Unbiased Implicit Variational Inference

Das, R. and Thompson, C.

Titsias, M. K. and Ruiz, F. J. R.

Athens University of Economics and Business
and University of Cambridge & Columbia University

October 3, 2022

Overview

Background

Let x be our data, z latent variable, θ unknown parameters

- Goal of VI: Approximate $p(z|x)$ of given probabilistic model $p(x, z)$ by maximizing the ELBO

$$\mathcal{L}(\theta) = E_{q_\theta(z)} \left[\log p(x, z) - \log q_\theta(z)\right]$$

- Assumptions of VI: (i) the model must be conditionally conjugate and (ii) variational family must have simplified form such as to be factorized across z (mean-field VI)

For UIVI, $q_\theta(z)$ is implicit i.e. we can draw samples from it but we **cannot** evaluate the density

UIVI: Semi-Implicit Variational Distribution

Define $q_\theta(z)$ hierarchically with mixing parameter ϵ as

$$\epsilon \sim q(\epsilon), z \sim q(z|\epsilon)$$

such that

$$q_\theta(z) = \int q_\theta(z|\epsilon) q(\epsilon) d\epsilon.$$

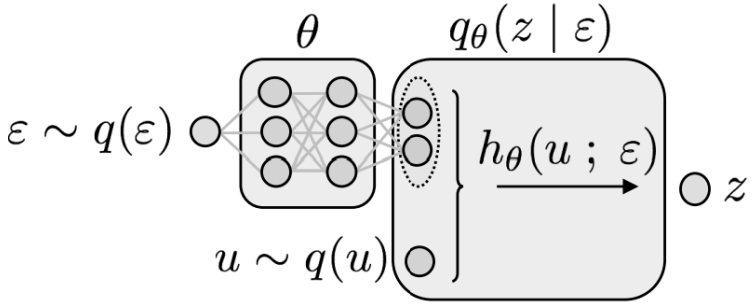
Assumptions on $q_\theta(z|\epsilon)$:

- $q_\theta(z|\epsilon)$ must be reparamaterizable (see next slide)
- It must be possible to evaluate the log-density $\log q_\theta(z|\epsilon)$ and its gradient with respect to z

These two properties help derive the unbiased estimate of the gradient of the ELBO.

UIVI: Semi-Implicit Variational Distribution

In UIVI, a sample ϵ is pushed through a neural network (parameterized by θ) and outputs parameters of the conditional distribution $q_\theta(z|\epsilon)$.



To draw samples of z,

- Sample $u \sim q(u)$ where the auxiliary distribution $q(u)$ has no parameters
- Set z as a deterministic function $h_\theta(\cdot)$ of the sampled u,

$$u \sim q(u), z = h_\theta(u; \epsilon) \equiv z \sim q_\theta(z|\epsilon)$$

IUIVI: Semi-Implicit Variational Distribution

Example: Consider a multivariate Gaussian distribution for $q_\theta(z|\epsilon)$ with parameters $\mu_\theta(\epsilon)$ and $\Sigma_\theta(\epsilon)$ for its mean and covariance. Both parameters are given by neural networks with parameters θ and ϵ.
Assumption 1: The Gaussian is reparameterizable since to generate a sample z, the sampling process is

$$u \sim q(u) = N(u|0, \mathrm{I}),$$

$$z = h_\theta(u; \epsilon) = \mu_\theta(\epsilon) + \Sigma_\theta(\epsilon)^{1/2} u$$

Assumption 2: The gradient of the log-density of the Gaussian can be evaluated as

$$\log q_\theta(z|\epsilon) \propto -\frac{1}{2}(z - \mu_\theta(\epsilon))^T \Sigma_\theta(\epsilon)^{-1}(z - \mu_\theta(\epsilon))$$
$$\nabla_z \log q_\theta(z|\epsilon) = -\Sigma_\theta(\epsilon)^{-1}(z - \mu_\theta(\epsilon)).$$

UIVI: Unbiased Gradient Estimator

We can rewrite the ELBO as

$$\mathcal{L}(\theta) = E_{q(\epsilon)q(u)}\left[\log p(x,z) - \log q_\theta(z)|_{z=h_\theta(u;\epsilon)}\right] .$$

Further, we can obtain the gradient of the ELBO as

$$\nabla_\theta \mathcal{L}(\theta) = E_{q(\epsilon)q(u)}\left[g_\theta^{\text{mod}}(\epsilon,u) + g_\theta^{\text{ent}}(\epsilon,u)\right] .$$

