Ex: Unsupervised POS Tagging

Given: sentences only (concatenated together)
Goal: infer tags for unlabeled sentences
Model: Bayesian HMM

\[ \log p(x|z) = \log \mathbb{E}_z p(x,z|\alpha) = \log \mathbb{E}_z \mathbb{E}_r p(\omega, \xi, \phi_e, \phi_r | \beta_e, \beta_r) \]

**Taming Objective**

Using the ELBO:

\[ \log p(x|\alpha) \geq \text{ELBO}(q_\theta) = E_{q_\theta} [\log p(x,z)] - E_{q_\theta} [\log q_\theta(z)] \]

\[ = E_{q_\theta} [\log p(\omega, \xi, \phi_e, \phi_r | \beta_e, \beta_r)] - E_{q_\theta} [\log q_\theta(t, \phi_e, \phi_r)] \]

**Variational Inference gives:**

\[ q_\theta(z) \approx p(x|z) \Rightarrow q_\theta(\xi) \approx p(\xi|\beta) \quad q_\theta(\phi_e) \approx p(\phi_e|\beta) \quad q_\theta(\phi_r) \approx p(\phi_r|\beta) \]

**Problem:** How to estimate \( \alpha \)?

**New Idea:** Jointly find \( p(x) \) and \( q_\theta \) to make ELBO as large as possible.

**Two Approximation:**

1. **Approximate Learning:** choosing \( \alpha \) for \( p \) can't compute / by assumption
Approximations:

1. Approximate Learning: choosing $\alpha$ for $p$
   - really want to maximize $\log p(x | \alpha)$
   - instead maximize a variational lower bound

2. Approximate Inference: choosing $\Theta$ for $q$
   - after reaching a (local) maximum of $(p_x, q_\theta)$
     query $q_\theta$ approximately about $z$
     \( \text{because directly querying } p_x \text{ about } z \text{ is intractable} \)

   Key Idea: jointly optimize ELBO as a function of both $\alpha$ and $\Theta$

Variational EM

Apply Block Coordinate Ascent to ELBO

Algorithm:
- While not converged:
  1. Variational E-step:
     - adjust $q_\Theta$ given current $p_x$
     - i.e. run VI to minimize $KL(q_\theta \| p_x)$ for $\Theta$ only
     - maximize ELBO($q_\theta; p_x$)
     - $\Theta = \arg\max_\Theta \text{ELBO}(q_\theta; p_x) = \text{run-variational-inf}(q_\theta; p_x)$
  2. Variational M-step:
     - adjust $p_x$ given current $q_\theta$
     - i.e. improve $\text{ELBO}(q_\theta; p_x)$ for $\alpha$ only
     - only one term involves $\alpha$
     - $\alpha = \arg\max_\alpha \text{ELBO}(q_\theta; p_x) = \text{argmax}_\alpha E_{q_\theta}[\log p_x(x, z)]$

Key Difference between Standard EM and Variational EM

We say that the "variational gap" in Standard EM is zero,

i.e. $\alpha(z)$ perfectly estimates $p_z(z | x)$
Standard EM is zero.

i.e. \( q_{\Theta}(z) \) perfectly estimates \( p_{\Theta}(z|x) \).