CS5340:
Tutorial 8

Asst. Prof. Harold Soh

TAs: Chen Kaiqi

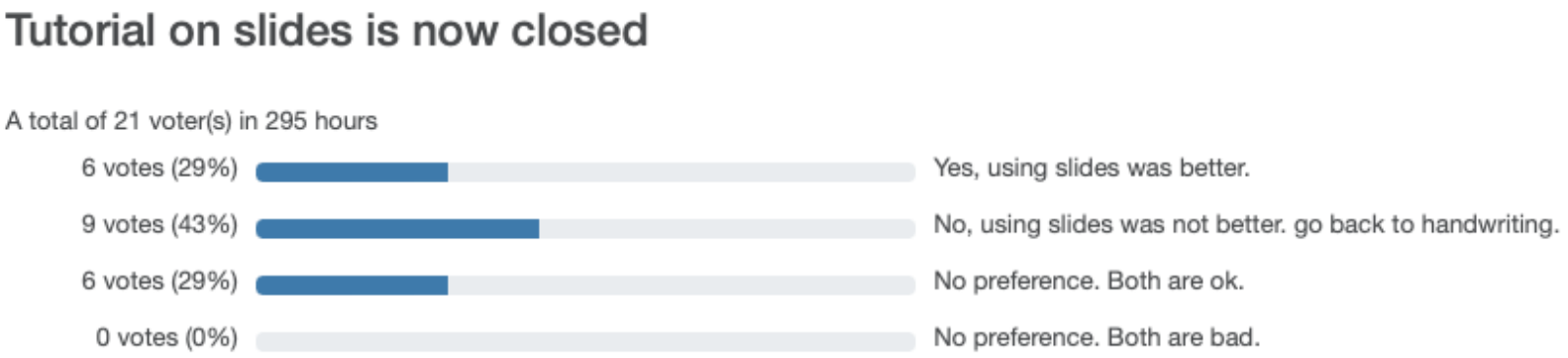
Course Schedule

Week	Date	Lecture Topic	Tutorial Topic
1	12 Jan	Introduction to Uncertainty Modeling + Probability Basics	Introduction
2	19 Jan	Simple Probabilistic Models	Probability Basics
3	26 Jan	Bayesian networks (Directed graphical models)	More Basic Probability
4	2 Feb	Markov random Fields (Undirected graphical models)	DGM modelling and d-separation
5	9 Feb	Variable elimination and belief propagation	MRF + Sum/Max Product
6	16 Feb	Factor graph and the junction tree algorithm	**Quiz 1**
-	-	RECESS WEEK	
7	2 Mar	Mixture Models and Expectation Maximization (EM)	Linear Gaussian Models
8	9 Mar	Hidden Markov Models (HMM)	Probabilistic PCA
9	**16** Mar	Monte-Carlo Inference (Sampling)	Linear Gaussian Dynamical System
10	23 Mar	Variational Inference	MCMC + Sequential VAE
11	30 Mar	Inference and Decision-Making (Special Topic)	**Quiz 2**
12	6 Apr	Gaussian Processes (Special Topic)	Wellness Day
13	13 Apr	**Project Presentations**	Closing

Quiz 2

- Covers everything up to MCMC.
  - More focus on second part of semester (GMM/EM onwards)
  - **NO** Variational Inference
- Similar to Quiz 1
  - Open-Book
  - On Canvas
  - Fully Multiple Choice

Poll Results

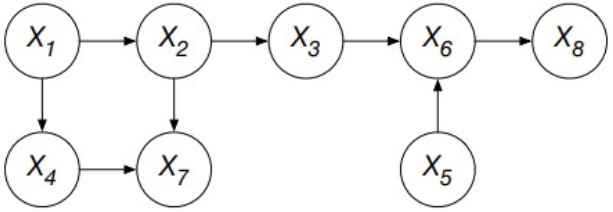


Questions?

https://pollev.com/haroldsohsoo986

Section 1: Gibbs Sampling

You want to run Gibbs sampling on the following graphical model. For each of the random variables below, what is the correct conditional to sample from? **Note:** If there are multiple correct answers, select the one that conditions upon the fewest number of random variables.



Gibbs Sampling

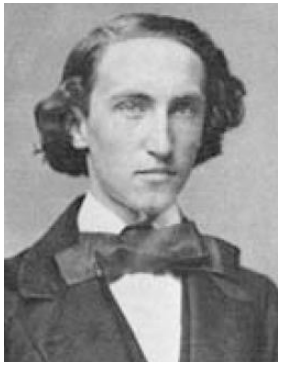
Algorithm : Gibbs Sampling

1. Initialize $\{x_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:

3. Sample $x_1^{\tau+1} \sim p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$.
4. Sample $x_2^{\tau+1} \sim p(x_2 | x_1^{(\tau+1)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$.

5. Sample $x_j^{\tau+1} \sim p\left(x_j \middle| x_1^{(\tau+1)}, \ldots, x_{j-1}^{(\tau+1)}, x_{j+1}^{(\tau)}, \ldots, x_M^{(\tau)}\right)$.

6. Sample $x_M^{\tau+1} \sim p(x_M | x_1^{(\tau+1)}, x_2^{(\tau+1)}, \ldots, x_{M-1}^{(\tau+1)})$



Josiah Willard Gibbs
1839–1903

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Gibbs Sampling: Markov Blankets

- The conditional $p(\, x_j \,|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_N)$ looks intimidating, but recall Markov Blankets:

$$p(\, x_j \,|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_N) = p(\, x_j \mid \underbrace{\mathrm{MB}(x_j)}_{\text{Markov blanket of } x_j} \,).$$

- Bayesian network: the Markov blanket of X_j is the set containing its parents, children, and co-parents.
- MRF: the Markov Blanket of X_j is its immediate neighbors.

