HOMEWORK 4
MONTE CARLO METHODS AND BAYESIAN MODELING[1]

10-708 PROBABILISTIC GRAPHICAL MODELS (SPRING 2021)
http://708.mlcourse.org

OUT: March 24, 2021
DUE: April 7, 2021 at 11:59 PM
TAs: Alex, Helen, Xiang

START HERE: Instructions

Summary In this assignment, you will implement a Metropolis Hastings algorithm on English Premier League data. Section A will help you develop a better understanding of similar sampling methods through some warm-up problems. Then, in Section B, you will build on this knowledge to build a topic model that discovers topics underlying a set of documents.

- **Collaboration policy:** The purpose of student collaboration is to facilitate learning, not to circumvent it. Studying the material in groups is strongly encouraged. It is also allowed to seek help from other students in understanding the material needed to solve a particular homework problem, provided no written notes (including code) are shared, or are taken at that time, and provided learning is facilitated, not circumvented. The actual solution must be done by each student alone. The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved. See the Academic Integrity Section on the course site for more information: http://www.cs.cmu.edu/~mgormley/courses/10708/about.html#7-academic-integrity-policies
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10708/about.html#6-general-policies
- **Submitting your work to Gradescope:** We use Gradescope to collect PDF submissions of open-ended questions on the homework (e.g. mathematical derivations, plots, short answers). The course staff will manually grade your submission, and you'll receive feedback explaining your final marks. You will also submit your code for programming questions on the homework to Gradescope (https://www.gradescope.com/courses/228238). We will manually grade your code for completeness.
- For **multiple choice** or **select all that apply questions**, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeX users, replace `\choice` with `\CorrectChoice` to obtain a shaded box/circle, and don't change anything else.

For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeX users, replace `\choice` with `\CorrectChoice` to obtain a shaded box/circle, and don't change anything else.

[1] Compiled on Friday 26th March, 2021 at 05:14

A Written Questions [103 pts]

Answer the following questions in the template provided. Then upload your solutions to Gradescope. You may use LaTeX or print the template and hand-write your answers then scan it in. Failure to use the template may result in a penalty. There are 103 points and 21 questions.

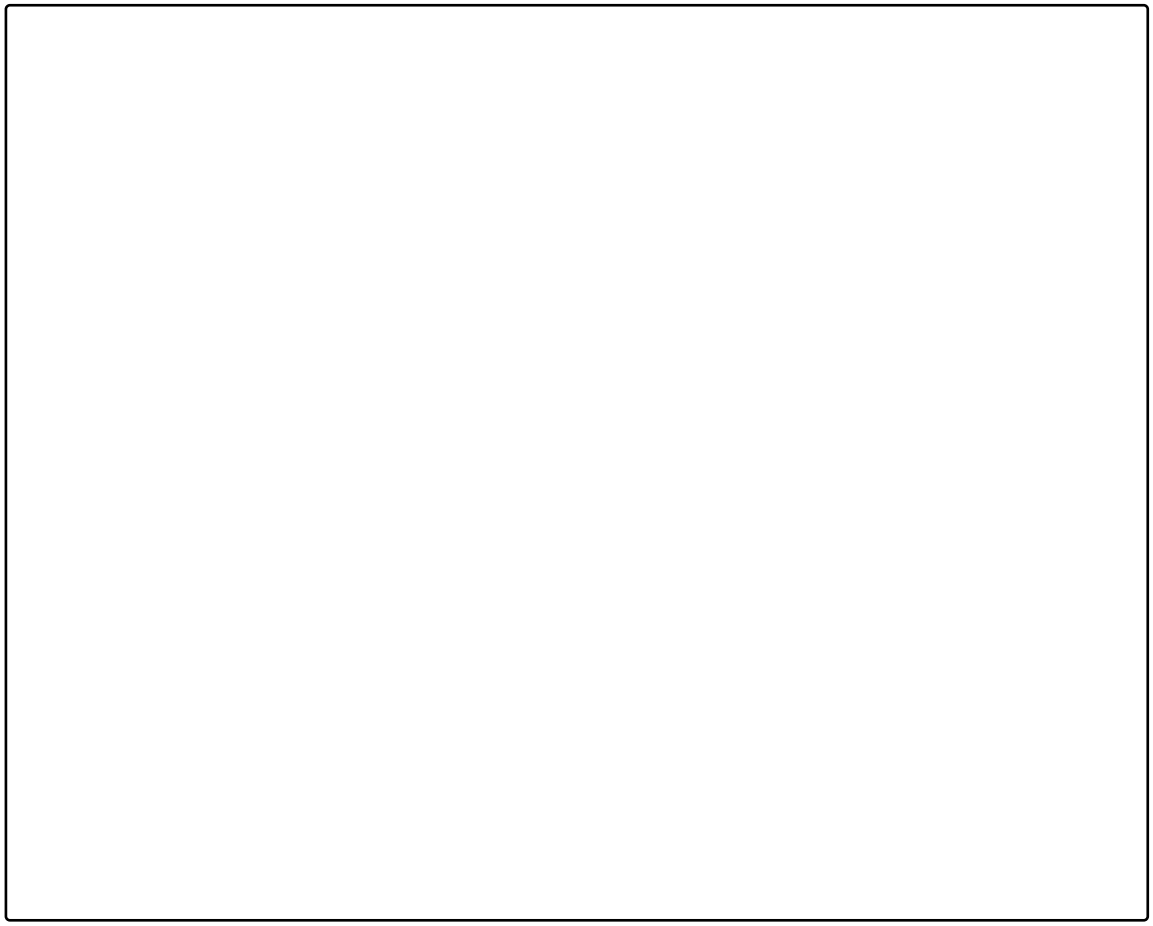
A.1 Computational Complexity of Marginal Inference

