Bayesian Theory and Computation, Problem Set 3

May 14, 2021 Due 05/28/2021

Problem 1.
Consider a stochastic process given by the following SDE

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t$$

Denote the probability density for X_t as $p(x, t)$.

(1) Derive the Fokker-Planck equation the governs the evolution of $p(x, t)$.

(2) Verify that first order langevin dynamics, second order langevin dynamics and Nosé-Hoover thermostat all have the target distribution as its stationary distribution.

Problem 2.
Consider the following banana-shaped distribution with normal priors

$$y_i \sim \mathcal{N}(\theta_1 + \theta_2^2, \sigma_y^2), \quad i = 1, \ldots, n, \quad \theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$$

where $\sigma_\theta = 1, \sigma_y = 5$. Generate $N = 10000$ iid data points from $\mathcal{N}(1, \sigma_y^2)$.

(1) Implement a Hamiltonian Monte Carlo sampler to collect 500 samples (with 500 discarded as burn-in), show the scatter plot. Test the following two strategies for the number of leapfrog steps L: (1) use a fixed L; (2) use a random one, say Uniform$(1, L_{\max})$. Do you find any difference? Explain it.

(2) Run HMC for 100000 iterations and discard the first 50000 samples as burn-in to form the ground truth. Implement stochastic gradient MCMC algorithms including SGLD, SGHMC and SGNHT. Show the convergence rate of different SGMCMC algorithms in terms of KL divergence to the ground truth as a function of iterations. You may want to use the ITE package `https://bitbucket.org/szzoli/ite-in-python/src/` to compute the KL divergence between two samples.

Problem 3.
In the mixture of experts model, we have observed variables $X_n \in \mathbb{R}^p$, $Y_n \in \mathbb{R}$ and latent variable $Z_n \in \{1, \ldots, K\}$, where the subscript n denotes the n-th data instance. The generative process for this model is as follows.

$$Z_n | X_n = x_n \sim \text{Categorial}(\pi_n), \quad \pi_n = \text{Softmax}(\gamma^T x_n)$$
$$Y_n | X_n = x_n, Z_n = k \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad \mu_k = \theta_k^T x_n$$

Note that when $Z_n = k$ is observed, $Y_n | X_n$ follows a simple regression model with feature weights θ_k and error variance σ_k^2. In the remainder of this problem, we will use

the indicator vector notation $Z_n = (Z_{n1}, \ldots, Z_{nK})$ where $Z_{nk} = 1$ if $Z_n = k$ and $Z_{nk} = 0$ otherwise. Assume that the parameter γ is fixed.

(1) If all of the variables in this model were fully observed, derive the MLE of the model parameters by maximizing the conditional log likelihood of the data, given by

$$\ell(\theta, \sigma^2) = \log p(y, z|x, \theta, \sigma^2, \gamma)$$

where the model parameters are denoted by $\theta = \{\theta_k\}_{k=1}^K, \sigma^2 = \{\sigma_k^2\}_{k=1}^K, \gamma = \{\gamma_k\}_{k=1}^K$ and the data is $x = \{x_n\}_{n=1}^N, y = \{y_n\}_{n=1}^N, z = \{z_n\}_{n=1}^N$.

(2) However, since the Z's are actually latent variables, the marginal log likelihood optimization does not have closed form solutions. In the EM algorithm, we instead work with a lower bound on the marginal log likelihood. Show how to construct such a lower bound and prove that it is locally optimal.

(3) Derive the E-step and M-step update equations for the model parameters.

(4) Set $p = 10, K = 5, N = 100$. Simulate data with $x_n \sim \mathcal{N}(0, I_p)$ and $\gamma_k = \mathbf{1}_p, \theta_k = \frac{1}{k}\mathbf{1}_p, \sigma_k = 1, \ k = 1, \ldots, K$. Run EM for 100 iterations and report the marginal log-likelihood as a function of iterations.