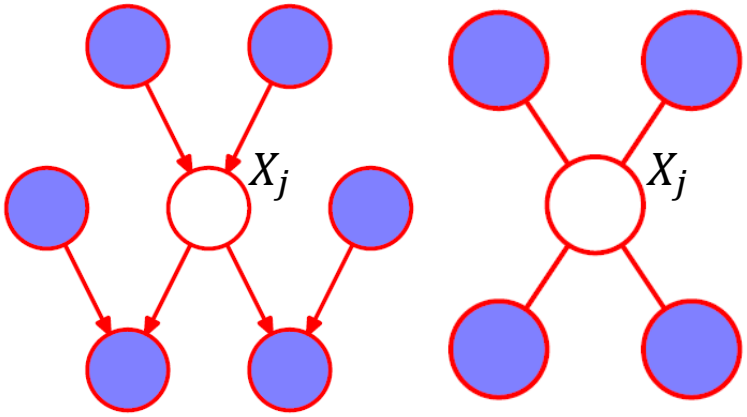
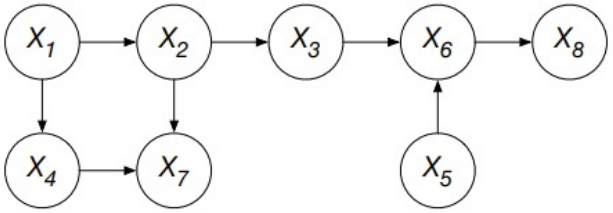


Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Section 1: Gibbs Sampling

You want to run Gibbs sampling on the following graphical model. For each of the random variables below, what is the correct conditional to sample from? **Note:** If there are multiple correct answers, select the one that conditions upon the fewest number of random variables.



Problem 1. Sample x_1.

A. $p(X_1)$ (sample from the prior)

B. $p(X_1|X_2, X_4, X_7)$

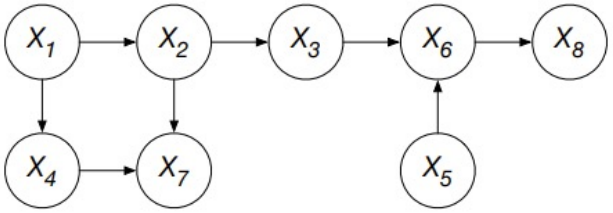
C. $p(X_1|X_2, X_4)$

D. $p(X_1|X_2, X_3, X_4, X_7)$

E. $p(X_1|X_2, X_3, X_4, X_7, X_5)$

Section 1: Gibbs Sampling

You want to run Gibbs sampling on the following graphical model. For each of the random variables below, what is the correct conditional to sample from? **Note:** If there are multiple correct answers, select the one that conditions upon the fewest number of random variables.



Problem 2. Sample x_2.

A. $p(X_2|X_1, X_3)$

B. $p(X_2|X_4, X_7)$

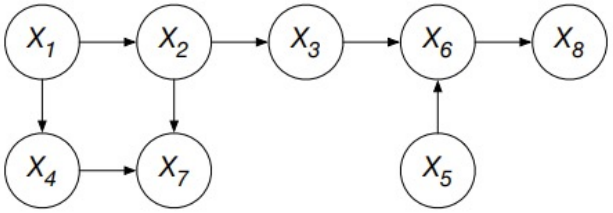
C. $p(X_2|X_1, X_3, X_4, X_7)$

D. $p(X_2|X_3, X_7)$

E. $p(X_2|X_1)$

Section 1: Gibbs Sampling

You want to run Gibbs sampling on the following graphical model. For each of the random variables below, what is the correct conditional to sample from? **Note:** If there are multiple correct answers, select the one that conditions upon the fewest number of random variables.



Problem 3. Sample x_3.

A. $p(X_3|X_2, X_6, X_7)$

B. $p(X_3|X_2, X_6)$

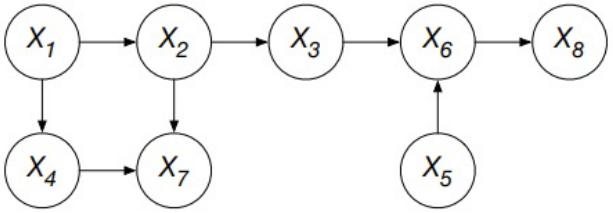
C. $p(X_3|X_2)$

D. $p(X_3|X_2, X_5, X_6)$

E. $p(X_3|X_2, X_5, X_6, X_8)$

Section 1: Gibbs Sampling

You want to run Gibbs sampling on the following graphical model. For each of the random variables below, what is the correct conditional to sample from? **Note:** If there are multiple correct answers, select the one that conditions upon the fewest number of random variables.



Problem 4. Sample x_4.

A. $p(X_4|X_1, X_2, X_7)$

B. $p(X_4|X_2, X_7)$

C. $p(X_4|X_1)$

D. $p(X_4|X_1, X_2, X_3, X_7)$

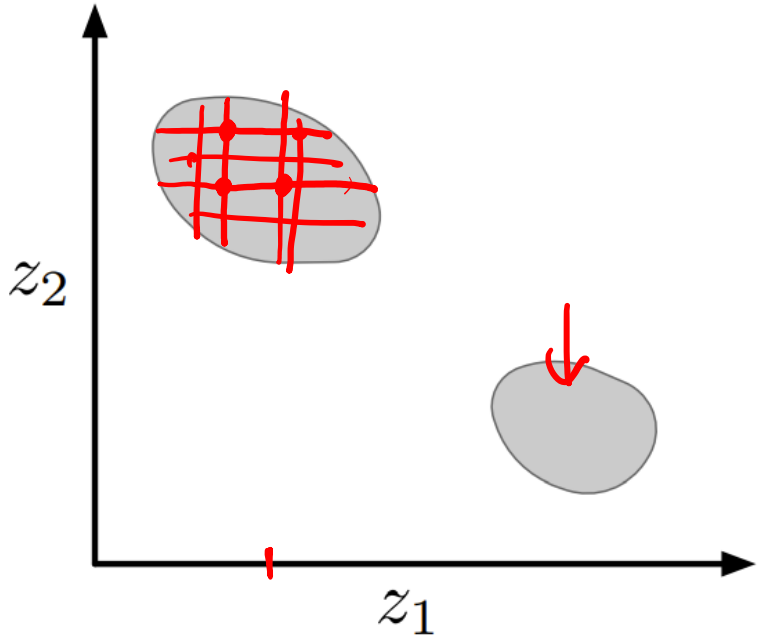
E. $p(X_4|X_1, X_2, X_3, X_7, X_6, X_8)$

Questions?