1. (5 points) Show that for marginal inference problems, computing the marginals of variables and cliques

$$p(x_i) = \sum_{\mathbf{x}' : x'_i = x_i} p(x'|\theta)$$

$$p(x_C) = \sum_{\mathbf{x}' : x'_C = x_C} p(x'|\theta)$$

is a #P-hard problem. (Hint: reduce #-SAT to the marginal inference problem)



A.2 Markov Chain Monte Carlo Methods

2. In class, we studied two Monte Carlo estimation methods: rejection sampling and importance sampling. Given a proposal distribution $Q(x)$, answer the following questions:

   (a) (1 point) If sampling from $Q(x)$ is computationally expensive, which of the following methods is likely to be more efficient?

   ◯ Rejection Sampling

   ◯ Importance Sampling

   ◯ Both are equally inefficient

   (b) (1 point) If $Q(x)$ is high-dimensional, which of the following methods is more efficient?

   ◯ Rejection Sampling

   ◯ Importance Sampling

   ◯ Both are equally inefficient

   (c) (1 point) For high-dimensional distributions, MCMC methods such as Metropolis Hastings are more efficient than rejection sampling and importance sampling.

   ◯ True

   ◯ False

   (d) (1 point) For low-dimensional distributions, MCMC methods such as Metropolis Hastings produce better samples than rejection sampling and importance sampling.

   ◯ True

   ◯ False

3. (2 points) Suppose you are using MCMC methods to sample from a distribution with multiple modes. Briefly explain what complications may arise while using MCMC.

A.3 Monte Carlo

Let $p(x)$ be a distribution on $x = [x_1, ..., x_D]^T \in \mathbb{R}^D$. Suppose we want to perform inference $\mathbb{E}_{p(x)}[f(x)]$ using importance sampling, with $q(x)$ as the proposal distribution. We draw L i.i.d. samples $x^{(1)}, ..., x^{(L)}$ from $q(x)$, and can estimate using importance sampling:

$$\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{\sum_{i=1}^{L} u_i} \sum_{i=1}^{L} f(x^{(i)}) u_i$$

where $u_i = \frac{p(x^{(i)})}{q(x^{(i)})}$ are the (unnormalized) importance weights.

4. For the following questions, your solution should be in its simplest form. Show your work.

   (a) (2 points) Find the mean of the unnormalized importance weights $\mathbb{E}_{q(x)}[u_i]$.

   (b) (3 points) Find the variance of the unnormalized importance weights $\text{Var}_{q(x)}[u_i]$.

5. (5 points) Prove the following lemma: $\mathbb{E}_{p(x)}\left[\frac{p(x)}{q(x)}\right] \geq 1$, and the equality holds only when $p = q$. (**Hint**: you may use without proof the inequality $y \geq \log(y) + 1$ for all $y > 0$. You may also use the fact that when the KL divergence between two distributions is 0, the two distributions are identical.)

