Continuous-depth Bayesian Neural Networks

Winnie Xu[1], Ricky T.Q. Chen[1], Xuechen Li[3], David Duvenaud[1, 2, 3]

1. University of Toronto 2. Vector Institute for Artificial Intelligence 3. Google Research

Contributions

We construct Bayesian Neural ODEs

- Prior and approx posterior over *continuous-depth weights* are defined using stochastic differential equations.
- This leads to flexible marginal posterior distributions, and can be trained using variational inference [3].
- We derive a low-variance ELBO estimator that has zero variance at the optimum based on [4].
- Benefits from low-memory gradients and adaptive compute.

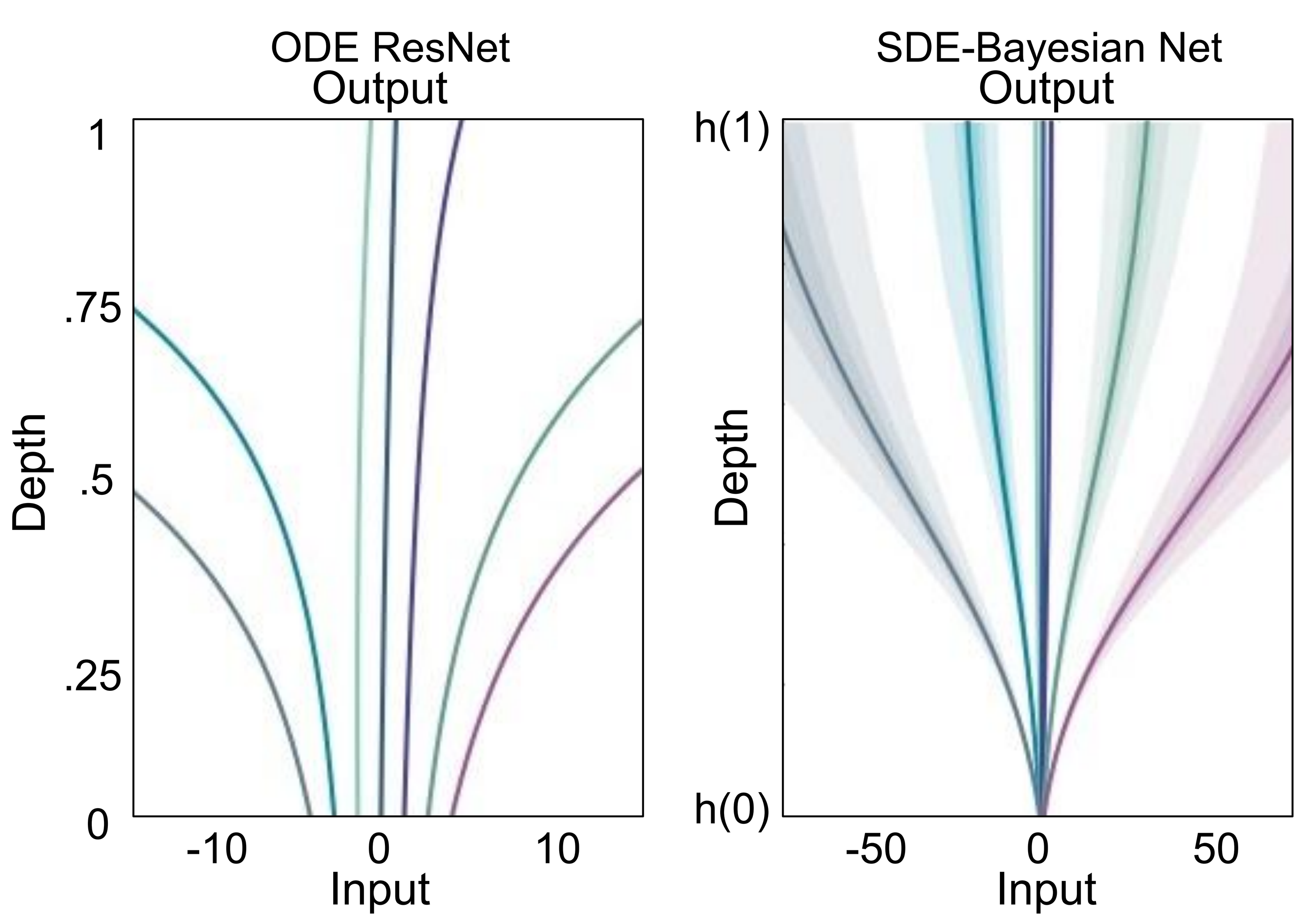
Infinitely Deep Bayesian Residual Networks

SDE-BNN replaces ResNet blocks with **SDESolve(*f, g, s(t_0), t_0, t_1*)** where *f* is a drift neural net (fh with parameters ϕ), g is the diffusion shared by the prior and posterior processes, *s(t_0)* is the initial state.