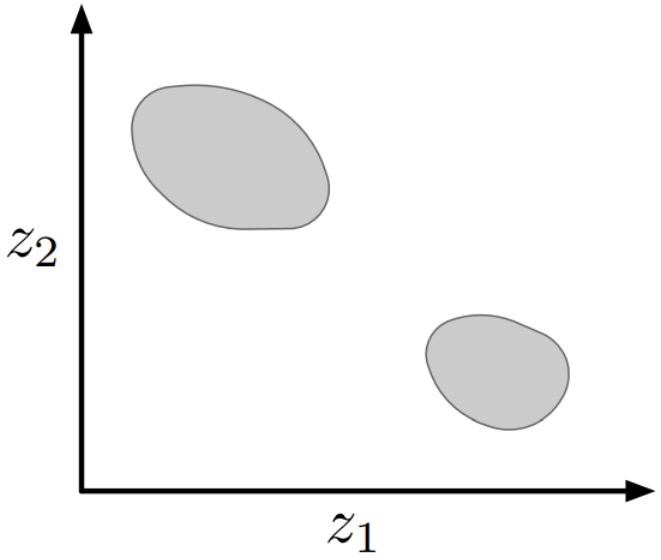
https://pollev.com/haroldsohsoo986

Section 2: Discussion Questions

Problem 5. Consider the following distribution of two variables z_1 and z_2 that is uniform over the shaded regions and that is zero everywhere else. **Discuss whether the standard Gibbs sampling procedure would sample correctly from this distribution.**



Section 2: Discussion Questions



1. Initialize $\{x_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:

3. Sample $x_1^{\tau+1} \sim p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$.
4. Sample $x_2^{\tau+1} \sim p(x_2 | x_1^{(\tau+1)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$.

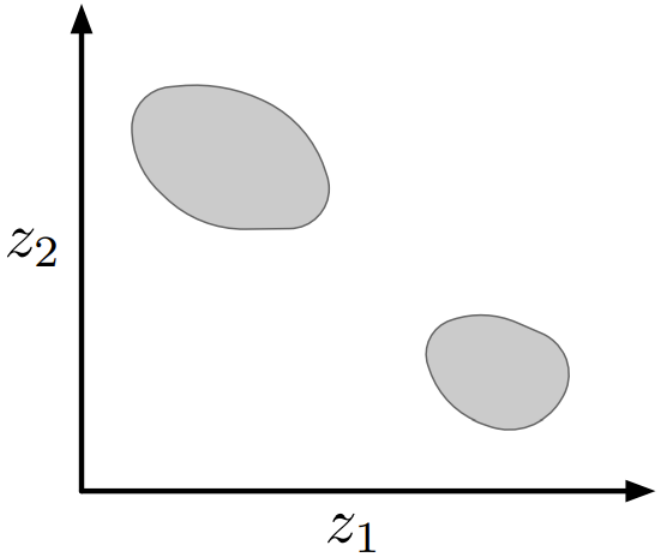
5. Sample $x_j^{\tau+1} \sim p\left(x_j \middle| x_1^{(\tau+1)}, \ldots, x_{j-1}^{(\tau+1)}, x_{j+1}^{(\tau)}, \ldots, x_M^{(\tau)}\right)$.

6. Sample $x_M^{\tau+1} \sim p(x_M | x_1^{(\tau+1)}, x_2^{(\tau+1)}, \ldots, x_{M-1}^{(\tau+1)})$

Ergodic Theorem for Markov Chains

- A Markov chain is ergodic if it is irreducible and aperiodic.
- **Ergodicity is important**: it implies we can reach the stationary/limiting distribution π, no matter the initial distribution π_0.
- All good MCMC algorithms must satisfy ergodicity, so that we cannot initialize in a way that will never converge.

Section 2: Discussion Questions



1. Initialize $\{x_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:

3. Sample $x_1^{\tau+1} \sim p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$.
4. Sample $x_2^{\tau+1} \sim p(x_2 | x_1^{(\tau+1)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$.

5. Sample $x_j^{\tau+1} \sim p(x_j | x_1^{(\tau+1)}, \ldots, x_{j-1}^{(\tau+1)}, x_{j+1}^{(\tau)}, \ldots, x_M^{(\tau)})$.

6. Sample $x_M^{\tau+1} \sim p(x_M | x_1^{(\tau+1)}, x_2^{(\tau+1)}, \ldots, x_{M-1}^{(\tau+1)})$

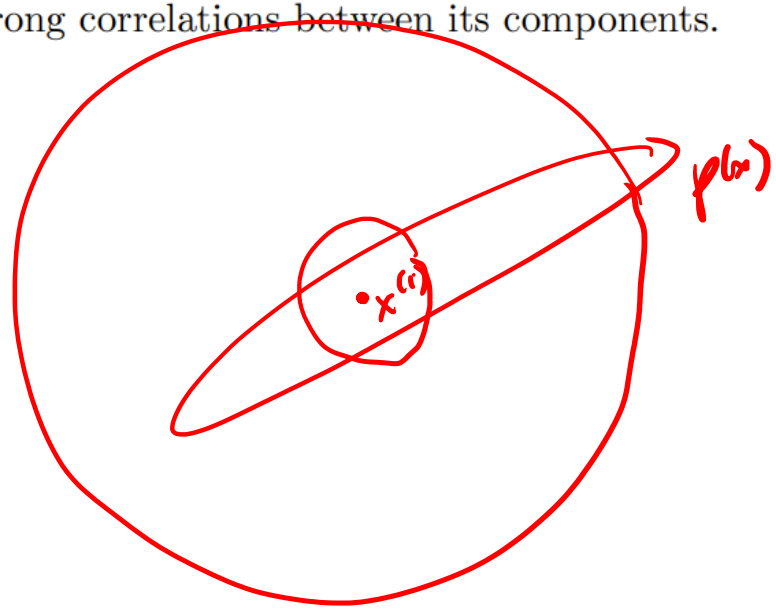
No. Gibbs Sampling would not work!

