NODE 2: Approximation Capabilities of Neural ODEs and Invertible Residual Networks

CS 598: Deep Generative and Dynamical Models

Instructor: Arindam Banerjee

November 4, 2021

Invertible Residual Networks

- ResNet: $x_{t+1} = x_t + f_t(x_t, \theta_t)$
- Usually the same function form in every layer, use $f_\Theta(x_t, t)$
- i-ResNets and Residual Flows
  - f_Θ is Lipschitz as a function of x for fixed t
  - Lipschitz constant is less than 1, i.e., $\text{Lip}(f_\Theta) < 1$
- Constraint is sufficient to ensure invertibility of the ResNet
  - $x_t \mapsto x_{t+1}$ is a one-to-one mapping
- For a mapping $x \mapsto 2x$, one layer i-ResNet is insufficient
  - Need two layers $x \mapsto x + (\sqrt{2} - 1)x$, due to Lipshitz constraint
- In general, i-ResNets have Lipschitz constant $\text{Lip}(I + f_\Theta) < 2$
- Can we approximate any invertible mapping with Lipschitz constant K using i-ResNets?

Neural ODEs

- Quick recap:

$$\frac{dx_t}{dt} = f_\Theta(x_t, t)$$

$$x_T = \mathrm{x}_0 + \int_0^T f_\Theta(x_t, t)dt$$

- p-dimensional ODE-Net (neural ODE)
  - Input / output must be p-dimensional
  - Inner layers can potentially use higher dimensions
- ODE-Nets are invertible by design, i.e., reverse limits of integral
- Adjoint sensitivity method based reverse time integration helps gradient descent
- Neural ODEs on its own are not universal approximators

Background: Flows

- A mapping $h : \mathcal{X} \mapsto \mathcal{X}$ is a *homeomorphism* if h is one-to-one, onto, and both h and h^{-1} are continuous
- A *topological transformation group* or *flow* is a triple $(\mathcal{X}, \mathbb{G}, \Phi)$
  - $\mathbb{G}$ is an additive group with neural element 0
  - $\Phi : \mathcal{X} \times \mathbb{G} \mapsto \mathcal{X}$, $\Phi(x, 0) = x, \Phi(\Phi(x, s), t) = \Phi(x, s + t)$
  - Φ is continuous w.r.t. the first argument
- We consider $\mathcal{X} \subset \mathbb{R}^p$, so p-homeomorphisms, p-flows
- Given a flow, an *orbit* or *trajectory* associated with $x \in \mathcal{X}$ is a subspace $G(x) = \{\Phi(x, t) : t \in \mathbb{G}\}$
- Given $x, y \in \mathcal{X}$, either $G(x) = G(y)$ or $G(x) \cap G(y) = \emptyset$

Background: Discrete and Continuous Flows

- A *discrete flow* is defined by setting $\mathbb{G} = \mathbb{Z}$
- For arbitrary homeomorphism h, the corresponding discrete flow is a discrete dynamical system:
$\phi_0(x) = x$, $\phi_{t+1} = h(\phi_t(x))$, $\phi_{t-1}(x) = h^{-1}(\phi_t(x))$
- Setting $f(x) = h(x) - x$ gives a ResNet: $x_{t+1} = x_t + f(x_t)$
- A *continuous flow* is defined by setting $\mathbb{G} = \mathbb{R}$
- Neural ODEs are continuous flows with continuous $d\Phi/dt$
- Continuous flows orbits are continuous
  - Implications on what homeomorphisms ϕ_t can result from a flow

Background: Continuous Flows and ODEs

- For a continuous flow $(\mathcal{X}, \mathbb{R}, \Phi)$, consider $V(x) = d\Phi(x,t)/dt|_{t=0}$
- ODE $dx/dt = V(x)$ corresponds to continuous flow $(\mathcal{X}, \mathbb{R}, \Phi)$
- Note: $\Phi(x_0, T) = x_0 + \int_0^T V(x_t)dt$, $\phi_{(S+T)}(x_0) = \phi_T(\phi_S(x_0))$
- $V(x)$ is continuous over $x \in \mathcal{X}$, constant over t: *autonomous* ODE

- Time dependent ODE can be converted to autonomous ODE
  - Rewrite $f_\Theta(x_t, t)$ by augmenting x by one dimension $x[p+1] = t$
  - We also have $dx[p+1]/dt = 1$ and $x_0[p+1] = 0$

Background: Flow Embedding Problem

- Given a p-flow, we can always find an ODE
- Given an ODE, under some conditions, we can find a flow, and the flow is necessarily a homeomorphism

