First we generate the global (corpus-wide) variables. There are two main sets of global variables: those that describe the topics people talk about and those that describe how people interact (interaction patterns). These variables are linked by a third set of variables that associate each topic with the pattern that best describes how people interact when talking about that topic.

There are T topics. Each topic $\boldsymbol{\phi}^{(t)}$ is a discrete distribution over V word types.

```
for $t = 1$ to $T$ do
    draw $\boldsymbol{\phi}^{(t)} \sim \mathrm{Dir}(\beta, \boldsymbol{m})$
end for
```

There are C interaction patterns. Each interaction pattern consists of an intercept $b^{(c)} \in \mathbb{R}$, a coefficient vector $\boldsymbol{\gamma}^{(c)} \in \mathbb{R}^P$, and a set of A positions $\{\boldsymbol{s}_a^{(c)} \in \mathbb{R}^K\}_{a=1}^A$—one for each person. We also associate each sender–recipient pair with an observed P-dimensional vector of covariates $\boldsymbol{x}^{(ar)}$; however, we assume that our generative process is conditioned on these covariates.

```
for $c = 1$ to $C$ do
    draw $b^{(c)} \sim \mathcal{N}(\mu, \sigma_1^2)$
    draw $\boldsymbol{\gamma}^{(c)} \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \, \mathrm{I}_P)$
    for $a = 1$ to $A$ do
        draw $\boldsymbol{s}_a^{(c)} \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \, \mathrm{I}_K)$
    end for
    for $a = 1$ to $A$ do
        for $r = 1$ to $A$ do
            if $r \neq a$ then
                set $p_{ar}^{(c)} = \sigma(b^{(c)} + \boldsymbol{\gamma}^{(c)\top} \boldsymbol{x}^{(ar)} - ||\boldsymbol{s}_a^{(c)} - \boldsymbol{s}_r^{(c)}||)$
            else
                set $p_{ar}^{(c)} = 0$
            end if
        end for
    end for
end for
```

The topics and interaction patterns are tied together via a set of T categorical variables (one per topic). These variables associate each topic with a single interaction pattern.

```
for $t = 1$ to $T$ do
    draw $l_t \sim \mathrm{Unif}(1, C)$
end for
```

Then, we generate the local variables. There are D emails. We assume that each email's sender $a^{(d)} \in [A]$ and length $N^{(d)} \in \mathbb{N}_0$ are observed; we do not generate these variables.

```
for $d = 1$ to $D$ do
    draw $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha, \boldsymbol{m})$
    set $\bar{N}^{(d)} = \max(1, N^{(d)})$
    for $n = 1$ to $\bar{N}^{(d)}$ do
        draw $z_n^{(d)} \sim \boldsymbol{\theta}^{(d)}$
        if $N^{(d)} \neq 0$ then
            draw $w_n^{(d)} \sim \boldsymbol{\phi}^{(z_n^{(d)})}$
        end if
    end for
    for $t = 1$ to $T$ do
        set $\bar{N}^{(t|d)} = \sum_{n=1}^{\bar{N}^{(d)}} \delta(z_n^{(d)} = t)$
    end for
    for $r = 1$ to $A$ do
        draw $y_r^{(d)} \sim \text{Bern}(\sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} p_{a^{(d)} r}^{(l_t)})$
    end for
end for
```

This generative process implies a particular factorization of the joint distribution over $\Phi = \{\boldsymbol{\phi}^{(t)}\}_{t=1}^{T}$, $\mathcal{B} = \{b^{(c)}\}_{c=1}^{C}$, $\Gamma = \{\boldsymbol{\gamma}^{(c)}\}_{c=1}^{C}$, $\mathcal{S} = \{\{\boldsymbol{s}_a^{(c)}\}\}_{c=1}^{C}$, $\mathcal{L} = \{l_t\}_{t=1}^{T}$, $\Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}$, $\mathcal{Z} = \{\boldsymbol{z}^{(d)}\}_{d=1}^{D}$, $\mathcal{W} = \{\boldsymbol{w}^{(d)}\}_{d=1}^{D}$, and $\mathcal{Y} = \{\boldsymbol{y}^{(d)}\}_{d=1}^{D}$ given $\mathcal{X} = \{\{\boldsymbol{x}^{(ar)}\}_{r=1}^{A}\}_{a=1}^{A}$ and $\mathcal{A} = \{a^{(d)}\}_{d=1}^{D}$:

$$\begin{aligned} &P(\Phi, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \Theta, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A}) \\ &\quad = P(\Phi)P(\mathcal{B})P(\Gamma)P(\mathcal{S})P(\mathcal{L})P(\Theta)P(\mathcal{Z} \,|\, \Theta)P(\mathcal{W} \,|\, \mathcal{Z}, \Phi)P(\mathcal{Y} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{X}, \mathcal{A}). \end{aligned} \tag{1}$$

We can simplify this further by integrating out Φ and Θ using Dirichlet–multinomial conjugacy:

$$P(\mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A}) = P(\mathcal{B})P(\Gamma)P(\mathcal{S})P(\mathcal{L})P(\mathcal{Z})P(\mathcal{W} \,|\, \mathcal{Z})P(\mathcal{Y} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{X}, \mathcal{A}). \tag{2}$$

Our inference goal is to draw samples from the posterior distribution

$$P(\mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z} \,|\, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \propto P(\mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A}). \tag{3}$$

In practice, we can achieve this inference goal by sequentially resampling the value of each latent variable (i.e., $b^{(c)}$, $\boldsymbol{\gamma}^{(c)}$, $\boldsymbol{s}_a^{(c)}$, l_t, or $z_n^{(d)}$) from its conditional posterior distribution.

Let's first consider $z_n^{(d)}$, whose conditional posterior probability of being topic t is

$$\begin{aligned} & P(z_n^{(d)} = t \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}_{\setminus d,n}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \\ & \quad \propto P(z_n^{(d)} = t, w_n^{(d)}, \boldsymbol{y}^{(d)} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}_{\setminus d,n}, \mathcal{W}_{\setminus d,n}, \mathcal{Y}_{\setminus d}, \mathcal{X}, \mathcal{A}) && (4) \\ & \quad = P(z_n^{(d)} = t \,|\, \mathcal{Z}_{\setminus d,n}) P(w_n^{(d)} \,|\, z_n^{(d)} = t, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}) P(\boldsymbol{y}^{(d)} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, z_n^{(d)} = t, \mathcal{Z}_{\setminus d,n}, \mathcal{X}, \mathcal{A}). && (5) \end{aligned}$$

We have dropped the conditional dependence on $\mathcal{Y}_{\setminus d}$ because $\boldsymbol{y}^{(d)} \perp \mathcal{Y}_{\setminus d}$. We know that

$$P(z_n^{(d)} = t \,|\, \mathcal{Z}_{\setminus d,n}) = \frac{P(z_n^{(d)} = t, \mathcal{Z}_{\setminus d,n})}{P(\mathcal{Z}_{\setminus d,n})} = \frac{\bar{N}_{\setminus d,n}^{(t|d)} + \frac{\alpha}{T}}{\bar{N}^{(d)} - 1 + \alpha}. \tag{6}$$

(The derivation is the same as for LDA.) We also know that

$$P(w_n^{(d)} \,|\, z_n^{(d)} = t, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}) = \frac{P(w_n^{(d)}, \mathcal{W}_{\setminus d,n} \,|\, z_n^{(d)} = t, \mathcal{Z}_{\setminus d,n})}{P(\mathcal{W}_{\setminus d,n} \,|\, \mathcal{Z}_{\setminus d,n})} = \frac{N_{\setminus d,n}^{(w_n^{(d)}|t)} + \frac{\beta}{V}}{N_{\setminus d,n}^{(t)} + \beta}, \tag{7}$$

