Chalmers University of Technology

SSY316

Take Home Exam

Advanced probabilistic machine learning

Author 1:
Hu Xiaolingzi
980326-8805
huxia@student.chalmers.se

Author 2:
Liang Zhitao
zhitao@chalmers.se

January 12, 2024

1 Principal Component Analysis of Genomes

1.1

995. The dimension would be the minimum of the numbers of samples and features and here the data shape is $(995, 10101)$.

1.2

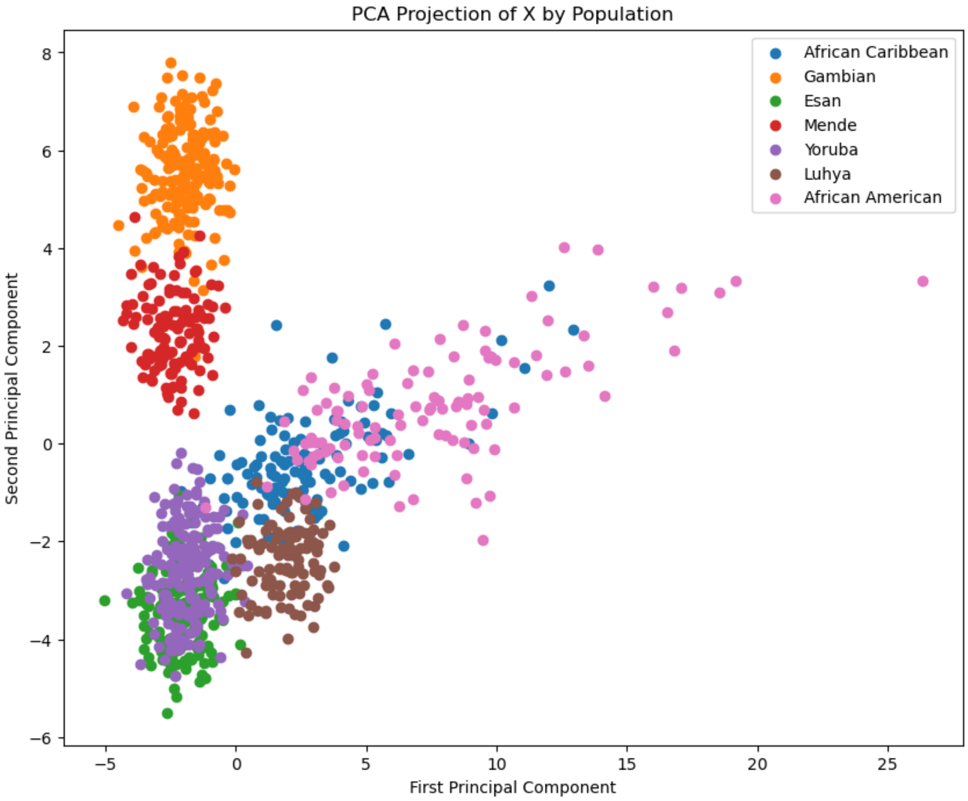


Figure 1

1.3

1.3.1

From the plot, we can tell that Gambian and Mende are quite genetically similar, forming a cluster here, which corresponds to their geographical closeness. Moreover, both African Carribean and African American hold a large genetic variance, which might be caused from the adaptation and mutation from their migration away from Africa.

1.3.2

The first two component which capture the greatest variance over the dataset could be their ancestral differences, reflecting the geography distribution of different populations. The second component shall distinguish the difference within the population, i.e., subregional variations. These two componenets should tell the geographical migration and separations or other historical influences which led to the form of the populations distribution.