The term corresponding to the model is

$$g_\theta^{\text{mod}}(\epsilon,u) \triangleq \nabla_z \log p(x,z)|_{z=h_\theta(u;\epsilon)} \nabla_\theta h_\theta(u;\epsilon) .$$

The term corresponding to the entropy is

$$g_\theta^{\text{ent}}(\epsilon,u) \triangleq -\nabla_z \log q_\theta(z)|_{z=h_\theta(u;\epsilon)} \nabla_\theta h_\theta(u;\epsilon) .$$

Note: The term $\nabla_z \log q_\theta(z)$ cannot be evaluated!

UIVI: Unbiased Gradient Estimator

This is where our second assumption comes in! We can write

$$\nabla_z \log q_\theta(z) = E_{q_\theta(\epsilon|z)} \left[\nabla_z \log q_\theta(z|\epsilon) \right] .$$

$$\begin{aligned}
\underline{\text{Proof}}: \ \nabla_z \log q_\theta(z) &= \frac{1}{q_\theta(z)} \nabla_z q_\theta(z) \\
&= \frac{1}{q_\theta(z)} \nabla_z \int q_\theta(z|\epsilon) q(\epsilon) d\epsilon \\
&= \frac{1}{q_\theta(z)} \int \nabla_z q_\theta(z|\epsilon) q(\epsilon) d\epsilon \\
&= \frac{1}{q_\theta(z)} \int q_\theta(z|\epsilon) q(\epsilon) \nabla_z \log q_\theta(z|\epsilon) d\epsilon
\end{aligned}$$

Thus, $g_\theta^{\text{ent}}(\epsilon, u) = -E_{q_\theta(\epsilon'|z)} \left[\nabla_z \log q_\theta(z|\epsilon') \right] |_{z=h_\theta(u;\epsilon)} \nabla_\theta h_\theta(u;\epsilon)$.

UIVI: Unbiased Gradient Estimator

Example 1: Consider the multivariate Gaussian distribution again. We can write the entropy component of the gradient as

$$g_\theta^{\text{ent}}(\epsilon, u) = E_{q_\theta(\epsilon'|z)}\left[\Sigma_\theta(\epsilon')(z - \mu_\theta(\epsilon'))\right]|_{z=h_\theta(u;\epsilon)} \nabla_\theta h_\theta(u;\epsilon)$$

Example 2: Consider a general example of a reparameterizable exponential family conditional distribution $q_\theta(z|\epsilon)$ with sufficient statistic $t(z)$ and natural parameter $\eta_\theta(\epsilon)$ such that

$$q_\theta(z|\epsilon) \propto exp\{t(z)^T \eta_\theta(\epsilon)\} .$$

Then we can write the entropy component of the gradient as

$$\log q_\theta(z|\epsilon) \propto t(z)^T \eta_\theta(\epsilon)$$

$$g_\theta^{\text{ent}}(\epsilon, u) = -\nabla_z t(z)^T E_{q_\theta(\epsilon'|z)}\left[\eta_\theta(\epsilon')\right]|_{z=h_\theta(u;\epsilon)} \nabla_\theta h_\theta(u;\epsilon)$$

Full Algorithm: Sampling from reverse conditional

To take the expectation of the reverse conditional $q_\theta(\epsilon|\theta)$, UIVI forms a Monte Carlo estimator using samples ϵ'_s from the reverse conditional.

- Note that each pair of samples (z_s, ϵ_s) comes from the joint distribution $q_\theta(z, \epsilon)$. Thus ϵ_s that generated z_s is a valid sample from the reverse conditional
- However, setting $\epsilon'_s = \epsilon_s$ in the entropy component breaks the assumption that ϵ'_s and ϵ are independent

Thus, UIVI runs a short MCMC method to draw from the reverse conditional that is initialized at ϵ_s to avoid a burn-in period.

