Bayesian Theory and Computation, Problem Set 2

April 21, 2021 Yutao He 1700012189

Problem 1.

(1)

$$\begin{aligned}\mathbb{E}[\mu(X_t,t)\Delta t|X_t=z] &= \mu(z,t)\Delta t\\ \mathbb{E}[\sigma(X_t,t)\Delta B_t|X_t=z] &= \sigma(X_t,t)\mathbb{E}[\Delta B_t]=0\\ \mathbb{E}[\sigma(X_t,t)^2(\Delta B_t)^2|X_t=z] &= \sigma(z,t)^2\mathbb{E}[(\Delta B_t)^2]=\sigma(z,t)^2\Delta t\end{aligned}$$

So

$$\begin{aligned}\mathbb{E}[\Delta X_t|X_t=z] &= \mathbb{E}[\mu(X_t,t)\Delta t+\sigma(X_t,t)\Delta B_t|X_t=z]=\mu(z,t)\Delta t\\ \mathbb{E}[(\Delta X_t)^2|X_t=z] &= \mathbb{E}[\mu(X_t,t)^2(\Delta t)^2+2\mu(X_t,t)\sigma(X_t,t)\Delta t\Delta B_t+\sigma(X_t,t)^2(\Delta B_t)^2|X_t=z]\\ &= \sigma(z,t)^2\Delta t+o(\Delta t)\end{aligned}$$

note that these formulas can be rewritten as

$$\int (x-z)\rho(x,t+\Delta t|z,t)dx=\mathbb{E}[X_{t+\Delta t}-X_t|X_t=z]=\mathbb{E}[\Delta X_t|X_t=z]=\mu(z,t)\Delta t$$

and similarly,

$$\int (x-z)^2\rho(x,t+\Delta t|z,t)dx=\mathbb{E}[(\Delta X_t)^2|X_t=z]=\sigma(z,t)^2\Delta t+o(\Delta t)$$

when $\Delta t>0$

$$\rho(x,t+\Delta t|y,s)=\int \rho(x,t+\Delta t|z,t)\rho(z,t|y,s)dz$$

Multiply the equation above by a smooth test function R(x) and integrate both sides with respect to x:

$$\int dx\ R(x)\ \rho(x,t+\Delta t|y,s)=\int dx\ R(x)\int \rho(x,t+\Delta t|z,t)\rho(z,t|y,s)dz$$

Talor expansion of R(x) around z:

$$R(x)=R(z)+R'(z)(x-z)+\frac{1}{2}R''(z)(x-z)^2+\cdots$$

Then we will have:

$$\int R(x)\rho(x,t+\Delta t|z,t)dx$$

$$\begin{aligned}
&= \int \left\{ R(z) + R'(z)(x-z) + \frac{1}{2}R''(z)(x-z)^2 + \cdots \right\} \rho(x,t+\Delta t|z,t)dx \\
&= R(z)\int \rho(x,t+\Delta t|z,t)dx \\
&\quad + R'(z)\int (x-z)\rho(x,t+\Delta t|z,t)dx \\
&\quad + \frac{1}{2}R''(z)\int (x-z)^2\rho(x,t+\Delta t|z,t)dx \\
&= R(z) + R'(z)\mu(z,t)\Delta t + \frac{1}{2}R''(z)\sigma(z,t)^2\Delta t + o(\Delta t)
\end{aligned}$$

We can expand the left-hand side of the equation above:

$$\begin{aligned}
\int R(x)\rho(x,t+\Delta t|y,s)dx &= \int R(x)[\rho(x,t|y,s) + \partial_t\rho(x,t|y,s)\Delta t + o(\Delta t)]dx \\
&= \int R(z)\rho(z,t|y,s)dz + \Delta t\int R(z)\partial_t\rho(z,t|y,s)dz + o(\Delta t).
\end{aligned}$$

Then we obtain:

$$\begin{aligned}
&\int R(z)\rho(z,t|y,s)dz + \Delta t\int R(z)\ \partial_t\rho(z,t|y,s)dz + o(\Delta t) \\
&\quad = \int R(z)\rho(z,t|y,s)dz \\
&\qquad + \Delta t\int \left\{ R'(z)\mu(z,t) + \frac{1}{2}R''(z)\sigma(z,t)^2 \right\} \rho(z,t|y,s)dz + o(\Delta t)
\end{aligned}$$