Addition of **continuous adjoint** with the diffusion term in ODESolve:

$$s_{t+1} = s_t + h(t) f(h(t), w(t), t) + \sqrt{h(t)}\epsilon g(w(t), t)$$

where $\epsilon \sim \mathcal{N}(0,1)$.



$dh(t) = f(h(t), t)\, dt$ $\qquad$ $dw(t) = f(w(t), t)\, dt + g(w(t), t)\, dB(t)$

$$\begin{aligned}
&\textbf{def}\ \ \text{SDE–BNN}(\phi,\ f,\ g,\ t):\\
&\quad B_t \sim \text{Brownian motion}\\
&\quad s_0 = \begin{bmatrix} x_0 \\ w_0 \\ 0 \end{bmatrix}\\
&\quad \text{dS} = \begin{bmatrix} w_t \\ h_t \\ KL \end{bmatrix} = \begin{bmatrix} f_w(w_t, t, \phi) \\ f_h(h_t, w_t, t) \\ KL \end{bmatrix} dt + \begin{bmatrix} g_w(w_t, t) \\ 0 \\ 0 \end{bmatrix} dB_t\\
&\quad \textbf{return}\ \ \text{SDESolve}\ (s_0, dS, t_0, t_1, B_t)
\end{aligned}$$

Stochastic Differential Adjoint

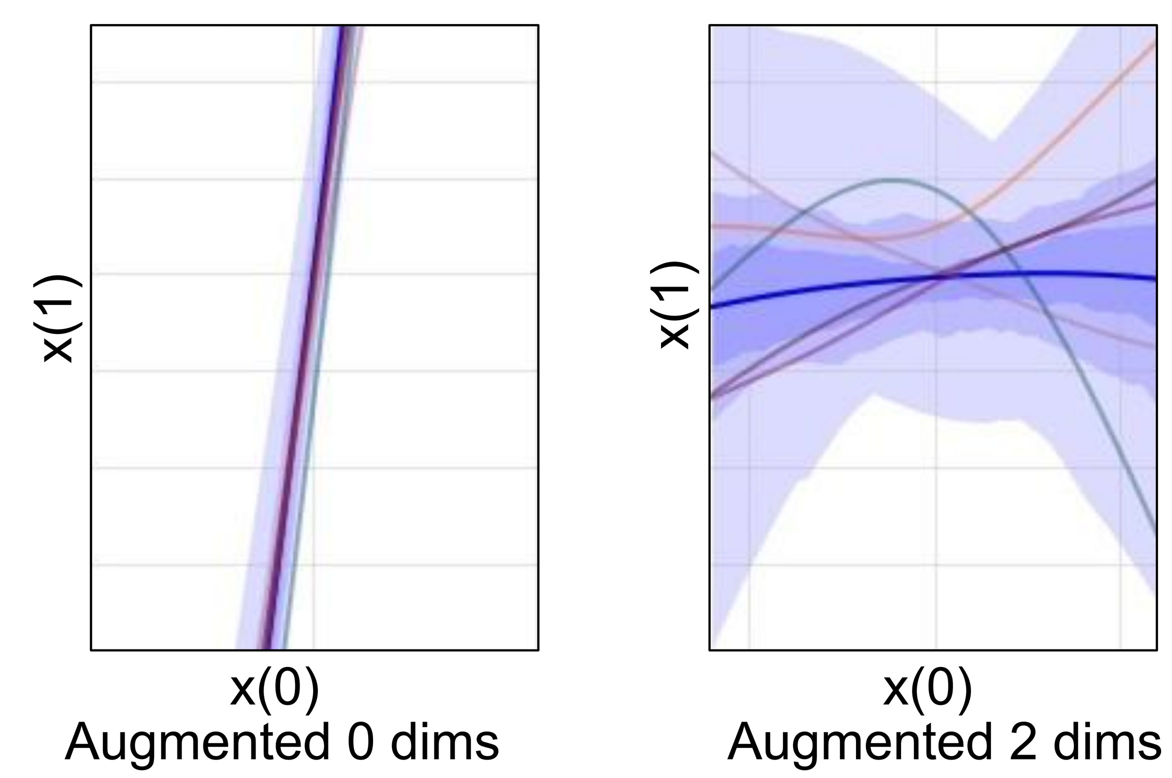
Stochastic differential equations (SDEs) generalizes ODEs to include a noise component steered by a Brownian/Wiener process

$$h_T = h_0 + \underbrace{\int_0^T f(h_t, t)dt}_{\text{drift}} + \underbrace{\int_0^T g(h_t, t)dB_t}_{\text{diffusion (Ito Integral)}}$$

