Stochastic Variational Inference for DGPs

- Intractable posterior over model parameters:
  \[ p(\theta | X) = \frac{p(Y | X, \theta) p(\theta)}{p(Y | X)} \]

- Lower bound on marginal likelihood with mini-batch Stochastic Gradient optimization:
  \[ \log p(Y | X, \theta) \geq \frac{1}{N} \sum_{i=1}^{N} E_{q_\phi(\theta)} \left[ \log p(y_i | \theta) \right] - D_{KL}(q_\phi(\theta) \| p(\theta | X)) \]

- Estimate the expectation using Monte Carlo:
  \[ E_{q_\phi(\theta)} \left[ \log p(y_i | \theta) \right] = \frac{1}{NMC} \sum_{m=1}^{NMC} \log p(y_i | \theta_m) \text{ with } \theta_m \sim q_\phi(\theta) \]

Deep Gaussian Processes (DGPs)

- Deep probabilistic models;
- Composition of functions:
  \[ f(x) = \left( h^{(N-1)} \left( h^{(N-2)} \left( \ldots h^{(1)} \left( h^{(0)}(x) \right) \ldots \right) \ldots \right) \right) \]

- DGPs with random feature expansions:
  - Example of RBF kernel approximated with trigonometric functions:
    \[ \Phi_H = \sqrt{\frac{\pi}{2N}} \cos \left( F \theta \right) \sin \left( F \theta \right) \]
    with \[ F = \Phi_{H1}, \quad p(\theta | y) \sim N(0, \Sigma^{-1}), \quad \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2) \]
  - DGPs become equivalent to Deep Neural Networks with low-rank weight matrices.

References