Asymptotically exact inference in differentiable generative models

MATT GRAHAM (m.m.graham@ed.ac.uk) arxiv.org/abs/1605.07826 git.io/dgm AMOS STORKEY (a.storkey@ed.ac.uk)

Problem: Inference in generative models

Given: Probabilistic model of

$\mathbf{y}$: observed variables $\in \mathcal{Y} = \mathbb{R}^{N_y}$,

$\mathbf{z}$: latent variables $\in \mathcal{Z} = \mathbb{R}^{N_z}$,

where we can only generate $(\mathbf{y}, \mathbf{z})$ pairs.

Task: Estimate conditional expectations

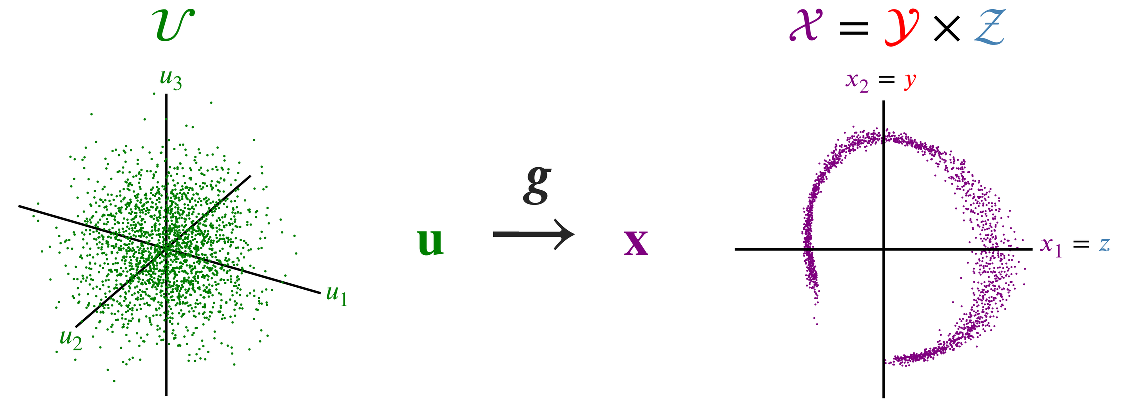
$$\mathbb{E}\left[f(\mathbf{z}) \mid \mathbf{y} = \mathbf{y}_{\text{obs}}\right] = \int_{\mathcal{Z}} f(\mathbf{z})\, \mathbb{p}_{\mathbf{z}\mid\mathbf{y}}\left[\mathbf{z} \mid \mathbf{y}_{\text{obs}}\right] \mathrm{d}\mathbf{z},$$

of latent variables given observations $\mathbf{y}_{\text{obs}}$.

Differentiable generative models

Generative model, a.k.a. implicit or simulator model :

A model where we can independently sample $(\mathbf{y}, \mathbf{z})$ but not necessarily evaluate $\mathbb{p}_{\mathbf{y},\mathbf{z}}$.



A generative model can be expressed in the form $\mathbf{u} \sim \rho$, $\mathbf{x} = \mathbf{g}(\mathbf{u})$ with

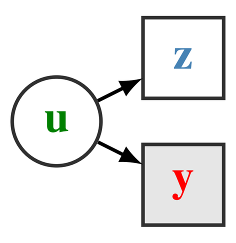
$\mathbf{u}$: random inputs $\in \mathcal{U}$, $\rho : \mathcal{U} \to \mathbb{R}^+$: input density,

$\mathbf{x} = (\mathbf{y}, \mathbf{z})$: generated outputs $\in \mathcal{X} = \mathcal{Y} \times \mathcal{Z}$, $\mathbf{g} : \mathcal{U} \to \mathcal{X}$: generator function.

Here we will consider models where $\mathcal{U} = \mathbb{R}^M$ and $\mathbf{g}$ is differentiable i.e. $\frac{\partial \mathbf{g}}{\partial \mathbf{u}}$ exists a.e.