For gradient based optimization with SDEs, we must solve sample paths/dynamics in reverse time. To reproduce Brownian noise, given a seed, the *virtual Brownian tree* algorithm [3] can be used to fetch time-specific values without storing activations.

Non-monotonic Prior Processes

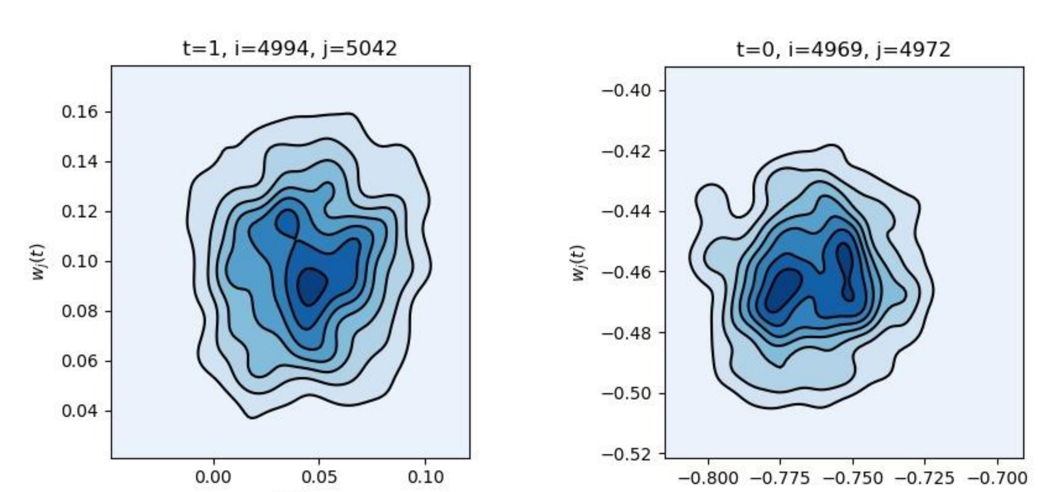
SDEs are not fit to data via maximum likelihood estimates but instead viewed as latent variables.



Solutions no longer constrained to monotonic trajectories.

Arbitrarily Expressive Approximate Posteriors

Non-Gaussian posterior samples given an initial, marginally Gaussian prior (Brownian motion).



The estimate on the weight processes are parameterized by the difference between continuous SDE dynamics and an Ornstein-Uhlenbeck prior. The expressive capacity of the approximate posterior can be larger by increasing the complexity of the drift f_w.

Variational Objective

Continuous ELBO (fully Monte Carlo estimator)

$$\underbrace{\log p(Y|X, \{x_t\}_{t\in[0,T]})}_{\text{neg. reconstruction loss}} - \underbrace{\int_{t_0}^{t_1} \frac{1}{2}|u(x_t, t, \phi)|^2 \,\mathrm{d}t}_{\text{KL divergence } f_p \,||\, f_w} - \underbrace{\int_{t_0}^{t_1} u(x_t, t, \phi)\, \mathrm{d}B_t}_{\text{score function}}$$

where $g(w(t), t)u(w(t), t) = f_w(w(t), t) - f_p(w(t), t)$.

