# Infinitely Deep Bayesian Neural Networks with Stochastic Differential Equations

Winnie Xu, Ricky T.Q. Chen, Xuechen Li, David Duvenaud

VECTOR INSTITUTE

# Continuous-time generative models

**Motivation**: going beyond point estimation

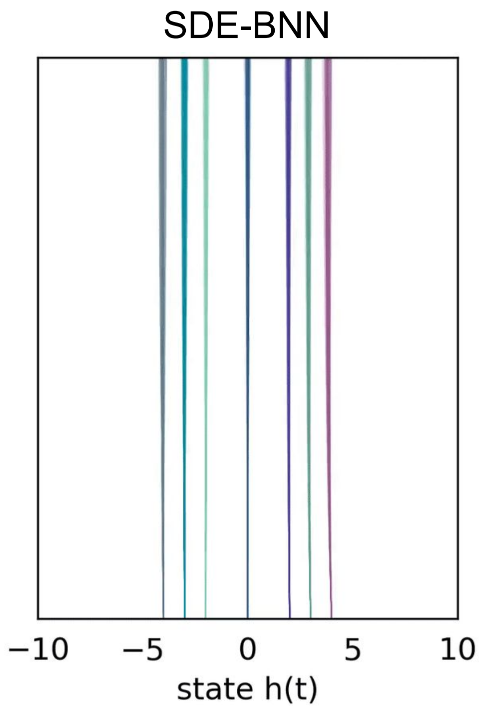
- Implicit specification of model hyperparameters

**Approach**: scalable gradients x practical architecture

- Adaptive computation + O(1) memory training
- Tunable speed vs precision

**Result**: expressive models that quantify uncertainty

- Arbitrarily parameterized dynamics and likelihoods
- Low-variance gradients
- Uncertainty quantification



# Stochastic transition dynamics

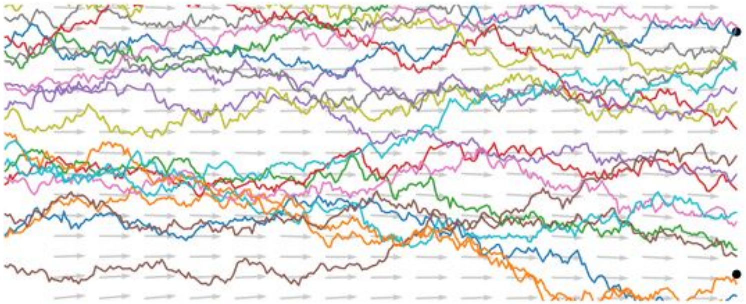
**Fixed state** size with **arbitrary step** size

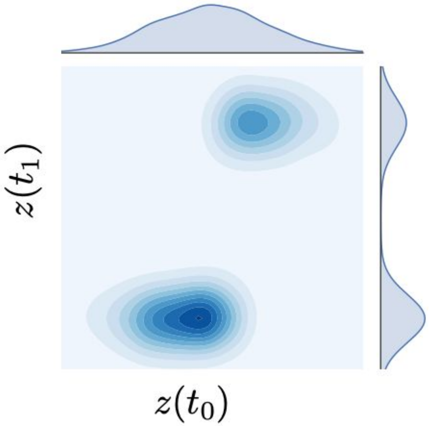
- Non-linear latent variable dynamics with noise at each step

Expressivity from an **implicit distribution** over functions

- Get different trajectories by sampling noise and integrating through time

$$dz = \underbrace{f_\theta(z(t))dt}_{\text{drift}} + \underbrace{\sigma_\theta(z(t))dB(t)}_{\text{diffusion}}$$





# Technicalities of going infinitely deep

Put neural nets in SDE dynamics functions to fit data!

Sample weights from approx. posterior and evaluate output in <u>one</u> SDESolve:

$$d \begin{bmatrix} w_t \\ h_t \end{bmatrix} = \begin{bmatrix} f_w(w_t, \phi) \\ f_h(h_t, w_t) \end{bmatrix} dt + \begin{bmatrix} I \\ \mathbf{0} \end{bmatrix} dB_t$$

Leverage advancements in numerical methods for stochastic control

- Stochastic adjoint sensitivity method and Brownian Tree (*Li et al. 2020*)

