The aim of Variational Inference is to provide a deterministic approximation to an intractable posterior distribution by finding parameters ϕ such that $\mathrm{D}_{KL}\left(q_\phi\left(\theta\right) \mid\mid p\left(\theta \mid D\right)\right)$ is minimised. This is achieved by noting that

$$\begin{aligned}\mathrm{D}_{KL}\left(q_\phi\left(\theta\right) \mid\mid p\left(\theta \mid D\right)\right) &= \log p\left(D\right) + \mathbb{E}_{q_\phi(\theta)}\left[\log q_\phi\left(\theta\right) - \log p\left(\theta, D\right)\right] \\ &=: \log p\left(X\right) - \mathcal{L}\left(\phi; D\right). \end{aligned} \tag{1}$$

Noting that $\log p\left(D\right)$ is constant w.r.t. ϕ, we can now minimise the KL-divergence by maximising the evidence lower bound (ELBO) $\mathcal{L}$ (that this is indeed a lower bound follows from the non-negativity of the KL-divergence). Aside from some notable exceptions (eg. [**?**]) this quantity is not tractably point-wise evaluable. However, if $q_\phi\left(\theta\right)$ and $\log p\left(\theta, D\right)$ are point-wise evaluable, it can be approximated using Monte Carlo as

$$\mathcal{L}\left(\phi; D\right) \approx \frac{1}{L}\sum_{l=1}^{L} \log p\left(\theta_l, D\right) - \log q_\phi\left(\theta_l\right), \quad \theta_l \sim q_\phi\left(\theta\right) \tag{2}$$

This stochastic approximation to the ELBO is not differentiable w.r.t. ϕ as the distribution from which each θ_l is sampled itself depends upon ϕ, meaning that the gradient of the log likelihood cannot be exploited to perform inference. One of the primary contributions of the paper being reviewed is to provide a differentiable estimator for $\mathcal{L}$ that allows gradient information to be exploited. In particular it notes that if there exists a tractable reparameterisation of the random variable $\tilde{\theta} \sim q_\phi\left(\theta\right)$ such that

$$\tilde{\theta} = g_\phi\left(\epsilon\right), \quad \epsilon \sim p\left(\epsilon\right), \tag{3}$$

then we can approximate the gradient of the ELBO as

$$\mathcal{L}\left(\phi; D\right) = \mathbb{E}_{p(\epsilon)}\left[\log p\left(\theta, X\right) - q_\phi\left(\theta\right)\right] \approx \frac{1}{L}\sum_{l=1}^{L} \log p\left(\theta_l, X\right) - \log q_\phi\left(\theta_l\right) =: \tilde{\mathcal{L}}^1\left(\phi; X\right), \tag{4}$$

where $\theta_l = g_\phi\left(\epsilon_l\right)$ and $\epsilon_l \sim p\left(\epsilon\right)$. Thus the dependence of the sampled parameters θ on ϕ has been removed, yielding a differentiable estimator provided that both q_ϕ and $\log p\left(\theta, D\right)$ are themselves differentiable. Approxiate inference can now be performed by computing the gradient of $\tilde{\mathcal{L}}^1$ w.r.t. ϕ either by hand or using one's favourite reverse-mode automatic differentiation package (eg. Autograd [**?**]) and performing gradient-based stochastic optimisation to maximise the elbo using, for example, AdaGrad [**?**].

The authors also point out that one can re-express the elbo in the following manner

$$\mathcal{L}\left(\phi; D\right) = \mathbb{E}_{q_\phi(\theta)}\left[\log p\left(D \mid \theta\right)\right] - \mathrm{D}_{KL}\left(q_\phi\left(\theta\right) \mid\mid p\left(\theta\right)\right). \tag{5}$$

This is useful as the KL-divergence between the variational approximation $q_\phi\left(\theta\right)$ and the prior over the parameters θ has a tractable closed-form expression in a number of useful cases. This leads to a second estimator for the elbo:

$$\tilde{\mathcal{L}}^2\left(\phi; D\right) := \frac{1}{L}\sum_{l=1}^{L} \log p\left(D \mid \theta_l\right) - \mathrm{D}_{KL}\left(q_\phi\left(\theta\right) \mid\mid p\left(\theta\right)\right). \tag{6}$$

It seems probable that this estimator will in general have lower variance than $\tilde{\mathcal{L}}^1$.