1.4

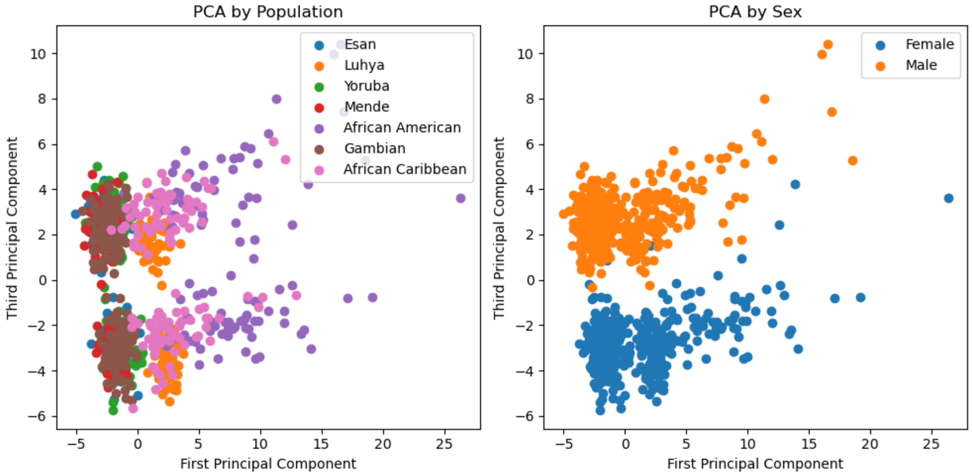


Figure 2

1.5

The third principal component should determines or be highly related with gender, since PCA by sex has a clear cluster-seperation according to it while PCA by population does not.

1.6

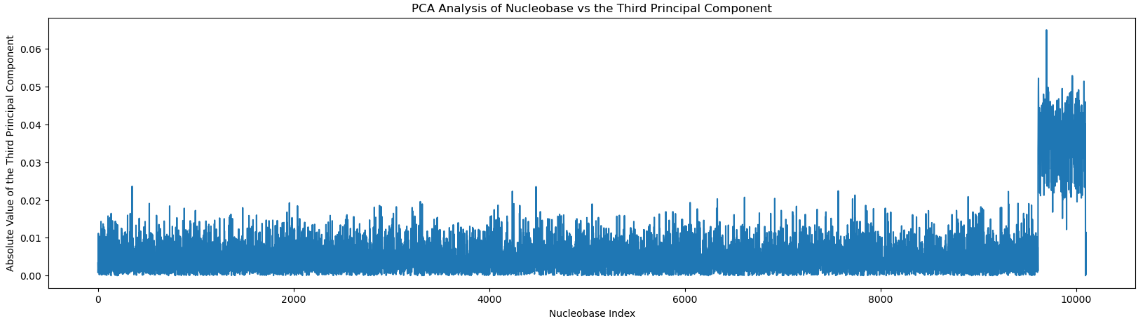


Figure 3

From the plot we can find the rise on the end of the nucleobase index, of the amount around 500, which should be chromosomes determined by the gender as what we inferred before.

2 Markov Chain Monte Carlo

2.1

For the joint likelihood $P(y, \theta, \eta)$, we could derive it as below:

$$\begin{aligned} P(y, \theta, \eta) &= P(y, \theta|\eta)P(\eta) \\ &= P(y|\theta, \eta)P(\theta|\eta)P(\eta) \\ &= P(y|\theta)P(\theta|\eta)P(\eta) \end{aligned} \tag{1}$$

And we could have:

$$\begin{aligned} P(y|\theta) &= \prod_{g=0,1,\ldots,379;j=0,1} P(y_{gj}|\theta_{gj}) \\ &= \prod_{g=0,1,\ldots,379;j=0,1} Poisson(y_{gj};\theta_{gj}) \end{aligned} \tag{2}$$

$$\begin{aligned} P(\theta|\eta) &= P(home) \prod_{t=0,1,\ldots,19} P(att_t|\mu_{att}, \tau_{att}) \prod_{t=0,1,\ldots,19} P(def_t|\mu_{def}, \tau_{def}) \\ &= N(home; 0, \tau_0^{-1}) \prod_{t=0,1,\ldots,19} N(att_t; \mu_{att}, \tau_{att}^{-1})N(def_t; \mu_{def}, \tau_{def}^{-1}) \end{aligned} \tag{3}$$

$$\begin{aligned} P(\eta) &= P(\mu_{att})P(\tau_{att})P(\mu_{def})P(\tau_{def}) \\ &= N(\mu_{att}; 0, \tau_1^-1)N(\mu_{def}; 0, \tau_1^-1)Gamma(\tau_{att}; \alpha, \beta)Gamma(\tau_{def}; \alpha, \beta) \end{aligned} \tag{4}$$

