3.2 Preconditioning

When the posterior distribution of \( \{\theta_\alpha\} \) has different scales along different variables, the original LMC with a common step size for all \( \theta_\alpha \)'s will traverse the parameter space slowly. We adopt a preconditioning matrix \( C \) to speed up the mixing, where \( C \) satisfies \( CC^T = H \) with \( H \) is the Hessian matrix of \( \log P(\theta_{MAP}|D) \), computed as:

\[
H(\theta_{MAP}) = \text{COV}_{P(\theta|\theta_{MAP})} f(x) + \sigma_0^{-2} \tag{6}
\]

This is reminiscent of the observed Fisher information matrix in Girolami and Calderhead (2011) except that we use the MAP estimate with the prior. We approximate \( H(\theta_{MAP}) \) by averaging over \( H(\theta_t) \) during a burn-in period and estimate \( \text{COV}_{P(\theta|\theta_t)} f \) with the set of state samples from the persistent Markov chains. The adoption of a preconditioning matrix also helps us pick a common step size parameter \( \varepsilon \) suitable for different training sets.

3.3 Partial Momentum Refreshment

The momentum term \( p \) in the leapfrog step represents the update direction of the parameter. Langevin dynamics is known to explore the parameter space through inefficient random walk behavior because it draws an independent sample for \( p \) at every iteration. We can achieve a better mixing rate with the partial momentum refreshment method proposed in Horowitz (1991). When \( p \) is updated at every step by:

\[
p_t \leftarrow \alpha p_t + \beta n_t \tag{7}
\]

where \( n_t \sim \mathcal{N}(0, I) \), and \( \alpha, \beta \) satisfy \( \alpha^2 + \beta^2 = 1 \), the momentum is partially preserved from the previous iteration and thereby suppresses the random-walk behavior in a similar fashion as HMC with multiple leapfrog steps.

\( \alpha \) controls how much momentum to be carried over from the previous iteration. With a large value of \( \alpha \), LMC reduces the auto-correlation between samples significantly relative to LMC without partial momentum refreshment. The improved mixing rate is illustrated in Figure 2. We also show that the mean and standard deviation of the posterior distribution does not change. However, caution should be exercised especially when the step size \( \eta \) is large because a value of \( \alpha \) that is too large would increase the error in the update equation which we do not correct with a Metropolis-Hastings step because that is intractable.

4 Sampling Edges by Reversible Jump MCMC

Langevin dynamics handles the continuous change in the parameter value \( A_\alpha \). As for discrete changes in the model structure, \( Y_\alpha \), we propose an approximate reversible jump MCMC (RJMCMC) step (Green, 1995) to sample \( Y_\alpha \) and \( A_\alpha \) jointly from the conditional distribution \( P(A, Y|p_0, \sigma_0, D) \). The proposed Markov chain adds/deletes one clique (or simply one edge when \( \alpha = (i, j) \)) at a time. When an edge does not exist, i.e., \( Y_\alpha = 0 \), the variable \( A_\alpha \) can be excluded from the model, and therefore we consider the jump between a full model with \( Y_\alpha \neq 0, A_\alpha = \alpha \) and a degenerate model with \( Y_\alpha = 0 \).

The proposed RJMCMC is as follows: when \( Y_\alpha = 0 \), propose adding an edge with probability \( P_{add} \) and sample \( A_\alpha = \alpha \) from a proposal distribution \( q(A) \) with support on \( [-\Delta_\alpha, \Delta_\alpha] \); when \( Y_\alpha = 1 \) and \( |A_\alpha| \leq \Delta_\alpha \), propose deleting an edge with \( P_{del} \). The reason of restricting the proposed move within \( [-\Delta_\alpha, \Delta_\alpha] \) will be explained later. It is easy to see that the Jacobian is 1. The jump is then accepted by the Metropolis-Hastings algorithm with a probability:

\[
Q_{add} = \min\{1, Q^*(\alpha)\}, \quad Q_{del} = \min\{1, 1/Q^*(\alpha)\}
\]

\[
Q^*(\alpha) = \exp\left(\alpha \sum_m f_\alpha(x^{(m)}) \right) \frac{Z(Y_\alpha = 0)}{Z(Y_\alpha = 1, A_\alpha = \alpha)} \right)^N \frac{p_0 N(A_\alpha = a|0, \sigma_0^2)}{(1 - p_0)} \frac{P_{del}}{P_{add} q(A_\alpha = \alpha)} \tag{8}
\]

The factors in the first line of \( Q^* \) represent the ratio of the model likelihoods while the other two are respectively the ratio of the prior distributions and the ratio of the proposal distributions.

4.1 Unbiased Estimate to \( Q^* \) and \( 1/Q^* \)

Computing the partition functions in equation 8 is generally intractable. However, noticing that \( Z(Y_\alpha = 1, A_\alpha = \alpha) \rightarrow Z(Y_\alpha = 0) \), as \( \alpha \rightarrow 0 \), the log-ratio of the two partition functions should be well approximated by a quadratic approximation at the origin when \( \alpha \) is small. In this way