed, if $\lambda = 1$, then (10) gives $Aw = 0$, which implies $w = 0$ since A is invertible. Then, (11) gives $Bu = 0$, which also implies $u = 0$. This contradicts the fact that (u, v) is an eigenvector (which is non-zero by definition).

Then, the first equation (10) gives $Aw = (\lambda - 1)u$, and multiplying (11) by A on the left gives

$$\lambda ABu = (\lambda - 1)^2 u \tag{12}$$

We also cannot have $\lambda = 0$, since it would imply $u = 0$. Then, dividing (12) by λ shows that $\frac{(\lambda-1)^2}{\lambda}$ is an eigenvalue of AB.

Next, we let $\mu \neq 0$ the eigenvalue of AB such that $\mu = \frac{(\lambda-1)^2}{\lambda}$. The equation can be rewritten as the second order equation

$$\lambda^2 - (2+\mu)\lambda + 1 = 0$$

This equation has two solutions $\lambda_1(\mu)$, $\lambda_2(\mu)$, and since the constant term is 1, we have $\lambda_1(\mu)\lambda_2(\mu) = 1$. Taking modulus, we get $|\lambda_1(\mu)||\lambda_2(\mu)| = 1$, which shows that necessarily, either $|\lambda_1(\mu)| \geq 1$ or $|\lambda_1(\mu)| \geq 1$.

Now, the previous reasoning is only a necessary condition on the eigenvalues, but we can now prove the advertised result by going backwards: we let $\mu \neq 0$ an eigenvalue of AB, and $u \in \mathbb{C}^d$ the associated eigenvector. We consider λ a solution of $\lambda^2 - (2+\mu)\lambda + 1 = 0$ such that $|\lambda| \geq 1$ and $\lambda \neq 1$. Then, we consider $w = (\lambda - 1)A^{-1}u$. We just have to verify that (u, v) is an eigenvector of $J(A, B)$. By construction, (10) holds. Next, we have

$$A(w + Bu + BAw) = (\lambda - 1)u + ABu + (\lambda - 1)ABu = (\lambda - 1)u + \lambda ABu$$

Leveraging the fact that u is an eigenvector of AB, we have $\lambda ABu = \lambda\mu u$, and finally:

$$A(w + Bu + BAw) = (\lambda - 1 + \lambda\mu)u = \lambda(\lambda - 1)u = \lambda Aw$$

Which recovers exactly (11): λ is indeed an eigenvalue of $J(A, B)$. □

A.1. Momentum ResNets in the limit $\varepsilon \to 0$ – Proof of Proposition 2

Proof. We take $T = 1$ without loss of generality. We are going to use the implicit function theorem. Note that x_ε is solution of (6) if and only if $(x_\varepsilon, v_\varepsilon = \dot{x_\varepsilon})$ is solution of

$$\begin{cases} \dot{x} &= v, \quad x(0) = x_0 \\ \varepsilon\dot{v} &= f(x, \theta) - v, \quad v(0) = v_0. \end{cases}$$

Consider for $u = (x, v) \in (x_0, v_0) + C_0^\infty([0, 1], \mathbb{R}^d)^2$

$$\Psi(u, \varepsilon) = \left(x_0 - x + \int_0^t v, \int_0^t (f(x, \theta) - v) - \varepsilon v + \varepsilon v_0\right),$$

so that x_ε is solution of (6) if and only if $u_\varepsilon = (x_\varepsilon, v_\varepsilon = \dot{x_\varepsilon})$ satisfies $\Psi(u_\varepsilon, \varepsilon) = 0$. Let $u^* = (x^*, \dot{x^*})$. One has $\Psi(u^*, 0) = 0$. Ψ is differentiable everywhere, and at $(u^*, 0)$ we have

$$\partial_u\Psi(u^*, 0)(x, v) = \left((\int_0^t v) - x, \int_0^t (\partial_x f(x*, \theta).x - v)\right).$$

$\partial_u\Psi(u^*, 0)$ is continuous, and it is invertible with continuous inverse because it is linear and continuous, and because $\partial_u\Psi(u^*, 0)(x, v) = 0$ if and only if

$$\begin{cases} \forall t \in [0, 1], x(t) = \int_0^t v \\ \forall t \in [0, 1], v(t) = \partial_x f(x^*(t), \theta(t)).x(t) \end{cases}$$

which is equivalent to

$$\begin{cases} \dot{x} = \partial f(x^*, \theta).x \\ x(0) = 0 \\ v = \dot{x}, \end{cases}$$

which is equivalent, because this equation is linear to $(x, v) = (0, 0)$. Using the implicit function theorem, we know that there exists two neighbourhoods $U \subset \mathbb{R}$ and $V \subset (x_0, v_0) + C_0^\infty([0, 1], \mathbb{R}^d)^2$ of 0 and u^* and a continuous function $\zeta : U \to V$ such that

$$\forall (u, \varepsilon) \in U \times V, \Psi(u, \varepsilon) = 0 \Leftrightarrow u = \zeta(\varepsilon)$$

This in particular ensures that x_ε converges uniformly to x^* as ε goes to 0 □

A.2. Momentum ResNets are more general than neural ODEs – Proof of Proposition 3

Proof. If x satisfies (5) we get by derivation that

$$\ddot{x} = \partial_x f(x,\theta) f(x,\theta) + \partial_\theta f(x,\theta)\dot{\theta}$$

Then, if we define $\hat{f}(x,\theta) = \varepsilon[\partial_x f(x,\theta) f(x,\theta) + \partial_\theta f(x,\theta)\dot{\theta}] + f(x,\theta)$, we get that x is also solution of the second-order model $\varepsilon\ddot{x} + \dot{x} = \hat{f}(x,\theta)$ with $(x(0), \dot{x}(0)) = (x_0, f(x_0, \theta_0))$. □

