Discretely Relaxing Continuous Variables for tractable Variational Inference

Trefor W. Evans & Prasanth B. Nair

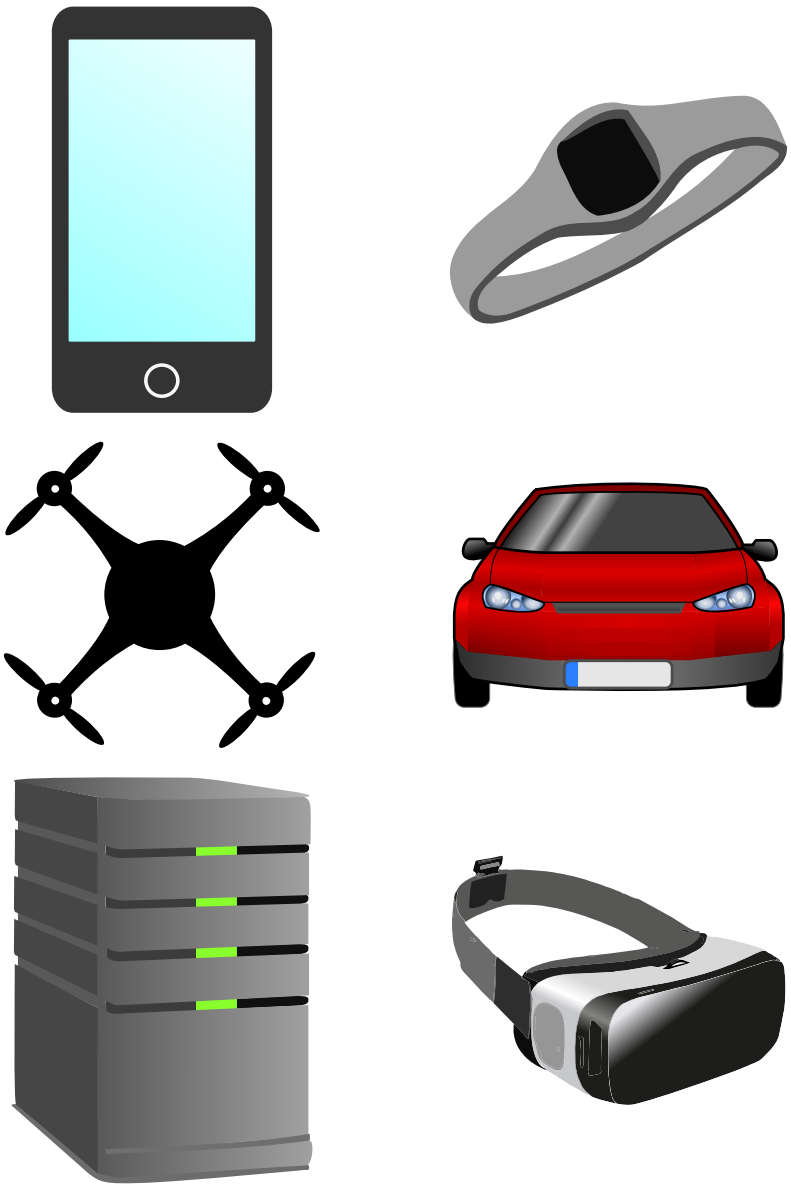
University of Toronto

Neural Information Processing Systems (NeurIPS)
Montréal, Canada
December, 2018

The Need for Efficient Inference

Memory and energy efficiency are critical for mobile devices performing on-board inference, as well as large-scale deployed models.

We introduce a new technique to perform approximate Bayesian inference with discrete variables.
This can dramatically reduce computational requirements.

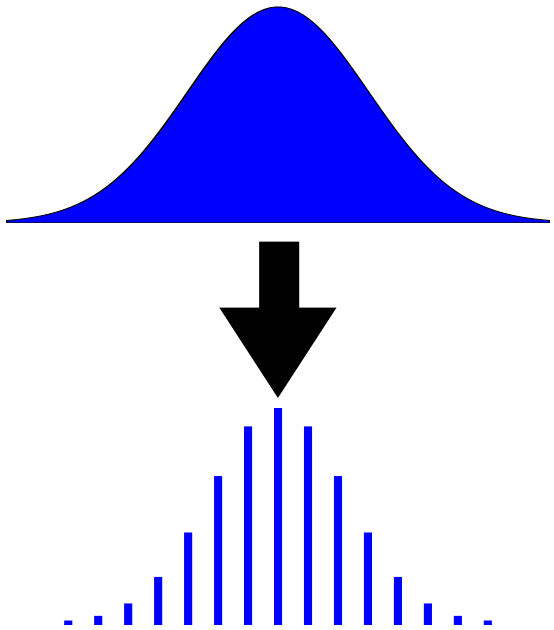


Discretely Relaxing Continuous Variables (DIRECT)

Continuous priors are typically used for approximate Bayesian inference due to computationally tractable training strategies (e.g. reparameterization trick).

