Quiz: Variational Inference

STA414/2104 - Winter 2026

1. **ELBO Properties**
   Which of the following statements about the Evidence Lower Bound (ELBO) is true?

   (a) Maximizing the ELBO requires computing the exact marginal likelihood.

   (b) Maximizing the ELBO is equivalent to maximizing the Kullback-Leibler divergence between the approximate distribution and the true posterior.

   (c) The ELBO is strictly greater than the log marginal likelihood.

   (d) The ELBO can be expressed as the sum of the expected log-joint probability and the entropy of the variational distribution.

   Correct Answer: (d)
   Rationale: The primary advantage of the ELBO is that it can be optimized without needing to compute the intractable marginal likelihood. Maximizing the ELBO minimizes the KL divergence, and because KL divergence is non-negative, the ELBO is always less than or equal to the log marginal likelihood. Mathematically, it is formulated as the expected log-joint probability plus the entropy of the variational distribution.

2. **I-projection vs. M-projection**

   What is a key characteristic of the I-projection ($KL(q||p)$) commonly used in Variational Inference, as opposed to the M-projection ($KL(p||q)$)?

   (a) It is generally computationally intractable because it requires taking expectations with respect to the true posterior.

   (b) It tends to underestimate the support of the true target distribution or concentrates the probability mass on smaller regions.

   (c) It heavily penalizes the approximate distribution for missing mass where the true distribution has mass.

   (d) It guarantees matching the exact moments, such as the mean and covariance, of the true posterior.

   Correct Answer: (b)
   Rationale: The I-projection heavily penalizes the approximate distribution for placing mass where the true distribution has none. This leads it to seek local modes and generally underestimate the overall support or variance of the target distribution. Moment matching and penalizing missed mass are properties of the M-projection.

3. **Mean-Field Approximation Geometry**
Suppose the true posterior $p(z_1, z_2|x)$ for two latent variables is a highly correlated bivariate Gaussian. If you apply a Mean-Field variational approximation $q(z_1, z_2) = q_1(z_1)q_2(z_2)$ and optimize using the standard KL divergence, what is the most likely geometric result of the optimal q?

   (a) The approximation will perfectly match the marginal distributions $p(z_1|x)$ and $p(z_2|x)$ but fail to capture their correlation.

   (b) The approximation will perfectly match the covariance structure of the true posterior but have incorrect means.

   (c) The approximation will be an axis-aligned Gaussian that concentrates heavily in the center, strictly underestimating the marginal variances of both z_1 and z_2.

   (d) The approximation will be an axis-aligned Gaussian that significantly overestimates the marginal variances to ensure it covers the entire support of the true posterior.

   Correct Answer: (c)
   Rationale: The Mean-Field assumption forces the contours of the approximation to be axis-aligned. Because standard Variational Inference minimizes the I-projection ($KL(q||p)$), the approximation strongly avoids regions where the true posterior has low probability, resulting in an underestimation of the marginal variances.

4. **Reparameterization Trick**
When applying the reparameterization trick to optimize the ELBO using stochastic gradient descent, why is it mathematically necessary to express the latent variable as $z = T(\epsilon, \phi)$?

   (a) To decouple the randomness of the sampling process from the parameters, allowing the gradient operator to be safely moved inside the expectation.

   (b) To transform the true, intractable posterior into a standard Normal distribution.

   (c) To analytically compute the exact Kullback-Leibler divergence between the variational distribution and the prior.

   (d) To ensure that the variational distribution always belongs to the exponential family.

   Correct Answer: (a)
   Rationale: Taking the derivative of an expectation is difficult when the distribution being integrated over depends on the parameters being optimized. By drawing from a parameter-free base distribution ($\epsilon \sim p_0$) and applying a deterministic transformation, the expectation no longer depends on the parameters, making Monte Carlo gradient estimation straightforward.