Canceling the equal terms in the left- and the right-hand side, collecting all the terms of order Δt and neglecting the terms of order $o(\Delta t)$, we obtain

$$0 = \int \left\{ R(z)\ \partial_t\rho(z,t|y,s) - \left[R'(z)\mu(z,t) + \frac{1}{2}R''(z)\sigma(z,t)^2 \right] \rho(z,t|y,s) \right\} dz$$

Integrate the terms containing derivatives of R(z) by parts to obtain

$$0 = \int R(z) \left\{ \partial_t\rho(z,t|y,s) + \partial_z[\mu(z,t)\rho(z,t|y,s)] - \frac{1}{2}\partial_{zz}[\sigma(z,t)^2\rho(z,t|y,s)] \right\} dz$$

Since this equation holds for any choice of test function R(z), we obtain the following equation for the transition density, which is called the Fokker-Planck equation:

$$\partial_t\rho(z,t|y,s) = -\partial_z[\mu(z,t)\rho(z,t|y,s)] - \frac{1}{2}\partial_{zz}[\sigma(z,t)^2\rho(z,t|y,s)]$$

Similarly we can derive the Fokker-Planck equation in arbitary dimensions:

$$\frac{\partial p(x,t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}(\mu_i(x,t)p(x,t)) + \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}(D_{ij}(x,t)p(x,t))$$

where $D = \frac{1}{2}\sigma\sigma^T$ is the diffuse tensor.

(2) First order Langevin dynamics:

$$\mu(X_t, t) = \frac{1}{2}\nabla log\ p(\theta_t|X), \quad \sigma(X_t, t) = I$$

According to the Fokker-Planck equation,

$$\begin{aligned}\frac{\partial \rho(\theta_t, t)}{\partial t} &= -\frac{1}{2}\sum_i \frac{\partial}{\partial \theta_i}\nabla log\ p(\theta_t|X)\rho(\theta_t, t) + \frac{1}{2}\sum_i \frac{\partial^2 \rho(\theta_t, t)}{\partial \theta_i^2} \\ &= -\frac{1}{2}\sum_i \frac{\partial}{\partial \theta_i}\left(\frac{\partial p(\theta_t|X)}{\partial \theta_i}\frac{\rho(\theta_t, t)}{p(\theta_t|X)}\right) + \frac{1}{2}\sum_i \frac{\partial^2 \rho(\theta_t, t)}{\partial \theta_i^2}\end{aligned}$$

When $t \to \infty$, $\rho(\theta_t, t) \to \rho(\theta)$, the left-hand side of the equation is equal to 0, so:

$$-\frac{1}{2}\sum_i \frac{\partial}{\partial \theta_i}\left(\frac{\partial p(\theta|X)}{\partial \theta_i}\frac{\rho(\theta)}{p(\theta|X)}\right) + \frac{1}{2}\sum_i \frac{\partial^2 \rho(\theta)}{\partial \theta_i^2} = 0$$

It is obvious that $\rho(\theta) = p(\theta|X)$ is the solution to the partial equation above. So we prove first order Langevin dynamics has the target distribution as its stationary distribution.
Second order Langevin dynamics:

$$\mu(z_t, t) = (D + G)\nabla_z H(z) = [-\partial_r H(z), \partial_x H(z) + C\partial_r H(z)]^T$$
$$\sigma(X_t, t) = \sqrt{2D}$$

where

$$G = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix},\ D = \begin{bmatrix} 0 & 0 \\ 0 & C \end{bmatrix}$$

Similar to the previous proof, $t \to \infty,\ \rho(z_t, t) \to \rho(z)$. According to the second order Langevin dynamics we have:

$$\partial_x(\partial_r H(z)\rho(z)) - \partial_r(\partial_x H(z)\rho(z)) - C\partial_r(\partial_r H(z)\rho(z)) + C\partial_r^2 \rho(z) = 0$$

We can show that $\rho(z) \sim exp(-H(z))$ is the solution to the equation above:

$$\partial_r \rho(z) = \partial_r H(z)\rho(z), \quad \partial_x \rho(z) = \partial_x H(z)\rho(z)$$

So

$$\begin{aligned}&\partial_x(\partial_r H(z)\rho(z)) - \partial_r(\partial_x H(z)\rho(z)) - C\partial_r(\partial_r H(z)\rho(z)) + C\partial_{rr}\rho(z) \\ &= \partial_{rx}H(z)\rho(z) + \partial_r H(z)\partial_x H(z)\rho(z) - \partial_{xr}H(z)\rho(z) - \partial_x H(z)\partial_r H(z)\rho(z) \\ &\quad - C\partial_r\partial_r\rho(z) + C\partial_{rr}\rho(z) = 0\end{aligned}$$

So the stationary distribution $\rho(z) \sim exp(-H(z))$, and $p(z, t) = p(\theta|X)p(r) \sim exp(-H(z))$, so the target distribution is the joint distribution of $\theta|X$ and r.

