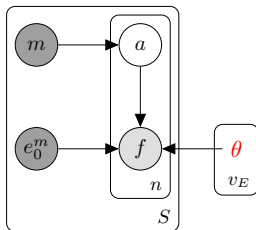


# Dirichlet priors for IBM model 1

Wilker Aziz

April 26, 2018

# MLE IBM 1



## Global variables

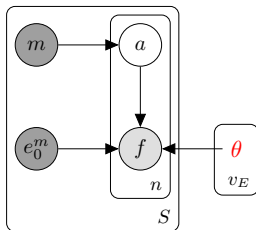
- ▶ For each English type  $e$ , we have a vector  $\theta_e$  of categorical parameters

- ▶  $0 < \theta_e < 1$

- ▶  $\sum_{f \in \mathcal{F}} \theta_{e,f} = 1$

and  $P_{F|E}(f|e) = \text{Cat}(F = f|\theta_e) = \theta_{e,f}$

# MLE IBM 1



## Global variables

- ▶ For each English type  $e$ , we have a vector  $\theta_e$  of categorical parameters
    - ▶  $0 < \theta_e < 1$
    - ▶  $\sum_{f \in \mathcal{F}} \theta_{e,f} = 1$
- and  $P_{F|E}(f|e) = \text{Cat}(F = f|\theta_e) = \theta_{e,f}$

## Local assignments

- ▶ For each French word position  $j$ ,

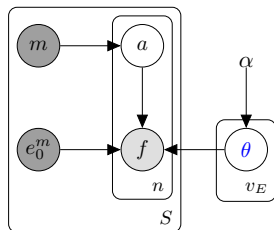
$$A_j \sim \mathcal{U}(0 \dots m)$$

$$F_j|e_{a_j} \sim \text{Cat}(\theta_{e_{a_j}})$$

# Bayesian IBM 1

## Global assignments

- For each English type  $e$ ,  
sample categorical parameters



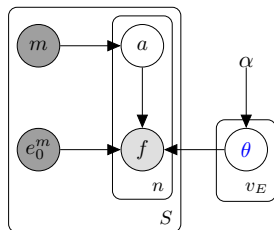
$$\theta_e \sim \text{Dir}(\alpha)$$

# Bayesian IBM 1

## Global assignments

- For each English type  $e$ ,  
sample categorical parameters

$$\theta_e \sim \text{Dir}(\alpha)$$



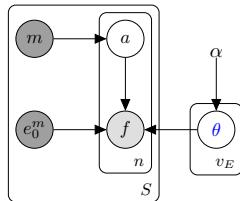
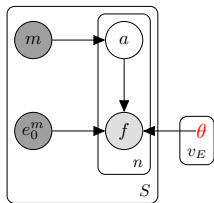
## Local assignments

- For each French word position  $j$ ,

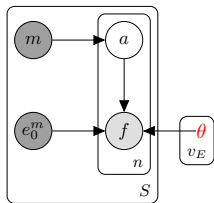
$$A_j \sim \mathcal{U}(0 \dots m)$$

$$F_j | e_{a_j} \sim \text{Cat}(\theta_{e_{a_j}})$$

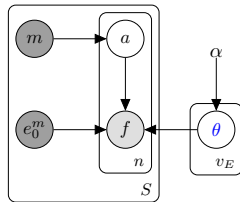
# MLE vs Bayesian IBM1



# MLE vs Bayesian IBM1

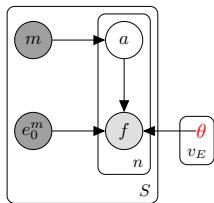


Incomplete data likelihood

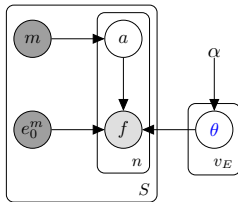


$$P(f_1^n | e_1^m, \theta_1^{v_E}) = \prod_{j=1}^n \underbrace{\sum_{a_j=0}^m \overbrace{P(a_j | m) P(f_j | e_{a_j}, \theta_1^{v_E})}^{P(f_j, a_j | e_1^m, \theta_1^{v_E})}}_{P(f_j | e_1^m, \theta_1^{v_E})} \quad (1)$$