Undirected and directed models

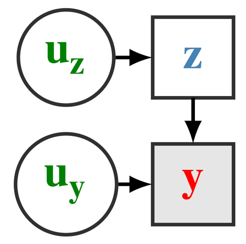
Undirected generative model



$\mathbf{z} = \mathbf{g}_{\mathbf{z}}(\mathbf{u})$

$\mathbf{y} = \mathbf{g}_{\mathbf{y}}(\mathbf{u})$

Directed generative model

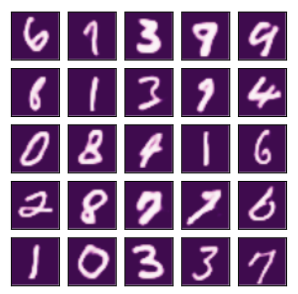


$\mathbf{z} = \mathbf{g}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z}})$

$\mathbf{y} = \mathbf{g}_{\mathbf{y}}(\mathbf{z}, \mathbf{u}_{\mathbf{y}})$

Examples of differentiable generative models

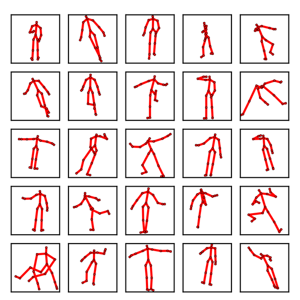
In all the examples below the input density is Gaussian $\rho(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{0}, \boldsymbol{I})$.



Decoder of a Gaussian Variational Autoencoder

$$\mathbf{x} = \boldsymbol{m}(\mathbf{u}_1) + \boldsymbol{s}(\mathbf{u}_1) \odot \mathbf{u}_2,$$

with $\boldsymbol{m}$ and $\boldsymbol{s}$ parameteric functions trained to match $\mathbb{p}_{\mathbf{x}}$ to the distribution of a dataset e.g. MNIST digit images.

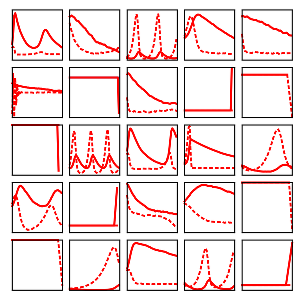


Pose generator, with learnt 3D pose parameter generators

joint angles $\mathbf{z}_a = \boldsymbol{f}_a(\mathbf{u}_a)$ bone lengths $\mathbf{z}_b = \boldsymbol{f}_b(\mathbf{u}_b)$ camera parameters $\mathbf{z}_c = \boldsymbol{f}_c(\mathbf{u}_c)$

and a pin-hole camera model to generate observed 2D projections

$$\underbrace{\mathbf{y}_j}_{\text{2D proj.}} = \underbrace{\boldsymbol{C}(\mathbf{z}_c)}_{\text{camera matrix}}\, \underbrace{\boldsymbol{r}_j(\mathbf{z}_a, \mathbf{z}_b)}_{\text{3D pos.}} + \underbrace{\sigma\, \mathbf{u}_j}_{\text{obs. noise}} \quad \forall j \in \{1 \dots J\}.$$