So the joint likelihood could be derived as:

$$\begin{aligned}P(y,\theta,\eta) &= P(y|\theta)P(\theta|\eta)P(\eta)\\ &= \prod_{g=0,1,\ldots,379;j=0,1} Poisson(y_{gj};\theta_{gj})N(home;0,\tau_0^{-1})\\ &\prod_{t=0,1,\ldots,19} N(att_t;\mu_{att},\tau_{att}^{-1})N(def_t;\mu_{def},\tau_{def}^{-1})N(\mu_{att};0,\tau_1^{-}1)\\ &N(\mu_{def};0,\tau_1^{-}1)Gamma(\tau_{att};\alpha,\beta)Gamma(\tau_{def};\alpha,\beta)\end{aligned} \tag{5}$$

2.2

The steps of Metropolis-Hastings algorithm for sampling from posterior $p(\theta,\eta|y)$ are shown as below:

Initialize: Choose an initial value for (θ,η).

For $t = 1$ **to** T**:**

- **Propose:** Sample a proposal candidate (θ',η') from a proposal distribution $q(\theta',\eta'|\theta_t,\eta_t)$.
- **Calculate Acceptance Ratio:**

$$\alpha = \frac{p(\theta',\eta'|y)}{p(\theta_t,\eta_t|y)} \times \frac{q(\theta_t,\eta_t|\theta',\eta')}{q(\theta',\eta'|\theta_t,\eta_t)}$$

- **Accept or Reject:** Sample $u \sim$ Uniform$(0,1)$. If $u < \alpha$, set $(\theta_{t+1},\eta_{t+1}) = (\theta',\eta')$; otherwise, set $(\theta_{t+1},\eta_{t+1}) = (\theta_t,\eta_t)$.

The acceptance function, α, is given by:

$$\alpha = \frac{p(\theta',\eta'|y)}{p(\theta_t,\eta_t|y)} \times \frac{q(\theta_t,\eta_t|\theta',\eta')}{q(\theta',\eta'|\theta_t,\eta_t)}$$

Given that the proposal distribution is an isotropic Gaussian, the proposal density $q(\theta',\eta'|\theta,\eta)$ is symmetric, meaning $q(\theta_t,\eta_t|\theta',\eta') = q(\theta',\eta'|\theta_t,\eta_t)$. Therefore, the acceptance function simplifies to:

$$\alpha = \frac{p(\theta',\eta'|y)}{p(\theta_t,\eta_t|y)}$$

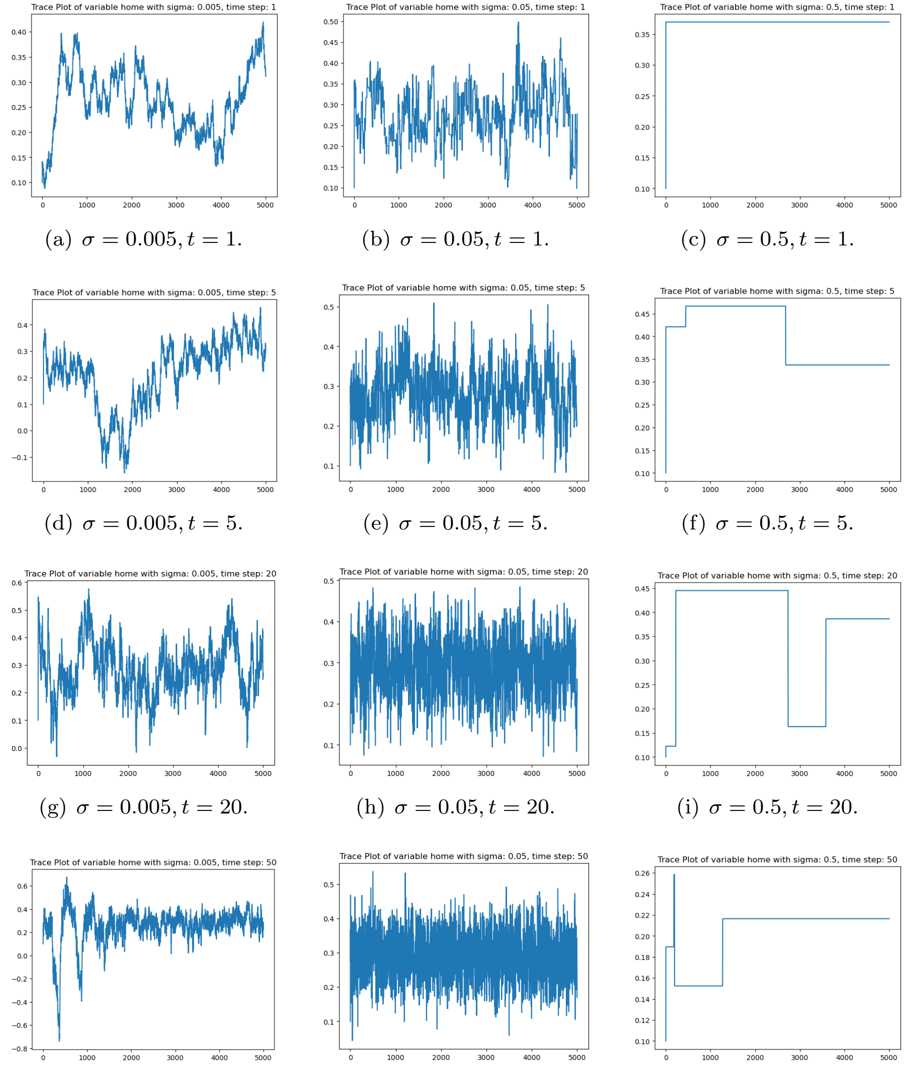
And we could get

$$p(\theta,\eta|y) = \frac{p(y,\theta,\eta)}{p(y)}$$

So

$$\alpha = \frac{p(\theta',\eta',y)}{p(\theta_t,\eta_t,y)}$$

2.3



(a) $\sigma = 0.005, t = 1$.

(b) $\sigma = 0.05, t = 1$.

(c) $\sigma = 0.5, t = 1$.

(d) $\sigma = 0.005, t = 5$.

(e) $\sigma = 0.05, t = 5$.

(f) $\sigma = 0.5, t = 5$.

(g) $\sigma = 0.005, t = 20$.

(h) $\sigma = 0.05, t = 20$.

(i) $\sigma = 0.5, t = 20$.

(j) $\sigma = 0.005, t = 50$.

(k) $\sigma = 0.05, t = 50$.

(l) $\sigma = 0.5, t = 50$.

Figure 4: The trace plot for home variant with different parameters in the MCMC sampling phrase.

parameters	rejection rate
$\sigma = 0.005, t = 1$	0.065
$\sigma = 0.005, t = 5$	0.063
$\sigma = 0.005, t = 20$	0.079
$\sigma = 0.005, t = 50$	0.087
$\sigma = 0.05, t = 1$	0.834
$\sigma = 0.05, t = 5$	0.868
$\sigma = 0.05, t = 20$	0.864
$\sigma = 0.05, t = 50$	0.865
$\sigma = 0.5, t = 1$	1.0
$\sigma = 0.5, t = 5$	1.0
$\sigma = 0.5, t = 20$	1.0
$\sigma = 0.5, t = 50$	1.0

Table 1: The rejection rate in sampling phrase for different parameter settings.