# MLE vs Bayesian IBM1



Incomplete data likelihood



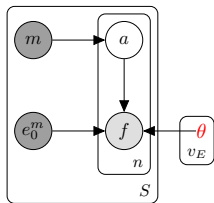
$$P(f_1^n | e_1^m, \theta_1^{v_E}) = \prod_{j=1}^n \underbrace{\sum_{a_j=0}^m \overbrace{P(a_j|m)P(f_j|e_{a_j}, \theta_1^{v_E})}^{P(f_j, a_j | e_1^m, \theta_1^{v_E})}}_{P(f_j | e_1^m, \theta_1^{v_E})} \quad (1)$$

Marginal likelihood (evidence)

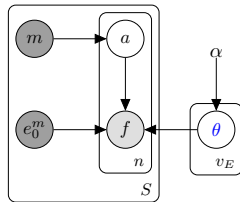
$$P(f_1^n | e_1^m, \alpha) = \int p(\theta_1^{v_E} | \alpha) P(f_1^n | e_1^m, \theta_1^{v_E}) d\theta_1^{v_E}$$



# MLE vs Bayesian IBM1



Incomplete data likelihood



$$P(f_1^n | e_1^m, \theta_1^{v_E}) = \prod_{j=1}^n \underbrace{\sum_{a_j=0}^m \overbrace{P(a_j|m)P(f_j|e_{a_j}, \theta_1^{v_E})}^{P(f_j, a_j | e_1^m, \theta_1^{v_E})}}_{P(f_j | e_1^m, \theta_1^{v_E})} \quad (1)$$

Marginal likelihood (evidence)

$$\begin{aligned} P(f_1^n | e_1^m, \alpha) &= \int p(\theta_1^{v_E} | \alpha) P(f_1^n | e_1^m, \theta_1^{v_E}) d\theta_1^{v_E} \\ &= \int p(\theta_1^{v_E} | \alpha) \prod_{j=1}^n \sum_{a_j=0}^m P(a_j|m) P(f_j | e_{a_j}, \theta_{e_{a_j}}) d\theta_1^{v_E} \end{aligned} \quad (2)$$

# What is a Dirichlet distribution?

Dirichlet:  $\theta_e \sim \text{Dir}(\alpha)$  with  $\alpha \in \mathbb{R}_{>0}^{v_F}$

$$\text{Dir}(\theta_e|\alpha) = \frac{\Gamma(\sum_{f \in \mathcal{F}} \alpha_f)}{\prod_{f \in \mathcal{F}} \Gamma(\alpha_f)} \prod_{f \in \mathcal{F}} \theta_{e,f}^{\alpha_f-1} \quad (3)$$

- ▶ an exponential family distribution over probability vectors
- ▶ each outcome is a  $v_F$ -dimensional vector of probability values that sum to 1
- ▶ can be used as a prior over the parameters of a Categorical distribution
- ▶ that is, a Dirichlet sample can be used to specify a Categorical distribution  
e.g.  $F|E = e \sim \text{Cat}(\theta_e)$

Use this [notebook](#) and this [wikipedia](#) to learn more

# Why a Dirichlet prior on parameters?

If we set the components of  $\alpha$  to the same value, we get a symmetric Dirichlet, if that value is small the Dirichlet will prefer

- ▶ samples that are very peaked
- ▶ in other words, categorical distributions that concentrate on few outcomes

# Why a Dirichlet prior on parameters?

If we set the components of  $\alpha$  to the same value, we get a symmetric Dirichlet, if that value is small the Dirichlet will prefer

- ▶ samples that are very peaked
- ▶ in other words, categorical distributions that concentrate on few outcomes

In MLE we choose one fixed set of parameters (via EM)

# Why a Dirichlet prior on parameters?

If we set the components of  $\alpha$  to the same value, we get a symmetric Dirichlet, if that value is small the Dirichlet will prefer