However, **discrete priors** offer many advantages at inference time since posterior samples will be sparse and low-precision quantized integers.

We introduce a variational inference technique that enables extremely fast and efficient training with discrete priors.

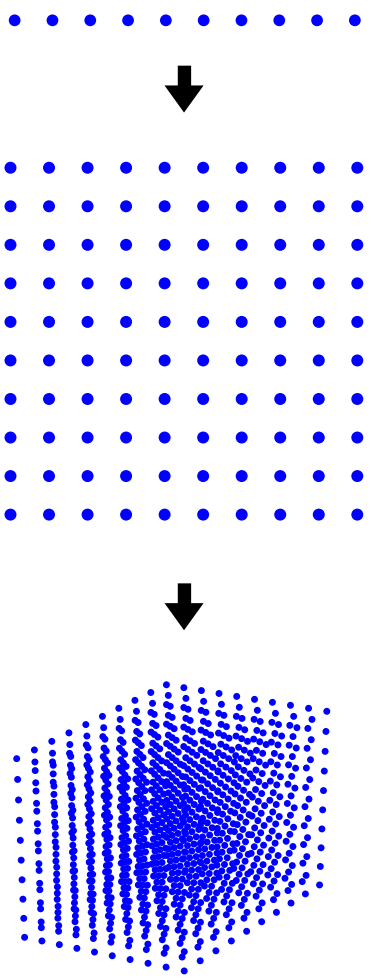


The Curse of Dimensionality

The size of the discretized hypothesis space increases exponentially with the number of variables in the model.

The ELBO (that we maximize during training) requires evaluating the log-likelihood at *each point* in the hypothesis space.
This quickly becomes computationally intractable!

For this reason, we historically resort to high-variance stochastic gradient estimators for training.

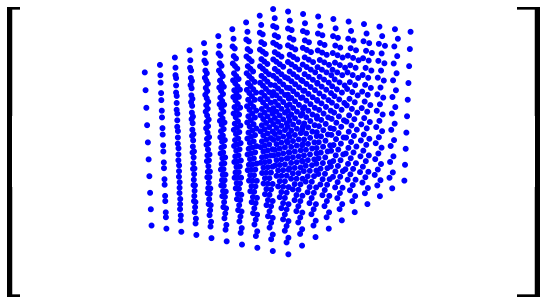


How we Compute the ELBO Exactly

Viewing the log-prior, log-likelihood, and variational distributions over the hypothesis space as tensors, we exploit the low-rank structure of these tensors to rewrite the ELBO in a compact form using Kronecker matrix algebra.

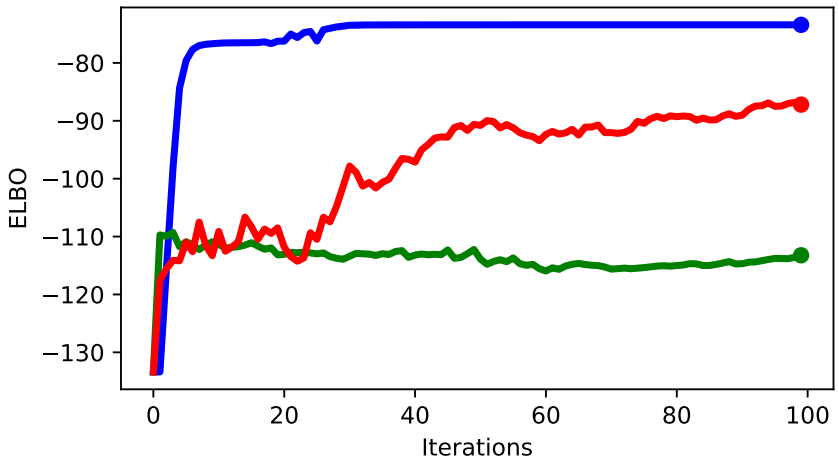
The cost of evaluating the ELBO in this compact form is independent of the number of training points!