From Table 1, we could see that the parameter combinations with $\sigma = 0.005$ got a quite low rejection rate(lower than 0.1) in the 5000 sampling steps, which indicates the new proposals were accepted highly frequently. The very low rejection rate may suggest that the algorithm is not exploring the parameter space efficiently. On the other hand, the combinations with $\sigma = 0.5$ have a rejection rate equal to 1, which might indicate that the proposal distribution is too conservative. Therefore, we selected a parameter setting with a medium rejection rate, $\sigma = 0.05, andt = 5$.

2.4

a

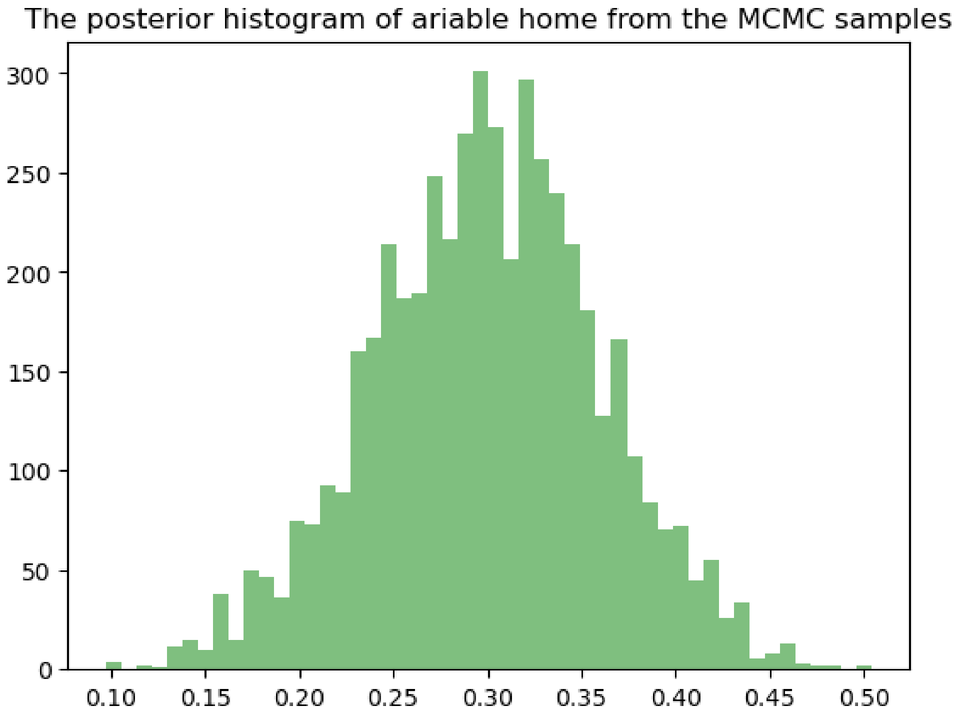


Figure 5: The posterior histogram of variable home from the MCMC samples with $\sigma = 0.05, t = 5$.

