Unbiased Implicit Variational Inference

Michalis K. Titsias* Francisco J. R. Ruiz†‡

*Athens University of Economics and Business †University of Cambridge ‡Columbia University

Summary

- **Goal:** Expand the flexibility of variational approximations through an expressive distribution
- We use an *implicit* variational distribution obtained in a hierarchical manner
- We develop UIVI, a method to obtain unbiased Monte Carlo estimates of the gradient of the ELBO
- The variational parameters are the parameters of a neural network
- Experiments: Bayesian multinomial logistic regression

Introduction

- Probabilistic model $p(x,z)$ (data x, latent variables z)
- Variational inference (VI) approximates the posterior $p(z\,|\,x)$ by maximizing the ELBO
$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(z)}[\log p(x,z) - \log q_\theta(z)]$$
- Classical VI:
  - Fully factorized distribution $q_\theta(z)$ (mean-field VI)
  - Coordinate-wise ascent
  - Limited to a certain class of models
- Goal: Extend the flexibility of $q_\theta(z)$ using an *implicit* distribution
  - It is easy to sample from $q_\theta(z)$
  - It is not possible to evaluate $q_\theta(z)$
- Advantages:
  - Generic inference for any (differentiable) model
  - Expressive distribution (beyond mean-field)
- Method:
  - Stochastic optimization of the ELBO
  - Obtain estimates of the gradients $\nabla_\theta \mathcal{L}(\theta)$
- Technical challenge: The entropy term in the ELBO and its gradient are intractable
  - It is not possible to evaluate $q_\theta(z)$
- Our approach (UIVI):
  - Define the implicit distribution $q_\theta(z)$ through an infinite mixture
  - Rewrite the gradient of the ELBO as an expectation
  - Obtain unbiased estimates of the gradient
- Key ideas:
  - Implicit variational distribution
  - Use a semi-implicit distribution [Yin & Zhou, 2018]
  - Rewrite the gradient as an expectation w.r.t. the *reverse conditional*
  - Use MCMC initialized at stationarity

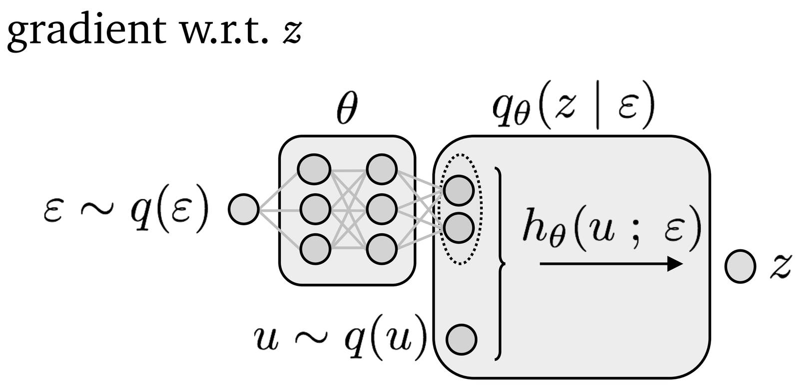
Unbiased Implicit Variational Inference

Variational distribution

- The distribution $q_\theta(z)$ is defined through an infinite mixture,
$$\varepsilon \sim q(\varepsilon), \quad z \sim q_\theta(z\,|\,\varepsilon),$$
or equivalently
$$q_\theta(z) = \int q_\theta(z\,|\,\varepsilon) q(\varepsilon) d\varepsilon$$
- The dependence of the conditional $q_\theta(z\,|\,\varepsilon)$ on ε is arbitrarily complex
- We use a deep neural network with parameters θ that takes ε as input

Assumptions on the conditional $q_\theta(z\,|\,\varepsilon)$

- Reparameterizable distribution
$$u \sim q(u), \;\; z = h_\theta(u\,;\,\varepsilon) \iff z \sim q_\theta(z\,|\,\varepsilon)$$
- It is possible to evaluate $\log q_\theta(z\,|\,\varepsilon)$ and its gradient w.r.t. z



Examples

1. Gaussian conditional
   - The conditional $q_\theta(z\,|\,\varepsilon)$ is multivariate Gaussian
   - Its parameters are $\mu_\theta(\varepsilon)$ and $\Sigma_\theta(\varepsilon)$ (given by neural networks with parameters θ and input ε)
   - Reparameterizable
$$u \sim q(u) = \mathcal{N}(u\,|\,0, I),$$
$$z = h_\theta(u\,;\,\varepsilon) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2} u$$
   - The Gaussian log-density and its gradient are available,
$$\nabla_z \log q_\theta(z\,|\,\varepsilon) = -\Sigma_\theta(\varepsilon)^{-1}(z - \mu_\theta(\varepsilon))$$
2. Reparameterizable exponential family distribution
   - The conditional $q_\theta(z\,|\,\varepsilon)$ is in the exponential family,
$$q_\theta(z\,|\,\varepsilon) \propto \exp\{t(z)^\top \eta_\theta(\varepsilon)\}.$$
   - The log-density and its gradient are available,
$$\nabla_z \log q_\theta(z\,|\,\varepsilon) = \nabla_z t(z)^\top \eta_\theta(\varepsilon)$$