- ▶ samples that are very peaked
- ▶ in other words, categorical distributions that concentrate on few outcomes

In MLE we choose one fixed set of parameters (via EM)

In Bayesian modelling we average over all possible parameters

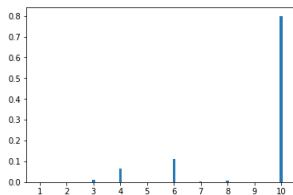
- ▶ where each parameter set is weighted by a prior belief
- ▶ we can use this as an opportunity to, for example, express our preferences towards “peaked models”

# Contrast the Dirichlet samples

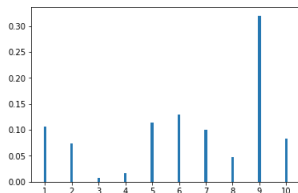
Top: sparse Dirichlet prior (small alpha)

- configurations that are this sparse will be roughly as likely
- less sparse configurations will be less likely
- “the prior doesn’t care where the tall bars are, as long as they are few”

```
plot_dirichlet_samples(alpha=0.1, nb_samples=1)
```

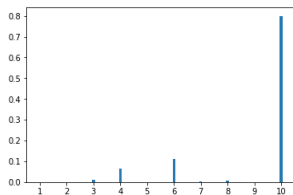


```
plot_dirichlet_samples(alpha=1, nb_samples=1)
```

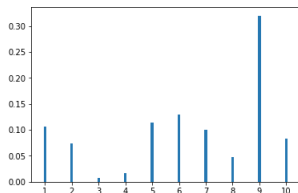


# Contrast the Dirichlet samples

```
plot_dirichlet_samples(alpha=0.1, nb_samples=1)
```



```
plot_dirichlet_samples(alpha=1, nb_samples=1)
```



Top: sparse Dirichlet prior (small alpha)

- configurations that are this sparse will be roughly as likely
- less sparse configurations will be less likely
- “the prior doesn’t care where the tall bars are, as long as they are few”

Take samples from the top Dirichlet to parameterise a Categorical distribution conditioning on English word “dog”

- locations of the bars correspond to French words in the vocabulary
- the prior basically expresses the belief that whatever “dog” translates to, there shouldn’t be many likely options available in French

## An alternative way to write the likelihood

We can write a likelihood based on Categorical events as follows

$$\begin{aligned} P(f_1^n, a_1^n | e_1^m, \theta_1^{v_E}) &= \prod_{j=1}^n \underbrace{P(a_j | m)}_{\frac{1}{m+1}} \underbrace{P(f_j | e_{a_j}, \theta_1^{v_E})}_{\theta_{f_j | e_{a_j}}} \\ &= \frac{1}{(m+1)^n} \prod_{j=1}^n \theta_{f_j | e_{a_j}} \end{aligned} \tag{4}$$

---

I use  $\theta_{e,f}$ ,  $\theta_{e \rightarrow f}$ , and  $\theta_{f|e}$  interchangeably



## An alternative way to write the likelihood

We can write a likelihood based on Categorical events as follows

$$\begin{aligned} P(f_1^n, a_1^n | e_1^m, \theta_1^{v_E}) &= \prod_{j=1}^n \underbrace{P(a_j | m)}_{\frac{1}{m+1}} \underbrace{P(f_j | e_{a_j}, \theta_1^{v_E})}_{\theta_{f_j | e_{a_j}}} \\ &= \frac{1}{(m+1)^n} \prod_{j=1}^n \theta_{f_j | e_{a_j}} \end{aligned} \quad (4)$$

an alternative way iterates over the vocabulary of pairs, rather than over the sentence

$$P(f_1^n, a_1^n | e_1^m, \theta_1^{v_E}) \propto \prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)} \quad (5)$$

where  $\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)$  counts how many times  $e$  and  $f$  are aligned in the sentence pair  $f_1^n, e_1^m$  given the alignments  $a_1^n$

---

I use  $\theta_{e,f}$ ,  $\theta_{e \rightarrow f}$ , and  $\theta_{f|e}$  interchangeably