# Parameter uncertainty with SDEs

Activations h follow a random ODE

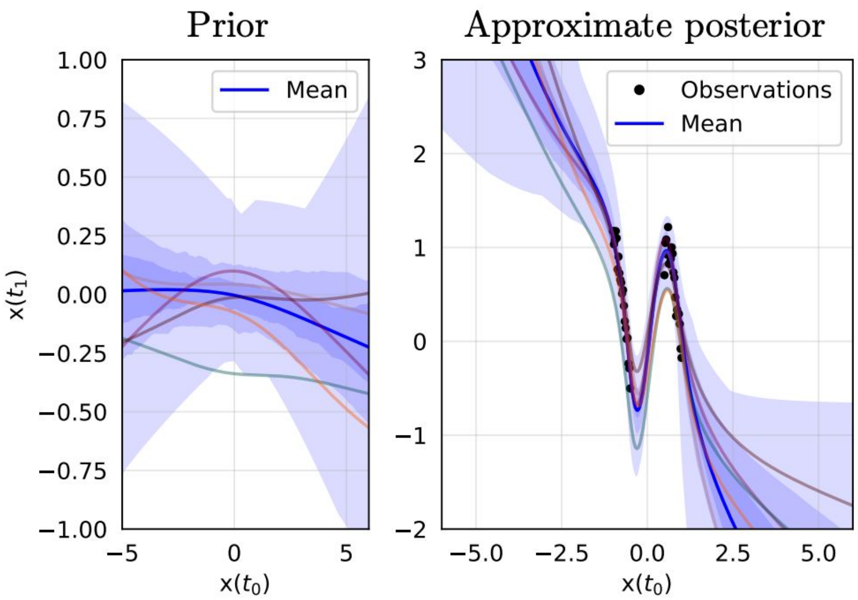
$$dh_t = f_h(h_t, w_t)$$

Prior on weights is an OU process

$$dw_t = -w_t dt + dB_t$$

Likelihood depends on activation at time 1

$$p(y \mid x, w) = \mathcal{N}(y \mid h_1, w)$$




Select expressiveness of the approximate posterior on weights

$$dw_t = f_w(w_t, \phi)dt + dB_t$$

# SVI gradient variance reduction

To do variational inference in this **non-parametric model** class, we need an **unbiased** estimate between the prior and approx. posterior

**Prior**:

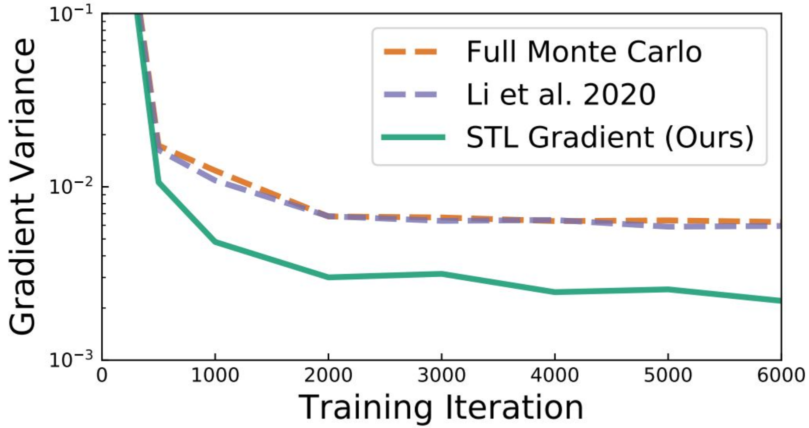
$$dz_p = f_\theta(z_t)dt + \sigma_\theta(z_t)dB(t)$$

**Approximate posterior:**

$$dz_q = f_\phi(z_t)dt + \sigma_\theta(z_t)dB(t)$$

**ELBO(q):**

$$\log p(y_i \mid x_i) - \mathbb{E}_{q(w|\phi)}\left[\frac{1}{2}\int \left|\frac{f_\theta(w_t) - f_\phi(w_t)}{\sigma_\theta(w_t)}\right|_2^2 dw\right]$$



# "Sticking the Landing" for SDEs

- Approximate posterior q(z) can be arbitrarily close to true posterior p(z|x)
- Prevent moving in suboptimal direction when optimizing parameters w.r.t. ELBO now that information is not lost to gradient noise

$$\widehat{\mathrm{KL}}_{\mathrm{STL}} = \int_0^1 \frac{1}{2} \|u(w_t, t, \phi)\|_2^2 \, \mathrm{d}t + \boxed{\int_0^1 u(w_t, t, \perp(\phi)) \, \mathrm{d}B_t}$$

`stop_gradient`

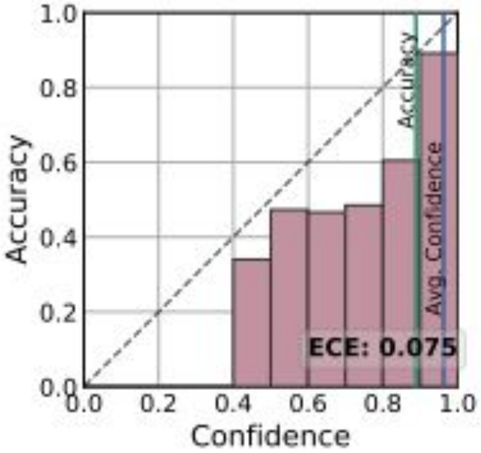
| Method | Accuracy (%) | Negative Log-likelihood ($\times 10^{-4}$) | ELBO |
|---|---|---|---|
| SDE BNN | $95.91 \pm 0.2$ | $1.17 \pm 0.309$ | $1.40 \pm 0.2$ |
| SDE BNN (+STL) | $96.89 \pm 0.2$ | $0.309 \pm 0.15$ | $1.183 \pm 0.2$ |

# One ultimate architecture at scale

**Variational posterior**: a highly parameterized, nonlinear SDE that can recover the true posterior with sufficient model capacity.