6. (6 points) A measure of the variability of two components in vector $u = [u_1, ..., u_L]^T$ is given by $\mathbb{E}_{q(x)}\left[(u_i - u_j)^2\right]$. Assume that both p and q can be factorized, i.e. $p(x) = \prod_{i=1}^{D} p_i(x_i)$, and $q(x) = \prod_{i=1}^{D} q_i(x_i)$. Explaining each step with words, show that $\mathbb{E}_{q(x)}\left[(u_i - u_j)^2\right]$ has exponential growth with respect to D.

   Specifically, show that $\mathbb{E}_{q(x)}\left[(u_i - u_j)^2\right] = 2\prod_{d=1}^{D} \mathbb{E}_{p_d(x)}\left[\frac{p_d(x)}{q_d(x)}\right] - 2$.

7. (1 point) Use the conclusion in (c) to explain why the standard importance sampling does not scale well with dimensionality and would blow up in high-dimensional cases.

A.4 Gibbs Sampling

Suppose you wish to build a Gibbs sampler for a Bayesian hidden Markov model (HMM). For a sequence of length L, we have observations $x_1, \ldots, x_L$ with $x_i \in \{1, \ldots, W\}$ and latent states $y_1, \ldots, y_L$ with $y_i \in \{1, \ldots, T\}$. The model has the usual emission parameters $\mathbf{A} \in \mathbb{R}^{T \times T}$ and transition parameters $\mathbf{B} \in \mathbb{R}^{T \times W}$; we make it a touch more Bayesian by placing Dirichlet priors over the rows of each with respective hyperpameters $\boldsymbol{\alpha} \in \mathbb{R}^T$ and $\boldsymbol{\beta} \in \mathbb{R}^W$. The data consists of a single sequence[2] with a fixed start symbol $y_0 = \text{START}$ so that the data likelihood can be written as,

$$p(\mathbf{x}, \mathbf{y} \mid \mathbf{A}, \mathbf{B}) = \prod_{i=1}^{L} p(y_i | y_{i-1}, \mathbf{A}) p(x_i | y_i, \mathbf{B})$$

The generative story for our Bayesian HMM is as follows.

$$\begin{aligned}
&\text{for } t \in \{1, \ldots, T\} \\
&\quad \mathbf{A}_{t,\cdot} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
&\quad \mathbf{B}_{t,\cdot} \sim \text{Dirichlet}(\boldsymbol{\beta}) \\
&\text{for } i \in \{1, \ldots, L\} \\
&\quad y_i \sim \text{Categorical}(\mathbf{A}_{y_{i-1},\cdot}) \\
&\quad x_i \sim \text{Categorical}(\mathbf{B}_{y_i,\cdot})
\end{aligned}$$

You wish to implement an explicit blocked Gibbs sampler that performs unsupervised inference of the latent states $\mathbf{y}$ and the parameters $\mathbf{A}$ and $\mathbf{B}$. The hyperameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed. You sketch out the high level function below.

```
1: procedure BayesHmmGibbs(x, α, β)
2:     y, A, B ← initialize(x, α, β)                    ▷ Initialize the random variables
3:     for t ∈ shuffle({1, ..., T}) do                  ▷ For each tag type
4:         A_{t,·} ← sample_transition(t, x, y, A, B, α, β)   ▷ Resample the tth row of the transition matrix
5:         B_{t,·} ← sample_emission(t, x, y, A, B, α, β)     ▷ Resample the tth row of the emission matrix
6:     for i ∈ shuffle({1, ..., L}) do                  ▷ For each token
7:         y_i ← sample_state(i, x, y, A, B, α, β)          ▷ Resample the ith token
```

You now proceed to implement the remaining methods. Some notes: Your solutions below may assume access to the function sample_dir($\mathbf{a}$) which draws a sample from a Dirichlet parameterized by $\mathbf{a}$ and the function sample_cat($\mathbf{p}$) which draws a sample from a Categorical parameterized by the probabilities $\mathbf{p}$ (and it elegantly handles proportions $\mathbf{p}$ as well). For full credit, your solutions should use the minimal amount of computation necessary.

8. (3 points) Write pseudocode for the function sample_transition$(t, \mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

9. (3 points) Write pseudocode for the function sample_emission$(t, \mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

10. (4 points) Write pseudocode for the function sample_state$(i, \mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

A.5 Metropolis Hastings

11. (4 points) Recall that conjugate priors lead to the posterior belonging to the same probability distribution family as the prior. Show that the beta distribution is the conjugate prior for the binomial distribution.