A.3. Solution of (7) – Proof of Proposition 4

(7) writes

$$\begin{cases} \dot{x} &= v, \quad x(0) = x_0 \\ \dot{v} &= \frac{\theta x - v}{\varepsilon}, \quad v(0) = 0. \end{cases}$$

For which the solution at time t writes

$$\begin{pmatrix} x(t) \\ v(t) \end{pmatrix} = \exp\begin{pmatrix} 0 & \mathrm{Id}_d t \\ \frac{\theta t}{\varepsilon} & -\frac{\mathrm{Id}_d t}{\varepsilon} \end{pmatrix} \cdot \begin{pmatrix} x_0 \\ 0 \end{pmatrix}.$$

The calculation of this exponential gives

$$x(t) = e^{-\frac{t}{2\varepsilon}} \left(\sum_{n=0}^{+\infty} \frac{1}{(2n)!} (\frac{\theta}{\varepsilon} + \frac{\mathrm{Id}_d}{4\varepsilon^2})^n t^{2n} + \sum_{n=0}^{+\infty} \frac{1}{2\varepsilon(2n+1)!} (\frac{\theta}{\varepsilon} + \frac{\mathrm{Id}_d}{4\varepsilon^2})^n t^{2n+1} \right) x_0.$$

Note that it can be checked directly that this expression satisfies (7) by derivations. At time 1 this effectively gives $x(1) = \Psi_\varepsilon(\theta) x_0$.

A.4. Representable mappings for a Momentum ResNet with linear residual functions – Proof of Theorem 1

In what follows, we denote by f_ε the function of matrices defined by

$$f_\varepsilon(\theta) = \Psi_\varepsilon(\varepsilon\theta - \frac{I}{4\varepsilon}) = e^{-\frac{1}{2\varepsilon}} \sum_{n=0}^{+\infty} \left(\frac{1}{(2n)!} + \frac{1}{2\varepsilon(2n+1)!} \right) \theta^n.$$

Because $\Psi_\varepsilon(\mathbb{R}^{d\times d}) = f_\varepsilon(\mathbb{R}^{d\times d})$, we choose to work on f_ε.

We first need to prove that f_ε is surjective on $\mathbb{C}$.

A.4.1. Surjectivity on $\mathbb{C}$ of f_ε

Lemma 1 (Surjectivity of f_ε)**.** *For $\varepsilon > 0$, f_ε is surjective on $\mathbb{C}$.*

Proof. Consider

$$\begin{aligned} F_\varepsilon \colon \mathbb{C} &\longrightarrow \mathbb{C} \\ z &\longmapsto e^{-\frac{1}{2\varepsilon}} (\cosh(z) + \frac{1}{2\varepsilon z}\sinh(z)). \end{aligned}$$

For $z \in \mathbb{C}$, we have $f_\varepsilon(z^2) = F_\varepsilon(z)$, and because $z \mapsto z^2$ is surjective on $\mathbb{C}$, it is sufficient to prove that F_ε is surjective on $\mathbb{C}$. Suppose by contradiction that there exists $w \in \mathbb{C}$ such that $\forall z \in \mathbb{C}$, $\exp(\frac{1}{2\varepsilon}) F_\varepsilon(z) \neq w$. Then $\exp(\frac{1}{2\varepsilon}) F_\varepsilon - w$ is an entire function (Levin, 1996) of order 1 with no zeros. Using Hadamard's factorization theorem (Conway, 2012), this implies that there exists $a, b \in \mathbb{C}$ such that $\forall z \in \mathbb{C}$,

$$\cosh(z) + \frac{\sinh(z)}{2\varepsilon z} - w = \exp(az + b).$$

However, since F_ε is an even function one has that $\forall z \in \mathbb{C}$

$$\exp(az+b) = \exp(-az+b)$$

so that $\forall z \in \mathbb{C}$, $2az \in 2i\pi\mathbb{Z}$. Necessarily, $a = 0$, which is absurd because F_ε is not constant.

□

We first prove Theorem 1 in the diagonalizable case.

A.4.2. Theorem 1 in the diagonalizable case

Proof. **Necessity** Suppose that D can be represented by a second-order model (7). This means that there exists a real matrix X such that $D = f_\varepsilon(X)$ with X real and

$$f_\varepsilon(X) = e^{-\frac{1}{2\varepsilon}}(\sum_{n=0}^{+\infty} a_n^\varepsilon X^n)$$

with

$$a_n^\varepsilon = \frac{1}{(2n)!} + \frac{1}{2\varepsilon(2n+1)!}.$$

X commutes with D so that there exists $P \in \mathrm{GL}_d(\mathbb{C})$ such that $P^{-1}DP$ is diagonal and $P^{-1}XP$ is triangular. Because $f_\varepsilon(P^{-1}XP) = P^{-1}DP$, we have that $\forall \lambda \in \mathrm{Sp}(D)$, there exists $z \in \mathrm{Sp}(X)$ such that $\lambda = f_\varepsilon(z)$. Because $\lambda < \lambda_\varepsilon$, necessarily, $z \in \mathbb{C} - \mathbb{R}$. In addition, $\lambda = f_\varepsilon(z) = \bar{\lambda} = f_\varepsilon(\bar{z})$. Because X is real, each $z \in \mathrm{Sp}(X)$ must be associated with $\bar{z}$ in $P^{-1}XP$. Thus, λ appears in pairs in $P^{-1}DP$.

