Week-4 Questions

Prof . Prathosh AP and Chandan J

May 2025

1. Consider two discrete probability distributions:

$$P = [0.5, 0.3, 0.2], \quad Q = [0.4, 0.4, 0.2]$$

Which of the following transport plans $\gamma \in \mathbb{R}^{3\times 3}$ is a **valid** transport plan from P to Q? That is,

(a)

$$\gamma = \begin{bmatrix} 0.4 & 0.1 & 0.0 \\ 0.0 & 0.2 & 0.1 \\ 0.0 & 0.1 & 0.1 \end{bmatrix}$$

(b)

$$\gamma = \begin{bmatrix} 0.3 & 0.2 & 0.0 \\ 0.1 & 0.1 & 0.1 \\ 0.0 & 0.1 & 0.1 \end{bmatrix}$$

(c)

$$\gamma = \begin{bmatrix} 0.4 & 0.0 & 0.1 \\ 0.0 & 0.3 & 0.0 \\ 0.0 & 0.1 & 0.1 \end{bmatrix}$$

(d)

$$\gamma = \begin{bmatrix} 0.2 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.0 \end{bmatrix}$$

Correct Answer: (c)

Explanation: Option (c) is the only plan where: - Row sums: $[0.4 + 0.0 + 0.1 = 0.5,\ 0.3,\ 0.2] = P$ - Column sums: $[0.4,\ 0.4,\ 0.2] = Q$ - All entries are non-negative

2. Which of the following best describes the **manifold hypothesis** commonly assumed in machine learning?

   (a) Real-world data in high-dimensional spaces is uniformly distributed across all directions of the space.

   (b) Data distributions lie on low-dimensional manifold embedded in the high-dimensional input space.

   (c) Neural networks create manifolds in feature space by projecting data to lower dimensions.

   (d) The data space can be fully explained using a finite number of axis-aligned hypercubes.

   Correct Answer: (b)

   Explanation: The manifold hypothesis suggests that although data like natural images live in a high-dimensional space (e.g., $\mathbb{R}^n$), they concentrate near a low-dimensional manifold due to inherent constraints in the data-generating process (e.g., lighting, pose, object type).

3. Why is the Earth Mover Distance (also called Wasserstein distance) preferred over Jensen-Shannon divergence in some GAN training setups?

   (a) It is symmetric and bounded between 0 and 1.

   (b) It yields meaningful gradients even when the generator and real data distributions have disjoint support.

   (c) It guarantees that the generator will eventually produce samples from a Gaussian distribution, regardless of the data distribution.

   (d) It slows down discriminator convergence, ensuring that the generator receives non-trivial gradients during training and avoids early saturation.

   Correct Answer: (b)

4. In WGAN, the Kantorovich-Rubinstein duality allows the Earth Mover distance to be expressed in terms of a supremum over which class of functions?

   (a) All measurable functions.

   (b) All convex functions.

   (c) All 1-Lipschitz functions.

   (d) All differentiable functions.

Correct Answer: (c)

5. Let $G(z)$ be the generator and $E(x)$ be the encoder in a BiGAN. The joint distribution of fake pairs is defined as $P_G(x, z)$, and real pairs as $P_E(x, z)$. The optimal discriminator $D^*(x, z)$ satisfies:

   (a) $D^*(x, z) = \frac{1}{1+\exp(-\|x-G(z)\|^2)}$

   (b) $D^*(x, z) = \frac{P_E(x,z)}{P_E(x,z)+P_G(x,z)}$

   (c) $D^*(x, z) = \frac{1}{2}$ when x and z are independent

   (d) $D^*(x, z) = \nabla_x \log P_E(x, z)$

   Correct Answer: (b)

6. In a PyTorch implementation of Wasserstein GAN (WGAN), which of the following lines is essential to enforce the Lipschitz constraint via weight clipping?

   (a) `for p in discriminator.parameters(): p.grad.clamp_(-0.01, 0.01)`

   (b) `for p in discriminator.parameters(): p.data.clamp_(-0.01, 0.01)`

   (c) `torch.nn.utils.clip_grad_norm_(discriminator.parameters(), max_norm=0.01)`

   (d) `optimizer_D.param_groups[0]['weight_decay'] = 0.01`

   Correct Answer: (b)

7. In a PyTorch BiGAN implementation, which part of the training loop is typically used to train the encoder E and generator G jointly?

   (a) `loss_EG = -torch.mean(discriminator(real_x, encoder(real_x)))`

   (b) `loss_EG = torch.mean(discriminator(fake_x, z))`

   (c) `loss_EG = -torch.mean(discriminator(fake_x.detach(), z.detach()))`

   (d) `loss_EG = F.mse_loss(encoder(G(z)), z)`

   Correct Answer: (a)

8. In a PyTorch implementation of Fréchet Inception Distance (FID), which of the following steps is **essential** for correctly computing the FID score between real and generated images?

   (a) Computing the pixel-wise mean squared error between real and generated images.

   (b) Extracting features from pretrained InceptionV3 network before classification head.

   (c) Calculating the L2 distance between the softmax outputs of real and fake images.

(d) Measuring the cross-entropy loss between the Inception predictions on real and generated images.

Correct Answer: (b)

9. In a BiGAN, the discriminator returns the following scores:

$$D(x_1, z_1) = 0.9, \quad D(x_2, z_2) = 0.8 \qquad \text{(real pairs)}$$
$$D(x_3, z_3) = 0.3, \quad D(x_4, z_4) = 0.2 \qquad \text{(fake pairs)}$$

Compute the total discriminator loss using binary cross-entropy:

$$\mathcal{L}_D = -\frac{1}{2}\sum_{i=1}^{2} \log D(x_i, z_i) - \frac{1}{2}\sum_{i=3}^{4} \log(1 - D(x_i, z_i))$$

(a) 0.22
(b) 0.35
(c) 0.51
(d) 0.57

Correct Answer: (d) (Approximate value from log terms)

10. In a WGAN, the critic outputs the following scores:

$$D(x_1) = 3.0, \quad D(x_2) = 2.5 \qquad \text{(real)}$$
$$D(G(z_1)) = -0.5, \quad D(G(z_2)) = -0.3 \qquad \text{(fake)}$$

Estimate the empirical Wasserstein-1 loss:

$$\mathcal{L}_{WGAN} = \frac{1}{2}\sum_{i=1}^{2} D(x_i) - \frac{1}{2}\sum_{i=1}^{2} D(G(z_i))$$

(a) 3.15
(b) 2.95
(c) 2.65
(d) 3.00

Correct Answer: (a) (Since $2.75 + 0.4 = 3.15$)