

✓ 恭喜！您通过了！

下一项



1 / 1 分

1。

Which of the following estimates are unbiased?

 $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i, X_i \stackrel{i.i.d.}{\sim} p \text{ for } \mathbb{E}X.$ 

正确

 $f\left(\frac{1}{N} \sum_{i=1}^N X_i\right), X_i \stackrel{i.i.d.}{\sim} p \text{ for } \mathbb{E}f(X), \text{ where } f \text{ is a linear function.}$ 

正确

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.)

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

未选择的是正确的



1 / 1 分

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  $x$ ?



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



正确

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



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



未选择的是正确的



$x$  is an observable variable in an arbitrary latent variable model  $p(x, z) = p(x | z)p(z)$



未选择的是正确的



1 / 1 分

3。

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

- 1)  $q(z_i | x_i, w) = \mathcal{N}(\mu(x_i, w), \text{diag}(\sigma^2(x_i, w)))$ , where  $\mu(\cdot, w)$  and  $\sigma(\cdot, w)$  are deep neural networks with  $w$ .
- 2) For each  $x_i$  the approximate posterior distribution over latent variable  $z_i$  is defined individually as a Gaussian distribution  $q_i(z_i | x_i, \mu_i, \sigma_i) = \mathcal{N}(z_i | \mu_i, \text{diag}(\sigma_i^2))$ .

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 for both training and test sets.

正确

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  $\mu, \sigma$  by solving an optimization problem).

- ☐ 1 is better for both training and test sets.
- ☐ 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 / 1 分

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  $q(z | x, w) = p(z | x)$ .

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

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

正确

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  $p(z | x)$ .
- ☐ A delta function concentrated in the mode of the posterior  $\text{argmax}_z p(z | x)$ .
- ☐ None of the above.



0 / 1 分

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  $q(z_j | x, w)$  after training?

☒ None of the above.

这个选项的答案不正确

- ☐ Distribution  $q(z_j | x, w)$  will not change during the training.
- ☐ The component will be distributed according to the prior distribution  $\mathcal{N}(z_j | 0, 1)$



1 / 1 分

## 6. Variational autoencoders

7/8 分 (87%)

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  $\nabla_w \mathbb{E}_{q(z|w)} f(z) = \int \nabla_w q(z|w) f(z) dz = \int \frac{q(z|w)}{q(z|w)} \nabla_w q(z|w) f(z) dz =$

$$\mathbb{E}_{q(z|w)} \nabla_w \log q(z|w) f(z)$$

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

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



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



未选择的是正确的



Density function  $q(z|w)$  must be differentiable with respect to the random variable value  $z$ .



未选择的是正确的



Density function  $q(z|w)$  must be differentiable with respect to the distribution parameters  $w$ .



正确

Since  $\nabla_w \log q(z|w)$  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.  $z$ .



未选择的是正确的



1 / 1 分

## 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  $q(z|w)$  where the distribution of  $\varepsilon$  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  $q(z|w)$  with a function  $g$  and a random variable  $\varepsilon$  such that  $z = g(\varepsilon, w)$ .



正确

This is the first step of the reparameterization trick.



Density function  $q(z|w)$  must be differentiable with respect to the random variable value  $z$ .



未选择的是正确的



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



正确

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.  $z$ .



正确

Yes, because the final expression  $\nabla_w f(g(\hat{\varepsilon}, w))$  includes the derivative of  $f$ . This means that the reparameterization trick is not applicable to discrete latent variables.

## Variational autoencoders

7/8 分 (87%)

✓ 测验 8 个问题 1/1 分

8。

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



Multivariate normal distribution with full covariance matrix.



正确

$z = g(\varepsilon, \mu, \Sigma) = \Sigma^{0.5} \varepsilon + \mu$ , where  $\varepsilon$  follows standard normal distribution and  $\Sigma^{0.5}$  is the square root of the matrix  $\Sigma$ , 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  $\mu$  and the parameter-matrix  $\Sigma$ .



Bernoulli distribution.



未选择的是正确的



Multivariate normal distribution with diagonal covariance matrix.



正确

$z = g(\varepsilon, \mu, \sigma) = \varepsilon \odot \sigma + \mu$ , where  $\varepsilon$  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  $\mu$  and  $\sigma$ .