Problem 2.

(1) HMC with fixed L:

Figure 1: HMC with fixed L=15

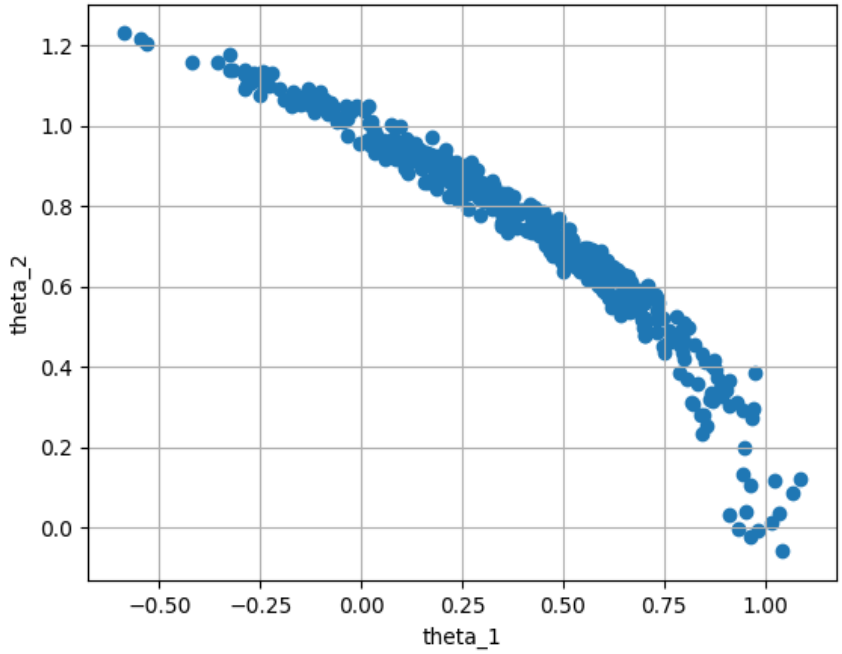


Figure 2: HMC with fixed L=20

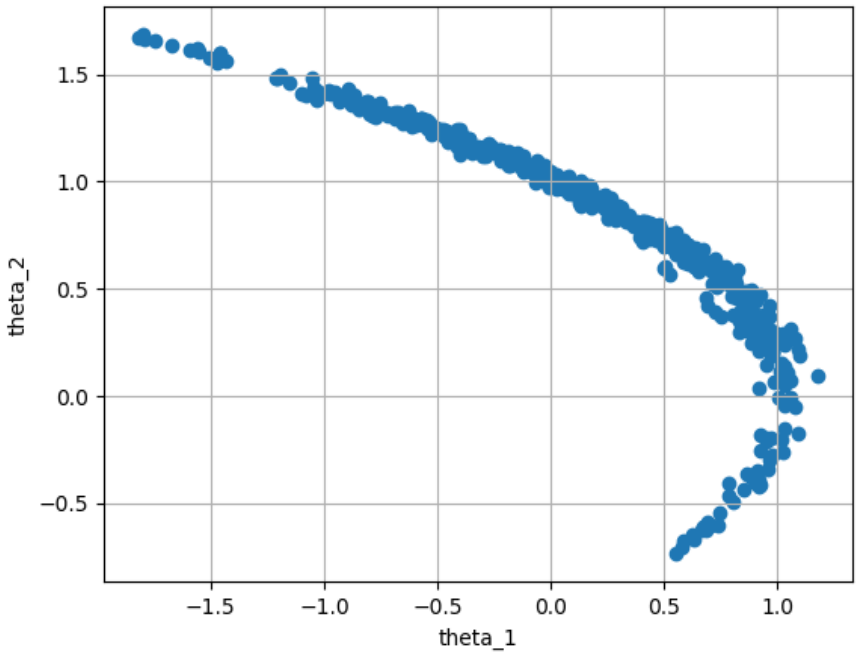
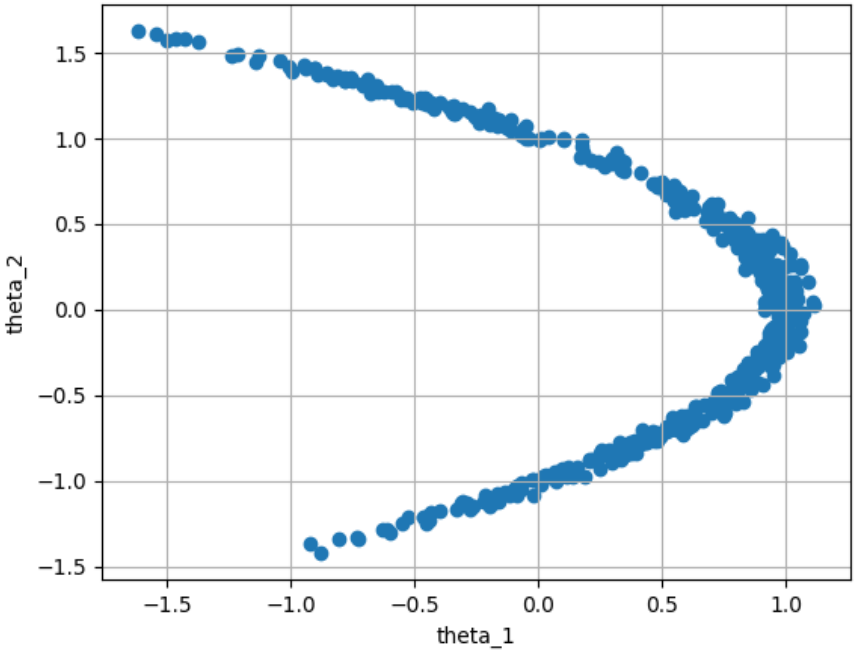