Sufficiency Now, suppose that $\forall \lambda \in \mathrm{Sp}(D)$ with $\lambda < \lambda_\varepsilon$, λ is of even multiplicity order. We are going to exhibit a X real such that $D = f_\varepsilon(X)$. Thanks to Lemma 1, we have that f_ε is surjective. Let $\lambda \in \mathrm{Sp}(D)$.

- If $\lambda \in \mathbb{R}$ and $\lambda < \lambda_\varepsilon$ or $\lambda \in \mathbb{C} - \mathbb{R}$ then there exists $z \in \mathbb{C} - \mathbb{R}$ by Lemma 1 such that $\lambda = f_\varepsilon(z)$.
- If $\lambda \in \mathbb{R}$ and $\lambda \geq \lambda_\varepsilon$, then because f_ε is continuous and goes to infinity when $x \in \mathbb{R}$ goes to infinity, there exists $x \in \mathbb{R}$ such that $\lambda = f_\varepsilon(x)$.

In addition, there exist $(\alpha_1, ..., \alpha_k) \in (\mathbb{C} - \mathbb{R})^k \cup [-\infty, \lambda_\varepsilon[^k$, $(\beta_1, ..., \beta_p) \in [\lambda_\varepsilon, +\infty]^p$ such that

$$D = Q^{-1}\Delta Q,$$

with $Q \in \mathrm{GL}_d(\mathbb{R})$, and

$$\Delta = \begin{pmatrix} P_1^{-1}D_{\alpha_1}P_1 & 0_2 & \cdots & \cdots & \cdots & 0_2 \\ 0_2 & \ddots & \cdots & \cdots & \cdots & 0_2 \\ \vdots & \vdots & P_k^{-1}D_{\alpha_k}P_k & 0_2 & \cdots & 0_2 \\ 0 & \cdots & \cdots & \beta_1 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \beta_p \end{pmatrix} \in \mathbb{R}^{d\times d}$$

with $P_j \in GL_2(\mathbb{C})$ and $D_{\alpha_j} = \begin{pmatrix} \alpha_j & 0 \\ 0 & \bar{\alpha_j} \end{pmatrix}$.

Let $(z_1, ..., z_k) \in (\mathbb{C} - \mathbb{R})^k$ and $(x_1, ..., x_p) \in \mathbb{R}^p$ be such that $f_\varepsilon(z_j) = \alpha_j$ and $f_\varepsilon(x_j) = \beta_j$. For $1 \leq j \leq k$, one has $P_j^{-1}D_{z_j}P_j \in \mathbb{R}^{2\times 2}$. Indeed, writing $\alpha_j = a_j + ib_j$ with $a_j, b_j \in \mathbb{R}$, the fact that $P_j^{-1}D_{\alpha_j}P_j \in \mathbb{R}^{2\times 2}$ implies that

$i\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix} \in i\mathbb{R}^{2\times 2}$. Writing $z_j = u_j + iv_j$ with $u_j, v_j \in \mathbb{R}$, we get that $P_j^{-1}D_{z_j}P_j \in \mathbb{R}^{2\times 2}$. Then

$$X = Q\begin{pmatrix} P_1^{-1}D_{z_1}P_1 & 0_2 & \cdots & \cdots & \cdots & 0_2 \\ 0_2 & \ddots & \cdots & \cdots & \cdots & 0_2 \\ \vdots & \vdots & P_k^{-1}D_{z_k}P_k & 0_2 & \cdots & 0_2 \\ 0 & \cdots & \cdots & x_1 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & x_p \end{pmatrix} Q^{-1} \in \mathbb{R}^{d\times d}$$

is such that $f_\varepsilon(X) = D$, and D is represented by a second-order model (7). $\square$

We now state and demonstrate the general version of Theorem 1.

First, we need to demonstrate properties of the complex derivatives of the entire function f_ε.

A.4.3. THE ENTIRE FUNCTION f_ε HAS A DERIVATIVE WITH NO-ZEROS ON $\mathbb{C} - \mathbb{R}$.

Lemma 2 (On the zeros of f'_ε)**.** *$\forall z \in \mathbb{C} - \mathbb{R}$ we have $f'_\varepsilon(z) \neq 0$.*

Proof. One has

$$G_\varepsilon(z) = e^{-\frac{1}{2\varepsilon}}(\cos(z) + \frac{1}{2\varepsilon z}\sin(z)) = f_\varepsilon(-z^2)$$

so that $G'_\varepsilon(z) = -2zf'_\varepsilon(-z^2)$ and it is sufficient to prove that the zeros of G'_ε are all real.

We first show that G_ε belongs to the Laguerre-Pólya class (Craven & Csordas, 2002). The Laguerre-Pólya class is the set of entire functions that are the uniform limits on compact sets of $\mathbb{C}$ of polynomials with only real zeros. To show that G_ε belongs to the Laguerre-Pólya class, it is sufficient to show (Dryanov & Rahman, 1999, p. 22) that:

- The zeros of G_ε are all real.
- If $(z_n)_{n\in\mathbb{N}}$ denotes the sequence of real zeros of G_ε, one has $\sum \frac{1}{|z_n|^2} < \infty$.
- G_ε is of order 1.

First, the zeros of G_ε are all real, as demonstrated in Runckel (1969). Second, if $(z_n)_{n\in\mathbb{N}}$ denotes the sequence of real zeros of G_ε, one has $z_n \sim n\pi + \frac{\pi}{2}$ as $n \to \infty$, so that $\sum \frac{1}{|z_n|^2} < \infty$. Third, G_ε is of order 1. Thus, we have that G_ε is indeed in the Laguerre-Pólya class.