12. (5 points) Show that the Metropolis Hastings algorithm satisfies detailed balance.

13. (5 points) Show that a variant of the Metropolis Hastings algorithm with the acceptance probability defined without the minimum, ie acceptance probability $a = A(x \leftarrow x_i) = \frac{p(x)q(x_i|x)}{p(x_i)q(x|x_i)}$, wouldn't satisfy the detailed balance.

B Programming [50 pts]

B.1 Task Background

Nowadays, statistical modelling of sport data has become an important part of sports analytics and is often a critical reference for the managers in their decision-making process. In this part, we will work on a real world example in professional sports. Specifically, we are going to use the data from the 2013-2014 Premier League, the top-flight English professional league for men's football (soccer, not American football) clubs, and build a predictive model on the number of goals scored in a single game by the two opponents. Bayesian hierarchical model is a good candidate for this kind of modeling task. We model each team's strength (both attacking and defending) as latent variables. Then in each game, the goals scored by the home team is a random variable conditioned on the attacking strength of the home team and the defending strength of the away team. Similarly, the goals scored by the away team is a random variable conditioned on the attack strength of the away team and the defense strength of the home team. Therefore, the distribution of the scoreline of a specific game is dependent on the relative strength between the home team A and the away team B, which also depends on the relative strength between those teams with their other opponents.

Table B.1: 2013-2014 Premier League teams

Index	0	1	2	3	4
Team	Arsenal	Aston Villa	Cardiff City	Chelsea	Crystal Palace
Index	5	6	7	8	9
Team	Everton	Fulham	Hull City	Liverpool	Manchester City
Index	10	11	12	13	14
Team	Manchester United	Newcastle United	Norwich City	Southampton	Stoke City
Index	15	16	17	18	19
Team	Sunderland	Swansea City	Tottenham Hotspur	West Bromwich Albion	West Ham United

Here we consider using the same model as described by Baio and Blangiardo (2010). The Premier League has 20 teams, and we index them as in Table B.1. Each team would play 38 matches every season (playing each of the other 19 teams home and away), which totals 380 games in the entire season. For the g-th game, assume that the index of home team is $h(g)$ and the index of the away team is $a(g)$. the observed number of goals is:

$$y_{gj} \mid \theta_{gj} = \text{Poisson}(\theta_{gj})$$

where the $\theta = (\theta_{g1}, \theta_{g2})$ represent the scoring intensity in the g-th game for the team playing at home ($j = 1$) and away ($j = 2$), respectively. We put a log-linear model for the θs:

$$\log \theta_{g1} = home + att_{h(g)} - def_{a(g)}$$
$$\log \theta_{g2} = att_{a(g)} - def_{h(g)}$$

Note that team strength is broken into attacking and defending strength. And $home$ represents home-team advantage, and in this model is assumed to be constant across teams. The prior on the home is a normal distribution

$$home \sim \mathcal{N}(0, \tau_0^{-1})$$

where the precision $\tau_0 = 0.0001$.

The team-specific attacking and defending effects are modeled as exchangeable:

$$att_t \sim \mathcal{N}(\mu_{att}, \tau_{att}^{-1})$$

$$def_t \sim \mathcal{N}(\mu_{def}, \tau_{def}^{-1})$$

We use conjugate priors as the hyper-priors on the attack and defense means and precisions:

$$\mu_{att} \sim \mathcal{N}(0, \tau_1^{-1})$$
$$\mu_{def} \sim \mathcal{N}(0, \tau_1^{-1})$$
$$\tau_{att} \sim \text{Gamma}(\alpha, \beta)$$
$$\tau_{def} \sim \text{Gamma}(\alpha, \beta)$$

where the precision $\tau_1 = 0.0001$, and we set parameters $\alpha = \beta = 0.1$.

