Hierarchical Implicit Models & Likelihood-Free Variational Inference

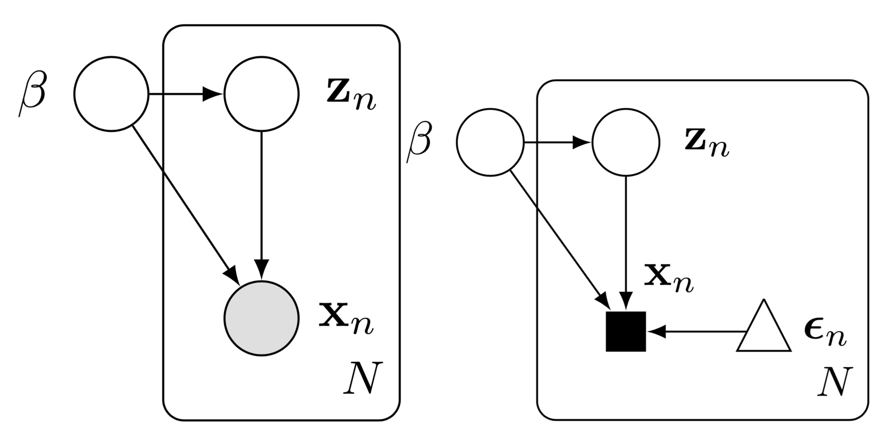
Dustin Tran[†], Rajesh Ranganath[*], David Blei[†]

[†]Columbia University, [*]Princeton University

TL;DR

- Implicit models encompass theories about the physical world.
- Implicit models are limited due to lack of latent structure and scalable inference.
- We develop *hierarchical implicit models* (HIMS). They combine the idea of implicit densities with hierarchical Bayesian models.
- We develop *likelihood-free variational inference* (LFVI). It is a scalable algorithm for HIMS and enables implicit densities as flexible posterior approximations.
- We scale simulators in ecology to unprecedented sizes.

Hierarchical Implicit Models



- Hierarchical models play an important role in sharing statistical strength across examples.
- A broad class of hierarchical Bayesian models can be written as a joint distribution,

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}) = p(\boldsymbol{\beta}) \prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\beta}) p(\mathbf{z}_n \mid \boldsymbol{\beta}). \tag{1}$$

  $\mathbf{x}_n$ is an observation, $\mathbf{z}_n$ are latent variables associated to that observation (local), $\boldsymbol{\beta}$ are latent variables shared across observations (global).
- HIM combine this idea with implicit densities: define a function g that takes in random noise $\boldsymbol{\epsilon}_n \sim s(\cdot)$ and outputs $\mathbf{x}_n$,

$$\mathbf{x}_n = g(\boldsymbol{\epsilon}_n \mid \mathbf{z}_n, \boldsymbol{\beta}), \quad \boldsymbol{\epsilon}_n \sim s(\cdot).$$

- The induced likelihood is

$$\Pr(\mathbf{x}_n \in A \mid \mathbf{z}_n, \boldsymbol{\beta}) = \int_{\{g(\boldsymbol{\epsilon}_n \mid \mathbf{z}_n, \boldsymbol{\beta}) = \mathbf{x}_n \in A\}} s(\boldsymbol{\epsilon}_n) \, \mathrm{d}\boldsymbol{\epsilon}_n.$$

  This integral is typically intractable.
- **Example: Physical Simulators.** For prey and predator populations $x_1, x_2 \in \mathbb{R}^+$ respectively, one process is

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = \beta_1 x_1 - \beta_2 x_1 x_2 + \epsilon_1, \qquad \epsilon_1 \sim \text{Normal}(0, 10),$$
$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = -\beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon_2, \qquad \epsilon_2 \sim \text{Normal}(0, 10),$$

  Lognormal priors are placed over $\boldsymbol{\beta}$.
- **Example: Bayesian Generative Adversarial Network.** The implicit model for a generative adversarial network (GAN) is

$$\mathbf{x}_n = g(\boldsymbol{\epsilon}_n; \boldsymbol{\theta}), \quad \boldsymbol{\epsilon}_n \sim s(\cdot), \tag{2}$$

  We make GANs amenable to Bayesian analysis by placing a prior on the parameters $\boldsymbol{\theta}$.