Unbiased Implicit VI

- The gradient of the ELBO is
$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)q(u)}\left[g_\theta^{\text{mod}}(\varepsilon,u) + g_\theta^{\text{ent}}(\varepsilon,u)\right],$$
where
$$g_\theta^{\text{mod}}(\varepsilon,u) \triangleq \nabla_z \log p(x,z)\Big|_{z=h_\theta(u\,;\,\varepsilon)} \nabla_\theta h_\theta(u\,;\,\varepsilon)$$
$$g_\theta^{\text{ent}}(\varepsilon,u) \triangleq -\nabla_z \log q_\theta(z)\Big|_{z=h_\theta(u\,;\,\varepsilon)} \nabla_\theta h_\theta(u\,;\,\varepsilon)$$
- The model component can be estimated as in standard reparameterization
- The entropy component is harder because $q_\theta(z)$ is implicit
- **Key idea:** Rewrite as an expectation
$$\nabla_z \log q_\theta(z) = \mathbb{E}_{q_\theta(\varepsilon'|z)}[\nabla_z \log q_\theta(z\,|\,\varepsilon')]$$
Monte Carlo gradient estimator:
$$g_\theta^{\text{ent}}(\varepsilon_s,u_s) \approx -\nabla_z \log q_\theta(z\,|\,\varepsilon_s')\nabla_\theta h_\theta(u_s\,;\,\varepsilon_s)$$
$$\varepsilon_s' \sim q_\theta(\varepsilon\,|\,z_s)$$
- $q_\theta(\varepsilon\,|\,z)$ is the *reverse conditional*

Sampling from the reverse conditional

- Each pair of samples (z_s, ε_s) comes from $q_\theta(z,\varepsilon)$
- Thus, ε_s is as a draw from $q_\theta(\varepsilon\,|\,z_s)$
- **Key idea:** To sample from the reverse conditional, initialize MCMC chain with ε_s
  - No burn-in period required (starts at stationarity)
  - Every subsequent sample is a sample from the reverse conditional
  - Discard a few samples to reduce correlation between ε and ε'

