1 / 1 points

1.

Which of the following estimates are unbiased?

- [x] $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad X_i \overset{i.i.d.}{\sim} p$ for $\mathbb{E}X$.

Correct

- [x] $f\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right), \quad X_i \overset{i.i.d.}{\sim} p$ for $\mathbb{E}f(X)$, where f is a linear function.

Correct
For a linear function, $f\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$ which is an unbiased estimate for $\mathbb{E}f(X)$. (See Lec 1 starting at 2:03.)

- [ ] $f\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right), \quad X_i \overset{i.i.d.}{\sim} p$ for $\mathbb{E}f(X)$, where f is not linear.

Un-selected is correct

1 / 1 points

2.

In which of the following scenarios probabilistic model has a tractable density function $p(x)$, i.e. it is computationally easy to compute the density at any point x?

- [ ] x is an observable variable in an arbitrary latent variable model $p(x, z) = p(x \mid z)p(z)$

Un-selected is correct

- [ ] x is defined by a smooth transformation $f : \mathbb{R}^d \to \mathbb{R}^d$ of a random vector $z \sim \mathcal{N}(0, I)$: $x = f(\varepsilon)$.

Un-selected is correct

- [x] Distribution is defined according to the chain rule $p(X) = \prod_{i=1}^{D} p_i(x_i \mid x_1, \ldots, x_{i-1})$, $X = (x_1, \ldots, x_D)$ with tractable conditional distributions.

Correct
Indeed in this case we can always compute the density using the chain rule.

1 / 1 points

3.

Consider two different choices for the family of variational distributions for training a VAE:

1) $q(z_i \mid x_i, w) = \mathcal{N}(z_i \mid \mu(x_i, w), \text{diag}(\sigma^2(x_i, w)))$, where $\mu(\cdot, w)$ and $\sigma(\cdot, w)$ are deep neural networks with parameters w.

2) For each x_i the approximate posterior distribution over latent variable z_i is defined individually as a Gaussian distribution $q_i(z_i \mid x_i, \mu_i, \sigma_i) = \mathcal{N}(z_i \mid \mu_i, \text{diag}(\sigma_i^2))$.

In which case, you would expect to get higher (better) variational lower bound on the training set? What about the test set?

- [ ] 2 is better on the training set, 1 is better on the test set.
- [ ] 1 is better on the training set, 2 is better on the test set.
- [ ] 1 is better for both training and test sets.
- [x] 2 is better for both training and test sets.

Correct
Yes! The more flexible the variational family is, the closer it approximates the posterior. And good posterior distribution approximation is beneficial both for the training and test performance (i.e. you should not overfit because of approximating the posterior more accurately).

However, on practice we prefer option 1 since it has much fewer parameters (only one neural network instead of a bunch of parameters per each object) and simplify working with new objects (with option 1 you can easily obtain the variational distribution of a new object, while with option 2 you should first find the parameters μ, σ by solving an optimization problem).

1 / 1 points

4.

Suppose the class of the approximate posterior distributions is flexible enough to capture any distribution. The evidence lower bound is known to achieve the optimal value with respect to the variational distribution when the variational distribution coincides with the true posterior distribution $q(z \mid x, w) = p(z \mid x)$.

Imagine that you train VAE by optimizing the following loss: $\sum_i \mathbb{E}_{q(z|x_i,w)} \log[p(x_i \mid z)p(z)]$, which is the usual variational lower bound $\sum_i \mathbb{E}_{q(z|x_i,w)} \log p(x_i \mid z) - \mathcal{KL}(q(z \mid x_i, w) \,\|\, p(z))$ without the entropy term $E_{q(z|x_i,w)} \log q(z \mid x_i, w)$. Which variational distribution you will obtain after training?

- A delta function concentrated in the mode of the joint distribution $\operatorname{argmax}_z p(z, x)$.

Correct
Indeed, without the entropy term, nothing will stop the variational distribution from collapsing to a delta function that maximizes the objective. If there are multiple modes, the variational distribution can become a mixture of them.

- True posterior distribution $p(z \mid x)$.
- A delta function concentrated in the mode of the posterior $\operatorname{argmax}_z p(z \mid x)$.
- None of the above.

1 / 1 points

5.

Suppose that a random variable z_j does not contribute to the value of decoder. That is, no matter what the value z_j takes, the output of decoder is the same. What will be the distribution $q(z_j \mid x, w)$ after training?