Continuous Lotka-Volterra predator-prey population model,

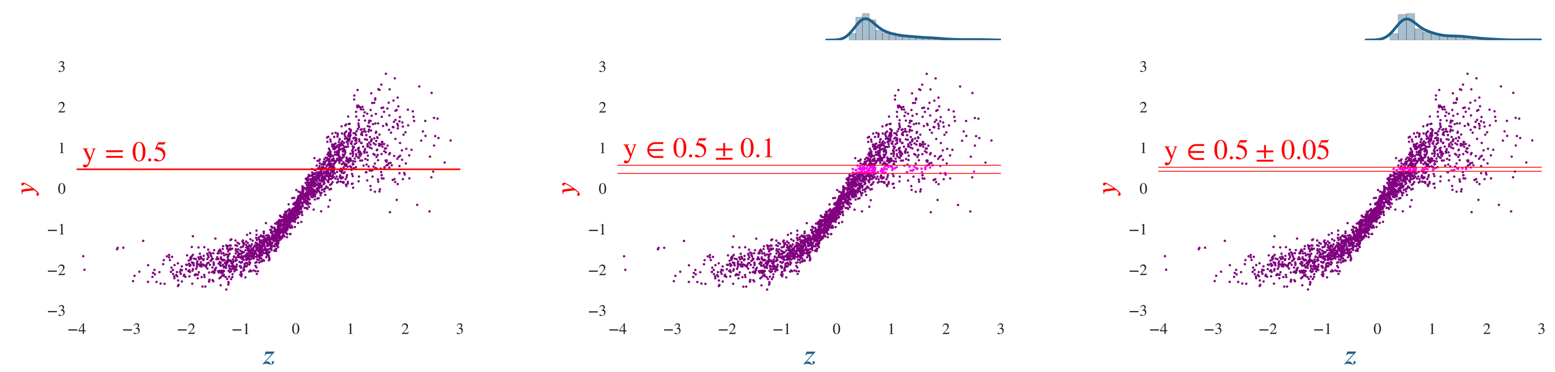
$$\mathrm{d}y_1 = (z_1 y_1 - z_2 y_1 y_2)\,\mathrm{d}t + \mathrm{d}n_1,$$
$$\mathrm{d}y_2 = (-z_3 y_2 + z_4 y_1 y_2)\,\mathrm{d}t + \mathrm{d}n_2.$$

Simulator for system can be expressed as a directed model

$\mathbf{z} = \mathbf{g}_{\mathbf{z}}(\mathbf{u}_{\mathbf{z}}) = \exp(\sigma \odot \mathbf{u}_{\mathbf{z}} - \boldsymbol{\mu})$: log-normal prior,

$\mathbf{y} = \mathbf{g}_{\mathbf{y}}(\mathbf{z}, \mathbf{u}_{\mathbf{y}})$: Euler-Maruyama integration of SDEs.

Approximate Bayesian Computation (ABC)



Family of approximate inference methods for generative models.
Typically applied to directed models where likelihood $\mathbb{p}_{\mathbf{y}\mid\mathbf{z}}$ is unavailable.

Key idea: true observations $\bar{\mathbf{y}}$ are decoupled from simulated observed values $\mathbf{y}$ by a *noise kernel* $\mathbb{p}_{\bar{\mathbf{y}}\mid\mathbf{y}}[\bar{\mathbf{y}} \mid \mathbf{y}] = k_\epsilon(\bar{\mathbf{y}}; \mathbf{y})$. Common choices for the kernel include

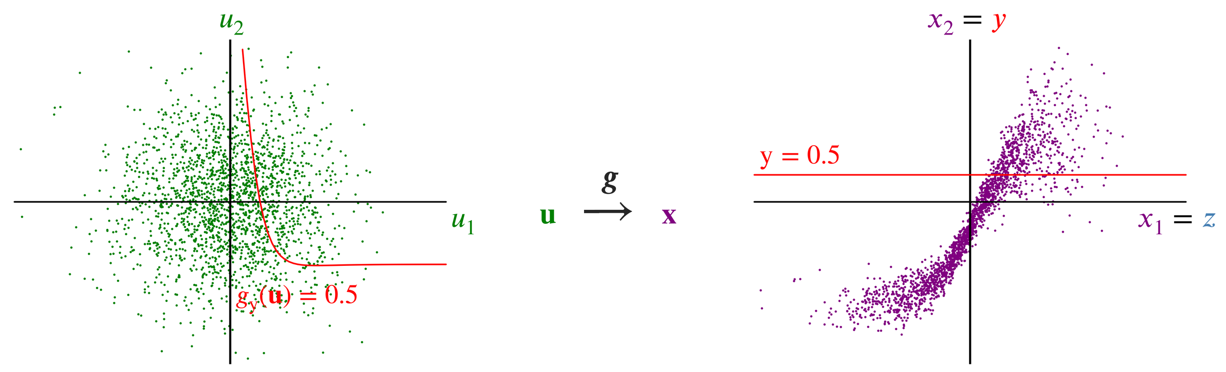
$$k_\epsilon(\bar{\mathbf{y}}; \mathbf{y}) \propto \mathbb{1}[|\bar{\mathbf{y}} - \mathbf{y}| < \epsilon] \text{ (uniform ball)}, \quad k_\epsilon(\bar{\mathbf{y}}; \mathbf{y}) = \mathcal{N}\left(\bar{\mathbf{y}}; \mathbf{y}, \epsilon^2 \boldsymbol{I}\right) \text{ (Gaussian)}.$$