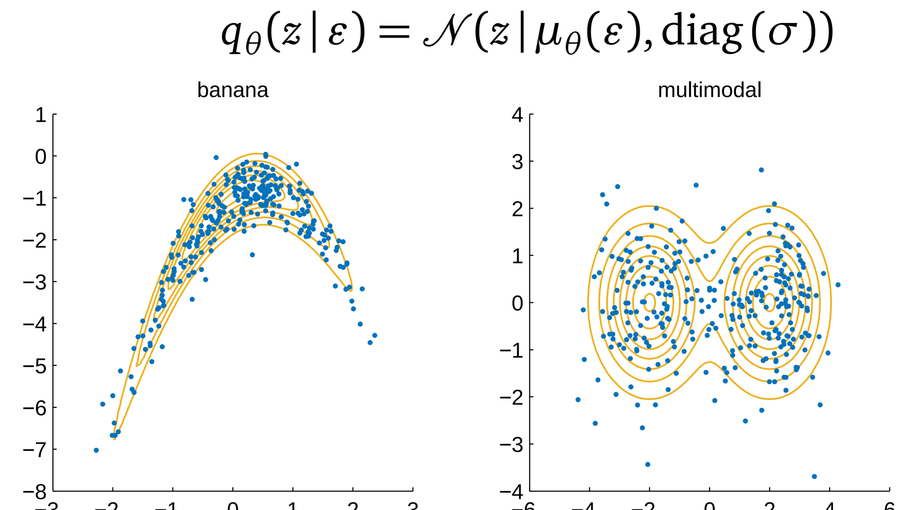
Full algorithm

Input: data x, semi-implicit variational family $q_\theta(z)$
Output: variational parameters θ
Initialize θ randomly
for iteration $t = 1, 2, \ldots,$ **do**
`# Sample from q:`
Sample $u_s \sim q(u)$ and $\varepsilon_s \sim q(\varepsilon)$
Set $z_s = h_\theta(u_s\,;\,\varepsilon_s)$
`# Sample from reverse conditional:`
Sample $\varepsilon_s' \sim q_\theta(\varepsilon\,|\,z_s)$ (HMC initialized at ε_s)
`# Estimate the gradient:`
Compute $g_\theta^{\text{mod}}(\varepsilon_s,u_s)$ (Eq. 6)
Compute $g_\theta^{\text{ent}}(\varepsilon_s,u_s)$ (Eq. 9, approximate using ε_s')
Compute $\widehat{\nabla}_\theta \mathcal{L} = g_\theta^{\text{mod}}(\varepsilon_s,u_s) + g_\theta^{\text{ent}}(\varepsilon_s,u_s)$
`# Take gradient step:`
Set $\theta \leftarrow \theta + \rho \cdot \widehat{\nabla}_\theta \mathcal{L}$
end for

Experiments

Toy experiments

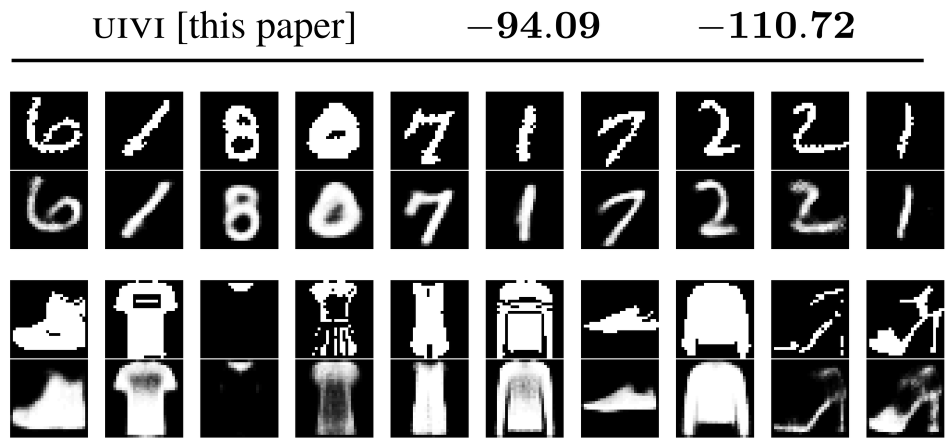
- Synthetic target distributions
- $q(\varepsilon)$ is Gaussian
- The variational conditional is Gaussian,
$$q_\theta(z\,|\,\varepsilon) = \mathcal{N}(z\,|\,\mu_\theta(\varepsilon), \text{diag}(\sigma))$$



Variational autoencoders

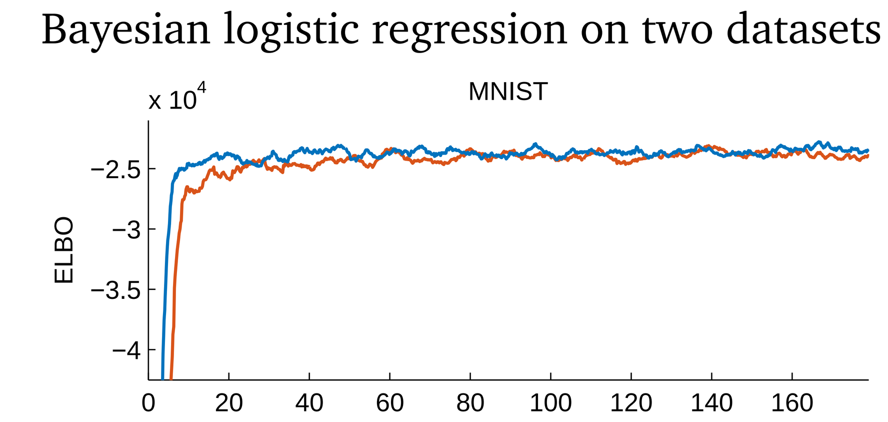
- Fitting a VAE on two datasets

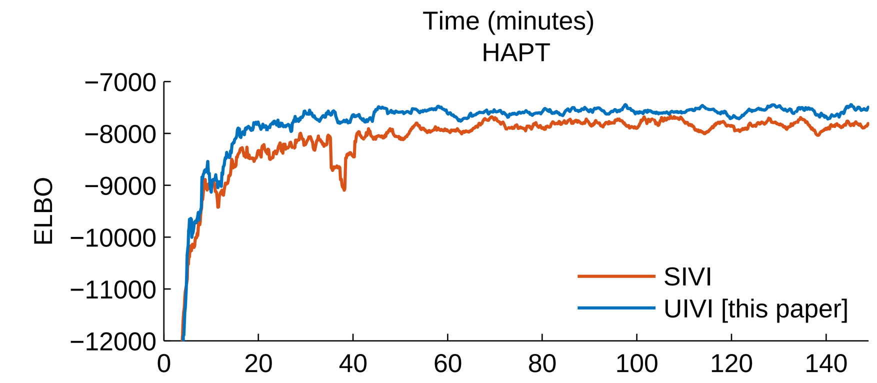
method	average test log-likelihood MNIST	Fashion-MNIST
Explicit (standard VAE)	−98.29	−126.73
SIVI	−97.77	−121.53
UIVI [this paper]	**−94.09**	**−110.72**