- The component will be distributed according to the prior distribution $\mathcal{N}(z_j \mid 0, 1)$

Correct
Indeed, since in this case the reconstruction term of the loss doesn't depend on the distribution $q(z_j \mid x, w)$, the KL term will push the distribution closer and closer to the prior.

- Distribution $q(z_j \mid x, w)$ will not change during the training.
- None of the above.

1 / 1 points

6.

Recall the log derivative trick (or as it is sometimes called, REINFORCE algorithm). It estimates the gradient $\nabla_w \mathbb{E}_{q(z|w)} f(z)$ by using the following derivation $\nabla_w \mathbb{E}_{q(z|w)} f(z) = \int \nabla_w q(z \mid w) f(z) dz = \int \frac{q(z|w)}{q(z|w)} \nabla_w q(z \mid w) f(z) dz = \mathbb{E}_{q(z|w)} \nabla_w \log q(z \mid w) f(z)$

and then builds an unbiased estimate $\mathbb{E}_{q(z|w)} \nabla_w \log q(z \mid w) f(z) \approx \nabla_w \log q(\hat{z} \mid w) f(\hat{z})$, where $\hat{z} \sim q(z \mid w)$.

Which conditions are necessary for the log derivative trick to work?

- [ ] One has to represent the variational approximation $q(z \mid w)$ with a function g and a random variable ε such that $z = g(\varepsilon, w)$.

Un-selected is correct

- [ ] Density function $q(z \mid w)$ must be differentiable with respect to the random variable value z.

Un-selected is correct

- [ ] Density function $q(z \mid w)$ must be differentiable with respect to the distribution parameters w.

Correct
Since $\nabla_w \log q(z \mid w)$ appears in the final expression, the density should be differentiable w.r.t. w. But this is a very mild condition, and log derivative trick can be applied almost to any model (the price to pay is the high variance of this method).

- [ ] Function $f(z)$ should be differentiable w.r.t. z.

Un-selected is correct

0.50 / 1 points

7.

Recall the derivation of the reparametrization trick. It estimates the gradient $\nabla_w \mathbb{E}_{q(z|w)} f(z)$ by introducing a function g such that $g(\varepsilon, w)$ is distributed according to $q(z \mid w)$ where the distribution of ε doesn't depend on parameters w. Then, $\nabla_w \mathbb{E}_{q(z|w)} f(z) = \nabla_w \mathbb{E}_{p(\varepsilon)} f(g(\varepsilon, w)) = \mathbb{E}_{p(\varepsilon)} \nabla_w f(g(\varepsilon, w))$ which can be easily estimated in an unbiased way as $\nabla_w f(g(\hat{\varepsilon}, w))$ where $\hat{\varepsilon} \sim p(\varepsilon)$.

Which conditions are necessary for the reparametrization trick to work?

- [ ] One has to represent the variational approximation $q(z \mid w)$ with a function g and a random variable ε such that $z = g(\varepsilon, w)$.

This should be selected

- [ ] Density function $q(z \mid w)$ must be differentiable with respect to the random variable value z.

Un-selected is correct

- [ ] $g(\varepsilon, w)$ must be differentiable with respect to w.

Correct
Yes, because this derivative is used in the final expression $\nabla_w f(g(\hat{\varepsilon}, w))$

- [ ] Function $f(z)$ should be differentiable w.r.t. z.

This should be selected

✔ 1 / 1 points

8.

Which of the following distribution families can be used in the reparametrization trick?

- [ ] Bernoulli distribution.

Un-selected is correct

- [ ] Multivariate normal distribution with diagonal covariance matrix.

Correct
$z = g(\varepsilon, \mu, \sigma) = \varepsilon \odot \sigma + \mu$, where ε follows standard normal distribution and $\odot$ is element-wise multiplication. Latent variable z expressed this way follows $z \sim \mathcal{N}(\mu, \text{diag}(\sigma^2))$.

Note that $g(\varepsilon, \mu, \sigma)$ is differentiable w.r.t. both parameter-vectors μ and σ.

- [x] Multivariate normal distribution with full convariance matrix.

Correct
$z = g(\varepsilon, \mu, \Sigma) = \Sigma^{0.5}\varepsilon + \mu$, where ε follows standard normal distribution and $\Sigma^{0.5}$ is the square root of the matrix Σ, i.e. a matrix B such that $BB = \Sigma$. Latent variable z expressed this way follows $z \sim \mathcal{N}(\mu, \Sigma)$.

Note that $g(\varepsilon, \mu, \Sigma)$ is differentiable w.r.t. both the parameter-vector μ and the parameter-matrix Σ.