**Ablation studies:** 1D regression, MNIST, and CIFAR-10 classification using our variance-reduction trick and novel neural architecture.

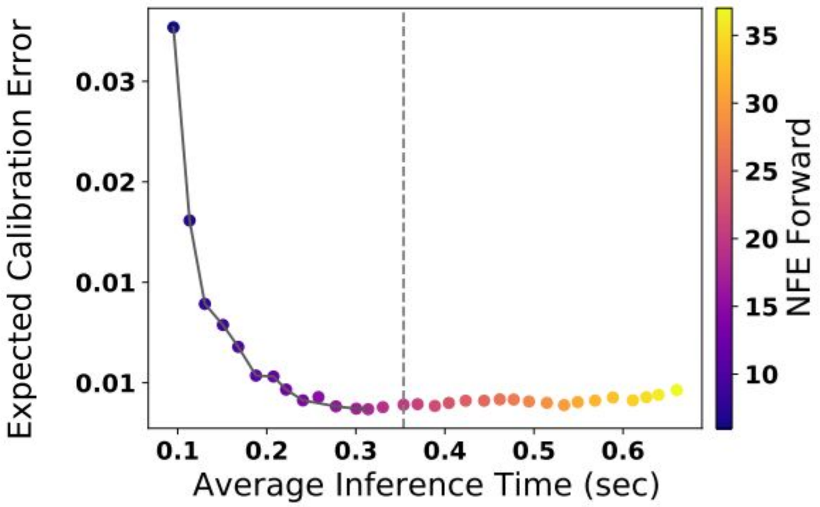
CIFAR-10

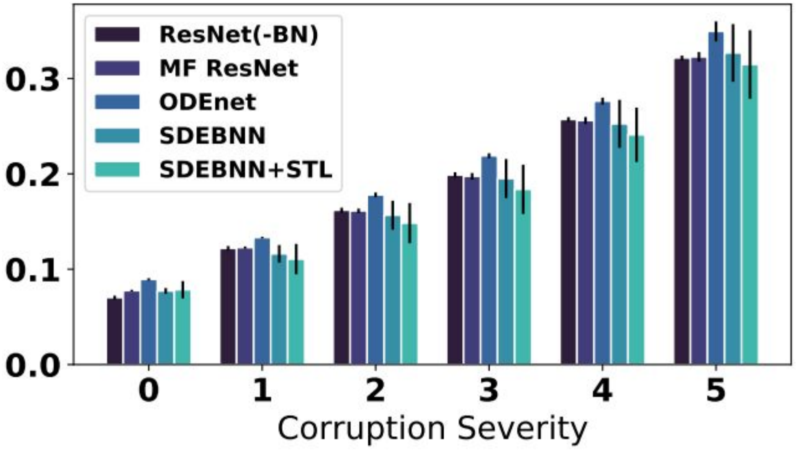



| Method | Posterior over Stochastic Process | Flexible Approximate Posterior | Adaptive Computation | References |
|---|---|---|---|---|
| Bayes by Backprop | ✗ | ✗ | ✗ | Blundell et al. (2015) |
| MCMC for BNNs | ✗ | ✓ | ✗ | (Neal, 1996; Wenzel et al., 2020; Izmailov |
| Bayesian Hypernets | ✗ | ✓ | ✗ | Krueger et al. (2018) |
| BBVI for SDEs | ✓ | ✗ | ✗ | Ryder et al. (2018) |
| Bayesian Neural ODEs | ✗ | ✗ | ✓ | Yıldız et al. (2019)<br>Dandekar et al. (2020) |
| SDE-BNN | ✓ | ✓ | ✓ | current work |

# Effectively learning reliability and robustness

Adjusting SDE-BNN solver tolerance at test time trades off computational speed for predictive performance




SDE-BNN evaluated on all 19 perturbations of CIFAR-10C benchmark.

# Thank You.

Winnie Xu

Xuechen Li

Ricky T.Q. Chen

David Duvenaud

Paper: https://arxiv.org/pdf/2102.06559.pdf
Code: https://github.com/xwinxu/bayeSDE
Tweet: https://twitter.com/DavidDuvenaud/status/1453424027180179464