- Given a homeomorphism h, does a p-flow such that $\phi_T = h$ exist?
- For a homeomorphism $h : \mathcal{X} \mapsto \mathcal{X}$, its *restricted embedding into a flow* is a flow $(\mathcal{X}, \mathbb{R}, \Phi)$ such that $h(x) = \Phi(x, T)$
  - Does not always exist (⇒ not universal approximator)
- An *unrestricted embedding into a flow* is a flow $(\mathcal{Y}, \mathbb{R}, \Phi)$ on $\mathcal{Y}$ of dimensionality higher than $\mathcal{X}$
- Involves a homeomorphism $g : \mathcal{X} \mapsto \mathcal{Z}$, where $\mathcal{Z} \subset \mathcal{Y}$ such that the flow on $\mathcal{Y}$ results in mappings on $\mathcal{Z}$ that are equivalent to h on $\mathcal{X}$, i.e., $g(h(x)) = \Phi(g(x), T)$

Approximating Homeomorphisms by Neural ODEs

- Assume $f_\Theta(x_t)$ is a universal approximator

  Theorem 1. *Let $\mathcal{X} = \mathbb{R}^p$, and let $\mathcal{Z} \subset \mathcal{X}$ be a set that partitions $\mathcal{X}$ into two or more disjoint, connected subsets C_i, for $i = [m]$; that is, $\mathcal{X} = \mathcal{Z} \cup (\bigcup_i C_i)$. Consider a mapping $h : \mathcal{X} \to \mathcal{X}$ that*

  - *is an identity transformation on $\mathcal{Z}$, that is, $\forall z \in \mathcal{Z}, h(z) = z$,*
  - *maps some $x \in C_i$ into $h(x) \in C_j$, for $i \neq j$.*

  Then, no p-ODE-Net can model h.

- Modeling with the same dimensionality is restricted
- E.g., mirror reflections cannot be handled

Neural ODEs with Extra Dimensions

- Let Neural ODEs operate in a higher dimensional space $q > p$
- For $h : \mathcal{X} \mapsto \mathcal{X}$, $q = 2p$ suffices
  - Uses an ODE which maps $[x, 0^{(p)}] \mapsto [h(x), 0^{(p)}]$

Theorem 2. *For any homeomorphism $h : \mathcal{X} \to \mathcal{X}$, $\mathcal{X} \subset \mathbb{R}^p$, if a feed-forward network for mapping $\delta(x) = h(x) - x$ can be constructed, then there exists a $2p$-ODE-Net $\phi_T : \mathbb{R}^{2p} \to \mathbb{R}^{2p}$ for $T = 1$ such that $\phi_T([x, 0^{(p)}]) = [h(x), 0^{(p)}]$ for any $x \in \mathcal{X}$.*

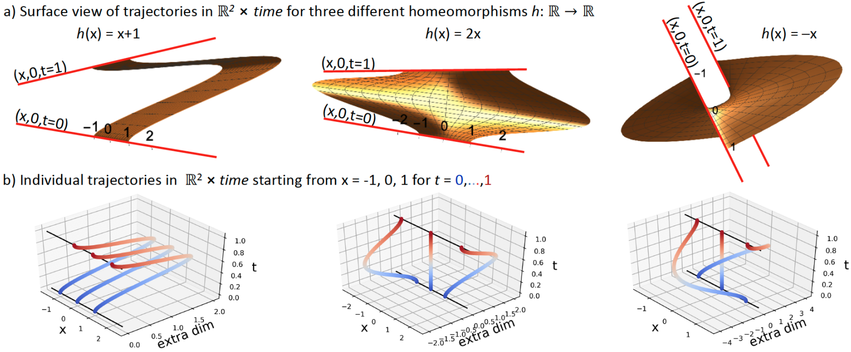

Figure 1. Trajectories in $\mathbb{R}^{2p}$ that embed an $\mathbb{R}^p \to \mathbb{R}^p$ homeomorphism, using $f(\tau) = (1 - \cos \pi\tau)/2$ and $g(\tau) = (1 - \cos 2\pi\tau)$. Three examples for $p = 1$ are shown, including the mapping $h(x) = -x$ that cannot be modeled by Neural ODE on $\mathbb{R}^p$, but can in $\mathbb{R}^{2p}$. In a), shading is used to represent 3D shapes, in b) trajectory color changes with time from blue ($t = 0$) to red ($t = 1$).

Recipe for Training Neural ODEs

- Simple approach to approximate continuous, invertible mapping h, and also get its inverse h^{-1}
- Pad the input $x \in \mathbb{R}^p$ with p zeroes
- Output is split into two parts
  - First p-dimensions use loss function w.r.t. $h(x)$
  - Remaining p-dimensions penalized deviation from 0
- Can approximate $(x, h(x))$ on the training set
- Generalization: Need not be invertible out-of-sample
  - Perhaps transport and Jacobian regularization can help

Approximating Homeomorphisms by i-ResNets

- Recall: Neural ODEs ($q = p$) cannot model reflections
- i-ResNets with same dimensions cannot model $f(x) = -x$

Theorem 3. *Let* $F_n(x) = (I + f_n) \circ (I + f_{n-1}) \circ \cdots \circ (I + f_1)(x)$ *be an* n*-layer i-ResNet, and let* $x_0 = x$ *and* $x_n = F_n(x_0)$. *If* $\mathrm{Lip}(f_i) < 1$ *for all* $i = 1, \ldots, n$, *then there are is no number* $n \geq 1$ *and no functions* f_i *for all* $i = 1, \ldots, n$ *such that* $x_n = -x_0$.