Problem 6. Consider the Metropolis-Hastings algorithm. At each step i, we sample a new point from the proposal distribution $q(\mathbf{x}|\mathbf{x}^{(i)})$. In the lectures, we used as an example the simple *isotropic* (spherical) Gaussian centered upon $\mathbf{x}^{(i)}$, i.e.,

$$q(\mathbf{x}|\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{x}^{(i)}, \sigma^2\mathbf{I})$$

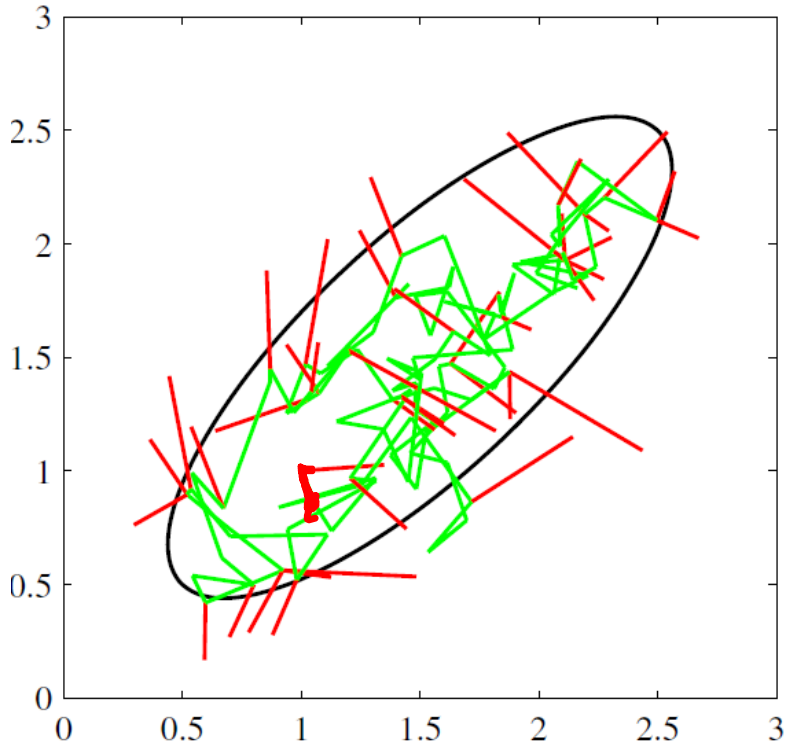
which is a common choice for continuous variables. The variance σ^2 is a parameter of the proposal distribution.

Discuss how sensitive you think MC sampling is to the parameter σ^2. What are the respective trade-offs when considering how to set σ^2? *Hint:* Consider a elongated bi-variate Gaussian having strong correlations between its components.



Metropolis Algorithm

- Illustration of using Metropolis algorithm (proposal distribution: isotropic Gaussian) to sample from a Gaussian distribution:

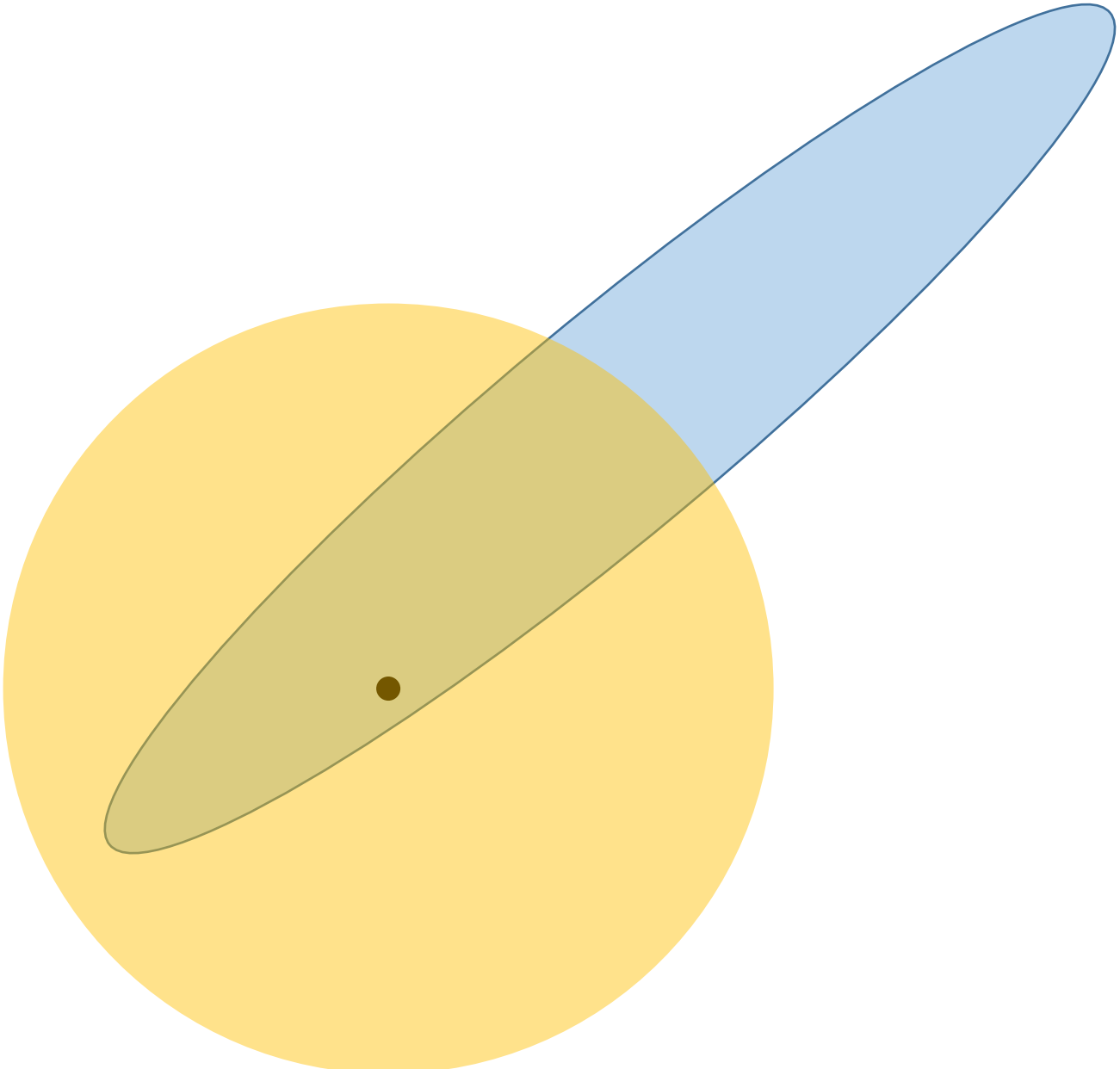


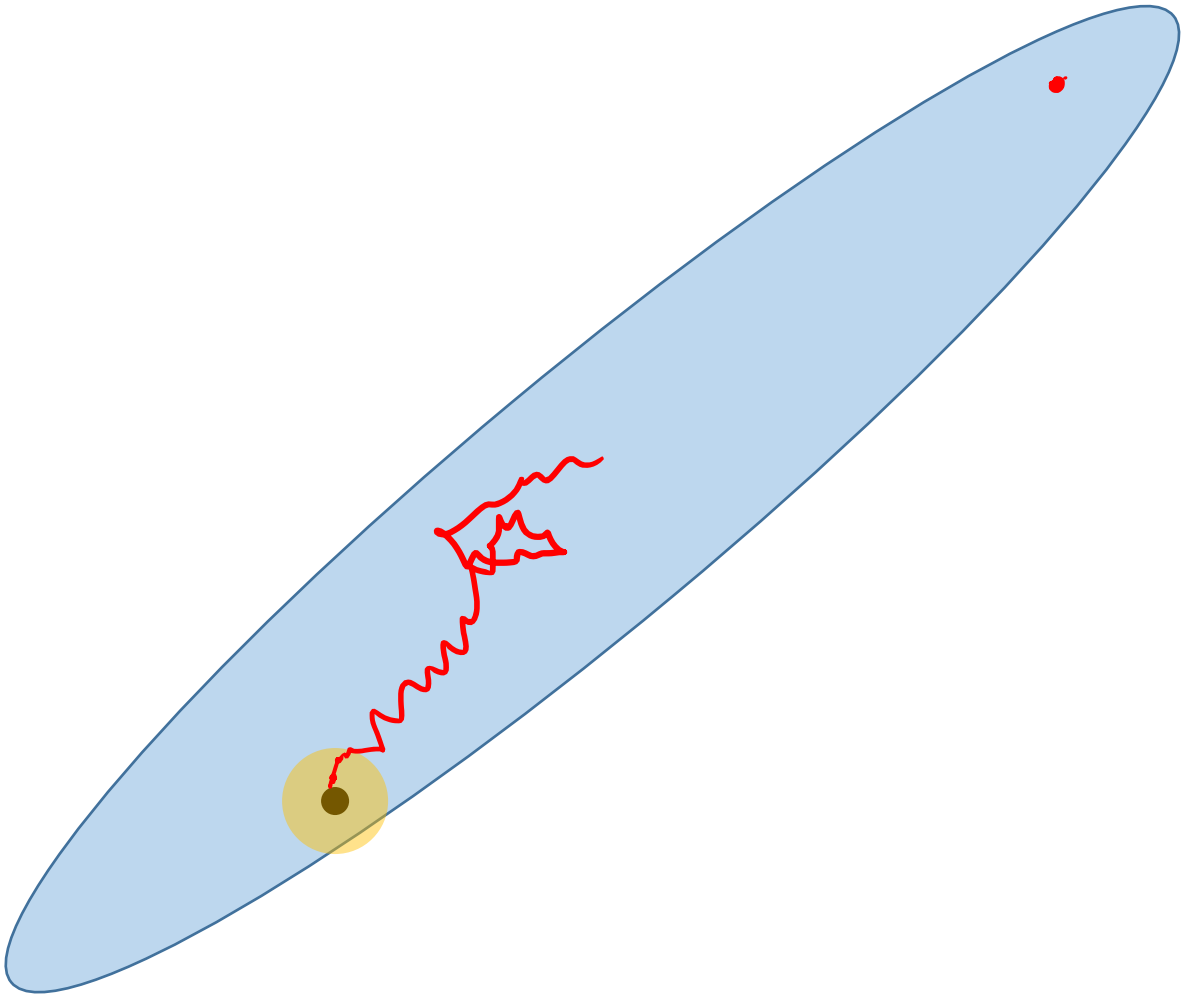
Accepted sample

Rejected sample

150 candidate samples are generated, 43 are rejected.

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop





Problem 6. Consider the Metropolis-Hastings algorithm. At each step i, we sample a new point from the proposal distribution $q(\mathbf{x}|\mathbf{x}^{(i)})$. In the lectures, we used as an example the simple *isotropic* (spherical) Gaussian centered upon $\mathbf{x}^{(i)}$, i.e.,

$$q(\mathbf{x}|\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{x}^{(i)}, \sigma^2\mathbf{I})$$

which is a common choice for continuous variables. The variance σ^2 is a parameter of the proposal distribution.

Discuss how sensitive you think MC sampling is to the parameter σ^2. What are the respective trade-offs when considering how to set σ^2? *Hint:* Consider a elongated bi-variate Gaussian having strong correlations between its components.

Metropolis Hastings is sensitive to the parameter!

Tradeoffs:

- **High σ^2 leads to high rejection rates**
- **Low σ^2 leads to slow exploration**

Can we fix this problem?

More Advanced Methods

- See Chp 11 of Bishop
- Slice Sampling (Neal, 2003)
- Hybrid Monte-Carlo (Neal, 1996)
- Extra: Langevin Dynamics

Questions?

https://pollev.com/haroldsohsoo986

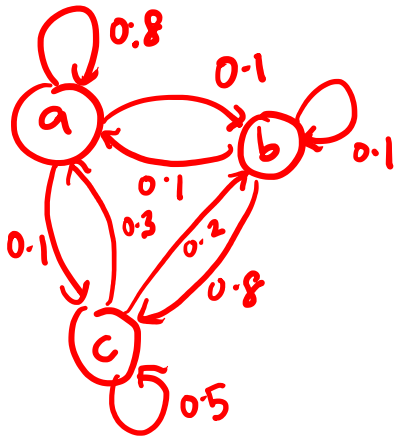
The Right Transitions

3 The Right Transitions

For each of the matrices below, select True if the matrix is valid *transition matrix* over the states for use in a MCMC algorithm. Select False otherwise. Justify your answer. *Hint:* Look up what properties are required for a transition matrix in an MCMC method.

