1 Mean-Field Variational Inference (MFVI)

Variational inference is widely used to approximate posterior densities for Bayesian models, an alternative strategy to Markov chain Monte Carlo (MCMC) sampling. Compared to MCMC, variational inference tends to be faster and easier to scale to large data. Rather than use sampling, the main idea behind variational inference is to use optimization. Consider a joint density of latent variable $\mathbf{z} = z_{1:m}$ and observations $\mathbf{x} = x_{1:n}$,

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z}).$$

A Bayesian model draws the latent variables from a prior density $p(\mathbf{z})$ and then relates them to the observations through the likelihood $p(\mathbf{x} \mid \mathbf{z})$. Inference in Bayesian model amounts to conditioning on data and computing the posterior $p(\mathbf{z}|\mathbf{x})$. In variational inference, we posit a family of approximate densities $\mathcal{D} = \{q(\mathbf{z}|\boldsymbol{\psi})\}$, which is a set of densities over the latent variables. Then, we try to find the member of that family that minimizes the Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(\mathbf{z}|\boldsymbol{\psi}_0) = \underset{q(\mathbf{z}|\boldsymbol{\psi}) \in \mathcal{D}}{\arg\min}\ \mathrm{KL}(q(\mathbf{z}|\boldsymbol{\psi}) || p(\mathbf{z}|\mathbf{x})),$$

which is equivalent to maximizing the evidence lower bound (ELBO),

$$\mathrm{ELBO}(q) = \mathrm{E}[\log\ p(\mathbf{z}, \mathbf{x})] - \mathrm{E}[\log\ q(\mathbf{z}|\boldsymbol{\psi})]. \tag{1}$$

Finally, we approximate the posterior with the optimized member of the family $q^*(\cdot)$.
The complexity of the family determines the complexity of the optimization; it is more difficult to optimize over a complex family than a simple family. In mean-field variational inference, we assume the latent variables are mutually independent and each governed by a distinct factor in the variational density. A generic member of the mean-field variational family is

$$q(\mathbf{z}|\boldsymbol{\psi}) = \prod_{j=1}^{m} q_j(z_j|\psi_j).$$

Each latent variable z_j is governed by its own variational factor, the density $q_j(z_j|\psi_j)$. In optimization, these variational factors are chosen to maximize the ELBO of (1).

Though it is flexible, the mean-field family makes strong independence assumptions. These assumptions help with scalable optimization, but they limit the expressibility of the variational family. Further, they can exacerbate issues with local optima of the objective and underestimating posterior variance.

Thus, many researchers are trying to develop better approximation while maintaining efficient optimization, such as adding dependencies between latent variables, constructing a hierarchical variational distribution, or transforming latent variables with complex deterministic and/or stochastic mappings. Normalizing flows help increase the flexibility of VI, but still require the mapping to be deterministic and invertible. Removing both restrictions, there have been several recent attempts to define highly flexible variational distributions with implicit models. A typical example is transforming random noise via a deep neural network, leading to a non-invertible highly nonlinear mapping and hence an implicit distribution, such as Semi-Implicit Variational Inference (SIVI).

2 Semi-Implicit Variational Inference (SIVI)

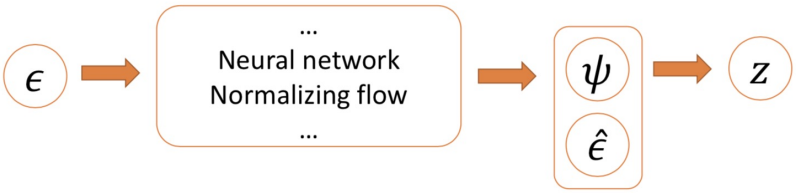
Rather that treating ψ as the variational parameter to be inferred, SIVI regards $\psi \sim q_\phi(\psi)$ as a random variable, where ϕ denotes the distribution parameter to be inferred. Then

$$z \sim q(z|\psi), \quad \psi \sim q_\phi(\psi).$$

We can view z as a random variable drawn from family $\mathcal{H}$, indexed by variational parameter ϕ, where

$$\mathcal{H} = \{h_\phi(z) :_\phi (z) = \int_\psi q(z|\psi) q_\phi(\psi) d\psi\}.$$

$q(z|\psi)$ is required to be explicit, and $q_\phi(\psi)$ is required to be reparameterizable but not necessarily explicit. SIVI draws from $q_\phi(\psi)$ by transforming random noise ϵ via a deep neural network, which generally leads to an implicit distribution for $q_\phi(\psi)$ due to a non-invertible transform.



Reparameterizable An implicit distribution consisting of a source of randomness $q(\epsilon)$ for $\epsilon \in \mathcal{R}^g$ and a deterministic transform $T_\phi : \mathcal{R}^g \to \mathcal{R}^d$, can be constructed as $\psi = T_\phi(\epsilon), \epsilon \sim q(\epsilon)$, with PDF

$$q_\phi(\psi) = \frac{\partial}{\partial \psi_1} \cdots \frac{\partial}{\partial \psi_d} \int_{T_\phi(\epsilon) \leqslant \psi} q(\epsilon) d\epsilon.$$

SIVI has a lower bound for its ELBO as

$$\mathcal{L} = E_{\psi \sim q_\phi(\psi)} E_{z \sim q(z|\psi)} \log \frac{p(x,z)}{q(z|\psi)} \leqslant \text{ELBO}(q) = E_{z \sim h_\phi(z)} \log \frac{p(x,z)}{h_\phi(z)}.$$

The PDF of $h_\phi(z)$ is often intractable, especially if $q_\phi(\psi)$ is implicit, prohibiting a Monte Carlo estimation of the ELBO. By contrast, a Monte Carlo estimation of $\mathcal{L}$ only require $q(z|\psi)$ to have an analytic PDF and $q_\phi(\psi)$ to be convenient to sample from. The analytic $q(z|\psi)$ is used to sidestep the hard problem of density ratio estimation, which is often transformed into a problem related to learning generative adversarial networks.