Kernel can be used to express approximate conditional expectations

$$\mathbb{E}\left[f(\mathbf{z}) \mid \bar{\mathbf{y}} = \mathbf{y}_{\text{obs}}; \epsilon\right] = \frac{1}{C} \iint_{\mathcal{Y}\times\mathcal{Z}} f(\mathbf{z})\, k_\epsilon(\mathbf{y}_{\text{obs}}; \mathbf{y})\, \mathbb{p}_{\mathbf{y},\mathbf{z}}[\mathbf{y}, \mathbf{z}]\, \mathrm{d}\mathbf{y}\, \mathrm{d}\mathbf{z},$$

with asymptotic consistence in that $\lim_{\epsilon \to 0} \mathbb{E}\left[f(\mathbf{z}) \mid \bar{\mathbf{y}} = \mathbf{y}_{\text{obs}}; \epsilon\right] = \mathbb{E}\left[f(\mathbf{z}) \mid \mathbf{y} = \mathbf{y}_{\text{obs}}\right]$.

Inference in the generator input space



The *Law of the Unconscious Statistician* can be used to rewrite the ABC conditional expectation as an integral over the generator input space $\mathcal{U}$

$$\mathbb{E}\left[f(\mathbf{z}) \mid \bar{\mathbf{y}} = \mathbf{y}_{\text{obs}}; \epsilon\right] = \frac{1}{C} \int_{\mathcal{U}} f \circ \mathbf{g}_{\mathbf{z}}(\boldsymbol{u})\, k_\epsilon\left(\mathbf{y}_{\text{obs}}; \mathbf{g}_{\mathbf{y}}(\boldsymbol{u})\right) \rho(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}.$$

By taking the $\epsilon \to 0$ limit of the above and applying *Federer's Co-Area Formula* the exact conditional expectation can be expressed as an integral

$$\mathbb{E}\left[f(\mathbf{z}) \mid \mathbf{y} = \mathbf{y}_{\text{obs}}\right] = \frac{1}{C} \int_{\mathcal{M}_{\mathbf{y}_{\text{obs}}}} f \circ \mathbf{g}_{\mathbf{z}}(\boldsymbol{u}) \left|\frac{\partial \mathbf{g}_{\mathbf{y}}}{\partial \boldsymbol{u}} \frac{\partial \mathbf{g}_{\mathbf{y}}}{\partial \boldsymbol{u}}^{\mathsf{T}}\right|^{-\frac{1}{2}} \rho(\boldsymbol{u})\, \mathcal{H}_{\mathcal{M}_{\mathbf{y}_{\text{obs}}}}\{\mathrm{d}\boldsymbol{u}\}.$$

over an implicitly defined manifold $\mathcal{M}_{\mathbf{y}_{\text{obs}}} = \left\{\boldsymbol{u} \in \mathcal{U} : \mathbf{g}_{\mathbf{y}}(\boldsymbol{u}) = \mathbf{y}_{\text{obs}}\right\}$ embedded in $\mathcal{U}$.