b

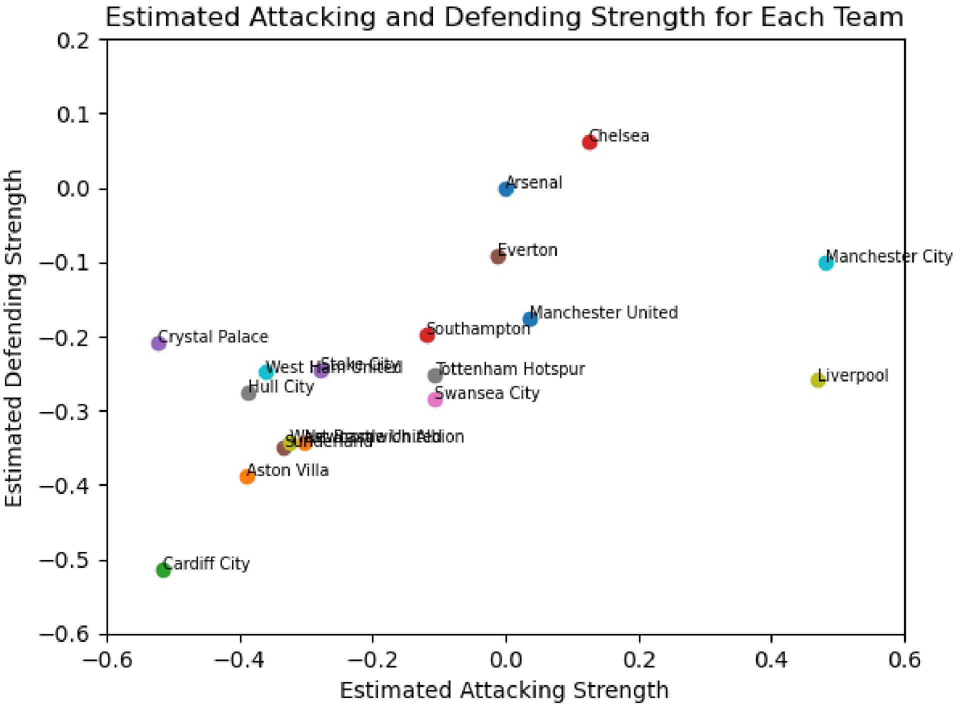


Figure 6: The estimated attacking strength against the estimated defending strength for each team from the MCMC samples with $\sigma = 0.05, t = 5$.

Figure 7

3 Variational Inference

3.1 A Review of Vanilla LDA

3.1.1

False.

3.1.2

True.

3.1.3

$K(V-1) + D(K-1)$

3.1.4

$O(I(DNK + KV + KD)) = O(IDNK)$

3.1.5

$O(DNK)$

3.1.6

False.

3.2 More HMMs

3.2.1

(a) The observed variables here are the words within the document, ω_{di}.
(b) The hidden variables here include topics z_{di}, and binary indicators ϵ_{di}.
(c) Parameters to be estimated contain: T, θ_d, α, β_k, α, γ, p_d, where:

- T, the transition matrix.
- θ_d, the distribution of word for every topic.
- α, the corresponding Dirichlet parameters for θ.
- β_k, the distribution of topic for every document.
- α, the corresponding Dirichlet parameters for β.
- γ, the parameter of Beta distribution for δ_{di}.
- p_d, the parameter of Bernoulli distribution for the documents.

3.2.2 HMM-enhanced LDA

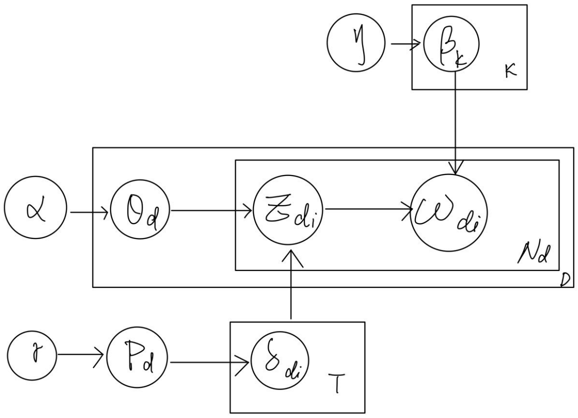


Figure 8

3.2.3

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{p}, \boldsymbol{\beta}; \alpha, \eta, T)$$
$$= \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} \left[p(\theta_d|\alpha) \prod_{i=1}^{N_d} \left(p(w_{di}|\beta_{z_{di}}) p(z_{di}|\theta_d, \delta_{di}, T, z_{d,i-1}) p(\delta_{di}|p_d) \right) p(p_d) \right]$$