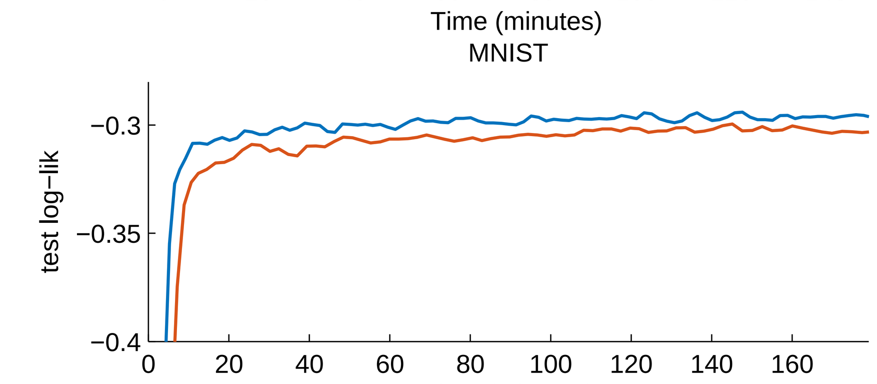


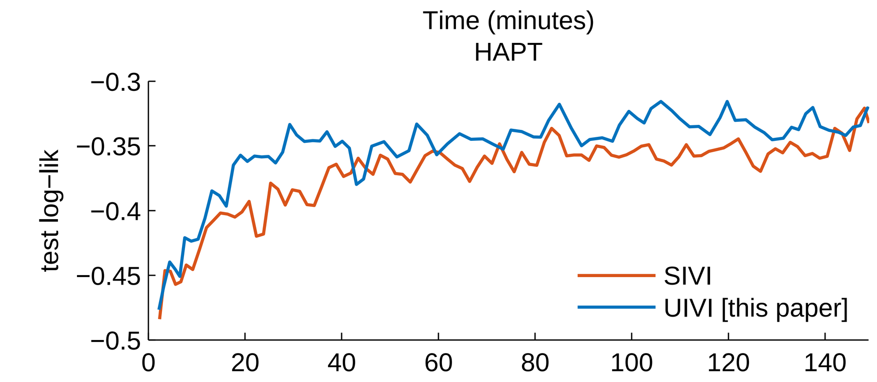
Bayesian logistic regression

- Bayesian logistic regression on two datasets









Proof for the Entropy Component

- Goal: Prove that
$$\nabla_z \log q_\theta(z) = \mathbb{E}_{q_\theta(\varepsilon|z)}[\nabla_z \log q_\theta(z\,|\,\varepsilon)]$$
- Start with log-derivative identity,
$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \nabla_z q_\theta(z)$$
- Apply the definition of $q_\theta(z)$ through a mixture,
$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \int \nabla_z q_\theta(z\,|\,\varepsilon) q(\varepsilon) d\varepsilon$$
- Apply the log-derivative identity on $q_\theta(z\,|\,\varepsilon)$,
$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \int q_\theta(z\,|\,\varepsilon) q(\varepsilon) \nabla_z \log q_\theta(z\,|\,\varepsilon) d\varepsilon.$$
- Apply Bayes' theorem

Related Work

- Linear response estimates [Giordano+, 2017]
- Structured variational family [Saul & Jordan, 1996]
- Mixtures [Bishop+, 1998; Gershman+, 2012; Salimans & Knowles, 2013]
- Boosting VI [Guo+, 2016; Miller+, 2017; Locatello+, 2017]
- Copulas [Tran+, 2015; Han+, 2016]
- Hierarchical models [Ranganath+, 2016; Tran+, 2016; Maaløe+, 2016]
- Invertible transformations [Rezende+, 2014; Kucukelbir+, 2015]
- Normalizing flows [Rezende & Mohamed, 2015; Papamakarios+, 2017]
- Sampling mechanisms [Salimans+, 2015; Maddison+, 2017; Naesseth+, 2017, 2018; Le+, 2018; Grover+, 2018]
- Implicit distributions [Mohamed & Lakshminarayanan, 2016; Nowozin+, 2016; Huszár, 2017; Tran+, 2017; Yin & Zhou, 2018]