Approximating Homeomorphisms by i-ResNets

- Leads to more general conclusions in high dimensions

Corollary 4. *Let the straight line connecting* $x_t \in \mathbb{R}^p$ *to* $x_{t+1} = x_t + f(x_t) \in \mathbb{R}^p$ *be called an* extended path $x_t \to x_{t+1}$ *of a time-discrete topological transformation group on* $\mathcal{X} \in \mathbb{R}^p$*. In* p*-i-ResNet, for* $x_t \neq x'_t$*, extended paths* $x_t \to x_{t+1}$ *and* $x'_t \to x'_{t+1} = x'_t$ *do not intersect.*

Theorem 5. *Let* $\mathcal{X} = \mathbb{R}^p$*, and let* $\mathcal{Z} \subset \mathcal{X}$ *and* $h : \mathcal{X} \to \mathcal{X}$ *be the same as in Theorem 1. No* p*-i-ResNet can model* h*.*

- As before, using $q = 2p$ dimensions by zero-padding helps

Theorem 6. *For any homeomorphism* $h : \mathcal{X} \to \mathcal{X}$, $\mathcal{X} \subset \mathbb{R}^p$ *with* $\mathrm{Lip}(h) \leq k$, *if a feed-forward network for mapping* $\delta(x) = h(x) - x$ *can be constructed, then there exists a* $2p$*-i-ResNet* $\phi : \mathbb{R}^{2p} \to \mathbb{R}^{2p}$ *with* $\lfloor k + 4 \rfloor$ *residual layers such that* $\phi([x, 0^{(p)}]) = [h(x), 0^{(p)}]$ *for any* $x \in \mathcal{X}$.

Recipe for Training iResNets

- Order k layers may be needed to approximate homeomorphisms $h(x)$ with $\text{Lip}(h) \leq k$
- Only the first and last layer depends on $h(x)$ and need to be trained
- Middle layers are simple fixed linear layers
- Approach: As before, zero-padding to $2p$ dimensions
- Do not need differentiability in the time domain

Invertible Networks as Universal Approximators

- Neural ODE or i-ResNet followed by a simple linear layer
  - Universal approximator similar to wide networks
- Consider $f : \mathbb{R}^p \mapsto \mathbb{R}^r$, (x, y) such that $y = f(x)$
- The mapping $(x, 0) \mapsto (x, y)$ is a $(p + r)$-homoemorphism
  - Can be approximated by a $2(p + r)$ Neural ODE or i-ResNet
  - y can be extracted by a simple linear layer

Theorem 7. *Consider a neural network $F : \mathbb{R}^p \to \mathbb{R}^r$ that approximates function $f : \mathcal{X} \to \mathbb{R}^r$ that is Lebesgue integrable for each of the r output dimensions, with $\mathcal{X} \subset \mathbb{R}^p$ being a compact subset. For $q = p + r$, there exists a linear layer-capped q-ODE-Net that can perform the mapping F. If f is Lipschitz, there also is a linear layer-capped q-i-ResNet for F.*

Results: Approximations with increased Dimensionality

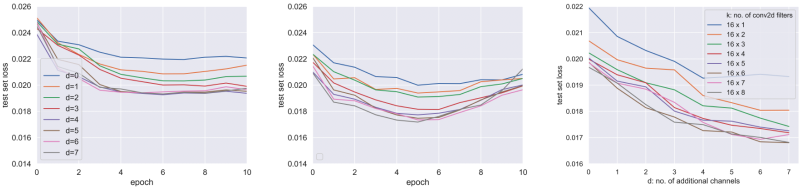

Figure 3. **Left and center:** test set cross-entropy loss, for increasing number d of null channels added to RGB images. For each d, the input images have dimensionality $32 \times 32 \times (3 + d)$. Left: ODE-Net with k=64 convolutional filters; center: k=128 filters. **Right:** Minimum of test set cross-entropy loss across all epochs as a function of d, the number of null channels added to input images, for ODE-Nets with different number of convolutional filters, k.

References

- H. Zhang, X. Gao, J. Unterman, T. Arodz. Approximation Capabilities of Neural Ordinary Differential Equations, ICML, 2020.
- R. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud. Neural ordinary differential equations, NeurIPS, 2018.
- J. Behrmann, W. Grathwohl, R. Chen, D. Duvenaud, J. Jacobsen, Invertible residual networks, ICML, 2019.
- E. Dupont, A. Doucet, Y. W. Teh. Augmented Neural ODEs NeurIPS, 2019.