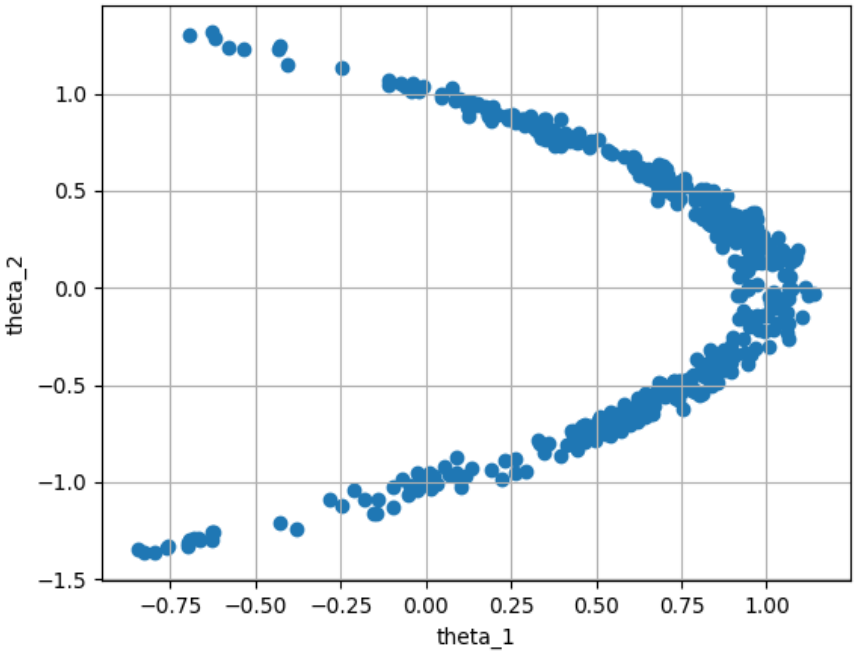
Figure 3: HMC with fixed L=15



As is shown in the three figures above, when L=25, HMC only explore the half of the solution space $\theta_1 + \theta_2^2 = 1$, and with the increase of L, the samples are gradually scattered over the full solution space, yet the time cost is increasing, too.

HMC with random L=Uniform(1,30):

Figure 4: HMC with random L = Uniform(1,30)



We can see that HMC with random L has better performance than the HMC with fixed L even when L=30 (the maximum of the range of uniform distribution), what's more, it costs much less time than the HMC with fixed L=30.
The reason for the difference in their performance is that HMC with fixed L has a rather fixed exploration track, and can be easily trapped at a certain point if the next point after L-step leap frog jump has an extremely high rejection rate, for example, $exp(-H(x', r') + H(x, r)) = 0.01$ or less. So clearly HMC with random L is much cleverer because its randomness can efficiently accelerate its exploration of the whole solution space with little worry about being trapped at certain point.