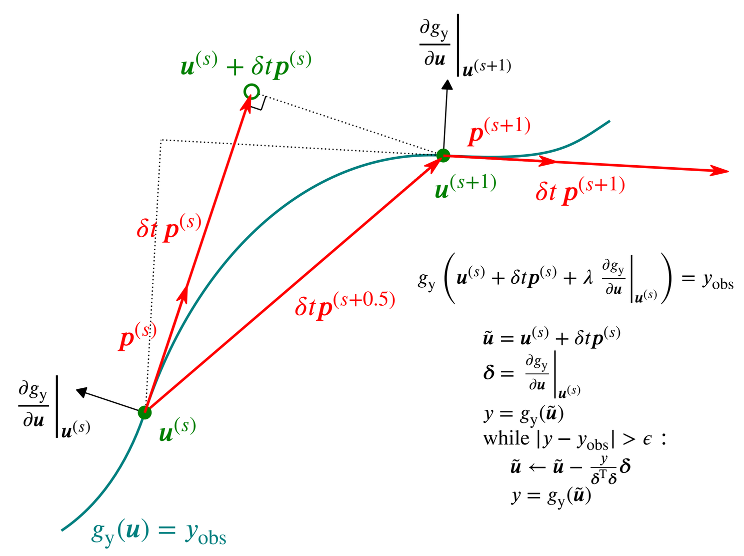
Constrained Hamiltonian Monte Carlo (CHMC)

Use simulated constrained Hamiltonian dynamic to propose moves on $\mathcal{M}_{\mathbf{y}_{\text{obs}}}$:

$$\pi(\boldsymbol{u}) = \frac{1}{C} \left|\frac{\partial \mathbf{g}_{\mathbf{y}}}{\partial \boldsymbol{u}} \frac{\partial \mathbf{g}_{\mathbf{y}}}{\partial \boldsymbol{u}}^{\mathsf{T}}\right|^{-\frac{1}{2}} \rho(\boldsymbol{u})$$

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \boldsymbol{p} \qquad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \frac{\partial \log \pi}{\partial \boldsymbol{u}} - \frac{\partial \mathbf{g}_{\mathbf{y}}}{\partial \boldsymbol{u}}^{\mathsf{T}} \boldsymbol{\lambda}$$

subject to $\mathbf{g}_{\mathbf{y}}(\boldsymbol{u}) = \mathbf{y}_{\text{obs}}$ and $\frac{\partial \mathbf{g}_{\mathbf{y}}}{\partial \boldsymbol{u}} \boldsymbol{p} = \boldsymbol{0}$.

