# Discretely Relaxing Continuous Variables for tractable Variational Inference

Trefor W. Evans and Prasanth B. Nair

University of Toronto

Code: `https://github.com/treforevans/direct`
Contact: `trefor.evans@mail.utoronto.ca`

## Overview

In the proposed "DIRECT" approach to variational inference, we discretely relax continuous variables such that posterior samples consist of sparse and low-precision quantized integers. This enables memory and energy efficient inference which is critical for on-board machine learning on mobile devices as well as large-scale deployed models. Variational inference for discrete latent variable models typically require the use of high variance stochastic gradient estimators, making training impractical for large-scale models. Instead, the DIRECT approach exploits algebraic structure of the ELBO, enabling

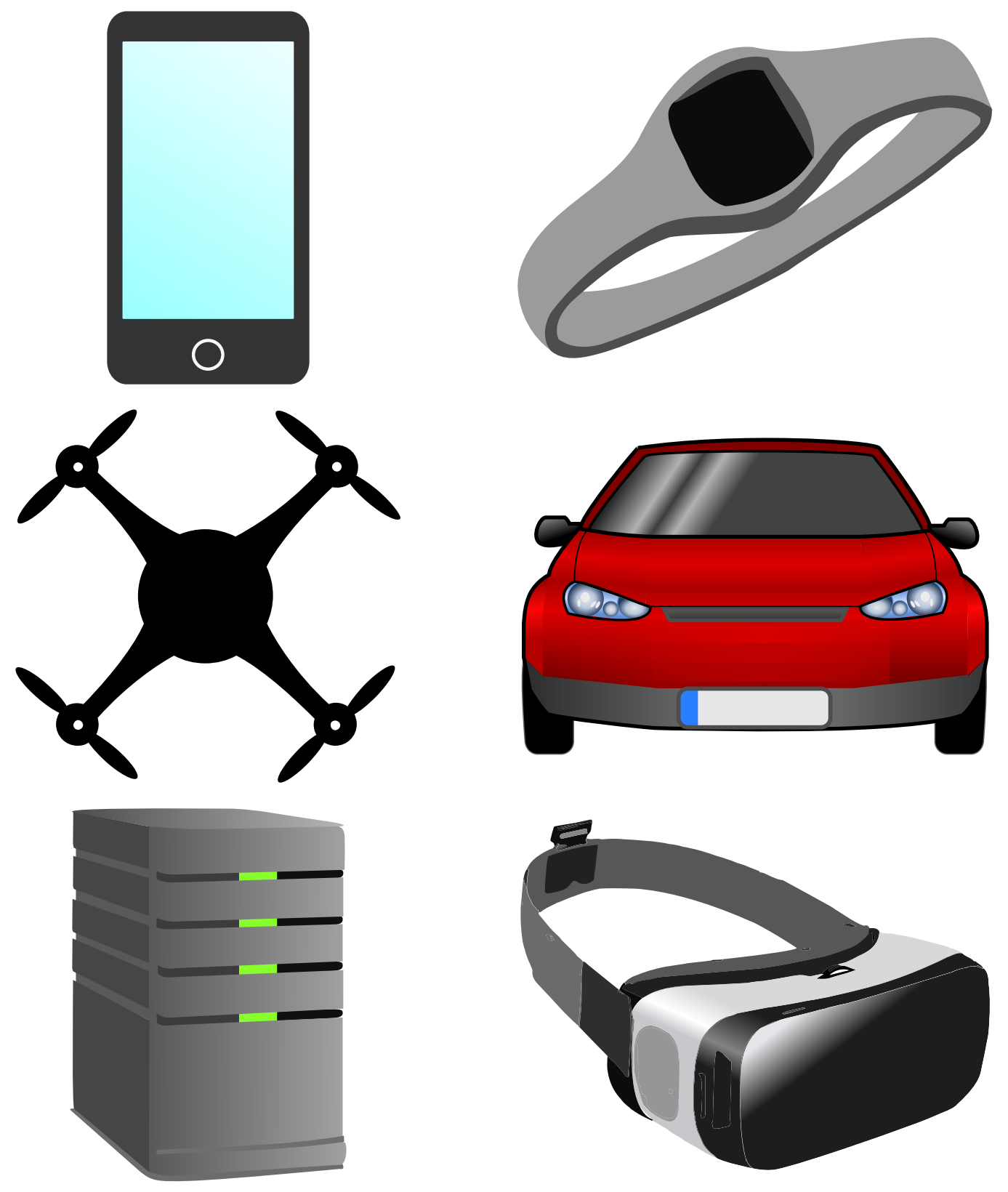
- exact computation of ELBO gradients, eliminating variance;
- its training complexity is independent of the number of training points; and
- posterior samples consist of sparse and low-precision quantized integers

We demonstrate accurate inference on huge datasets using 4-bit quantized integers and an ELBO summing over $10^{2352}$ log-likelihood evaluations.