Full Algorithm

UIVI estimates the expectation of the gradient of the ELBO by averaging the sum of the model and entropy component over S samples as follows:

$$\nabla_\theta \mathcal{L}(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} (g_\theta^{\text{mod}}(\epsilon_s, u_s) + g_\theta^{\text{ent}}(\epsilon_s, u_s)) \ ,$$

$$\epsilon_s \sim q(\epsilon) \ , u_s \sim q(u) \ .$$

An unbiased estimator of the entropy component using samples from $q_\theta(\epsilon|z_s)$ is

$$g_\theta^{\text{ent}}(\epsilon_s, u_s) \approx -\nabla_z \log q_\theta(z|\epsilon_s') \nabla_\theta h_\theta(u_s; \epsilon_s) \ ,$$

$$\epsilon_s' \sim q_\theta(\epsilon|z_s) \ , z_s = h_\theta(u_s; \epsilon_s) \ .$$

We average over a few samples of the entropy from ϵ' samples to approximate each internal expectation in the entropy component.

UIVI: Full Algorithm

Algorithm 1 Unbiased implicit variational inference

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 $\varepsilon_s' \sim q_\theta(\varepsilon \mid z_s)$ (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 Example

Experiments: Toy Example

name	$p(z)$
banana	$\mathcal{N}\left(\begin{bmatrix} z_1 \\ z_2 + z_1^2 + 1 \end{bmatrix} \middle\| \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$
multimodal	$0.5\mathcal{N}\left(z \middle\| \begin{bmatrix} -2 \\ 0 \end{bmatrix}, I\right) + 0.5\mathcal{N}\left(z \middle\| \begin{bmatrix} 2 \\ 0 \end{bmatrix}, I\right)$
x-shaped	$0.5\mathcal{N}\left(z \middle\| 0, \begin{bmatrix} 2 & 1.8 \\ 1.8 & 2 \end{bmatrix}\right) + 0.5\mathcal{N}\left(z \middle\| 0, \begin{bmatrix} 2 & -1.8 \\ -1.8 & 2 \end{bmatrix}\right)$

Table 1. Synthetic distributions used in the toy experiment.

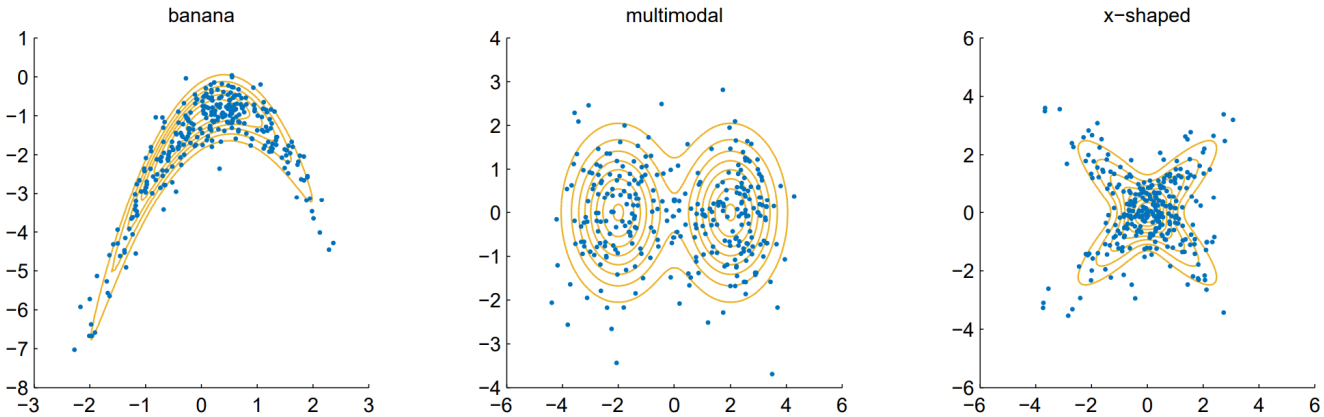


Figure 2. The samples from the variational distribution fitted with UIVI (blue) match the shape of the true synthetic target distributions (orange) considered in Section 4.1.