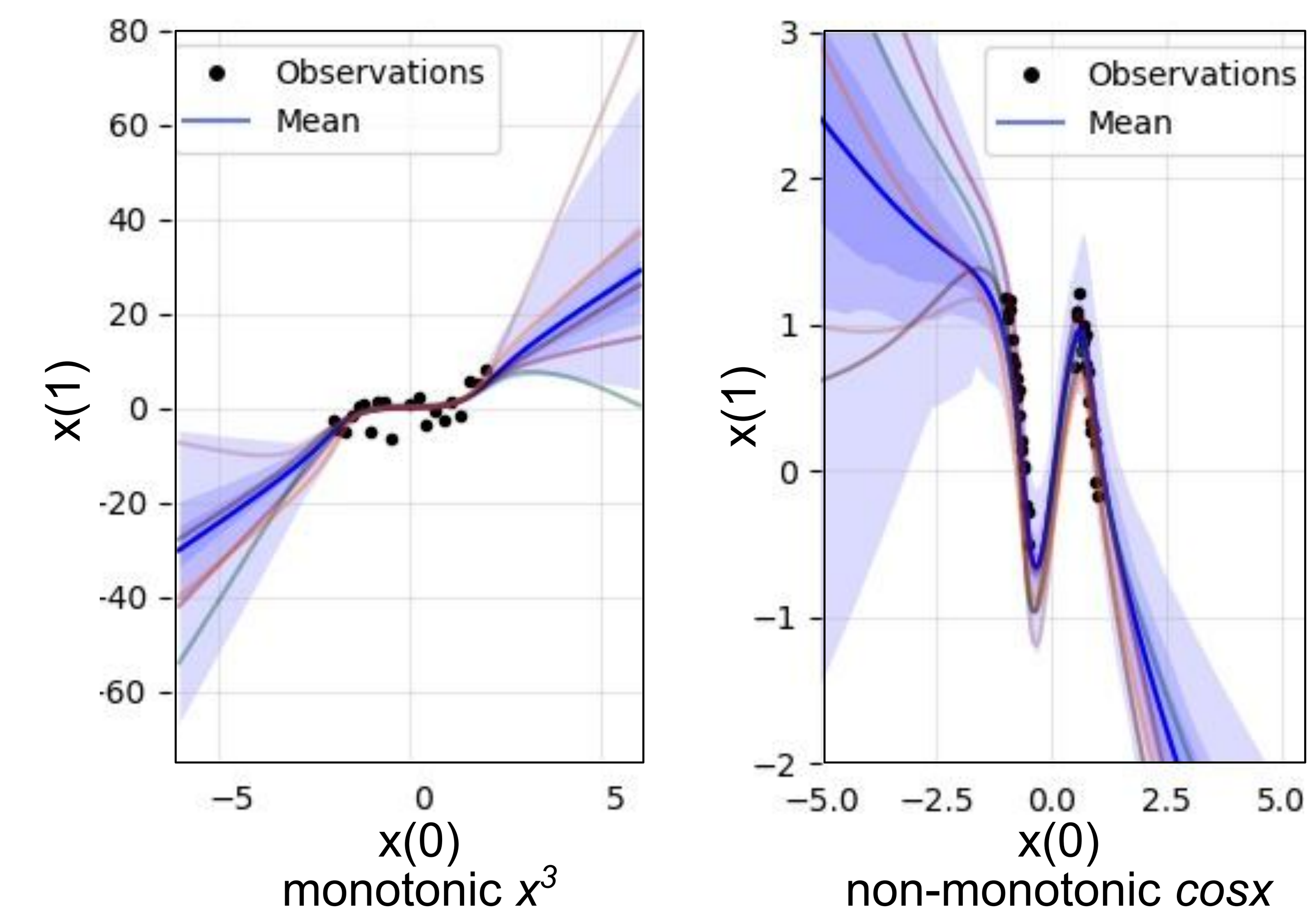
Variance Reduced Gradients

Continuous-time Sticking the Landing (STL) removes the score function term from the ELBO stochastic optimization procedure, remaining an unbiased estimator. $\mathbb{E}_{w_i}[\cdot] = 0$

$$\widehat{\mathrm{KL}} = \int_{t_0}^{t_1} \frac{1}{2}|u(w(t), t, \phi)|^2 dt + \boxed{\int_{t_0}^{t_1} u(w(t), t, \mathrm{sg}(\phi))dB(t)}$$

Learning Non-Monotonic Functions

Weights of the ODE dynamics are not single point estimates but a non-Gaussian, potentially multimodal, posterior density.



Our method demonstrates the utility of SDE-nets and reverse-mode autodiff for approximate inference in a Bayesian setting.

References

[1] Peluchetti. "Infinitely deep neural networks as diffusion processes." (2019)
[2] Chen et al. "Neural Ordinary Differential Equations." (2018)
[3] Li et al. "Scalable Gradients for Stochastic Differential Equations". (2020)
[4] Roeder et al. "Sticking the Landing: Simple, lower-variance gradient estimators for variational inference." (2017)