## 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 efficiently perform approximate Bayesian inference with discrete variables. These discrete models can dramatically reduce computational requirements at inference time.

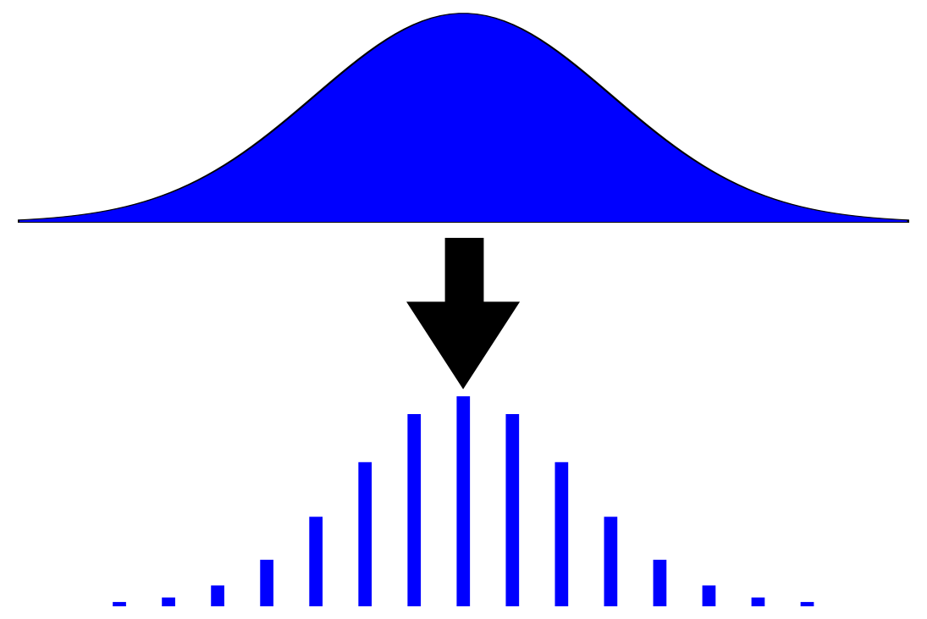


## Discretely Relaxing Continuous Variables (DIRECT)

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

Conversely, models with **discrete priors** are challenging to train, however, they offer many advantages at inference time since posterior samples will be sparse and low-precision quantized integers.

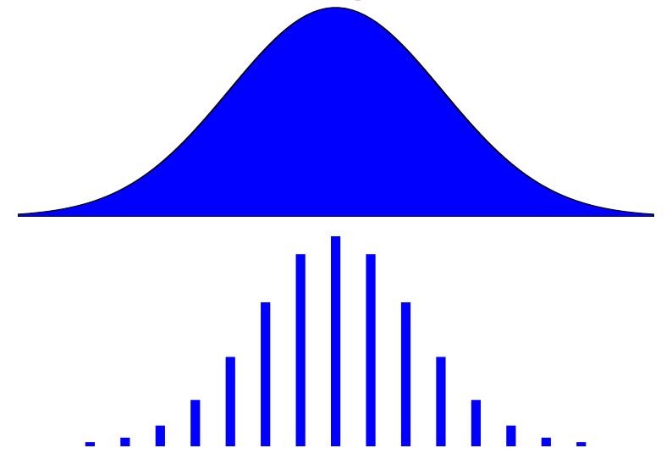
The DIRECT strategy we introduce allows these discrete models to be trained extremely efficiently, allowing us to discretely relax continuous priors to perform tractable variational inference.