## An alternative way to write the likelihood (cont)

The new form reveals similarities to the Dirichlet

Dirichlet prior

$$p(\theta_1^{v_E} | \alpha) = \overbrace{\prod_{e \in \mathcal{E}} \text{Dir}(\theta_e | \alpha)}^{\text{independent priors}} = \prod_{e \in \mathcal{E}} \frac{\Gamma(\sum_{f \in \mathcal{F}} \alpha_f)}{\prod_{f \in \mathcal{F}} \Gamma(\alpha_f)} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\alpha_f - 1} \quad (6)$$

Multinomial (or Categorical likelihood)

$$P(f_1^n, a_1^n | e_1^m, \theta) \propto \prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)} \quad (7)$$

## An alternative way to write the likelihood (cont)

The new form reveals similarities to the Dirichlet

Dirichlet prior

$$p(\theta_1^{v_E} | \alpha) = \overbrace{\prod_{e \in \mathcal{E}} \text{Dir}(\theta_e | \alpha)}^{\text{independent priors}} = \prod_{e \in \mathcal{E}} \frac{\Gamma(\sum_{f \in \mathcal{F}} \alpha_f)}{\prod_{f \in \mathcal{F}} \Gamma(\alpha_f)} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\alpha_f - 1} \quad (6)$$

Multinomial (or Categorical likelihood)

$$P(f_1^n, a_1^n | e_1^m, \theta) \propto \prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)} \quad (7)$$

Thus

$$\begin{aligned} p(\theta_1^{v_E}, f_1^n, a_1^n | e_1^m, \alpha) &= p(\theta_1^{v_E} | \alpha) p(f_1^n, a_1^n | e_1^m, \theta_1^{v_E}) \\ &\propto \prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \underbrace{\theta_{f|e}^{\alpha_f - 1} \times \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)}} \end{aligned}$$

## An alternative way to write the likelihood (cont)

The new form reveals similarities to the Dirichlet

Dirichlet prior

$$p(\theta_1^{v_E} | \alpha) = \overbrace{\prod_{e \in \mathcal{E}} \text{Dir}(\theta_e | \alpha)}^{\text{independent priors}} = \prod_{e \in \mathcal{E}} \frac{\Gamma(\sum_{f \in \mathcal{F}} \alpha_f)}{\prod_{f \in \mathcal{F}} \Gamma(\alpha_f)} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\alpha_f - 1} \quad (6)$$

Multinomial (or Categorical likelihood)

$$P(f_1^n, a_1^n | e_1^m, \theta) \propto \prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)} \quad (7)$$

Thus

$$\begin{aligned} p(\theta_1^{v_E}, f_1^n, a_1^n | e_1^m, \alpha) &= p(\theta_1^{v_E} | \alpha) p(f_1^n, a_1^n | e_1^m, \theta_1^{v_E}) \\ &\propto \prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \underbrace{\theta_{f|e}^{\alpha_f - 1} \times \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)}}_{\theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m) + \alpha_f - 1}} \end{aligned} \quad (8)$$

# Bayesian IBM 1: Joint Distribution

Sentence pair:  $(e_0^m, f_1^n)$

$$p(f_1^n, a_1^n, \theta_1^{vE} | e_0^m, \alpha) = \overbrace{P(a_1^n | m)}^{\text{constant}} \underbrace{\prod_{e \in \mathcal{E}}}_{\text{English types}} \overbrace{p(\theta_e | \alpha)}^{\text{Dir prior}} \overbrace{\prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)}}^{\text{likelihood}} \underbrace{\prod_{f \in \mathcal{F}}}_{\text{French types}}$$

