Modern Computational Statistics, Problem Set 4

December 2, 2019 Due 12/16/2019

Problem 1.

Given a corpus $w = \{w_1, \ldots, w_D\}$, consider a latent Dirichlet allocation (LDA) model with $K = 10$ topics

$$z_{dn}|\theta_d \sim \text{Discrete}(\theta_d), \quad w_{dn}|z_{dn}, \beta \sim \text{Discrete}(\beta_{z_{dn}}), \quad d = 1, \ldots, D,\ n = 1, \ldots, N$$

$$\theta_d \sim \text{Dirichlet}(\alpha), \quad \beta_k \sim \text{Dirichlet}(\eta), \quad d = 1, \ldots, D,\ k = 1, \ldots, K$$

Consider the following mean field approximation

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k|\lambda_k) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{d=1}^{D} \prod_{n=1}^{N} q(z_{dn}|\phi_{dn})$$

Download the data from the course website (data already processed as indices in the vocabulary).

(1) Derive the coordinate ascent algorithm for mean field variational inference.
(2) Derive the variational lower bound for mean field approximation.
(3) Find the vocabulary size V. Set the hyperparameters $\alpha = 1_K$ and $\eta = 1_V$. Run mean field VI for 100 iterations (be careful about your initialization of the variational parameters). Show the variational lower bound as a function of the number of processed documents.
(4) Implement the stochastic variational inference algorithm and run 100 epochs. Show the variational lower bound as a function of the number of processed documents. How does it compare to the standard VI?

Problem 2.

Consider a logistic regression model with normal priors

$$y_i \sim \text{Bernoulli}(p_i),\ p_i = \frac{1}{1 + \exp(-x_i^T \beta)}, \quad i = 1, \ldots, n. \quad \beta \sim \mathcal{N}(0, \sigma_\beta^2)$$

where $\sigma_\beta = 1$. Use the spherical Gaussian $q(\beta|\mu, \sigma) = \mathcal{N}(\mu, \sigma^2 I)$ as our variational distribution. Download the data from the course website (hw2 p3).

(1) Derive the score function estimator for the gradient of the ELBO with respect to the variational parameters.
(2) Implement black-box VI using control variates for variance reduction (you can adapt the baseline using an exponential moving average of the learning signals). Use your favorite stochastic gradient ascent method for training. Show the lower bound as a function of the number of processed observations. Hint: you can estimate the lower bound

using a large number (say, 1000) of samples.
(3) Use the reparameterization trick for the stochastic gradient estimate. Repeat and compare the result to (2). Try to use minibatch instead of full batch when computing the likelihood and its gradients. Does this affect the performances of the score function estimator and the reparameterization trick?
(4) Compare the variance of the stochastic gradient estimates obtained via three methods: score function estimator, score function estimator + control variates, and the reparameterization trick. Report your results for different numbers of Monte Carlo samples.
(5) Bonus points. Retrain your variational approximations using Rényi's α-divergence with $\alpha = -\infty, 0, 0.5, 1, +\infty$. Report and explain your findings.
(6) Bonus points. Can you find a better family of variational distributions? Fit your variational approximations and compare to the spherical Gaussian case.