## The ELBO

The following is the evidence lower bound (or ELBO) for discrete or continuous priors

**Prior**



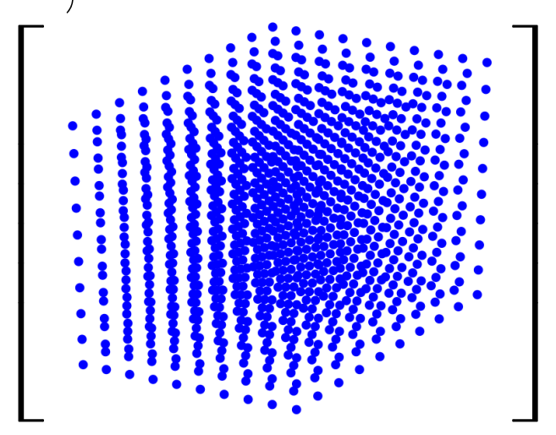
**ELBO**

$$\mathsf{ELBO}(\boldsymbol{\theta}) = \int q_{\boldsymbol\theta}(\mathbf{w})\Big(\log \Pr(\mathbf{y}|\mathbf{w}) + \log \Pr(\mathbf{w}) - \log q_{\boldsymbol\theta}(\mathbf{w})\Big)d\mathbf{w},$$

$$\mathsf{ELBO}(\boldsymbol{\theta}) = \mathbf{q}^T\Big(\log\boldsymbol{\ell} + \log\mathbf{p} - \log\mathbf{q}\Big),$$

where $\log\boldsymbol{\ell} = \{\log\Pr(\mathbf{y}|\mathbf{w}_i)\}_{i=1}^m$, $\log\mathbf{p} = \{\log\Pr(\mathbf{w}_i)\}_{i=1}^m$, $\mathbf{q} = \{q_{\boldsymbol\theta}(\mathbf{w}_i)\}_{i=1}^m$, and $\{\mathbf{w}_i\}_{i=1}^m = \mathbf{W} \in \mathbb{R}^{b\times m}$ is the entire support set of the discrete prior, written as follows,

$$\mathbf{W} = \begin{pmatrix} \overline{\mathbf{w}}_1^T \otimes \mathbf{1}_{\bar m}^T \otimes \cdots \otimes \mathbf{1}_{\bar m}^T \\ \mathbf{1}_{\bar m}^T \otimes \overline{\mathbf{w}}_2^T \otimes \cdots \otimes \mathbf{1}_{\bar m}^T \\ \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \\ \mathbf{1}_{\bar m}^T \otimes \mathbf{1}_{\bar m}^T \otimes \cdots \otimes \overline{\mathbf{w}}_b^T \end{pmatrix},$$



where $\mathbf{1}_{\bar m} \in \mathbb{R}^{\bar m}$ denotes a vector of ones, $\overline{\mathbf{w}}_i \in \mathbb{R}^{\bar m}$ contains the $\overline{m}$ discrete values that the $i$th latent variable $w_i$ can take, $m = \overline{m}^b$, and $\otimes$ denotes the Kronecker product. By observing that the vectors $\mathbf{q} \in \mathbb{R}^m$ and $\mathbf{p} \in \mathbb{R}^m$ can be written as Kronecker product vectors when the prior and variational distributions factorize, we can efficiently and exactly compute two terms in the ELBO as follows,

$$\mathsf{ELBO}(\boldsymbol{\theta}) = \mathbf{q}^T\log\boldsymbol{\ell} + \sum_{i=1}^b \mathbf{q}_i^T\log\mathbf{p}_i - \sum_{i=1}^b \mathbf{q}_i^T\log\mathbf{q}_i,$$

where we use the fact that $\mathbf{p}_i, \mathbf{q}_i \in \mathbb{R}^{\bar m}$ define valid probability distributions for the $i$th latent variable such that $\mathbf{p}_i, \mathbf{q}_i$ both sum to unity. We also extend these results for unfactorized prior and variational distributions. We next consider the likelihood term for a popular generalized linear model with a Gaussian likelihood.

## Theorem 1: Exact ELBO for a GLM

The ELBO can be exactly computed for a discretely relaxed generalized linear model (GLM) for regression as follows

$$\mathsf{ELBO}(\boldsymbol{\theta}) = -\frac{n}{2}\mathbf{q}_\sigma^T\log\boldsymbol{\sigma}^2 - \frac{1}{2}(\mathbf{q}_\sigma^T\boldsymbol{\sigma}^{-2})\Big(\mathbf{y}^T\mathbf{y} - 2\mathbf{s}^T(\boldsymbol{\Phi}^T\mathbf{y}) + \mathbf{s}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{s} - \mathsf{diag}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^T\mathbf{s}^2 + \sum_{j=1}^b \mathbf{q}_j^T\mathbf{h}_j\Big) + \sum_{i=1}^b(\mathbf{q}_i^T\log\mathbf{p}_i - \mathbf{q}_i^T\log\mathbf{q}_i) + \mathbf{q}_\sigma^T\log\mathbf{p}_\sigma - \mathbf{q}_\sigma^T\log\mathbf{q}_\sigma,$$