# Bayesian IBM 1: Joint Distribution

Sentence pair:  $(e_0^m, f_1^n)$

$$\begin{aligned}
 p(f_1^n, a_1^n, \theta_1^{vE} | e_0^m, \alpha) &= \overbrace{P(a_1^n | m)}^{\text{constant}} \underbrace{\prod_{e \in \mathcal{E}}}_{\text{English types}} \overbrace{p(\theta_e | \alpha)}^{\text{Dir prior}} \overbrace{\prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)}}^{\text{likelihood}} \\
 &= P(a_1^n | m) \underbrace{\prod_e \frac{\Gamma(\sum_f \alpha_f)}{\prod_f \Gamma(\alpha_f)}}_{\text{Dirichlet}} \underbrace{\prod_f \theta_{f|e}^{\alpha_f - 1} \prod_f \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)}}_{\text{Categorical}}
 \end{aligned}$$

# Bayesian IBM 1: Joint Distribution

Sentence pair:  $(e_0^m, f_1^n)$

$$\begin{aligned}
 p(f_1^n, a_1^n, \theta_1^{vE} | e_0^m, \alpha) &= \overbrace{P(a_1^n | m)}^{\text{constant}} \underbrace{\prod_{e \in \mathcal{E}}}_{\text{English types}} \overbrace{p(\theta_e | \alpha)}^{\text{Dir prior}} \overbrace{\prod_{e \in \mathcal{E}} \prod_{f \in \mathcal{F}} \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)}}^{\text{likelihood}} \\
 &= P(a_1^n | m) \underbrace{\prod_e \frac{\Gamma(\sum_f \alpha_f)}{\prod_f \Gamma(\alpha_f)}}_{\text{Dirichlet}} \underbrace{\prod_f \theta_{f|e}^{\alpha_f - 1} \prod_f \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m)}}_{\text{Categorical}} \\
 &\propto P(a_1^n | m) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | a_1^n) + \alpha_f - 1}
 \end{aligned}$$

# Bayesian IBM 1: Joint Distribution (II)

Sentence pair:  $(e_0^m, f_1^n)$

$$p(f_1^n, a_1^n, \theta_1^{vE} | e_0^m, \alpha) \propto P(a_1^n | m) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m) + \alpha_f - 1} \quad (9)$$

Corpus:  $(\mathbf{e}, \mathbf{f})$

$$\begin{aligned} p(\mathbf{f}, \mathbf{a}, \theta_1^{vE} | \mathbf{e}, \mathbf{m}, \alpha) &\propto \prod_{(e_0^m, f_1^n, a_1^n)} P(a_1^n | m) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | f_1^n, a_1^n, e_1^m) + \alpha_f - 1} \\ &= P(\mathbf{a} | \mathbf{m}) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | \mathbf{f}, \mathbf{a}, \mathbf{e}) + \alpha_f - 1} \end{aligned} \quad (10)$$

where I use boldface to indicate the collection



# Bayesian IBM 1: Inference

In Bayesian modelling **there is no optimisation**

- ▶ we **do not** pick one model
- ▶ instead, we infer a posterior distribution over unknowns and reason using all models (or a representative sample)

# Bayesian IBM 1: Posterior

Intractable marginalisation

$$p(\mathbf{a}, \theta_1^{vE} | \mathbf{e}, \mathbf{m}, \mathbf{f}, \alpha) = \frac{p(\mathbf{f}, \mathbf{a}, \theta | \mathbf{e}, \mathbf{m}, \alpha)}{\int \sum_{\mathbf{a}'} p(\mathbf{f}, \mathbf{a}', \theta' | \mathbf{e}, \mathbf{m}, \alpha) d\theta'} \quad (11)$$

- ▶  $\theta_1^{vE}$  are global variables: posterior depends on the entire corpus
- ▶ the summation goes over every possible alignment configuration for every possible parameter setting

# Bayesian IBM 1: Approximate inference

Traditionally, we would approach posterior inference with an approximate algorithm such as Markov chain Monte Carlo

- ▶ based on sampling from the posterior by sampling one variable at a time and forming a chain whose stationary distribution is the true posterior

# Bayesian IBM 1: Approximate inference

Traditionally, we would approach posterior inference with an approximate algorithm such as Markov chain Monte Carlo

- ▶ based on sampling from the posterior by sampling one variable at a time and forming a chain whose stationary distribution is the true posterior