Experiments: Bayesian Multinomial Logistic Regression

Experiments: Bayesian Multinomial Logistic Regression

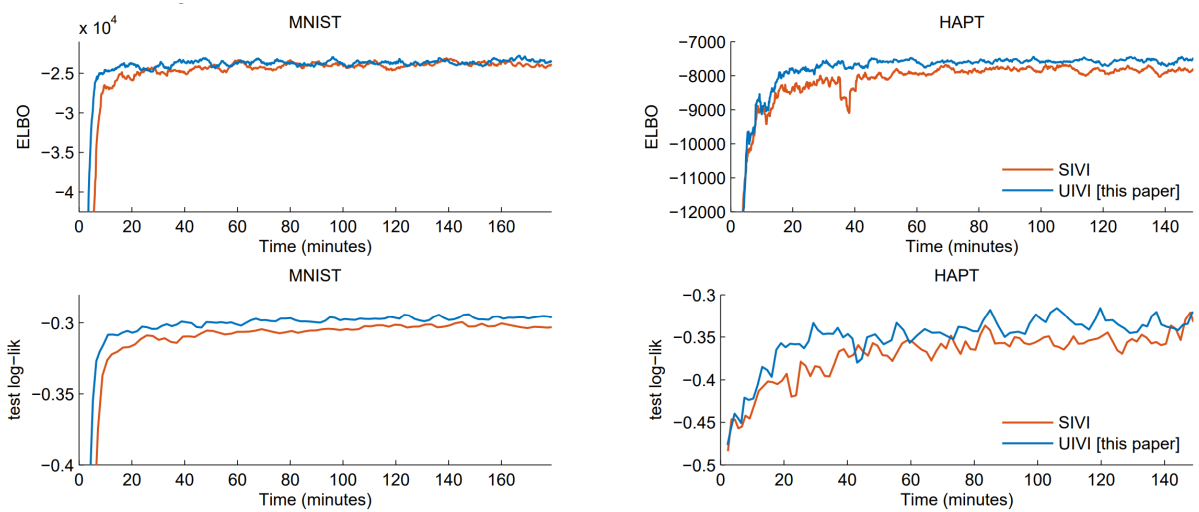


Figure 3. Estimates of the ELBO and the test log-likelihood as a function of wall-clock time for the Bayesian multinomial logistic regression model (Section 4.2). Compared to SIVI (red), UIVI (blue) achieves a better bound on the marginal likelihood and has better predictive performance.

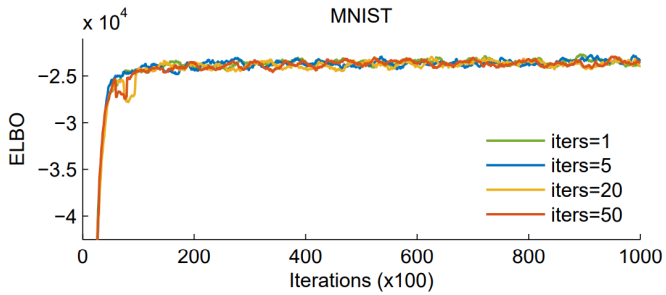
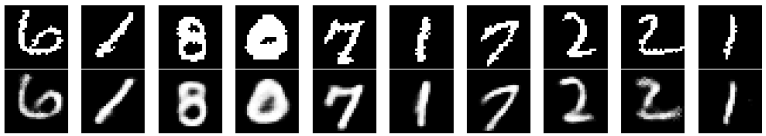


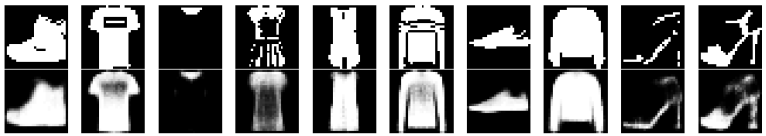
Figure 4. Estimates of the ELBO for the Bayesian multinomial logistic regression model (Section 4.2), obtained with UIVI under four different settings, which differ only in the number of HMC iterations. The number of HMC iterations in UIVI has a small impact on the results.

Experiments: Variational Autoencoders

Experiments: Variational Autoencoders



(a) MNIST images.



(b) Fashion-MNIST images.

Figure 5. Ten images reconstructed with the VAE model fitted with UIVI (Section 4.3). For each dataset, the top row shows training instances; the bottom row corresponds to the reconstructed images.

method	average test log-likelihood MNIST	Fashion-MNIST
Explicit (standard VAE)	-98.29	-126.73
SIVI	-97.77	-121.53
UIVI [this paper]	$\mathbf{-94.09}$	$\mathbf{-110.72}$

Table 2. Estimates of the marginal log-likelihood on the test set for the VAE (Section 4.3). UIVI gives better predictive performance than SIVI.

Conclusion and Future Direction

Blocks of Highlighted Text

Block 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Block 2

Pellentesque sed tellus purus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis magna at risus dictum tempor eu vitae velit.

Block 3

Suspendisse tincidunt sagittis gravida. Curabitur condimentum, enim sed venenatis rutrum, ipsum neque consectetur orci, sed blandit justo nisi ac lacus.

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table: Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)

Title of the publication

Journal Name 12(3), 45 – 678.

The End