Likelihood-Free Variational Inference

Variational inference posits an approximating family $q \in \mathcal{Q}$ and optimizes to find the member closest to $p(\mathbf{z}, \boldsymbol{\beta} \mid \mathbf{x})$.
There are many choices of objective functions. To choose one, we lay out desiderata:

1. *Scalability*. The objective should admit unbiased subsampling,

$$\sum_{n=1}^{N} f(\mathbf{x}_n) \approx \frac{N}{M} \sum_{m=1}^{M} f(\mathbf{x}_m),$$

2. *Implicit Local Approximations*. Implicit models specify flexible densities and induce complex posterior distributions. The objective should only require that one can sample $\mathbf{z}_n \sim q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})$ and not evaluate its density.

KL Variational Objective

Classical VI maximizes the ELBO,

$$\mathcal{L} = \mathbb{E}_{q(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}) - \log q(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{x})].$$

Substitute in factorizations,

$$\mathcal{L} = \mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\boldsymbol{\beta}) - \log q(\boldsymbol{\beta})] + \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{\beta}) q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})}[\log p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta}) - \log q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})].$$

This objective presents difficulties: the local densities $p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})$ and $q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})$ are both intractable.

Trick: Density Ratio Estimation

Let $q(\mathbf{x}_n)$ be the empirical distribution on $\mathbf{x}$. Subtract $\log q(\mathbf{x}_n)$ from the ELBO,

$$\mathcal{L} \propto \mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\boldsymbol{\beta}) - \log q(\boldsymbol{\beta})] + \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{\beta}) q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})}\left[\log \frac{p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}{q(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}\right].$$

Train $r(\cdot; \boldsymbol{\theta})$ by minimizing a loss function,

$$\mathcal{D} = \mathbb{E}_{p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}[-\log \sigma(r(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\beta}; \boldsymbol{\theta}))] + \mathbb{E}_{q(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}[-\log(1 - \sigma(r(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\beta}; \boldsymbol{\theta})))].$$

If $r(\cdot; \boldsymbol{\theta})$ is sufficiently expressive, minimizing the loss returns the optimal function,

$$r^*(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\beta}) = \log p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta}) - \log q(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta}).$$

As we minimize X, we use $r(\cdot; \boldsymbol{\theta})$ as a proxy to the log ratio in X. Note r estimates the log ratio; it's of direct interest and more numerically stable than the ratio.