MCMC is fully general, but can be hard to derive, and can be slow in practice

# Variational inference

Optimise an auxiliary model to perform inference

# Variational inference

Optimise an auxiliary model to perform inference

- ▶ postulate a family  $\mathcal{Q}$  of tractable approximations  $q(z)$  to true posterior  $p(z|x)$   
where  $z$  are latent variables and  $x$  are observations

# Variational inference

Optimise an auxiliary model to perform inference

- ▶ postulate a family  $\mathcal{Q}$  of tractable approximations  $q(z)$  to true posterior  $p(z|x)$   
where  $z$  are latent variables and  $x$  are observations
- ▶ pick the member  $q^*$  of  $\mathcal{Q}$  that is closest to  $p(z|x)$   
measure closeness with KL divergence [wikipedia](#)

# Variational inference

Optimise an auxiliary model to perform inference

- ▶ postulate a family  $\mathcal{Q}$  of tractable approximations  $q(z)$  to true posterior  $p(z|x)$   
where  $z$  are latent variables and  $x$  are observations
- ▶ pick the member  $q^*$  of  $\mathcal{Q}$  that is closest to  $p(z|x)$   
measure closeness with KL divergence [wikipedia](#)
- ▶ use tractable  $q^*$  instead of  $p$  for inference and predictions



# Variational inference

Optimise an auxiliary model to perform inference

- ▶ postulate a family  $\mathcal{Q}$  of tractable approximations  $q(z)$  to true posterior  $p(z|x)$   
where  $z$  are latent variables and  $x$  are observations
- ▶ pick the member  $q^*$  of  $\mathcal{Q}$  that is closest to  $p(z|x)$   
measure closeness with KL divergence [wikipedia](#)
- ▶ use tractable  $q^*$  instead of  $p$  for inference and predictions

Objective

$$q^* = \arg \min_{q \in \mathcal{Q}} \text{KL}(q(z) || p(z|x))$$

# Variational inference

Optimise an auxiliary model to perform inference

- ▶ postulate a family  $\mathcal{Q}$  of tractable approximations  $q(z)$  to true posterior  $p(z|x)$   
where  $z$  are latent variables and  $x$  are observations
- ▶ pick the member  $q^*$  of  $\mathcal{Q}$  that is closest to  $p(z|x)$   
measure closeness with KL divergence [wikipedia](#)
- ▶ use tractable  $q^*$  instead of  $p$  for inference and predictions

Objective

$$\begin{aligned} q^* &= \arg \min_{q \in \mathcal{Q}} \text{KL}(q(z) || p(z|x)) \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z|x)} \right] \end{aligned} \tag{12}$$

## Variational Inference - Objective

The original objective is intractable due to posterior

$$q^* = \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z|x)} \right]$$

# Variational Inference - Objective

The original objective is intractable due to posterior

$$\begin{aligned} q^* &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\textcolor{red}{p}(z|x)} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\frac{p(z,x)}{\textcolor{red}{p}(x)}} \right] \end{aligned}$$

# Variational Inference - Objective

The original objective is intractable due to posterior

$$\begin{aligned} q^* &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\textcolor{red}{p}(z|x)} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\frac{p(z,x)}{\textcolor{red}{p}(x)}} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z, x)} \right] + \underbrace{\textcolor{red}{\log p(x)}}_{\text{constant}} \end{aligned}$$

# Variational Inference - Objective

The original objective is intractable due to posterior

$$\begin{aligned} q^* &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\textcolor{red}{p}(z|x)} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\frac{p(z,x)}{\textcolor{red}{p}(x)}} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z, x)} \right] + \underbrace{\textcolor{red}{\log p(x)}}_{\text{constant}} \\ &= \arg \min_{q \in \mathcal{Q}} - \mathbb{E}_{q(z)} \left[ \log \frac{p(z, x)}{q(z)} \right] \end{aligned}$$