This class being stable under differentiation, we get that G'_ε also belongs to the Laguerre-Pólya class. So that the roots of G'_ε are all real, and hence those of f_ε as well.

$\square$

A.4.4. THEOREM 1 IN THE GENERAL CASE

When $\varepsilon = 0$, we have in the general case the following from Culver (1966):

Let $A \in \mathbb{R}^{d\times d}$. Then A can be represented by a first-order model (8) **if and only if** A is not singular and each Jordan block of A corresponding to an eigen value $\lambda < 0$ occurs an even number of time.

We now state and demonstrate the equivalent of this result for second order models (7).

Theorem 2 (Representable mappings for a Momentum ResNet with linear residual functions – General case)**.** *Let $A \in \mathbb{R}^{d\times d}$.*

If A can be represented by a second-order model (7), then each Jordan block of A corresponding to an eigen value $\lambda < \lambda_\varepsilon$ occurs an even number of time.

Reciprocally, if each Jordan block of A corresponding to an eigen value $\lambda \leq \lambda_\varepsilon$ occurs an even number of time, then A can be represented by a second-order model.

Proof. We refer to the arguments from Culver (1966) and use results from Gantmacher (1959) for the proof.

Suppose that A can be represented by a second-order model (7). This means that there exists $X \in \mathbb{R}^{d\times d}$ such that $A = f_\varepsilon(X)$. The fact that X is real implies that its Jordan blocks are:

$$(\lambda - z_k)^{a_k},\ z_k \in \mathbb{R}$$
$$(\lambda - z_k)^{b_k} \text{ and } (\lambda - \bar{z_k})^{b_k},\ z_k \in \mathbb{C} - \mathbb{R}.$$

Let $\lambda_k = f_\varepsilon(z_k)$ be an eigenvalue of A such that $\lambda_k < \lambda_\varepsilon$. Necessarily, $z_k \in \mathbb{C} - \mathbb{R}$, and $f'_\varepsilon(z_k) \neq 0$ thanks to Lemma 2. We then use Theroem 9 from Gantmacher (1959) (p. 158) to get that the Jordan blocks of A corresponding to λ_k are

$$(\lambda - f_\varepsilon(z_k))^{b_k} \text{ and } (\lambda - f_\varepsilon(\bar{z_k}))^{b_k}.$$

Since $f_\varepsilon(\bar{z_k}) = f_\varepsilon(z_k) = \lambda_k$, we can conclude that the Jordan blocks of A corresponding $\lambda_k < \lambda_\varepsilon$ occur an even number of time.

Now, suppose that each Jordan block of A corresponding to an eigen value $\lambda \leq \lambda_\varepsilon$ occurs an even number of times. Let λ_k be an eigenvalue of A.

- If $\lambda_k \in \mathbb{C} - \mathbb{R}$ we can write, because f_ε is surjective (proved in Lemma 1), $\lambda_k = f_\varepsilon(z_k)$ with $z_k \in \mathbb{C} - \mathbb{R}$. Necessarily, because A is real, the Jordan blocks of A corresponding to λ_k have to be associated to those corresponding to $\bar{\lambda_k}$. In addition, thanks to Lemma 2, $f'_\varepsilon(z_k) \neq 0$
- If $\lambda_k < \lambda_\varepsilon$, we can write, because f_ε is surjective, $\lambda_k = f_\varepsilon(z_k) = f_\varepsilon(\bar{z_k})$ with $z_k \in \mathbb{C} - \mathbb{R}$. In addition, $f'_\varepsilon(z_k) \neq 0$.
- If $\lambda_k > \lambda_\varepsilon$, then there exists $z_k \in \mathbb{R}$ such that $\lambda_k = f_\varepsilon(z_k)$ and $f'_\varepsilon(z_k) \neq 0$ because, if x_ε is such that $f_\varepsilon(x_\varepsilon) = \lambda_\varepsilon$, we have that $f'_\varepsilon > 0$ on $]x_\varepsilon, +\infty[$.
- If $\lambda_k = \lambda_\varepsilon$, there exists $z_k \in \mathbb{R}$ such that $\lambda_k = f_\varepsilon(z_k)$. Necessarily, $f'_\varepsilon(z_k) = 0$ but $f''_\varepsilon(z_k) \neq 0$.

This shows that the Jordan blocks of A are necessarily of the form

$$(\lambda - f_\varepsilon(z_k))^{b_k} \text{ and } (\lambda - f_\varepsilon(\bar{z_k}))^{b_k},\ z_k \in \mathbb{C} - \mathbb{R}$$
$$(\lambda - f_\varepsilon(z_k))^{a_k},\ z_k \in \mathbb{R},\ f_\varepsilon(z_k) \neq \lambda_\varepsilon$$
$$(\lambda - \lambda_\varepsilon)^{c_k} \text{ and } (\lambda - \lambda_\varepsilon)^{c_k}.$$

Let $Y \in \mathbb{R}^{d\times d}$ be such that its Jordan blocks are of the form

$$(\lambda - z_k)^{b_k} \text{ and } (\lambda - \bar{z_k})^{b_k},\ z_k \in \mathbb{C} - \mathbb{R},\ f'_\varepsilon(z_k) \neq 0$$
$$(\lambda - z_k)^{a_k},\ z_k \in \mathbb{R},\ f_\varepsilon(z_k) \neq \lambda_\varepsilon,\ f'_\varepsilon(z_k) \neq 0$$
$$(\lambda - z_k)^{2c_k},\ z_k \in \mathbb{R},\ f_\varepsilon(z_k) = \lambda_\varepsilon.$$