3.2.4

1. Set γ_d the parameter vector, then the variational distribution for θ_d is: $p(\theta_d|\gamma_d) = Dir(\theta_d|\gamma_d)$;

2. Set μ_{di} the parameter vector indicating the the probability of each topic for the i-th word in document d, then the variational distribution for Z_{di} is: $p(Z_{di}|\mu_{di}) = Multinomial(Z_{di}|\mu_{di})$;

3. Let η_{di} be the probability of $\delta_{di} = 1$, then the variational distribution for δ_{di} is: $p(\delta_{di}|\eta_{di}) = Bernoulli(\delta_{di}|\eta_{di})$;

4. Set λ_d the parameter vector, then the variational distribution for p_d is: $p(p_d|\lambda_d) = Beta(p_d|\lambda_d)$.

3.2.5

$$
\begin{aligned}
ELBO(\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\lambda}) =& E_q[\log p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{p}, \boldsymbol{\beta})] - E_q[\log q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{p})] \\
=& \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log p(w_{di}|z_{di}, \boldsymbol{\beta})] + \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log p(z_{di}|\theta_d, \delta_{di}, T)] \\
&+ \sum_{d=1}^{D} E_q[\log p(\theta_d|\alpha)] + \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log p(\delta_{di}|p_d)] \\
&+ \sum_{d=1}^{D} E_q[\log p(p_d)] - \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log q(z_{di}|\mu_{di})] \\
&- \sum_{d=1}^{D} \sum_{i=1}^{N_d} E_q[\log q(\delta_{di}|\eta_{di})] - \sum_{d=1}^{D} E_q[\log q(\theta_d|\gamma_d)] \\
&- \sum_{d=1}^{D} E_q[\log q(p_d|\lambda_d)]
\end{aligned}
\tag{6}
$$

3.2.6 Update rule

1. For the k-th parameter of the variational Dirichlet distribution in document d:
   $\gamma_{dk} = \alpha_k + \sum_{i=1}^{N_d} \mu_{dik}$

2. For the probability of topics for the i-th word in document d assigned to topic k:
   $\mu_{dik} = \frac{\exp\left(\Psi(\gamma_{dk})+\log(\beta_{kw_{di}})\right)}{\sum_{j=1}^{K} \exp\left(\Psi(\gamma_{dj})+\log(\beta_{jw_{di}})\right)}$, where Ψ is the digamma function

3. As for η_{di}, the probability of $\delta_{di} = 1$:
   $\eta_{di} = \frac{\exp\left(\log(\lambda_{d1})+\sum_{k=1}^{K} \mu_{di-1,k} \log(T_{k,z_{di}})\right)}{\exp\left(\log(\lambda_{d1})+\sum_{k=1}^{K} \mu_{di-1,k} \log(T_{k,z_{di}})\right)+\exp(\log(\lambda_{d2}))}$

4. The update rule for λ_d contains two parts, λ_{d1} and λ_{d2}, which represents the number of $\delta_{di} = 1$ and $\delta_{di} = 0$ respectively, that:
   $\lambda_{d1} = i_d + \sum_{i=2}^{N_d} \eta_{di}$, and $\lambda_{d2} = j_d + \sum_{i=2}^{N_d}(1 - \eta_{di})$, where i_d and j_d are the prior parameters of the Beta distribution.

3.2.7

1. For topic distribution β_k:$\beta_{kw} = \frac{\sum_{d=1}^{D} \sum_{i=1}^{N_d} 1(w_{di}=w)\cdot\mu_{dik}}{\sum_{w'} \sum_{d=1}^{D} \sum_{i=1}^{N_d} 1(w_{di}=w')\cdot\mu_{dik}}$;

2. For transition matrix T: $T_{kl} = \frac{\sum_{d=1}^{D} \sum_{i=2}^{N_d} \mu_{di-1,k}\cdot\mu_{dil}\cdot(1-\eta_{di})}{\sum_{l'} \sum_{d=1}^{D} \sum_{i=2}^{N_d} \mu_{di-1,k}\cdot\mu_{dil'}\cdot(1-\eta_{di})}$.