# Variational Inference - Objective

The original objective is intractable due to posterior

$$\begin{aligned} q^* &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z|x)} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\frac{p(z,x)}{p(x)}} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z,x)} \right] + \underbrace{\log p(x)}_{\text{constant}} \\ &= \arg \min_{q \in \mathcal{Q}} - \mathbb{E}_{q(z)} \left[ \log \frac{p(z,x)}{q(z)} \right] \\ &= \arg \max_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{p(z,x)}{q(z)} \right] \end{aligned}$$

# Variational Inference - Objective

The original objective is intractable due to posterior

$$\begin{aligned} q^* &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z|x)} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\frac{p(z,x)}{p(x)}} \right] \\ &= \arg \min_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{p(z,x)} \right] + \underbrace{\log p(x)}_{\text{constant}} \\ &= \arg \min_{q \in \mathcal{Q}} - \mathbb{E}_{q(z)} \left[ \log \frac{p(z,x)}{q(z)} \right] \\ &= \arg \max_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} \left[ \log \frac{p(z,x)}{q(z)} \right] \\ &= \arg \max_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} [\log p(z,x)] - \underbrace{\mathbb{E}_{q(z)} [\log q(z)]}_{\mathbb{H}(q(z))} \end{aligned}$$



## Evidence lowerbound (ELBO)

We've shown that minimising  $\text{KL}(q(z)||p(z|x))$  is equivalent to maximising a simpler objective

$$q^* = \arg \max_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} [\log p(z, x)] + \mathbb{H}(q(z))$$

known as the evidence lowerbound

---

The name ELBO has to do with the fact that  $\log p(x) \geq \text{ELBO}$

# Evidence lowerbound (ELBO)

We've shown that minimising  $\text{KL}(q(z)||p(z|x))$  is equivalent to maximising a simpler objective

$$q^* = \arg \max_{q \in \mathcal{Q}} \mathbb{E}_{q(z)} [\log p(z, x)] + \mathbb{H}(q(z))$$

known as the evidence lowerbound

For certain pairs of distributions in the exponential family, the quantities involved are both tractable

- ▶ e.g. the entropy of a Dirichlet variable is an analytical function of the parameter  $\alpha$
- ▶ e.g. check [this lecture script](#) for analytical results for the first term

---

The name ELBO has to do with the fact that  $\log p(x) \geq \text{ELBO}$

## How do we design $q$ for Bayesian IBM1?

Mean field assumption: make latent variables independent in  $q$

$$\begin{aligned} q(a_1^n, \theta_1^{vE}) &= q(\theta_1^{vE}) \times Q(a_1^n) \\ &= \prod_{\mathbf{e}} q(\theta_{\mathbf{e}}) \times \prod_{j=1}^n Q(a_j) \end{aligned} \tag{13}$$

## How do we design $q$ for Bayesian IBM1?

Mean field assumption: make latent variables independent in  $q$

$$\begin{aligned} q(a_1^n, \theta_1^{vE}) &= q(\theta_1^{vE}) \times Q(a_1^n) \\ &= \prod_{\mathbf{e}} q(\theta_{\mathbf{e}}) \times \prod_{j=1}^n Q(a_j) \end{aligned} \tag{13}$$

Pick convenient parametric families

$$\begin{aligned} q(a_1^n, \theta_1^{vE} | \phi, \lambda) &= \prod_{\mathbf{e}} q(\theta_{\mathbf{e}} | \lambda_{\mathbf{e}}) \times \prod_{j=1}^n Q(a_j | \phi_j) \\ &= \prod_{\mathbf{e}} \text{Dir}(\theta_{\mathbf{e}} | \lambda_{\mathbf{e}}) \times \prod_{j=1}^n \text{Cat}(a_j | \phi_j) \end{aligned} \tag{14}$$

## How do we design $q$ for Bayesian IBM1?

Mean field assumption: make latent variables independent in  $q$

$$\begin{aligned} q(a_1^n, \theta_1^{v_E}) &= q(\theta_1^{v_E}) \times Q(a_1^n) \\ &= \prod_{\mathbf{e}} q(\theta_{\mathbf{e}}) \times \prod_{j=1}^n Q(a_j) \end{aligned} \tag{13}$$