where $\mathbf{q}_\sigma, \mathbf{p}_\sigma \in \mathbb{R}^{\bar m}$ are factorized variational and prior distributions over the Gaussian noise variance $\sigma^2$ for which we consider the discrete positive values $\boldsymbol{\sigma}^2 \in \mathbb{R}^{\bar m}$, respectively. Also, $\boldsymbol{\Phi} = \{\phi_j(\mathbf{x}_i)\}_{i,j} \in \mathbb{R}^{n\times b}$ contains the evaluations of the basis functions on the training data, and we use the shorthand notation $\mathbf{H} = \{\overline{\mathbf{w}}_j^2\sum_{i=1}^n\phi_{ij}^2\}_{j=1}^b \in \mathbb{R}^{\bar m\times b}$, and $\mathbf{s} = \{\mathbf{q}_j^T\overline{\mathbf{w}}_j\}_{j=1}^b \in \mathbb{R}^b$.

Viewing the log-prior, log-likelihood, and variational distributions over the hypothesis space as tensors, this technique basically exploits the low-rank structure of these tensors to re-write the ELBO in a compact form.

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

A similar viewpoint can be used to show that statistical moments of the predictive posterior can be exactly computed for this model as well.

This "DIRECT" approach is not practical for all likelihoods, however, we identify a couple of models (including the GLM) that are practical.

## Theorem 3: Mixture Entropy Lower Bound

The following inequality holds when we consider a finite mixture of factorized categorical distributions for the variational distribution ($\mathbf{q} = \sum_{i=1}^r\alpha_i\bigotimes_{j=1}^b\mathbf{q}_j^{(i)}$),

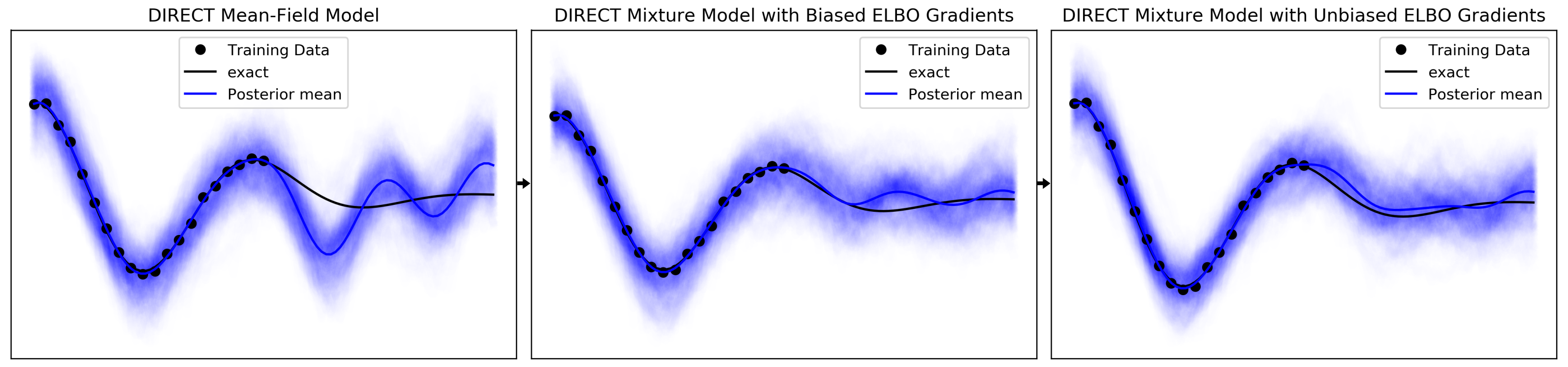
$$-\mathbf{q}^T\log\mathbf{q} \geqslant \max_{\{\mathbf{a}_i\in(0,1)^{\bar m}\}_{i=1}^b} 1 - \sum_{j=1}^r\alpha_j\Big(\sum_{i=1}^b\mathbf{q}_i^{(j)T}\log\mathbf{a}_i + \alpha_j\prod_{i=1}^b\mathbf{q}_i^{(j)T}\frac{\mathbf{q}_i^{(j)}}{\mathbf{a}_i} + 2\sum_{k=j+1}^r\alpha_k\prod_{i=1}^b\mathbf{q}_i^{(j)T}\frac{\mathbf{q}_i^{(k)}}{\mathbf{a}_i}\Big),$$

where $\mathbf{a} = \otimes_{i=1}^b\mathbf{a}_i$, $\mathbf{a}_i \in (0,1)^{\bar m}$ is the center of the Taylor series approximation of $\log\mathbf{q}$, and $\boldsymbol{\alpha} \in (0,1)^r$ is a vector of mixture probabilities for the $r$ components.