So far Stochastic Variational Inference has been discussed only in a general parametric setting. The paper's other primary contribution is to use a differentiable recognition network to learn to parameterise the posterior distribution over latent variables z_i local to each observation x_i in a parametric model. In particular, they assume that given some global parameters θ, $z_i \sim p_\theta\left(z\right)$ and $x_i \sim p_\theta\left(x_i \mid z_i\right)$. In the general case the posterior distribution over each z_i will be intractable. Furthermore, the number of latent variables z_i increases as the number of observations increases, meaning that under the framework discussed above we would have to optimise the variational objective with respect to each of them independently. This is potentially computationally intensive and quite wasteful as it completely disregards the any information about the posterior distribution over the z_i provided by the similarities between inputs locations $x_{\neq i}$and corresponding posteriors $z_{\neq i}$. To rectify this the recognition model $q_\phi\left(z \mid x\right)$ is introduced.

Given the recognition model and a point estimate for θ, the ELBO becomes

$$\begin{aligned}\mathcal{L}\left(\theta, \phi; D\right) &= \mathbb{E}_{q_\phi(z_1),\ldots,q_\phi(z_N)}\left[\log \prod_{i=1}^{N} p_\theta\left(x_i, z_i\right) - \log \prod_{i=1}^{N} q_\phi\left(z_i\right)\right] \\ &= \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_i)}\left[\log p_\theta\left(x_i, z_i\right) - \log q_\phi\left(z_i\right)\right] \end{aligned} \tag{7}$$

For this ELBO we can derive a similar result to $\tilde{\mathcal{L}}^1$, where we do no assume a closed-form solution for the KL divergence between distributions and include mini-batching to obtain an estimator for the ELBO for a mini-batch of observations

$$\tilde{\mathcal{L}}^A(\theta, \phi; D) \approx \frac{N}{LM} \sum_{i=1}^{M} \sum_{l=1}^{L} \log p_\theta(x_i, z_{i,l}) - \log q_\phi(z_{i,l}), \quad z_{i,l} = g_\phi(x_i, \epsilon_{i,l}), \ \epsilon_{i,l} \sim p(\epsilon) \tag{8}$$

where the M observations in the mini-batch are drawn uniformly from the data set comprised of N obnservations and for each observation we draw L samples from the approximate posterior $q_\phi(z_i \mid x_i)$. Similarly, if $q_\phi(z \mid x)$ and $p_\theta(z)$ are such that the KL-divergence between them has a tractable closed-form solution then we can use an approximate bound which we could reasonably expect to have a lower variance:

$$\begin{aligned}
\mathcal{L}(\theta, \phi; D) &= \mathbb{E}_{q_\phi(z_1), \ldots, q_\phi(z_N)} \left[\log \prod_{i=1}^{N} p_\theta(x_i \mid z_i) \right] - \mathrm{D}_{KL} \left(\prod_{i=1}^{N} q_\phi(z_i \mid x_i) \ || \ \prod_{i=1}^{N} p_\theta(z_i) \right) \\
&= \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_i)} \left[\log p_\theta(x_i \mid z_i) \right] - \mathrm{D}_{KL} \left(q_\phi(z_i \mid x_i) \ || \ p_\theta(z_i) \right) \\
&\approx \frac{N}{M} \sum_{i=1}^{M} \left[\frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x_i \mid z_{i,l}) - \mathrm{D}_{KL} \left(q_\phi(z_i \mid x_i) \ || \ p_\theta(z_i) \right) \right] =: \tilde{\mathcal{L}}^B(\theta, \phi; D),
\end{aligned} \tag{9}$$

where $z_{i,l} = g_\phi(x_i, \epsilon_{i,l})$ and $\epsilon_{i,l} \sim p(\epsilon)$.