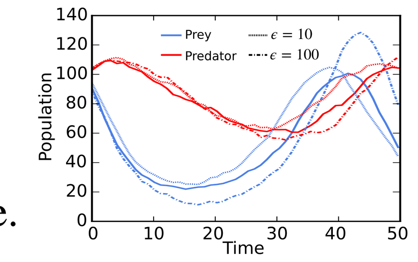


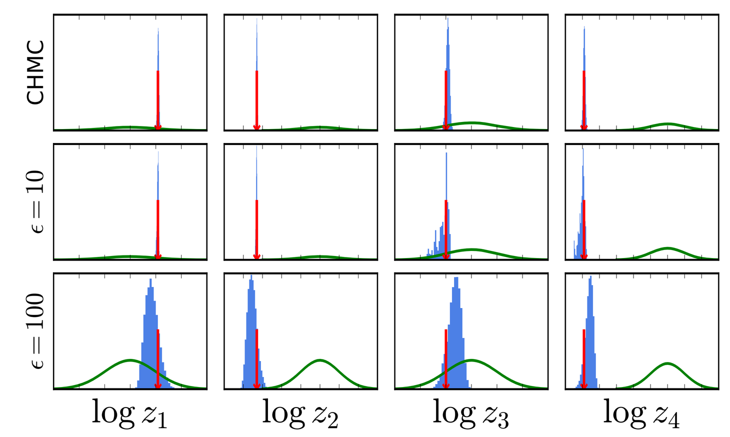
Experiments

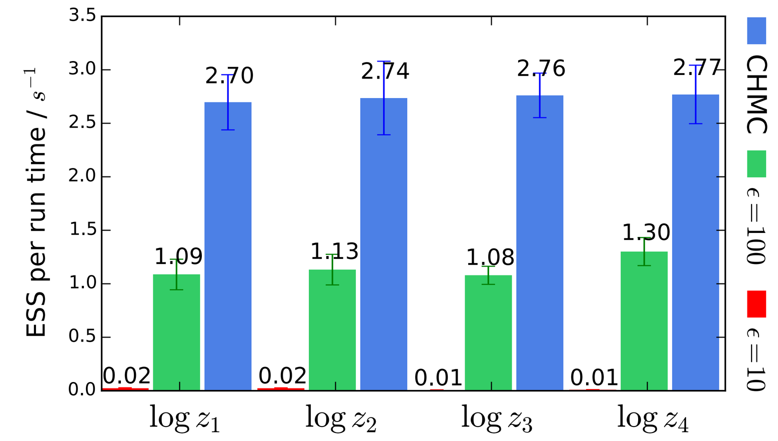
Lotka-Volterra parameter inference

Infer posterior on 4 system parameters given observations of predator and prey populations at 50 simulated time-steps.

Pseudo-marginal slice sampling in ABC posterior with uniform ball kernel with $\epsilon = 10$ and $\epsilon = 100$ used as baseline.