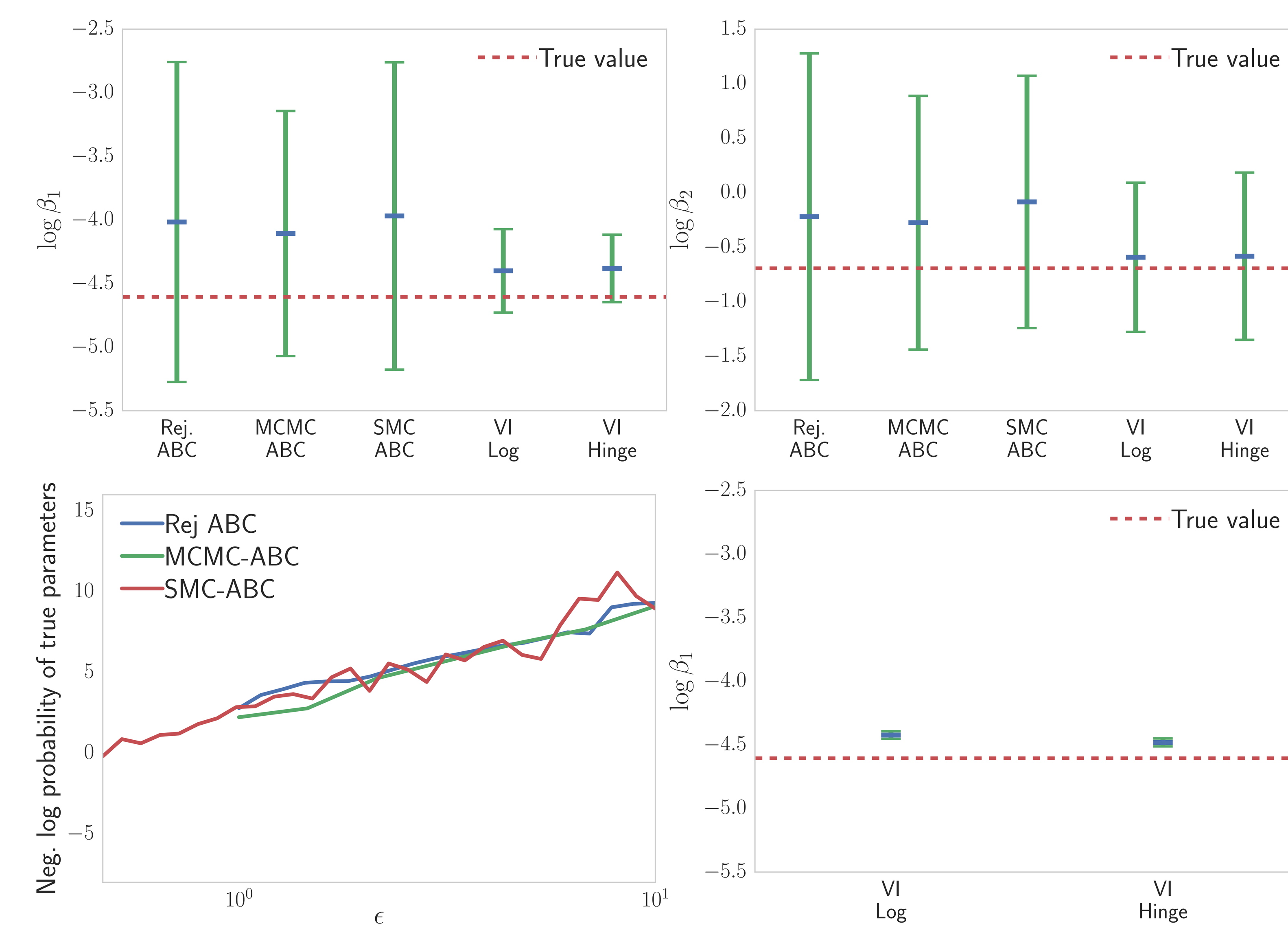
New KL Variational Objective

Optimizing the ELBO involves subsituting in the ratio estimator,

$$\mathcal{L} = \mathbb{E}_{q(\boldsymbol{\beta} \mid \mathbf{x})}[\log p(\boldsymbol{\beta}) - \log q(\boldsymbol{\beta})] + \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{\beta} \mid \mathbf{x}) q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})}[r(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\beta})].$$

All terms are tractable. We can calculate gradients to optimize the variational family q using reparameterization gradients.

Lotka-Volterra Predator-Prey Simulator



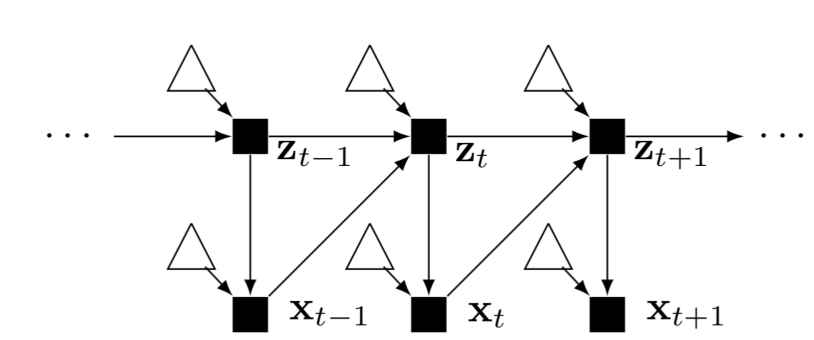
LFVI achieves more accurate results and scales to unprecedented sizes.

Bayesian GAN

	Test Set Error			
Model + Inference	Crabs	Pima	Covertype	MNIST
Bayesian GAN + VI	0.03	**0.232**	**0.154**	**0.0136**
Bayesian GAN + MAP	0.12	0.240	0.185	0.0283
Bayesian NN + VI	**0.02**	0.242	0.164	0.0311
Bayesian NN + MAP	0.05	0.320	0.188	0.0623

Classification accuracy across small/medium-size data. Bayesian GANs achieve comparable or better performance to their Bayesian neural net counterpart.

Recipe: Injecting Noise into Hidden Units



How do you build an implicit model? Inject noise!
For sequences $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, write an RNN,

$$\mathbf{z}_t = g_z(\mathbf{x}_{t-1}, \mathbf{z}_{t-1}, \epsilon_{t,z}), \qquad \epsilon_{t,z} \sim \mathcal{N}(0, 1),$$
$$\mathbf{x}_t = g_x(\mathbf{z}_t, \epsilon_{t,x}), \qquad \epsilon_{t,x} \sim \mathcal{N}(0, 1),$$

The g functions are dense layers with ReLUs and layer norm. Standard normal priors are placed over all weights and biases.

[1] Saatchi, Y. and Wilson, A. G. (2017). Bayesian GAN. In *Neural Information Processing Systems*.
[2] Tran, D. and Blei, D. M. (2017). Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*.