Problem 7.

$$T = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.3 & 0.2 & 0.1 \end{bmatrix} \end{array}$$



Ergodic Theorem for Markov Chains

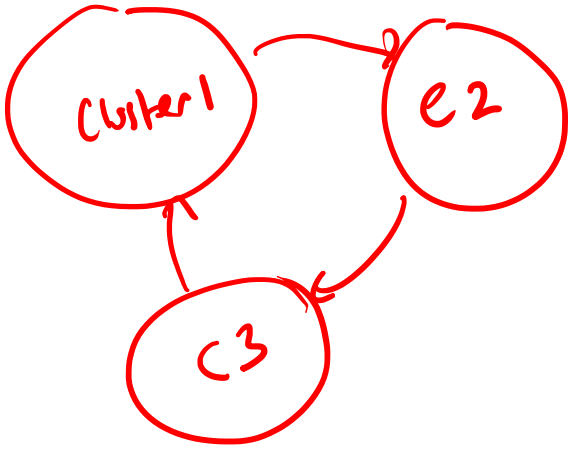
- A Markov chain is ergodic if it is irreducible and aperiodic.
- **Ergodicity is important**: it implies we can reach the stationary/limiting distribution π, no matter the initial distribution π_0.
- All good MCMC algorithms must satisfy ergodicity, so that we cannot initialize in a way that will never converge.

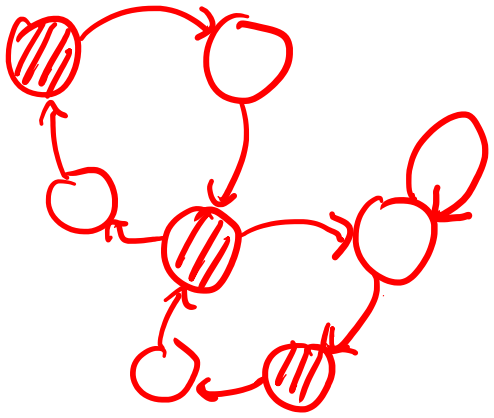
Checking Irreducible Property

- Check that a path exists from each node to every other node in the graph.

Checking Aperiodicity

- Cluster states into 2 or more groups such that one group transitions to another in a deterministic manner. If such a clustering exists, the Markov Chain is periodic.
- If there are self-loops, then the Markov chain is aperiodic.





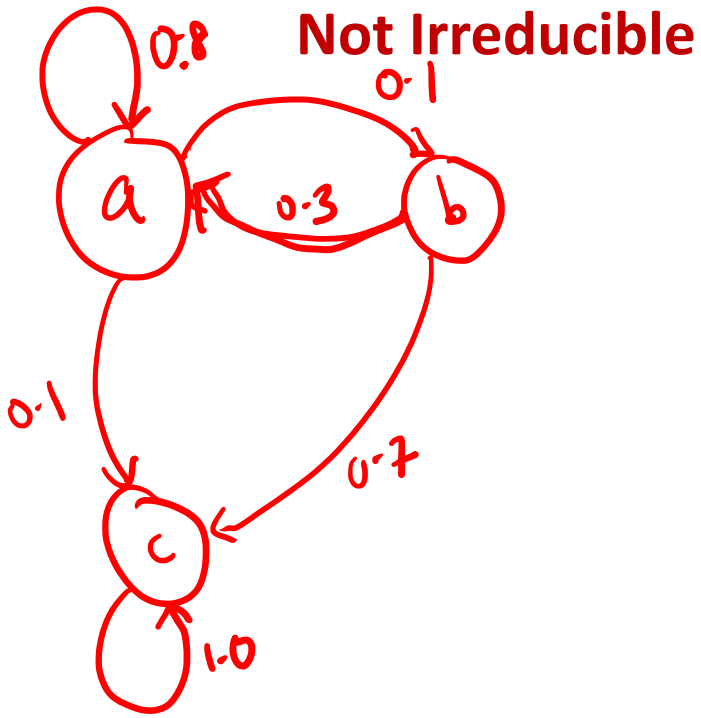
Problem 7

$$T = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.3 & 0.2 & 0.1 \end{bmatrix} \neq 1$$

Not a valid transition matrix.

Problem 8

$$T = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.3 & 0.0 & 0.7 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$



Problem 9

$$T = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.9 & 0.1 & 0.0 \\ 0.1 & 0.9 & 0.0 \\ 0.5 & 0.3 & 0.2 \end{bmatrix} \end{array}$$

Not Irreducible

Problem 10

$$T = \begin{bmatrix} 0.7 & 0.1 & 0.2 \\ 0.2 & 0.3 & 0.5 \\ 0.0 & 0.6 & 0.4 \end{bmatrix}$$

Yes. Aperiodic and Irreducible

Problem 11

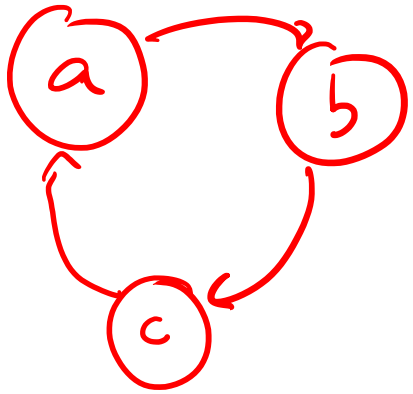
$$T = \begin{bmatrix} 0.0 & 0.8 & 0.2 \\ 0.2 & 0.3 & 0.5 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}$$