(2) I implement SGLD, SGHMC and SGNHT with Python, and here are the results.
Note: The performance of SGMCMC is greatly affected by the choice of the initial point, so we initialize $\theta = [-3, 2.2]$
Parameters in common: batch size=50, $e_t = \alpha * (1 + t)^{\beta}$

Figure 5: SGLD (100000 samples), $\alpha = 1e - 4$, $\beta = 1e - 4$

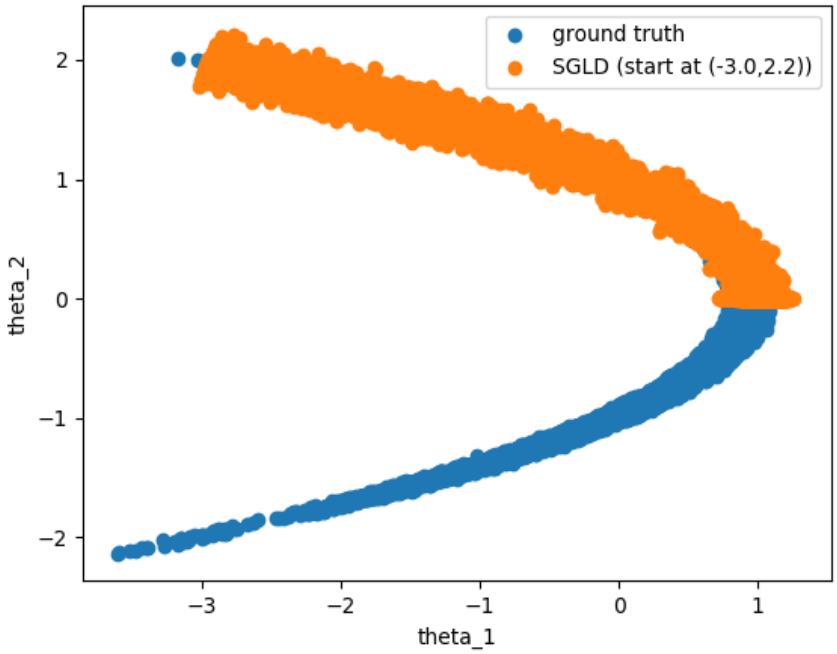


Figure 6: SGHMC (100000 samples), C=20, L=Uniform(0,30), $\alpha = 1e-3$, $\beta = 0.05$