This hierarchical Bayesian model can be represented using a directed acyclic graph as shown in Figure B.1. Where the goals of each game $\mathbf{y} = \{y_{gj} | g = 0, ..., 379, j = 1, 2\}$ are 760 observed variables, and parameters $\boldsymbol{\theta} = (home, att_0, ..., att_{19}, def_0, ..., def_{19})$ and hyper-parameters $\boldsymbol{\eta} = (\mu_{att}, \mu_{def}, \tau_{att}, \tau_{def})$ are unobserved variables that we need to make inference. To ensure identifiability, we enforce a corner constraint on the parameters (pinning one team's parameters to 0,0). Here we use the first team as reference and assign its attacking and defending strength to be 0:

$$att_0 = def_0 = 0$$

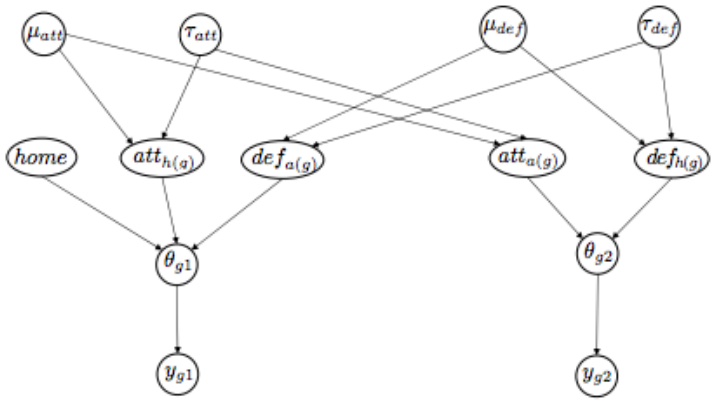


Figure B.1: The DAG representation of the hierarchical Bayesian model. Figure adapted from Baio and Blangiardo (2010).

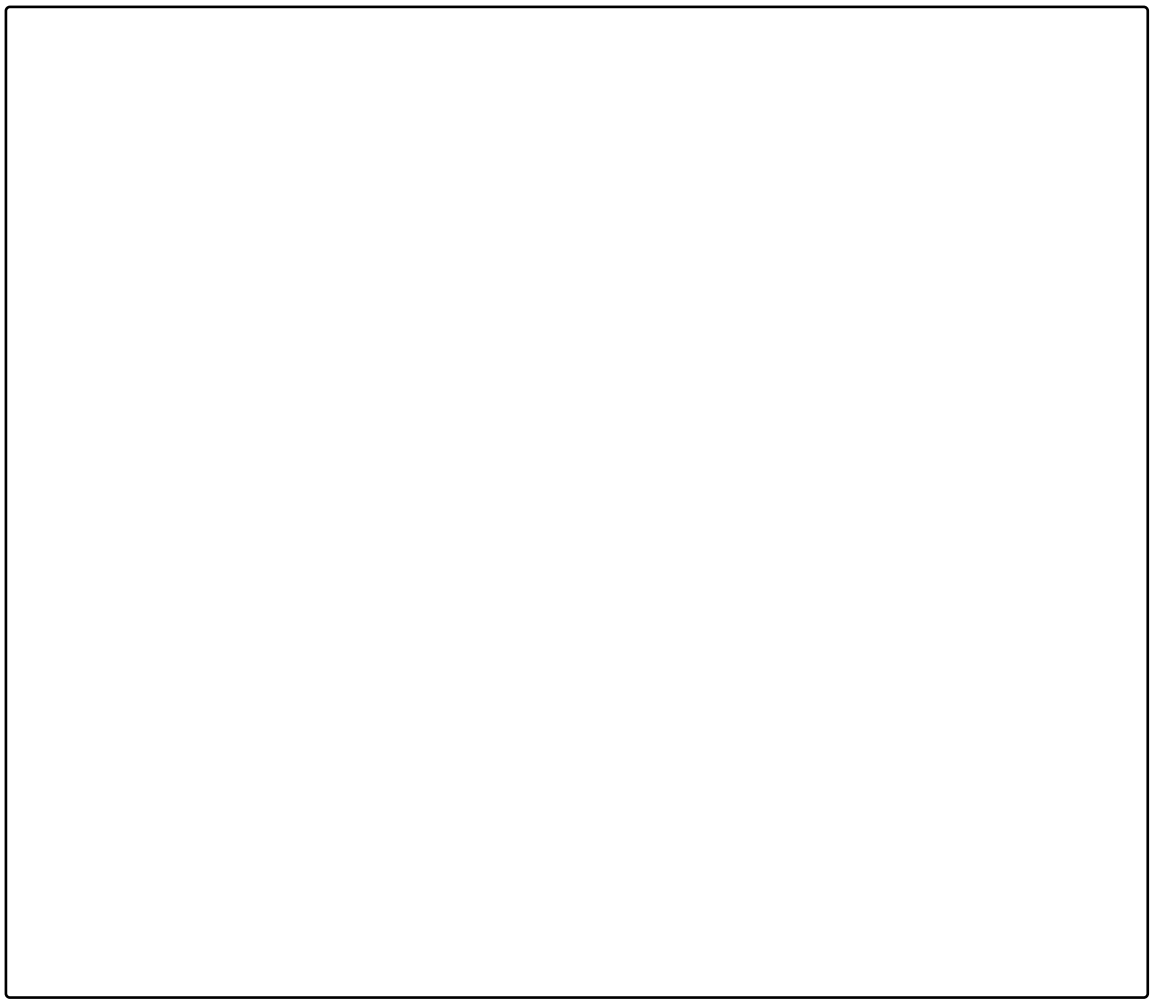
In this question, we want to estimate the posterior mean of the attacking and defending strength for each team, i.e. $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[att_i]$, $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[def_i]$, and $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[home]$.