3.2.8

Top 10 words under 10 topics obtained from the EM algorithm listed as the table below (abandoning the words absent in the vocabulary), under parameters: Dirichlet prior $\alpha = I$ and Beta prior parameters $a = 1, b = 1$. Process coded in the script.

Topic No.	Top 10 words
1	300 200 500 100 law caused cup winter love level
2	200 500 100 300 winter guns cup love level studio
3	500 200 100 300 winter guns cup love level studio
4	300 200 500 100 law caused cup winter love level
5	500 200 100 300 winter guns cup love level studio
6	100 200 500 300 winter guns cup love level studio
7	100 500 200 300 originally stage sir cup love level
8	500 200 100 300 winter guns cup love level studio
9	200 500 100 300 winter guns cup love level studio
10	200 500 100 300 winter guns cup love level studio

Table 2: Top 10 words under 10 topics obtained from the EM algorithm

3.2.9

We can obtain a convergence of the value of ELBO by tuning the prior parameters of Dirichlet and Beta, shown in the Q3.py file.

4 QWOP

4.1

We have tried three different algorithms to optimize the game plan, including Gradient Descent, Metropolis-Hastings Algorithm and Differential Evolution Algorithm. The best distances achieved by the three algorithms are shown in the table.

Metropolis_Hastings	Gradient Descent	Differential Evolution
6.0015	5.7139	9.6514

Table 3: The best distances achieved by three different algorithms.

The Differential Evolution algorithm got the highest outcome in this problem. The procedure of it is described below.

1. **Initialization.** A population of candidate plans is initialized randomly.

2. **Mutation.** For each individual in the population, there is a probability that the individual will be mutated. And the new individual would be created by adding a scaled difference between two randomly chosen individuals.

   $$new\_individual = target\_individual + scaling\_factor * (rand1 - rand2).$$

3. **Crossover.** The crossover operation is performed with a given probability. Crossover would randomly select a subset of the mutated individual, and the rest information of the individual is filled with the original values from the target individual.

4. **Selection.** The individual after mutation and crossover is compared with the original individual based on the fitness values, which is the distance achieved by the avatar in this case.

5. **Termination.** The algorithm would stop after repeating a given number of iterations.

The plan optimized by the Differential Evolution algorithm is:

[0.3881443 0.70216632 0.95475725 0.98389038 1. 0.99843408 -0.12645366 0.45902393 -0.77969429 -0.99052372 0.9584196 0.53890926 1. -0.97607534 1. 1. 0.0013154 -0.81757202 0.91351521 -0.5189027 0.99574655 0.99535224 -0.32297164 0.99606488 0.13524925 -0.97008509 0.95143395 -0.743552 -0.26656817 0.87585921 -0.98959096 0.99663264 -0.92106722 -0.55280325 0.99884727 -0.98823304 -0.89013668 -0.70721028 0.58456288 0.86516142].

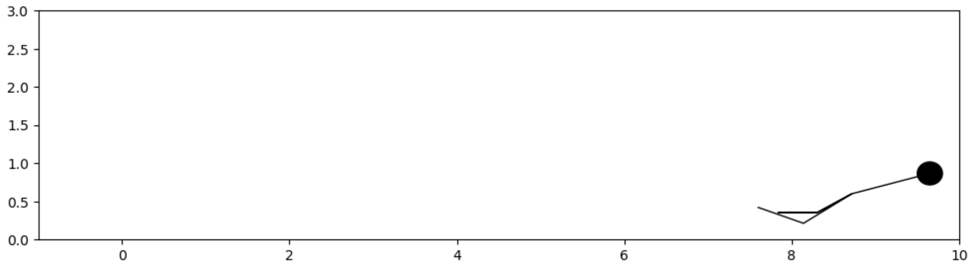


Figure 9: The last frame of the animation performing the best plan.