Then again by the use of Theorem 7 from Gantmacher (1959) (p. 158), because if $f_\varepsilon(z_k) = \lambda_\varepsilon$ with $z_k \in \mathbb{R}$, $f''_\varepsilon(z_k) \neq 0$, we have that $f_\varepsilon(Y)$ is similar to A. Thus A writes $A = P^{-1} f_\varepsilon(Y) P = f_\varepsilon(P^{-1}YP)$ with $P \in \mathrm{GL}_d(\mathbb{R})$. Then, $X = P^{-1}YP$ satisfies $X \in \mathbb{R}^{d\times d}$ and $f_\varepsilon(X) = A$. □

B. Additional theoretical results

B.1. On the convergence of the solution of a second order model when $\varepsilon \to \infty$

Proposition 5 (Convergence of the solution when $\varepsilon \to +\infty$)**.** *We let x^* (resp. x_ε) be the solution of $\ddot{x} = f(x, \theta)$ (resp. $\ddot{x} + \frac{1}{\varepsilon}\dot{x} = f(x, \theta)$) on $[0, T]$, with initial conditions $x^*(0) = x_\varepsilon(0) = x_0$ and $\dot{x}^*(0) = \dot{x}_\varepsilon(0) = v_0$. Then x_ε converges uniformly to x^* as $\varepsilon \to +\infty$.*

Proof. The equation $\ddot{x} + \frac{1}{\varepsilon}\dot{x} = f(x, \theta)$ with $x_\varepsilon(0) = x_0$, $\dot{x}_\varepsilon(0) = v_0$ writes in phase space (x, v)

$$\begin{cases} \dot{x} & = v, \quad x(0) = x_0 \\ \dot{v} & = f(x, \theta) - \frac{v}{\varepsilon}, \quad v(0) = v_0. \end{cases}$$

It then follows from the Cauchy-Lipschitz Theorem with parameters (Perko, 2013, Theorem 2, Chapter 2) that the solutions of this system are continuous in the parameter $\frac{1}{\varepsilon}$. That is x_ε converges uniformly to x^* as $\varepsilon \to +\infty$.

□

B.2. Universality of Momentum ResNets

Proposition 6 (When v_0 is free any mapping can be represented)**.** *Consider $h : \mathbb{R}^d \to \mathbb{R}^d$, and the ODE*

$$\ddot{x} + \dot{x} = 0$$
$$(x(0), \dot{x}(0)) = (x_0, \frac{h(x_0) - x_0}{1 - 1/e})$$

Then $\varphi_1(x_0) = h(x_0)$.

Proof. This is because the solution is $\varphi_t(x_0) = x_0 - v_0(e^{-t} - 1)$. □

B.3. Non-universality of Momentum ResNets when $v_0 = 0$

Proposition 7 (When $v_0 = 0$ there are mappings that cannot be learned if the equation is autonomous.)**.** *When $d = 1$, consider the autonomous ODE*

$$\begin{aligned} \varepsilon\ddot{x} + \dot{x} &= f(x) \\ (x(0), \dot{x}(0)) &= (x_0, 0) \end{aligned} \tag{13}$$

If there exists $x_0 \in \mathbb{R}+^*$ such that $h(x_0) \leq -x_0$ and $x_0 \leq h(-x_0)$ then h cannot be represented by (13).

This in particular proves that $x \mapsto \lambda x$ for $\lambda \leq -1$ cannot be represented by this ODE with initial conditions $(x_0, 0)$.

Proof. Consider such an x_0 and h. Since $\varphi_1(x_0) = h(x_0) \leq -x_0$, that $\varphi_0(x_0) = x_0$ and that $t \mapsto \varphi_t(x_0)$ is continuous, we know that there exists $t_0 \in [0, 1]$ such that $\varphi_{t_0}(x_0) = -x_0$. We denote $x(t) = \varphi_t(x_0)$, solution of

$$\ddot{x} + \frac{1}{\varepsilon}\dot{x} = f(x)$$

Since $d = 1$, one can write f as a derivative: $f = -E'$. The energy $E_m = \frac{1}{2}\dot{x}^2 + E$ satisfies:

$$\dot{E_m} = -\frac{1}{\varepsilon}\dot{x}^2$$

So that

$$E_m(t_0) - E_m(0) = -\frac{1}{\varepsilon}\int_0^{t_0} \dot{x}^2$$

In other words:

$$\frac{1}{2}v(t_0)^2 + \frac{1}{\varepsilon}\int_0^{t_0} \dot{x}^2 + E(-x_0) = E(x_0)$$

So that $E(-x_0) \leq E(x_0)$ We now apply the exact same argument to the solution starting at $x_1 = -x_0$. Since $x_0 \leq h(-x_0) = h(x_1)$ there exists $t_1 \in [0, 1]$ such that $\varphi_{t_1}(x_1) = x_0$. So that:

$$\frac{1}{2}v(t_1)^2 + \frac{1}{\varepsilon}\int_0^{t_1} \dot{x}^2 + E(x_0) = E(-x_0)$$

So that $E(x_0) \leq E(-x_0)$. We get that

$$E(x_0) = E(-x_0)$$

This implies that $\dot{x} = 0$ on $[0, t_0]$, so that the first solution is constant and $x_0 = -x_0$ which is absurd because $x_0 \in \mathbb{R}*$. □

B.4. When $v_0 = 0$ there are mappings that can be represented by a second-order model but not by a first-order one.