This allows for the use of unfactorized variational distributions through the use of a novel Taylor series approximation. This technique is very fast, however, since it does not allow us to compute the ELBO exactly, it introduces bias. Alternatively, since the entropy doesn't depend on the data, it is actually cheap to compute so we can instead use a low-variance stochastic estimator for the entropy as follows

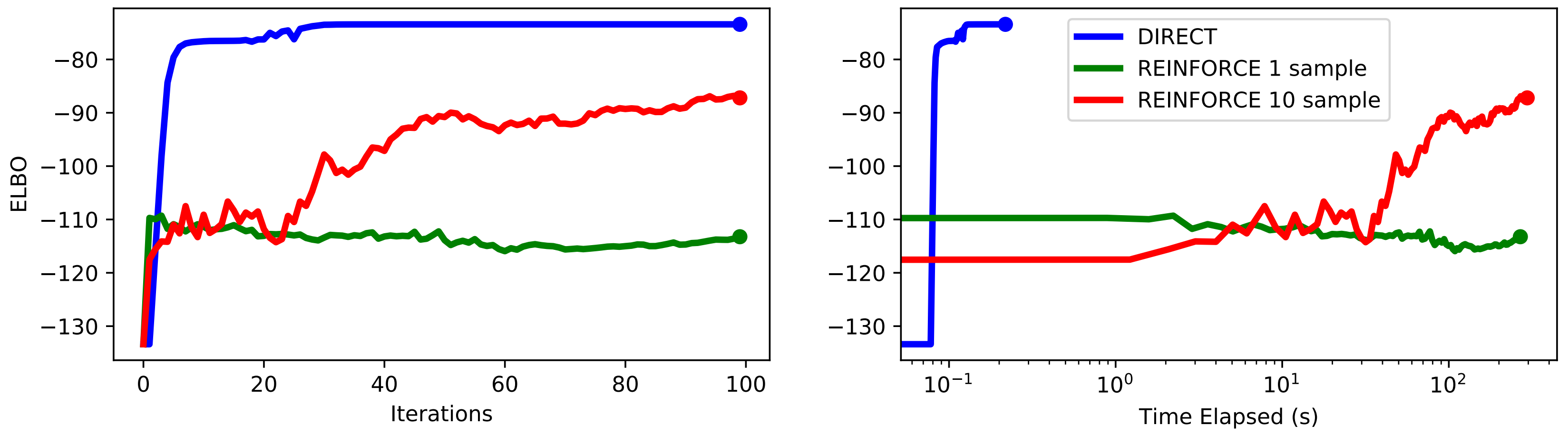
$$\frac{\partial}{\partial\theta}\mathbf{q}^T\log\mathbf{q} \approx \frac{\partial}{\partial\theta}\frac{1}{2t}\sum_{i=1}^t\big(\log q(\mathbf{s}_i) + 1\big)^2,$$

where $\mathbf{s}_i \in \mathbb{R}^b$ is the $i$th of $t$ samples, and we can achieve low variance by using many samples. We can see the effects of extending mean-field DIRECT inference to using a flexible mixture model with biased gradients, and then unbaised gradients.



## Comparison with REINFORCE

Training with DIRECT greatly outperforms REINFORCE in train time and iterations



## UCI Regression Datasets

On many datasets, DIRECT can outperform REPARAM in training time and accuracy, in addition to having sparse and quantized posterior samples.

| | | 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 | $0.425 \pm 0.2$ | 0% | $0.129 \pm 0.063$ | 51% | $\mathbf{0.122 \pm 0.056}$ | 51% |
| gas | 2.5K | $0.27 \pm 0.052$ | 0% | $0.211 \pm 0.058$ | 84% | $\mathbf{0.184 \pm 0.063}$ | 76% |
| protein | 45K | $0.642 \pm 0.006$ | 0% | $0.619 \pm 0.007$ | 76% | $\mathbf{0.618 \pm 0.007}$ | 60% |
| song | 515K | $0.537 \pm 0.002$ | 0% | $0.501 \pm 0.002$ | 32% | $\mathbf{0.498 \pm 0.002}$ | 28% |
| electric | 2M | $9.26 \pm 4.47$ | 0% | $0.575 \pm 0.032$ | 99.6% | $\mathbf{0.557 \pm 0.055}$ | 99.6% |

**Acknowledgements:** Work funded by an NSERC Discovery Grant and the Canada Research Chairs program.