

1. Consider the VAE training algorithm below:

Initialize θ and ϕ network parameters randomly.

for number of training iterations **do**

Sample minibatch of B examples $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^B$ from data D

Compute $\mu_{\mathbf{z}|\mathbf{x}^i} \leftarrow q_\phi(\mathbf{x}^i), \sigma_{\mathbf{z}|\mathbf{x}^i} \leftarrow q_\alpha(\mathbf{x}^i)$

Get B samples $\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^B$ from noise prior $P(\mathbf{z})$. Each is of K dims.

Compute $\mathbf{z}^i = \sigma_{\mathbf{z}|\mathbf{x}^i} \mathbf{v}^i + \mu_{\mathbf{z}|\mathbf{x}^i}$

Compute $\mu_{\mathbf{x}^i|\mathbf{z}}, \sigma_{\mathbf{x}^i|\mathbf{z}} \leftarrow P_\theta(\mathbf{z}^i)$

$\min_{\theta, \phi, \alpha} \sum_i \log N(\mathbf{x}^i | \mu_{\mathbf{x}^i|\mathbf{z}}, \sigma_{\mathbf{x}^i|\mathbf{z}}) + \sum_i \mu_{\mathbf{x}^i|\mathbf{z}}^2 + \sigma_{\mathbf{x}^i|\mathbf{z}}^2 - \log \sigma_{\mathbf{x}^i|\mathbf{z}}$

end for

Write the gradient of the training objective with respect to ϕ .

..2

2. Consider a 1-dimensional dataset D from a distribution $P_D(x)$ which is a mixture of three Gaussians with the three means at $\mu_1 = 10, \mu_2 = 20$, and $\mu_3 = 30$ each with variance of 1 and equal fraction of examples from each Gaussian. We will see how good GANs are in learning to generate samples from such a distribution.

- (a) First consider GANs. Say, as generator $G(z)$ we use a 1-d hidden variable $z \sim \mathcal{N}(0, 1)$ followed by a linear layer $\theta_1 z + \theta_2$ to generate an output x . Assume the discriminator $D_{\theta_d}(x)$ is all powerful and can assign exact Bayes probability $P(\text{real}|x)$ over the real distribution (from $D \sim P_D(x)$) and whatever generated distribution x it sees. Provide all values of θ_1, θ_2 for which the GAN objective will be maximized?

..2

- (b) Now, let us say that the generator is actually a mixture of three Gaussians $P_G(x) = \pi_1 \mathcal{N}(x; \mu_1, 1) + \pi_2 \mathcal{N}(x; \mu_2, 1) + \pi_3 \mathcal{N}(x; \mu_3, 1)$ where the generator parameters are $\theta_g = [\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3,]$, $\pi_1 + \pi_2 + \pi_3 = 1$. For this the hidden variable z will be a three-way multinomial variable with parameters π_1, π_2, π_3 and conditioned on z we sample a x from $\mathcal{N}(x; \mu_z, 1)$. The θ_g are trained using the GAN objective $\min_{\theta_g} \max_{\theta_d} [E_{x \sim P_D} \log D_{\theta_d}(x) + E_{x \sim P_G} \log(1 - D_{\theta_d}(x))]$. When the generator parameters θ_g are fixed at: $\pi_1 = 1, \mu_1 = 10$, what is the optimal solution to $\max_{D_{\theta_d}} [E_{x \sim P_D} \log D_{\theta_d}(x) + E_{x \sim P_G} \log(1 - D_{\theta_d}(x))]$? Choose all the correct answers and provide a brief explanation.

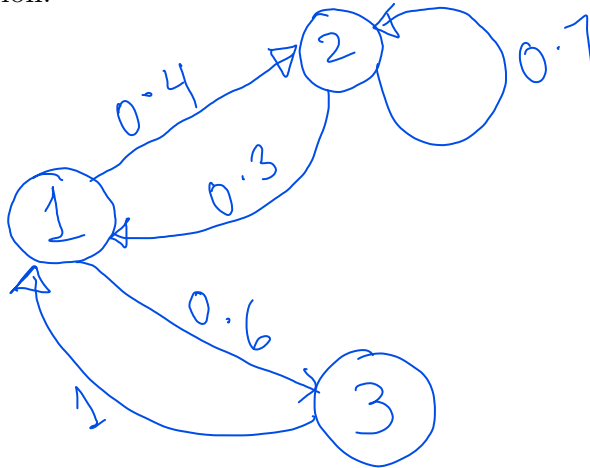
- i. $\log(1/3) + \log(1 - 0.5)$
- ii. $1/3 \log(0.5) + \log(0.5)$.
- iii. $1/3 \log(0.25) + \log(0.75)$
- iv. 0
- v. None of the above