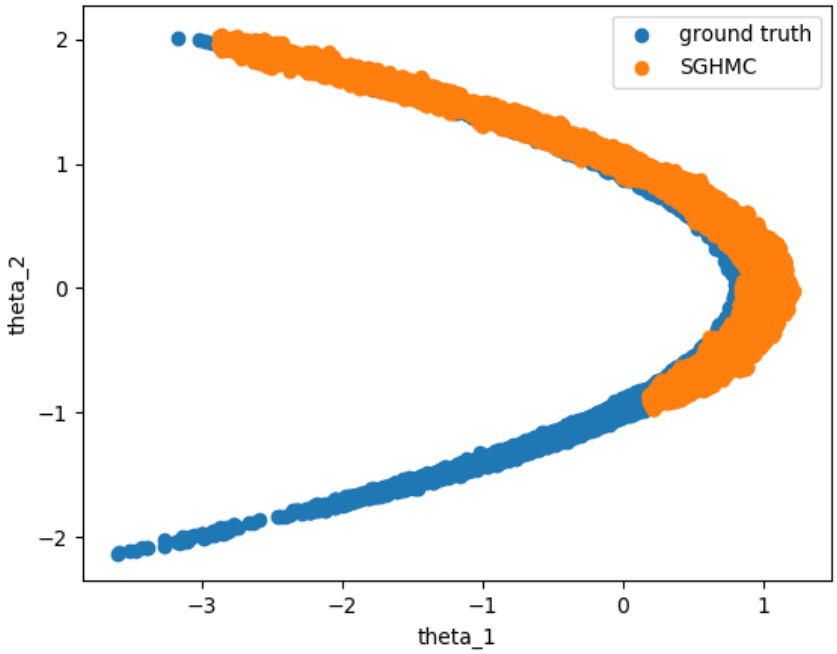


Figure 7: SGNHT (100000 samples), A=100, $\alpha = 1e-3$, $\beta = 0.05$

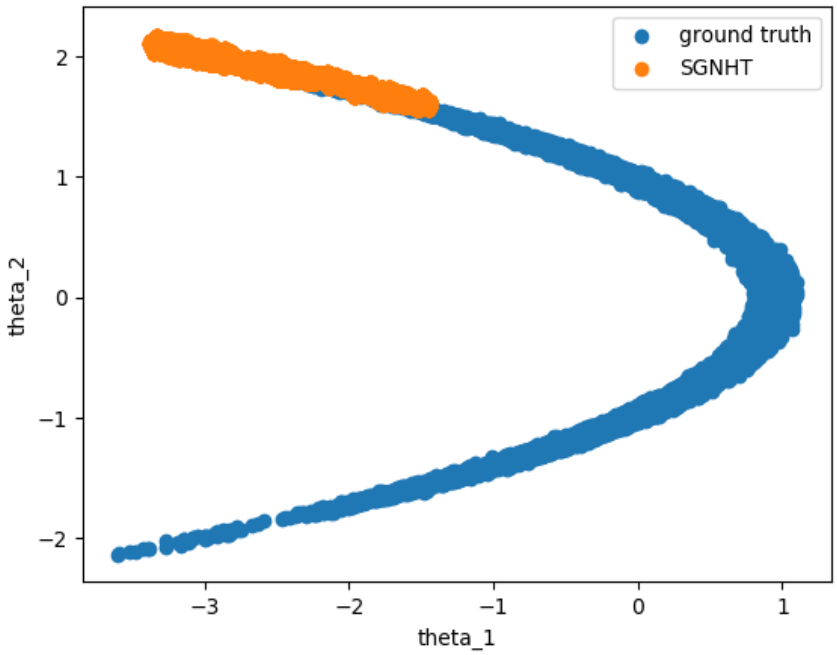
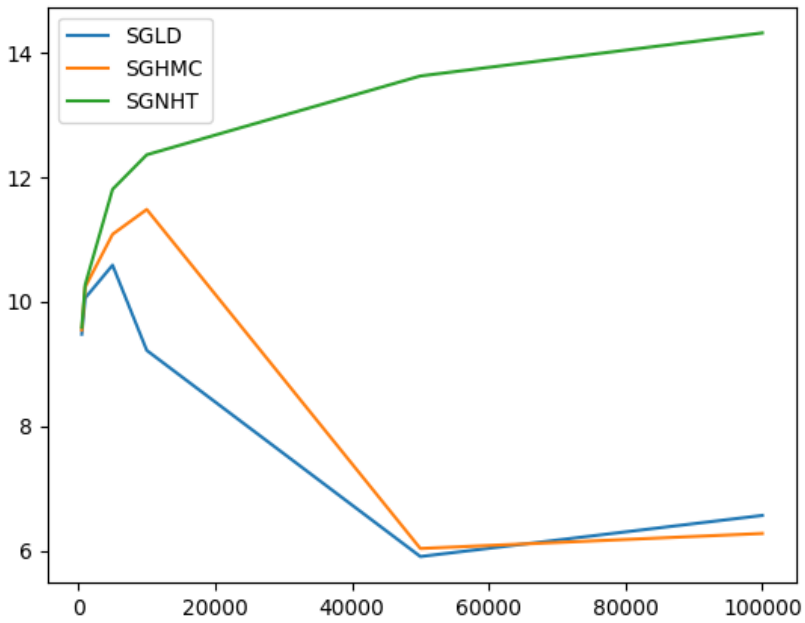


Figure 8: KL divergence vs iteration times



None of the SGMCMC explore the whole solution space, and SGHMC has the best performance among SGLD, SGHMC and SGNHT, yet SGLD and SGNHT are much faster than SGHMC. It might result from the poor choice of hyperparameters.

Problem 3.

