Biostats 270: HW 1

Andrew Holbrook

Spring 2024

For the following, please include code and output in a single pdf file. Assignment due by the beginning of class on Thursday 4/25.

1. We wish to obtain Monte Carlo estimates $\widehat{Ef(\theta)}$ for expectations with respect to the D dimensional Gaussian distribution $N_D(\mathbf{0}, \mathbf{I})$.

   (a) Using the proposal distribution $N_D(\mathbf{0}, 2\mathbf{I})$:

      i. Write down pseudocode for a rejection sampler. What do you choose for M?

      ii. Code your rejection sampler and **plot** acceptance rates (keeping the number of trials fixed at 10,000) and allowing the dimensionality D to increase from $D = 1, 5, 10, 15, \dots,$ until no proposals are accepted. **Plot** mean absolute errors of the empirical estimates $\widehat{E\theta}$ and $\widehat{E\theta^2}$ as a function of the number of accepted samples.

      iii. Write down pseudocode for an importance sampler.

      iv. Code your importance sampler and **plot** empirical entropies

      $$-\sum_i w_i \log w_i \quad \text{for} \quad w_i = \frac{w(\theta_i)}{\sum_s w(\theta_s)},\ i = 1, \dots, 10000.$$

      for dimensionalities $D = 1, 5, 10, 15, \dots,$ until entropies drop below 0.01. **Plot** mean absolute errors of the empirical estimates $\widehat{E\theta}$ and $\widehat{E\theta^2}$ as a function of these entropies.

   (b) Using the proposal distribution $N_D(\mathbf{1}, \mathbf{I})$:

      i. Write down pseudocode for a rejection sampler. What do you choose for M?

      ii. Code your rejection sampler and **plot** acceptance rates (keeping the number of trials fixed at 10,000) and allowing the dimensionality D to increase from $D = 1, 5, 10, 15, \dots,$ until no proposals are accepted. **Plot** mean absolute errors of the empirical estimates $\widehat{E\theta}$ and $\widehat{E\theta^2}$ as a function of the number of accepted samples.

      iii. Write down pseudocode for an importance sampler.

      iv. Code your importance sampler and **plot** empirical entropies

      $$-\sum_i w_i \log w_i \quad \text{for} \quad w_i = \frac{w(\theta_i)}{\sum_s w(\theta_s)},\ i = 1, \dots, 10000.$$

for dimensionalities $D = 1, 5, 10, 15, \ldots$, until entropies drop below 0.01. **Plot** mean absolute errors of the empirical estimates $\widehat{E\theta}$ and $\widehat{E\theta^2}$ as a function of these entropies.

(c) Write pseudocode for the Metropolis algorithm and code it up using proposal distribution $N_D(\mathbf{0}, \frac{2.38}{D}\mathbf{I})$. Calculate the effective sample sizes

$$S_{eff} = \frac{S}{\sum_{s=-\infty}^{\infty} \rho_s}$$

(where ρ_s the autocorrelation at lag s) for chain length $S = 10000$ and dimensions $D = 5, 10, 15, \ldots$. Stop when S_{eff} drops below 50. **Plot** S_{eff} as a function of D.

(d) Let $h = 0.02$ and let $T = 200$. Approximate a D dimensional, continuous time stochastic process starting at time $t = 0$ using the following discretization:

$$\boldsymbol{\theta}(t + h) = (1 - h)\boldsymbol{\theta}(t) + \sqrt{2h}\mathbf{z}_t\,, \quad \mathbf{z}_t \sim N_D(\mathbf{0}, \mathbf{I})\,.$$

Plot the empirical density for $D = 5$. For $D = 5, 10, 15, \ldots$, plot the effective sample sizes until S_{eff} drops below 60.

(e) Which method is best? Why? Are we missing any helpful information?