14. (5 points) Find the joint likelihood $p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta})$.

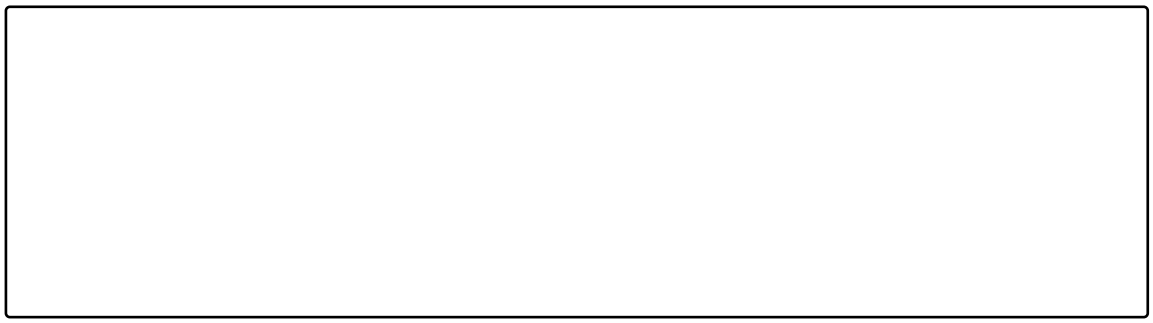
15. (5 points) Write down the Metropolis-Hastings algorithm for sampling from posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y})$, and derive the acceptance function for a proposal distribution of your choice (e.g. isotropic Gaussian).

16. (20 points) Implement the M-H algorithm to inference the posterior distribution. The data can be found from `premier_league_2013_2014.dat`, which contains a 380×4 matrix. The first column is the number of goals y_{g1} scored by the home team, the second column is the number of goals y_{g2} scored by the away team, the third column is the index for the home team $h(g)$, and the fourth column is the index for the away team $a(g)$. Use isotropic Gaussian as proposal distribution, $\mathcal{N}(0, \sigma^2 I)$, use 0 as the starting point. **Note: You are NOT allowed to use any existing implementations of M-H in this problem. Please submit your implementation to Gradescope separately.**

17. (12 points) Run the MCMC chain for 5000 steps to burn in and then collect 5000 samples with t steps in between (i.e., run M-H for $5000t$ steps and collect only each t-th sample). This is called *thinning*, which reduces the autocorrelation of the MCMC samples introduced by the Markovian process. The parameter sets are $\sigma = 0.005, 0.05, 0.5$, and $t = 1, 5, 20, 50$. Plot the trace plot of the burn-in phase and the MCMC samples for the latent variable $home$ using proposal distributions with different σ and t. (There should be 12 plots in total.)



18. (2 points) Comment on the results. Which parameter setting worked the best for the algorithm?



In the following questions, use the results from the optimal parameter setting.

19. (3 points) Plot the posterior histogram of $home$ from the MCMC samples.

20. (3 points) Plot the estimated attacking strength $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[att_i]$ against the estimated defending strength $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[def_i]$ for each the team in one scatter plot. Please make sure to identify the team index of each point on your scatter plot.

B.2 Wrap-up Questions

21. (1 point) **Multiple Choice:** Did you correctly submit your code to Autolab?

    ◯ Yes

    ◯ No

B.3 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies for this course.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.

References

[1] Baio, Gianluca, and Marta Blangiardo. ”Bayesian hierarchical model for the prediction of football results.” Journal of Applied Statistics 37.2 (2010): 253-264.