Lecture 6: Quiz (Variational Inference)

Instructor: Thibault Randrianarisoa

Questions

Question 1 In the context of Variational Inference, what is the primary difference in behavior between optimizing the Reverse KL divergence $\mathbb{D}_{\mathrm{KL}}(Q||P)$ versus the Forward KL divergence $\mathbb{D}_{\mathrm{KL}}(P||Q)$ over Q?

a) Reverse KL is "mode-covering" and forces the approximation to bridge gaps between modes, leading to an overestimation of variance.

b) Reverse KL is "mode-seeking" and forces the approximation to avoid regions where the target has low probability, often locking onto a single peak.

c) Forward KL ensures that the approximation Q will be perfectly symmetric, regardless of the shape of the target distribution P.

d) There is no practical difference: standard Variational Inference uses both interchangeably since they yield the same optimal parameters.

Question 2 Why is the objective function $\mathcal{L}(\phi)$ maximized in Variational Inference referred to as the Evidence Lower Bound (ELBO)?

a) It provides a strict lower bound on the true posterior variance, ensuring the approximation does not become too narrow.

b) It is a lower bound on the Kullback-Leibler divergence between the prior and the posterior distributions.

c) It is derived using Jensen's inequality and provides a mathematical lower bound on the marginal likelihood $\log p(\mathbf{X})$.

d) It minimizes the evidence required to make a statistically significant inference about the model parameters.

Question 3 What problem does the Reparameterization Trick solve in Stochastic Variational Inference?

a) It decouples the randomness from the variational parameters, allowing the gradient of an expectation to be computed as the expectation of a gradient using Monte Carlo methods.

b) It transforms an intractable multimodal posterior into a sequence of simpler, mutually independent Gaussian distributions.

c) It forces a mean-field approximation to account for the correlation between different parameters, fixing its tendency to underestimate variance.

d) It allows MCMC methods, like the Gibbs Sampler, to be directly embedded within the Variational Inference optimization loop.

Question 4 Which of the following describes a key characteristic and consequence of using a Mean-Field Approximation for the variational family?

a) It assumes all parameters are mutually independent, which, when coupled with the mode-seeking behavior of the Reverse KL, typically leads to an underestimation of the posterior variance.

b) It assumes the true posterior is perfectly Gaussian, which strictly limits its use to linear regression models.

c) It assumes that the expectation of the variational approximation is exactly equal to the expectation of the posterior.

d) It guarantees that the resulting ELBO is exactly equal to the log evidence, effectively making the approximation exact.

Question 5 How does Coordinate Ascent Variational Inference (CAVI) conceptually relate to Gibbs Sampling?

a) CAVI draws random samples from the exact complete conditional distributions, exactly like Gibbs Sampling, but runs in parallel.

b) Both methods rely on the complete conditional distributions: however, while Gibbs sampling draws random samples from them, CAVI iteratively updates each variational factor using the expected log of the conditional.

c) CAVI is used exclusively when the complete conditional distributions are unknown, whereas Gibbs sampling is used when they are available in closed form.

d) There is no relationship: CAVI is an optimization method based on gradients, while Gibbs Sampling is a Markov Chain technique.

Solutions

1. Correct Answer: b) Reverse KL is "mode-seeking" and forces the approximation to avoid regions where the target has low probability, often locking onto a single peak.
Reasoning: Standard Variational Inference minimizes the Reverse KL divergence. To minimize $\mathbb{E}_Q[\log(q/p)]$, $q(x)$ must be near zero wherever $p(x)$ is near zero to prevent the ratio from blowing up. This results in "mode-seeking" behavior where the approximation avoids low-probability regions of the target, often capturing only one mode of a multimodal distribution and underestimating the variance.

2. Correct Answer: c) It is derived using Jensen's inequality and provides a mathematical lower bound on the marginal likelihood (the log evidence), $\log p(\mathbf{X})$**.**
Reasoning: The ELBO is derived by applying Jensen's inequality to the log of the marginal likelihood (the evidence), utilizing the concavity of the logarithm. Because $\mathcal{L}(\phi) = \log p(\mathbf{X}) - \mathbb{D}_{\mathrm{KL}}(q_\phi || \pi(\theta|\mathbf{X}))$ and the KL divergence is always non-negative, $\mathcal{L}(\phi)$ acts as a tight lower bound on the log evidence $\log p(\mathbf{X})$.

3. Correct Answer: a) It decouples the randomness from the variational parameters, allowing the gradient of an expectation to be computed as the expectation of a gradient using Monte Carlo methods.
Reasoning: When computing the gradient of the ELBO, if the distribution q_ϕ depends on the parameters ϕ, passing the gradient inside the integral creates terms that are not expectations, preventing standard Monte Carlo estimation. The reparameterization trick solves this by expressing the random variable θ as a deterministic function of a parameter-free noise variable ϵ, allowing the gradient to safely pass inside the expectation.

4. Correct Answer: a) It assumes all parameters are mutually independent, which, when coupled with the mode-seeking behavior of the Reverse KL, typically leads to an underestimation of the posterior variance.
Reasoning: The mean-field family is defined by factorizing the distribution such that the coordinates of the parameters are mutually independent ($q(\theta) = \prod q_j(\theta_j)$). Because it cannot capture correlations between variables, and because it relies on the variance-underestimating Reverse KL divergence, it tends to severely underestimate the true variance of the posterior.

5. Correct Answer: b) Both methods rely on the complete conditional distributions; however, while Gibbs sampling draws random samples from them, CAVI iteratively updates each variational factor using the expected log of the conditional.
Reasoning: Both Gibbs sampling and CAVI isolate individual parameters and rely on the complete conditional distribution of that parameter given the others. The difference is mechanical: a Gibbs sampler takes a random realization (sample) from this conditional, whereas CAVI takes the expected value of the log of this complete conditional to deterministically update the variational factor.