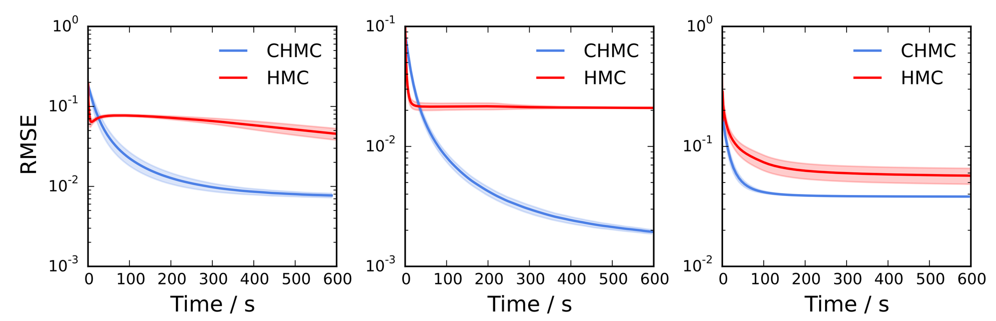




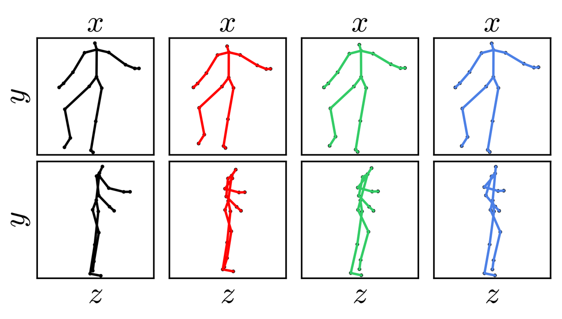


Human pose inference

3D pose given binocular 2D projections



3D pose given monocular 2D projection

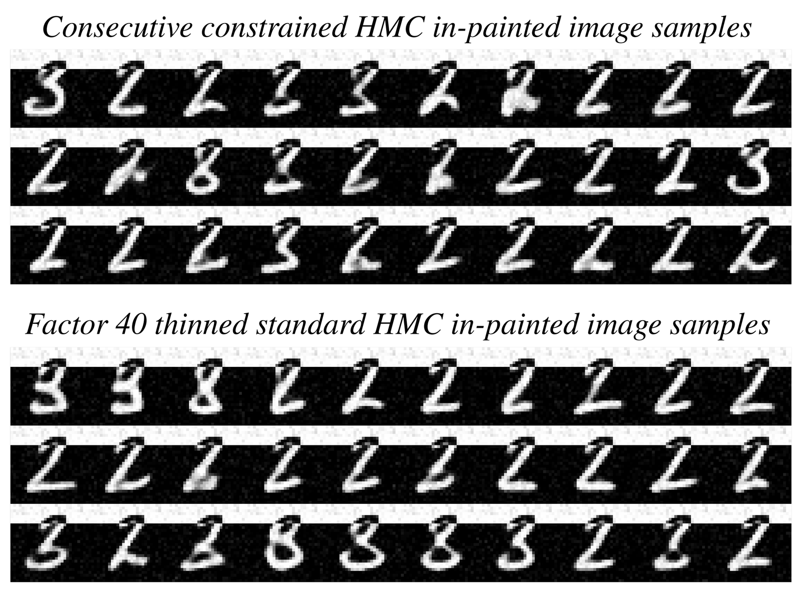


MNIST digit image in-painting

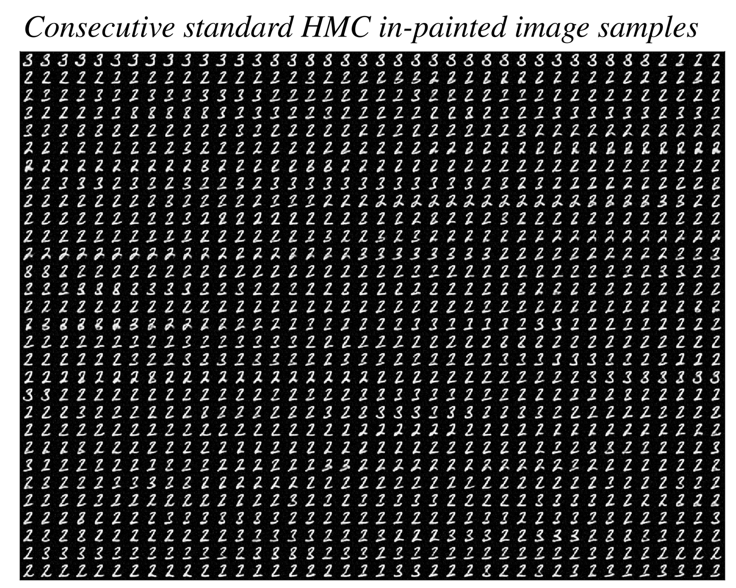
In-painting bottom 75% of digit images given observed top 25%. 60 *s* total sampling time.

Consecutive constrained HMC in-painted image samples

Factor 40 thinned standard HMC in-painted image samples



Consecutive standard HMC in-painted image samples



Conclusions

Generally applicable inference method for differentiable generative models.

Asymptotically exact alternative to ABC where applicable:
$\epsilon = 0$ / no summary statistics.

Key idea: consider conditioning as constraint on inputs to generator function.

Constrained HMC allow efficient gradient-based exploration of target density on constraint manifold corresponding to input consistent with observations.

References

1. H. C. Andersen. RATTLE: A 'velocity' version of the SHAKE algorithm for molecular dynamics calculations. *Journal of Computational Physics.*, 1983.
2. M. A. Brubaker, M. Salzmann, and R. Urtasun. A family of MCMC methods on implicitly defined manifolds. *AISTATS*, 2012.
3. P. Diaconis, S. Holmes and M. Shahshahani. Sampling from a Manifold. *Advances in Modern Statistical Theory and Applications*, 2013.
4. I. Murray and M. M. Graham. Pseudo-marginal slice sampling. *AISTATS*, 2016.
5. D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *ICLR*, 2014.

This work was supported in part by grants EP/F500385/1 and BB/F529254/1 for the University of Edinburgh School of Informatics Doctoral Training Centre in Neuroinformatics and Computational Neuroscience (www.anc.ac.uk/dtc) from the UK Engineering and Physical Sciences Research Council (EPSRC), UK Biotechnology and Biological Sciences Research Council (BBSRC), and the UK Medical Research Council (MRC).