..2

- (c) With the above discriminator parameter fixed, when the generator is retrained what are all configurations of θ_g values at which the generator objective is optimal? Choose all the correct answers and provide a brief explanation. ..2

- i. $\pi_1 = \pi_2 = \pi_3 = 1/3, \mu_1 = 10, \mu_2 = 20, \mu_3 = 30$
- ii. $\pi_1 = 0, \pi_2 = \pi_3 = 1/2, \mu_1 = 20, \mu_2 = 20, \mu_3 = 30$
- iii. $\pi_1 = 0, \pi_2 = 1, \pi_3 = 0, \mu_1 = 10, \mu_2 = 20, \mu_3 = 30$
- iv. $\pi_1 = 0.1, \pi_2 = 0.6, \pi_3 = 0.3, \mu_1 = 20, \mu_2 = 20, \mu_3 = 30$
- v. None of the above

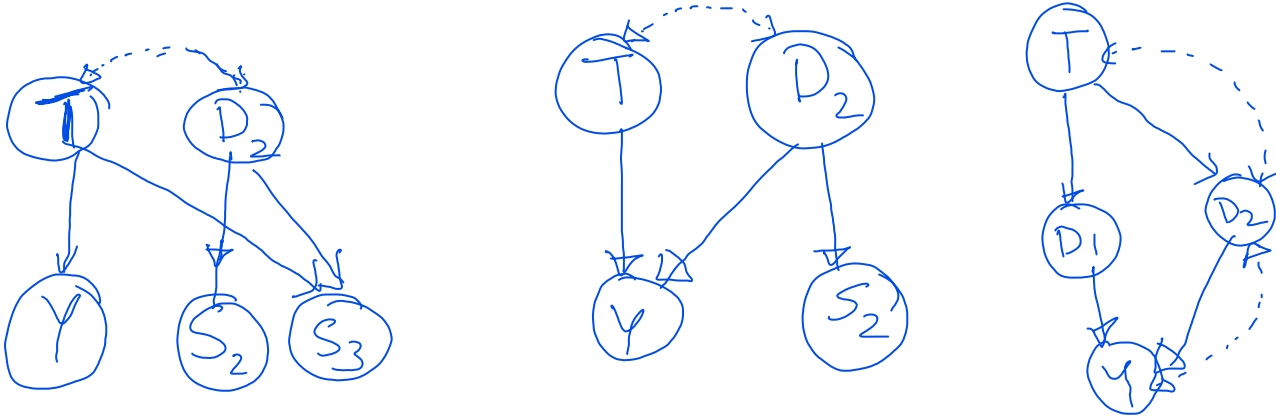
3. Consider a MCMC sampler over the following 3 variables with the transition probabilities given on the edge. State which of the following is true about this chain with a brief justification.



..2

- (a) The chain is not regular.
 - (b) The chain is regular since any state X' can be reached from any state X in exactly 2 steps.
 - (c) The chain is regular since any state X' can be reached from any state X in ≥ 2 steps.
 - (d) The chain is regular since any state X' can be reached from any state X in exactly 3 steps.
 - (e) The chain is regular since any state X' can be reached from any state X in ≤ 3 steps.
 - (f) The chain is regular but none of the reasons above are correct.
4. In the above graph, calculate the stationary distribution and write the value of $\pi_1 + \pi_2 - \pi_3$ below. Also, upload the derivation. ..2

5. State which of the following graphs are identifiable for answering the query $P(Y|do(T))$. The dotted lines denote presence of confounding factors between the connecting nodes. If possible write the expression that can be used to compute this value. ..3



6. Consider a Gaussian Process $f(x) \sim GP(m(x), K)$ where K is the polynomial kernel $K(x_1, x_2) = (1 + x_1 x_2)^2$ of degree p , where $x \in \mathbb{R}$ and mean $m(x) = x^2 - 2x + 2$ and we have no observation. What is the first point that we will acquire if $x' = \operatorname{argmin}_x \mu(x) - \sigma(x)$..2
7. In the Gaussian Copula paper, the log-likelihood of the training data is written as the sum of two terms:

$$\log p(\mathbf{z}; \mu, \Sigma) = \log \phi_{\mu, \Sigma}(\Phi^{-1}(\hat{F}(\mathbf{z}))) + \log \frac{d}{dz} \Phi^{-1}(\hat{F}(\mathbf{z}))$$

Explain the second term in the above objective using concepts that we have covered in the class? ..3

Total: 20