Yes. Aperiodic and Irreducible

Problem 12

$$T = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}$$

Not Aperiodic

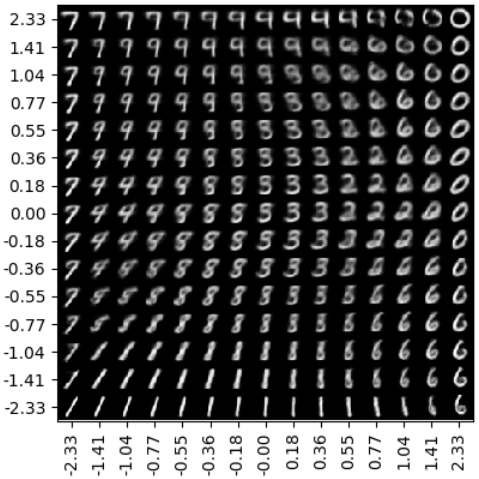


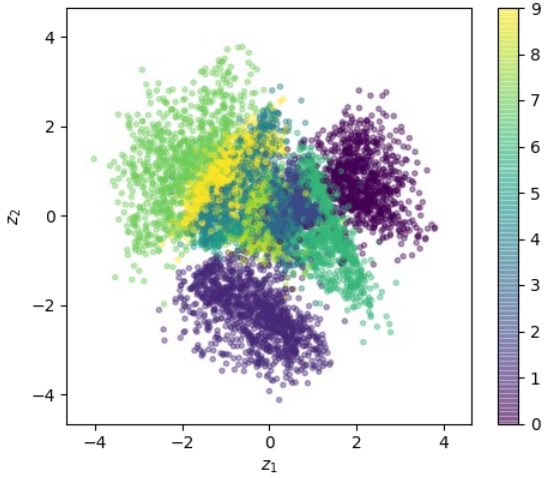
Questions?

https://pollev.com/haroldsohsoo986

Variational Inference

- Brief Intro to Variational Inference
- Variational Autoencoders
- If time permits: Sequential VAE Tutorial Question.





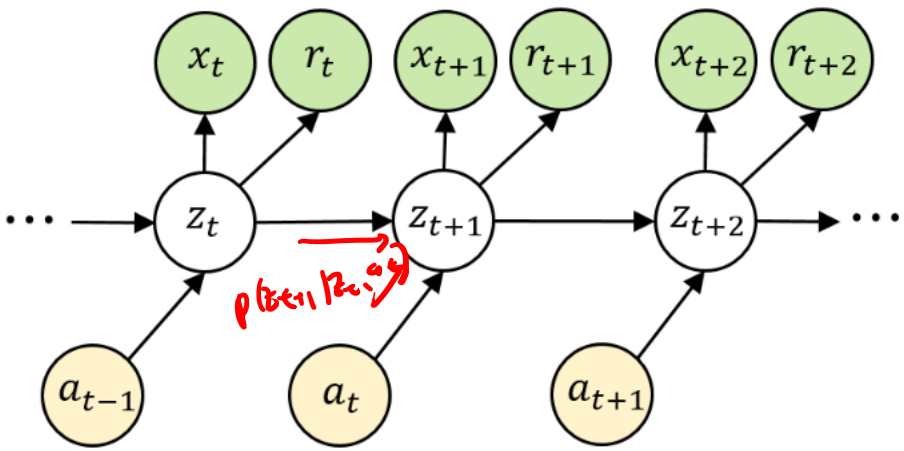
Fatir and Soh, AAAI 2019

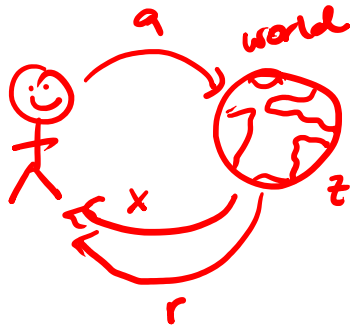
Image credit: https://tiao.io/post/tutorial-on-variational-autoencoders-with-a-concise-keraas-implementation/

Notebook for Derivations.

Sequential VAE

Problem 13. The problem above can be formulated as a Bayesian network:





Each of the factorized distributions are modelled using nonlinear functions:

- Transitions: $p_\theta(z_t|z_{t-1}, a_{t-1}) = p(z_t|f_\theta(z_{t-1}, a_{t-1}))$
- Observations: $p_\theta(x_t|z_t) = p(x_t|d_\theta(z_t))$
- Rewards: $p_\theta(r_t|z_t) = p(r_t|r_\theta(z_t))$

where f_θ, d_θ^m, r_θ are neural networks parameterized by θ. First, consider that the actions are always observed. Write out the factorization of the probability $p_\theta(x_{1:T}, r_{1:T}, z_{1:T}|a_{1:T-1})$ corresponding to the DGM above.

$$p(z_{1:T}, x_{1:T}, r_{1:T} \mid a_{1:T-1}) = p(z_1)\left[\prod_{t=2} p(z_t \mid z_{t-1}, a_{t-1})\right] \prod_{t=1} p(x_t|z_t)\, p(r_t|z_t)$$

Problem 14. Learning this model is intractable due to the nonlinear transition, observation, and reward functions. We will perform variational inference to learn the parameters of the model. Assume we observe trajectories τ sampled from data distribution $p_d(\tau)$. Each trajectory is an observation $\tau = \{(x_t, r_t, a_t)\}_{t=1}^{T}$.

To obtain the maximum likelihood estimate (MLE) of the parameters θ, which of the following functions should we optimize?

A. $\mathbb{E}_{p_d}[\log p(x_{1:T}, r_{1:T} | a_{1:T-1}; \theta)]$

B. $\mathbb{E}_{p_d}[\log p(x_{1:T}, r_{1:T}, z_{1:T} | a_{1:T-1}; \theta)]$

C. $\mathbb{E}_{p_d}[\log p(\theta | x_{1:T}, r_{1:T}, z_{1:T}, a_{1:T-1})]$

D. $\mathbb{E}_{p_d}[\log p(\theta, x_{1:T} | r_{1:T}, z_{1:T}, a_{1:T-1})]$

E. Any of the above would work.

Problem 15. Note that the maximum likelihood estimation requires us to marginalize out the latent variables $z^i_{1:T}$ for each trajectory τ^i in a dataset $\mathcal{D}$. We will need the variational posterior q. Consider three choices:

A. $q(z^i_{1:T}) = \prod_{t=1} q(z^i_t)$ where the q's are Gaussian distribution that share the same parameters (mean and covariance).