Pick convenient parametric families

$$\begin{aligned} q(a_1^n, \theta_1^{v_E} | \phi, \lambda) &= \prod_{\mathbf{e}} q(\theta_{\mathbf{e}} | \lambda_{\mathbf{e}}) \times \prod_{j=1}^n Q(a_j | \phi_j) \\ &= \prod_{\mathbf{e}} \text{Dir}(\theta_{\mathbf{e}} | \lambda_{\mathbf{e}}) \times \prod_{j=1}^n \text{Cat}(a_j | \phi_j) \end{aligned} \tag{14}$$

Find optimum parameters under the ELBO

- ▶ one Dirichlet parameter vector  $\lambda_{\mathbf{e}}$  per English type  
 $\lambda_{\mathbf{e}}$  consists of  $v_F$  strictly positive numbers
- ▶ one Categorical parameter vector  $\phi_j$  per alignment link  
 $\phi_j$  consists of a probability vector over  $m + 1$  positions

# ELBO for Bayesian IBM1

## Objective

$$\begin{aligned}(\hat{\lambda}, \hat{\phi}) &= \arg \max_{\lambda, \phi} \mathbb{E}_q[\log p(f_1^n, a_1^n, \theta_1^{vE} | e_1^m, \alpha)] + \mathbb{H}(q) \\&= \arg \max_{\lambda, \phi} \sum_{j=1}^m \mathbb{E}_q[\log P(a_j | m) P(f_j | e_{a_j}, \theta_1^{vE}) - \log Q(a_j | \phi_j)] \\&\quad + \sum_e \underbrace{\mathbb{E}_q[\log p(\theta_e | \alpha) - \log q(\theta_e | \lambda_e)]}_{- \text{KL}(q(\theta_e | \lambda_e) || p(\theta_e | \alpha))}\end{aligned}\tag{15}$$

# VB for IBM1

Optimal  $Q(a_j|\phi_j)$

$$\phi_{jk} = \frac{\exp \left( \Psi \left( \lambda_{f_j|e_k} \right) - \Psi \left( \sum_{\mathbf{f}} \lambda_{\mathbf{f}|e_k} \right) \right)}{\sum_{i=0}^m \exp \left( \Psi \left( \lambda_{f_j|e_i} \right) - \Psi \left( \sum_{\mathbf{f}} \lambda_{\mathbf{f}|e_i} \right) \right)} \quad (16)$$

where  $\Psi(\cdot)$  is the [digamma function](#)

Optimal  $q(\theta_e|\lambda_e)$

$$\lambda_{\mathbf{f}|\mathbf{e}} = \alpha_{\mathbf{f}} + \sum_{(e_0^m, f_1^n)} \sum_{j=1}^n \mathbb{E}_{Q(a_j|\phi_j)} [\#(\mathbf{e} \rightarrow \mathbf{f} | f_j, a_j, e_1^m)] \quad (17)$$

# Algorithmically

E-step as in MLE IBM1,

however, using  $Q(a_j|\phi_j)$  instead of  $P(a_j|e_0^m, f_j, \theta_1^{vE})$

- ▶ equivalent to using  $\theta \approx \hat{\theta}$  where
- ▶  $\hat{\theta}_{f|e} = \exp(\Psi(\lambda_{f|e}) - \Psi(\sum_{f'} \lambda_{f'|e}))$



# Algorithmically

E-step as in MLE IBM1,

however, using  $Q(a_j|\phi_j)$  instead of  $P(a_j|e_0^m, f_j, \theta_1^{vE})$

- ▶ equivalent to using  $\theta \approx \hat{\theta}$  where
- ▶  $\hat{\theta}_{f|e} = \exp(\Psi(\lambda_{f|e}) - \Psi(\sum_{f'} \lambda_{f'|e}))$

M-step

- ▶  $\lambda_{f|e} = \alpha_f + \mathbb{E}[\#(e \rightarrow f)]$   
where expected counts come from E-step

# References I