Proposition 8. *There exits f such that the solution of*

$$\ddot{x} + \frac{1}{\varepsilon}\dot{x} = f(x)$$

with initial condition $(x_0, 0)$ at time 1 is

$$x(1) = -x_0 \times \exp(-\frac{1}{2\varepsilon})$$

Proof. Consider the ODE

$$\ddot{x} + \frac{1}{\varepsilon}\dot{x} = (-\pi^2 - \frac{1}{4\varepsilon^2})x \tag{14}$$

with initial condition $(x_0, 0)$ The solution of this ODE is

$$x(t) = x_0 e^{-\frac{t}{2\varepsilon}}(\cos(\pi t) + \frac{1}{2\pi\varepsilon}\sin(\pi t))$$

which at time 1 gives:

$$x(1) = -x_0 e^{-\frac{1}{2\varepsilon}}$$

□

B.5. Orientation preservation of first-order ODEs

Proposition 9 (The homeomorphisms represented by (5) are orientation preserving.)**.** *If $K \subset \mathbb{R}^d$ is a compact set and $h : K \to \mathbb{R}^d$ is a homeomorphism represented by (5), then h is in the connected component of the identity function on K for the $\|.\|_\infty$ topology.*

We first prove the following:

Lemma 3. *Consider $K \subset \mathbb{R}^d$ a compact set. Suppose that $\forall x \in K$, $\Phi_t(x)$ is defined for all $t \in [0, 1]$. Then*

$$C = \{\Phi_t(x) \mid x \in K, t \in [0, 1]\}$$

is compact as well.

Proof. We consider $(\Phi_{t_n}(x_n))_{n\in\mathbb{N}}$ a sequence in C. Since $K \times [0, 1]$ is compact, we can extract sub sequences $(t_{\varphi(n)})_{n\in\mathbb{N}}$, $(x_{\varphi(n)})_{n\in\mathbb{N}}$ that converge respectively to t_0 and x_0. We denote them $(t_n)_{n\in\mathbb{N}}$ and $(x_n)_{n\in\mathbb{N}}$ again for simplicity of the notations. We have that:

$$\|\Phi_{t_n}(x_n) - \Phi_t(x)\| \leq \|\Phi_{t_n}(x_n) - \Phi_{t_n}(x)\| + \|\Phi_{t_n}(x) - \Phi_t(x)\|.$$

Thanks to Gronwall's lemma, we have

$$\|\Phi_{t_n}(x_n) - \Phi_{t_n}(x)\| \leq \|x_n - x\| \exp(kt_n),$$

where k is f's Lipschitz constant. So that $\|\Phi_{t_n}(x_n) - \Phi_{t_n}(x)\| \to 0$ as $n \to \infty$. In addition, it is obvious that $\|\Phi_{t_n}(x) - \Phi_t(x)\| \to 0$ as $n \to \infty$. We conclude that

$$\Phi_{t_n}(x_n) \to \Phi_t(x) \in C,$$

so that C is compact. □

Proof. Let's denote by H the set of homeomorphisms defined on K. The application

$$\Psi : [0, 1] \to H$$

defined by

$$\Psi(t) = \Phi_t$$

is continuous. Indeed, we have for any x_0 in $\mathbb{R}^d$ that

$$\|\Phi_{t+\varepsilon}(x_0) - \Phi_t(x_0)\| = \| \int_t^{t+\varepsilon} f(\Phi_s(x_0))ds\| \leq \varepsilon M_f,$$

where M_f bounds the continuous function f on C defined in lemma 3. Since M_f does not depend on x_0, we have that

$$\|\Phi_{t+\varepsilon} - \Phi_t\|_\infty \to 0$$

as $\varepsilon \to 0$, which proves that Ψ is continuous. Since $\Psi(0) = Id_K$, we get that $\forall t \in [0, 1]$, Φ_t is connected to Id_K. □

B.6. On the linear mappings represented by autonomous first order ODEs in dimension 1

Consider the autonomous ODE

$$\dot{x} = f(x), \tag{15}$$

Theorem 3 (Linearity)**.** *Suppose $d = 1$. If (15) represents a linear mapping $x \mapsto ax$ at time 1, we have that f is linear.*

Proof. If $a = 1$, consider some $x_0 \in \mathbb{R}$. Since $\Phi_1(x_0) = x_0 = \Phi_0(x_0)$, there exists, by Rolle's Theorem a $t_0 \in [0, 1]$ such that $\dot{x}(t_0) = 0$. Then $f(x(t_0)) = 0$. But since the constant solution $y = x(t_0)$ then solves $\dot{y} = f(y), y(0) = x(t_0)$, we get by the unicity of the solutions that $x(t_0) = y(0) = x(1) = y(1 - t_0) = x_0$. So that $f(x_0) = f(x(t_0)) = 0$. Since this is true for all x_0, we get that $f = 0$. We now consider the case where $a \neq 1$ and $a > 0$. Consider some $x_0 \in \mathbb{R}^*$. If $f(x_0) = 0$, then the solution constant to x_0 solves (3), and thus cannot reach ax_0 at time 1 because $a \neq 1$. Thus, $f(x_0) \neq 0$ if $x_0 \neq 0$. Second, if the trajectory starting at $x_0 \in \mathbb{R}^*$ crosses 0 and $f(0) = 0$, then by the same argument we know that $x_0 = 0$, which is absurd. So that, $\forall x_0 \in \mathbb{R}^*$, $\forall t \in [0, 1]$, $f(\Phi_t(x_0)) \neq 0$. We can thus rewrite (3) as

$$\frac{\dot{x}}{f(x)} = 1. \tag{16}$$

Consider F a primitive of $\frac{1}{f}$. Integrating (16), we get

$$F(ax_0) - F(x_0) = \int_0^1 F'(x(t))\dot{x}(t)\mathrm{d}t = 1.$$

In other words, $\forall x \in \mathbb{R}*$:

$$F(ax) = F(x) + 1.$$

We derive this equation and get:

$$af(x) = f(ax).$$

This proves that $f(0) = 0$. We now suppose that $a > 1$. We also have that

$$a^n f(\frac{x}{a^n}) = f(x).$$

But when $n \to \infty$, $f(\frac{x}{a^n}) = \frac{x}{a^n} f'(0) + o(\frac{1}{a^n})$ so that

$$f(x) = f'(0)x$$

and f is linear. The case $a < 1$ treats similarly by changing a^n to a^{-n}. □

B.7. There are mappings that are connected to the identity that cannot be represented by a first order autonomous ODE