B. $q(z^i_{1:T}) = \prod_{t=1} q^i_t(z^i_t)$ where each q^i_t is a Gaussian distribution with *different* parameters.

C. $q(z^i_{1:T}|x^i_{1:T}, a^i_{1:T-1}) = \prod_{t=1}^{T} q_\phi(z^i_t|g_\phi(x^i_{1:t}, a^i_{1:t-1}))$ where q_ϕ is a Gaussian distribution and the *inference network* $g_\phi(x_{1:t}, a_{1:t-1})$ is a neural network (usually a recurrent neural network like a LSTM or GRU) parameterized by ϕ that outputs the mean and covariance for each z^i_t. The inference networks provides the parameters for the mean and the covariance of the distributions.

Between A, B and C, which variational distribution is the least expressive? Which is the most expressive?

A is the least expressive.
B is the most expressive.

Problem 16. Consider the variational distribution given in C above, i.e.,

$$q(z^i_{1:T}|x^i_{1:T}, a^i_{1:T-1}) = \prod_{t=1}^{T} q_\phi(z^i_t|g_\phi(x^i_{1:t}, a^i_{1:t-1}))$$

where each q_ϕ is a Gaussian distribution and the *inference network* $g_\phi(x_t, z_{t-1}, a_{t-1})$ is a neural network parameterized by ϕ. The inference network provides the parameters for the mean and the covariance of the distributions. Is $q(z^i_{1:T}|x^i_{1:T}, a^i_{1:T-1})$ a multivariate Gaussian in general? Provide a brief justification.

Yes, it is a multivariate Gaussian.

Problem 17. Suppose we pick the inference network variational distribution given in C above. To simplify notation, we call this $q_\phi(z_t)$. Given all these distributions and trajectories $\tau \sim p_d(\tau)$, we seek to learn the parameters θ and ϕ. We optimize the evidence lower bound (ELBO) under the data distribution p_d using a variational distribution q_ϕ over the latent state variables z_t.

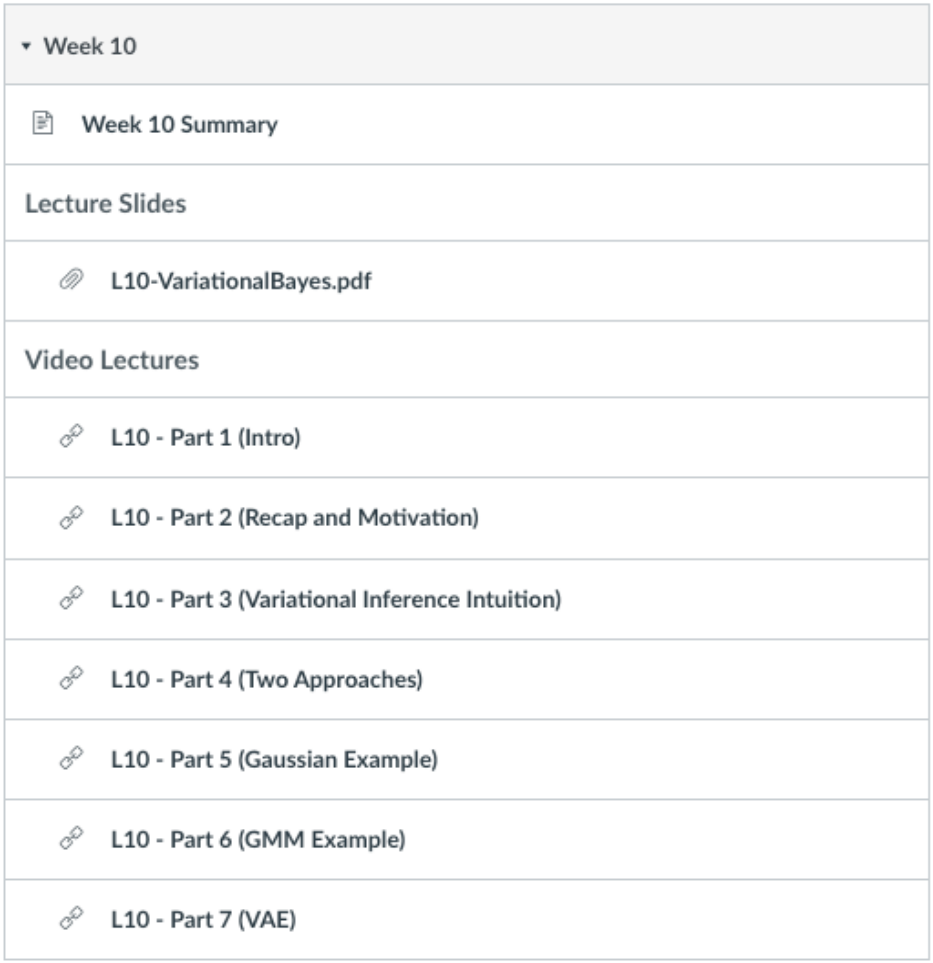
$$\mathbb{E}_{p_d}[\text{ELBO}] \leq \mathbb{E}_{p_d}[\log p_\theta(x_{1:T}, r_{1:T}|a_{1:T-1})] \tag{2}$$

where

$$\text{ELBO} = \sum_{t=1}^{T} \left(\mathop{\mathbb{E}}_{q_\phi(z_t)}[\log p_\theta(x_t|z_t)] + \mathop{\mathbb{E}}_{q_\phi(z_t)}[\log p_\theta(r_t|z_t)] \right) \tag{3}$$

$$- \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_\phi(z_{t-1})}[\text{KL}\left[q_\phi(z_t)\|p_\theta(z_t|z_{t-1}, a_{t-1})\right]] - \text{KL}\left[q_\phi(z_1)\|p_\theta(z_1)\right] \tag{4}$$

Homework


Week 10
Week 10 Summary
Lecture Slides
L10-VariationalBayes.pdf
Video Lectures
L10 - Part 1 (Intro)
L10 - Part 2 (Recap and Motivation)
L10 - Part 3 (Variational Inference Intuition)
L10 - Part 4 (Two Approaches)
L10 - Part 5 (Gaussian Example)
L10 - Part 6 (GMM Example)
L10 - Part 7 (VAE)