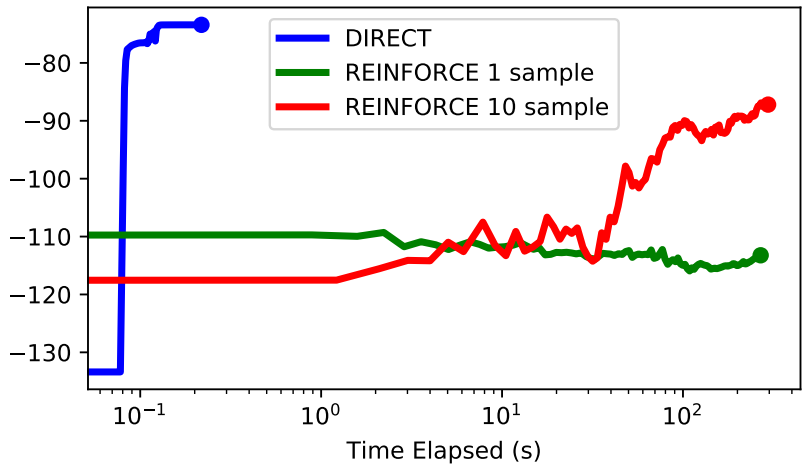
This “DIRECT” approach
is not practical for all likelihoods, however,
we identify a couple that are practical.



Experiments

Training with DIRECT greatly outperforms REINFORCE:



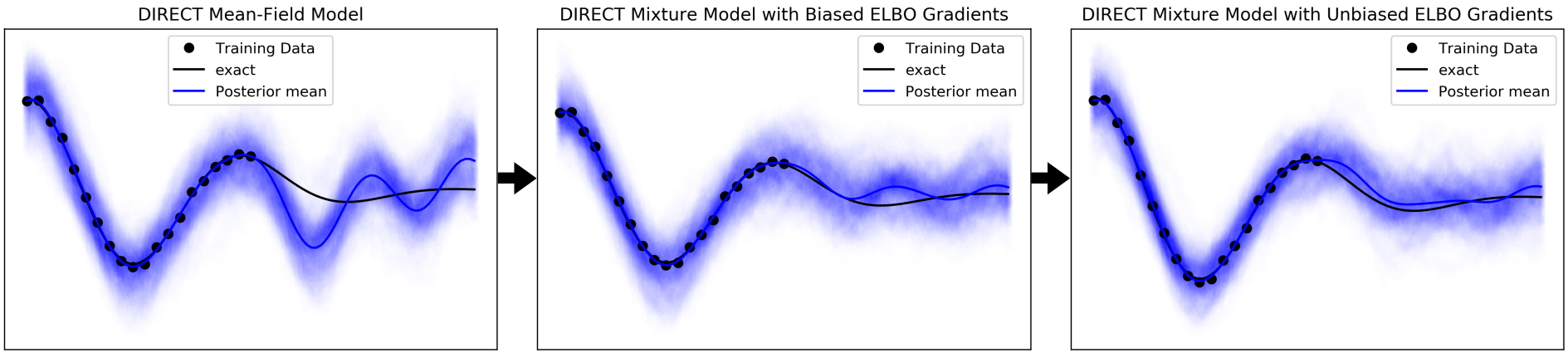


DIRECT can outperform REPARAM:

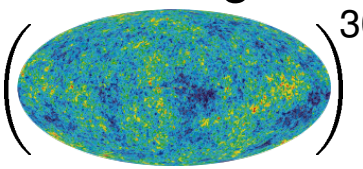
		Continuous Prior		Discrete 4-bit Prior			
		REPARAM Mean-Field		DIRECT Mean-Field		DIRECT 5-Mixture SGD	
Dataset	n	RMSE	Sparsity	RMSE	Sparsity	RMSE	Sparsity
auto	159	0.425 ± 0.2	0%	0.129 ± 0.063	51%	$\mathbf{0.122 \pm 0.056}$	51%
gas	2.5K	0.27 ± 0.052	0%	0.211 ± 0.058	84%	$\mathbf{0.184 \pm 0.063}$	76%
protein	45K	0.642 ± 0.006	0%	0.619 ± 0.007	76%	$\mathbf{0.618 \pm 0.007}$	60%
song	515K	0.537 ± 0.002	0%	0.501 ± 0.002	32%	$\mathbf{0.498 \pm 0.002}$	28%
electric	2M	9.26 ± 4.47	0%	0.575 ± 0.032	99.6%	$\mathbf{0.557 \pm 0.055}$	99.6%

Discussion

Poster #39



The proposed "DIRECT" approach can

1. exactly compute ELBO gradients, eliminating variance
2. its training complexity is independent of the number of training points
3. posterior samples consist of sparse and low-precision quantized integers
4. we demonstrate accurate inference using 4-bit quantized integers and an ELBO summing over $10^{2352} \approx$  ^{30} log-likelihood evaluations

Code: `https://github.com/treforevans/direct`

Contact: `trefor.evans@mail.utoronto.ca`