In bigger dimension, we can exhibit a matrix in $\mathrm{GL}_d^+(\mathbb{R})$ (and hence connected to the identity) that cannot be represented by the autonomous ODE (15).

Proposition 10 (A non-representable matrix)**.** *Consider the matrix*

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -\lambda \end{pmatrix},$$

where $\lambda > 0$ and $\lambda \neq 1$. Then $A \in GL_2^+(\mathbb{R}) - GL_2(\mathbb{R})^2$ and A cannot be represented by (15).

Proof. The fact that $A \in GL_2^+(\mathbb{R}) - GL_2(\mathbb{R})^2$ is because A has two single negative eigenvalues, and because $\det(A) = \lambda > 0$. We consider the point $(0, 1)$. At time 1, it has to be in $(0, -\lambda)$. Because the trajectory are continuous, there exists $0 < t_0 < 1$ such that the trajectory is at $(x, 0)$ at time t_0, and thus at $(-x, 0)$ at time $t_0 + 1$, and again at $(x, 0)$ at time $t_0 + 2$. However, the particle is at $(0, \lambda^2)$ at time 2. All of this is true because the equation is autonomous. Now, we showed that trajectories starting at $(0, 1)$ and $(0, \lambda^2)$ would intersect at time t_0 at $(x, 0)$, which is absurd. Figure 11 illustrates the paradox. $\square$

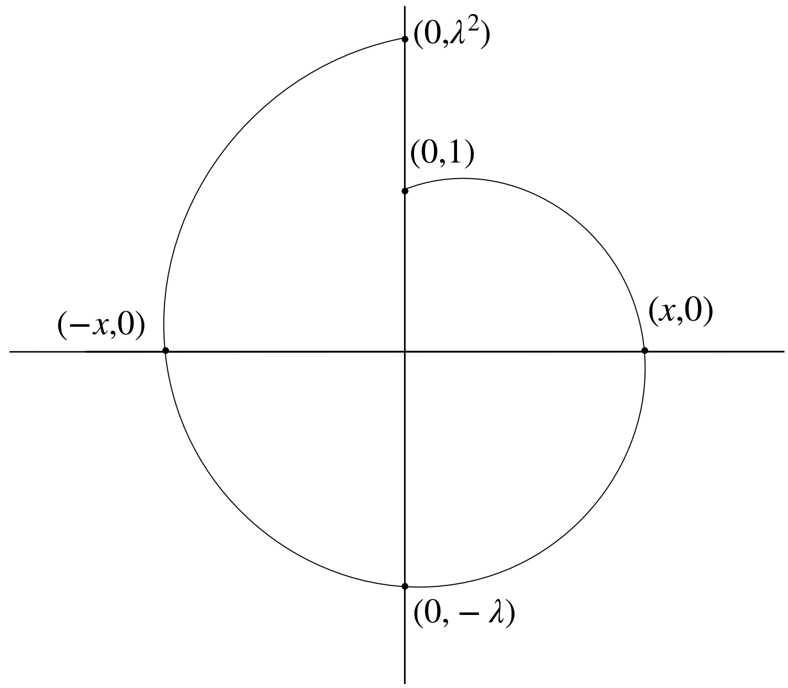

Figure 11. Illustration of Proposition 10. The points starting at $(0, 1)$ and $(0, \lambda^2)$ are distinct but their associated trajectories would have to intersect in $(x, 0)$, which is impossible.

C. Exact multiplication

Algorithm 1 Exactly reversible multiplication by a ratio, from Maclaurin et al. (2015)

1: **Input:** Information buffer i, value c, ratio n/d
2: $i = i \times d$
3: $i = i + (c \mod d)$
4: $c = c \div d$
5: $c = c \times n$
6: $c = c + (i \mod n)$
7: $i = i \div n$
8: **return** updated buffer i, updated value c

We here present the algorithm from Maclaurin et al. (2015). In their paper, the authors represent γ as a rational number, $\gamma = \frac{n}{d} \in \mathbb{Q}$. The information is lost during the integer division of v_n by d in (2). The store this information, it is sufficient to store the remainder r of this integer division. r is stored in an "information buffer" i. To update i, one has to left-shift the bits in i by multiplying it by n before adding r. The entire procedure is illustrated in Algorithm 1 from Maclaurin et al. (2015).

D. Implementation details

D.1. Creating a Momentum ResNet with a MLP

```
import torch
import torch.nn as nn
from momentumnet import MomentumNet

function = nn.Sequential(nn.Linear(2, 16), nn.Tanh(), nn.Linear(16, 2))

mom_net = MomentumNet([function, ], gamma=0.9, n_iters=15)
```

D.2. Drop-in replacement

To illustrate the fact that Momentum ResNets are a drop-in replacement for ResNets, we implement a function

```
transform(model, pretrained=False, gamma=0.9)
```

This function takes a torchvision model ResNet and returns its Momentum ResNet counterpart. The Momentum ResNet can be initialized with weights of a pretrained ResNet on ImageNet, and hence, as we show in this paper, quickly achieves great performances on new datasets.