(1)

$$l(\theta,\sigma^2) = logp(y,z|x,\theta,\sigma^2,\gamma) = log \prod_{n=1}^{N}\prod_{k=1}^{K}(\gamma_k^T x_n)^{Z_{nk}}\mathcal{N}(\mu_k,\sigma_k^2)^{Z_{nk}}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{nk} log\ \gamma_k^T x_n + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{nk}(log\ \frac{1}{\sigma_k} - \frac{(y_n-\theta_k^T x_n)^2}{2\sigma_k^2}) + C$$

$$\frac{\partial l}{\partial \theta_k} = \sum_{n=1}^{N}\frac{Z_{nk}}{\sigma_k^2}(y_n-\theta_k^T x_n)x_n,\ \ \frac{\partial l}{\partial \sigma_k} = \sum_{n=1}^{N}\frac{Z_{nk}}{\sigma_k^3}((y_n-\theta_k^T x_n)^2-\sigma_k^2)$$

So the MLE of the model parameters $\hat{\theta_k},\ \hat{\sigma_k}$ satisfy the equations:

$$\sum_{n=1}^{N} Z_{nk}\hat{\theta_k}^T x_n \cdot x_n = \sum_{n=1}^{N} Z_{nk} y_n \cdot x_n$$

$$\sum_{n=1}^{N} Z_{nk}\sigma_k^2 = \sum_{n=1}^{N} Z_{nk}(y_n - \hat{\theta_k}^T x_n)^2$$

As is shown above, $\hat{\theta_k}$ is the solution to the linear equation:

$$\sum_{i=1}^{p}\sum_{n=1}^{N} Z_{nk}x_{ni}x_{nj}\theta_{ki} = \sum_{n=1}^{N} Z_{nk}y_n x_{nj},\ j = 1, 2\cdots p$$

and if ($\sum_{n=1}^{N} Z_{nk} \neq 0$)

$$\hat{\sigma_k}^2 = \frac{\sum_{n=1}^{N} Z_{nk}(y_n - \hat{\theta_k}^T x_n)^2}{\sum_{n=1}^{N} Z_{nk}}$$

otherwise $\hat{\sigma_k}$ can be an arbitrary real number > 0

(2)

$$\begin{aligned} l(\theta, \sigma^2) &= log\ \sum_{n=1}^{N} p(y, z_n|x, \theta, \sigma^2, \gamma) = log\ \sum_{n=1}^{N} q(z_n)\frac{p(y, z_n|x, \theta, \sigma^2, \gamma)}{q(z_n)} \\ &\geq \sum_{n=1}^{N} q(z_n) log\ \frac{p(y, z_n|x, \theta, \sigma^2, \gamma)}{q(z)} = \sum_{n=1}^{N} q(z_n)p(y, z_n|x, \theta, \sigma^2, \gamma) - \sum_{n=1}^{N} q(z) log\ q(z) \end{aligned}$$

Let $\Theta = (\theta, \sigma^2, \gamma)$, and the lower bound is

$$\begin{aligned} \mathcal{F}(q, \Theta) &= \sum_{n=1}^{N} q(z_n)p(y, z_n|x, \Theta) - \sum_{n=1}^{N} q(z) log\ q(z) \\ &= \sum_{n=1}^{N} q(z_n) log\ \frac{p(z_n|y, x, \Theta)p(y|x, \Theta)}{q(z)} \\ &= \sum_{n=1}^{N} q(z_n) log\ \frac{p(z_n|y, x, \Theta)}{q(z)} + log\ p(y|x, \Theta) \\ &= \mathcal{L}(\Theta) - D_{KL}(q(z)||p(z|x, \Theta)) \\ &\leq \mathcal{L}(\Theta) \end{aligned}$$

when $q^{(t)}(z) = p(z|x, \Theta^{(t)})$, $\mathcal{F}(q, \Theta)$ is maximized and $\mathcal{F}(q, \Theta^{(t)}) = \mathcal{L}(\Theta^{(t)})$, so $\mathcal{F}(q, \Theta)$ is locally optimal.

(3) E-step:

$$p(z_n = k|y_n, x_n, \Theta^{(t)}) = \frac{\gamma_k^T x_n \mathcal{N}(y_n|\theta_k^{(t)T} x_n, \sigma_k^{(t)})}{\sum_k \gamma_k^T x_n \mathcal{N}(y_n|\theta_k^{(t)} x_n, \sigma_k^{(t)})}$$