This method can be used as follow:

```
mresnet152 = transform(resnet152(pretrained=True), pretrained=True)
```

and is made available in the code.

E. Experiment details

In all our image experiments, we use Nvidia Tesla V100 GPUs.

For our experiments on CIFAR-10 and 100, we used a batch-size of 128 and we employed SGD with a momentum of 0.9. The training was done over 220 epochs. The initial learning rate was 0.01 and was decayed by a factor 10 at epoch 180. A constant weight decay was set to 5×10^{-4}. Standard inputs preprocessing as proposed in Pytorch (Paszke et al., 2017) was performed.

For our experiments on ImageNet, we used a batch-size of 256 and we employed SGD with a momentum of 0.9. The training was done over 100 epochs. The initial learning rate was 0.1 and was decayed by a factor 10 every 30 epochs. A constant weight decay was set to 10^{-4}. Standard inputs preprocessing as proposed in Pytorch (Paszke et al., 2017) was performed: normalization, random croping of size 224×224 pixels, random horizontal flip.

For our experiments in the continuous framework, we adapted the code made available by Chen et al. (2018) to work on the CIFAR-10 data set and to solve second order ODEs. We used a batch-size of 128, and used SGD with a momentum of 0.9. The initial learning rate was set to 0.1 and reduced by a factor 10 at iteration 60. The training was done over 120 epochs.

For the learning to optimize experiment, we generate a random Gaussian matrix D of size 16×32. The columns are then normalized to unit variance. We train the networks by stochastic gradient descent for 10000 iterations, with a batch-size of 1000 and a learning rate of 0.001. The samples y_q are generated as follows: we first sample a random Gaussian vector $\tilde{y}_q$, and then we use $y_q = \frac{\tilde{y}_q}{\|D^\top \tilde{y}_q\|_\infty}$, which ensures that every sample verify $\|D^\top y_q\|_\infty = 1$. This way, we know that the solution x^* is zero if and only if $\lambda \geq 1$. The regularization is set to $\lambda = 0.1$.

F. Backpropagation for Momentum ResNets

In order to backpropagate the gradient of some loss in a Momentum ResNet, we need to formulate an explicit version of (2). Indeed, (2) writes explicitly

$$\begin{aligned} v_{n+1} &= \gamma v_n + (1-\gamma) f(x_n, \theta_n) \\ x_{n+1} &= x_n + (\gamma v_n + (1-\gamma) f(x_n, \theta_n)). \end{aligned} \tag{17}$$

Writing $z = (x, v)$, the backpropagation for Momentum ResNets then writes, for some loss L

$$\nabla_{z_{k-1}} L = \begin{bmatrix} I + (1-\gamma)\partial_x f(x_{k-1}, \theta_{k-1}) & \gamma I \\ (1-\gamma)\partial_x f(x_{k-1}, \theta_{k-1}) & \gamma I \end{bmatrix}^T \nabla_{z_k} L$$

$$\nabla_{\theta_{k-1}} L = (1-\gamma) \begin{bmatrix} \partial_\theta f(x_{k-1}, \theta_{k-1}) \\ \partial_\theta f(x_{k-1}, \theta_{k-1}) \end{bmatrix}^T \nabla_{z_k} L.$$

We implement these formula to obtain a custom Jacobian-vector product in Pytorch.

G. Additional figures

G.1. Learning curves on CIFAR-10

We here show the learning curves when training a ResNet-101 and a Momentum ResNet-101 on CIFAR-10.

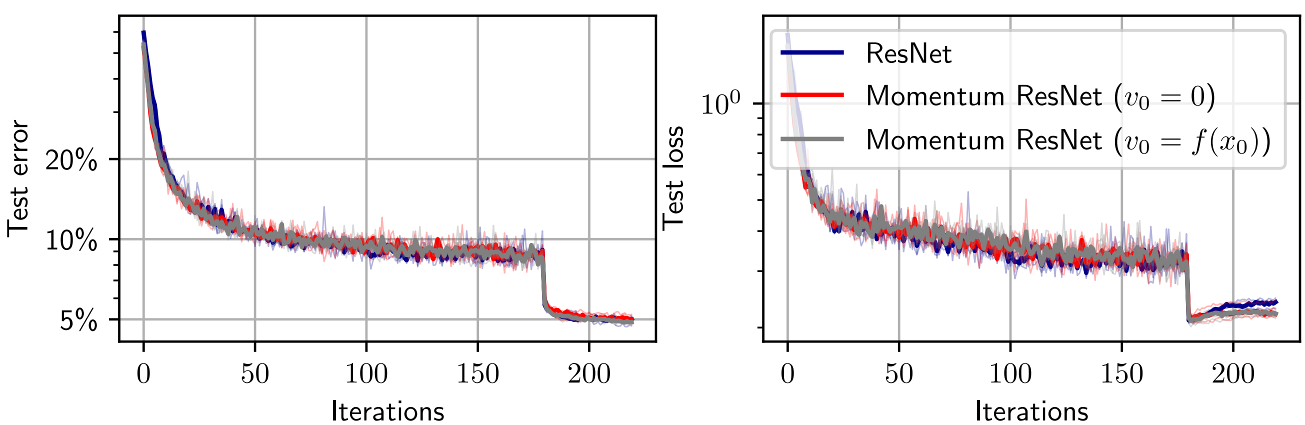

Figure 12. Test error and test loss as a function of depth on CIFAR-10 with a ResNet-101 and two Momentum ResNets-101.