Denote $\phi_{n,k}^{(t)} = p(z_n = k|y_n, x_n, \Theta^{(t)})$, $\sum_k \phi_{n,k}^{(t)} = 1$

$$\begin{aligned} Q^{(t)}(\Theta) &= \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|y_n, x_n, \Theta^{(t)})\ log\ p(y_n, z_n = k|\Theta^{(t)}) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{n,k}^{(t)}\ (log\ \gamma_k^T x_n + log\ \mathcal{N}(y_n|\theta_k^{(t)T} x_n, \sigma_k^{(t)})) \\ &= \sum_{k=1}^{K}\sum_{n=1}^{N} \phi_{n,k}^{(t)}(\ log\ \gamma_k^T x_n - \frac{1}{2}log(2\pi) - log\ \sigma_k - \frac{(y_n - \theta_k^T x_n)^2}{2\sigma_k^2}) \end{aligned}$$

M-step:

Denote $\theta_k = [\theta_{k1}, \theta_{k2}, \cdots \theta_{ki} \cdots \theta_{kp}]$

$$\frac{\partial Q^{(t)}(\Theta)}{\partial \theta_{ki}} = \sum_{n=1}^{N} \phi_{n,k}^{(t)} \frac{1}{\sigma_k^2} x_{ni}(y_n - \sum_{j=1}^{p} x_{nj}\theta_{kj})$$

So $\theta_k^{(t+1)}$ is the solution to the linear equation:

$$\sum_{j=1}^{p}(\sum_{n=1}^{N} \phi_{n,k}^{(t)} x_{ni} x_{nj})\theta_{kj} = \sum_{n=1}^{N} \phi_{n,k}^{(t)} x_{ni} y_n, \quad i = 1, 2, \cdots p$$

$$\frac{\partial Q^{(t)}(\Theta)}{\partial \sigma_k} = \sum_{n=1}^{N} \phi_{n,k}(-\frac{1}{\sigma_k} + \frac{(y_n - \theta_k^T x_n)^2}{\sigma_k^3})$$

So

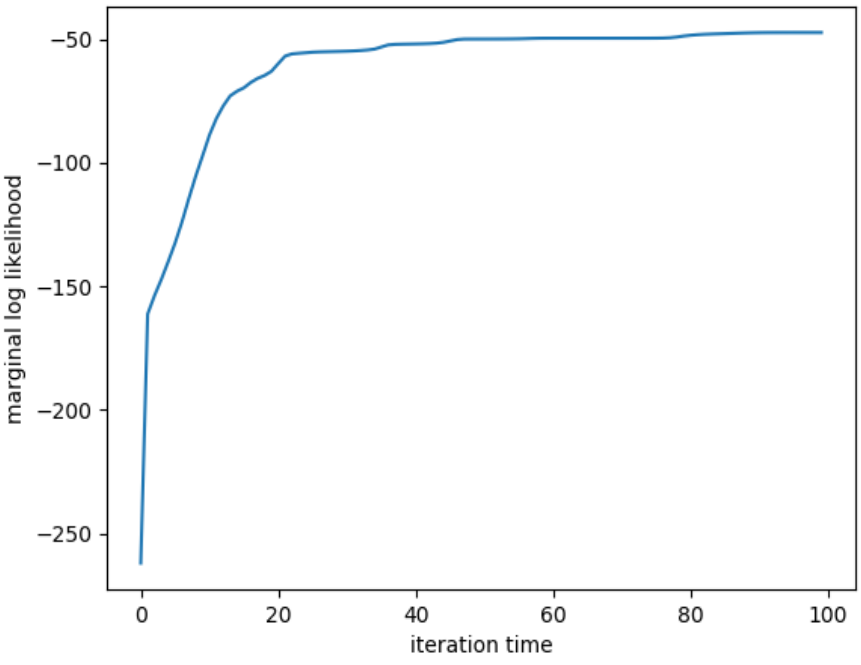
$$(\sigma_k^{(t+1)})^2 = \frac{\sum_{n=1}^{N} \phi_{n,k}(y_n - \theta_k^{(t+1)T} x_n)^2}{\sum_{n=1}^{N} \phi_{n,k}}$$

(4) We use the maximized lower boundary of marginal log likelihood

$$F(q, \Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|x_n, y_n, \Theta) log \ \frac{p(y_n, z_n|x_n, \Theta)}{p(z_n = k|x_n, y_n, \Theta)}$$

to calculate $p(y|x, \Theta)$

Figure 9: Marginal log likelihood vs t



As shown in figure 9, $p(y^{(t)}|x, \Theta^{(t)})$ increased with iteration time, and after about 40 iterations $p(y^{(t)}|x, \Theta^{(t)})$ converged to -50, indicating that EM had converged to a local maximum and failed to find the global maximum.