where $N_{\setminus d,n}^{(w_n^{(d)}|t)}$ is the number of tokens assigned to topic t whose type is the same as that of $w_n^{(d)}$, excluding $w_n^{(d)}$ itself, and $N_{\setminus d,n}^{(t)} = \sum_{v=1}^{V} N_{\setminus d,n}^{(v|t)}$. Finally, we know that

$$\begin{aligned} & P(\boldsymbol{y}^{(d)} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, z_n^{(d)} = t, \mathcal{Z}_{\setminus d,n}, \mathcal{X}, \mathcal{A}) \\ & \quad = \prod_{r=1}^{A} \left(\sum_{t'=1}^{T} \frac{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t' = t)}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(l_{t'})} \right)^{y_r^{(d)}} \left(1 - \sum_{t'=1}^{T} \frac{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t' = t)}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(l_{t'})} \right)^{1 - y_r^{(d)}} . && (8) \end{aligned}$$

. Therefore, if $N^{(d)} > 0$, then

$$\begin{aligned} & P(z_n^{(d)} = t \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}_{\setminus d,n}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \\ & \quad \propto \left(\bar{N}_{\setminus d,n}^{(t|d)} + \frac{\alpha}{T} \right) \frac{N_{\setminus d,n}^{(w_n^{(d)}|t)} + \frac{\beta}{V}}{N_{\setminus d,n}^{(t)} + \beta} \times \\ & \quad \prod_{r=1}^{A} \left(\sum_{t'=1}^{T} \frac{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t' = t)}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(l_{t'})} \right)^{y_r^{(d)}} \left(1 - \sum_{t'=1}^{T} \frac{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t' = t)}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(l_{t'})} \right)^{1 - y_r^{(d)}} , && (9) \end{aligned}$$

If $N^{(d)} = 0$, the first term becomes $\frac{\alpha}{T}$ and disappears because it is a constant. The second term disappears because there are no tokens. Since $\bar{N}^{(t'|d)}_{\backslash d,n} = 0$, we therefore have

$$P(z_1^{(d)} = t \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}_{\backslash d,n}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \propto \prod_{r=1}^{A} \left(p^{(l_t)}_{a^{(d)}r} \right)^{y_r^{(d)}} \left(1 - p^{(l_t)}_{a^{(d)}r} \right)^{1 - y_r^{(d)}}. \tag{10}$$

Next, let's consider l_t, whose conditional posterior probability of being latent space c

$$\begin{aligned} &P(l_t = c \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}_{\backslash t}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \\ &\qquad \propto P(l_t, \mathcal{Y} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}_{\backslash t}, \mathcal{Z}, \mathcal{W}, \mathcal{X}, \mathcal{A}) &(11) \\ &\qquad = P(l_t = c) P(\mathcal{Y} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, l_t = c, \mathcal{L}_{\backslash t}, \mathcal{Z}, \mathcal{X}, \mathcal{A}). &(12) \end{aligned}$$

We know that $P(l_t = c) = \frac{1}{C}$. Since this is a constant, this term disappears. We also know that

$$\begin{aligned} &P(\mathcal{Y} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, l_t = c, \mathcal{L}_{\backslash t}, \mathcal{Z}, \mathcal{X}, \mathcal{A}) \\ &\qquad = \prod_{d=1}^{D} \prod_{r=1}^{A} \left(\sum_{t'=1}^{T} \frac{\bar{N}^{(t'|d)}}{\bar{N}^{(d)}} p^{(l_{t'})}_{a^{(d)}r} \right)^{y_r^{(d)}} \left(1 - \sum_{t'=1}^{T} \frac{\bar{N}^{(t'|d)}}{\bar{N}^{(d)}} p^{(l_{t'})}_{a^{(d)}r} \right)^{1 - y_r^{(d)}}, \end{aligned} \tag{13}$$

with $l_t = c$ throughout. (Note that in the product over d, I think we need only consider those emails that actually use topic t; the others will have no terms involving l_t.) Therefore,

$$\begin{aligned} &P(l_t = c \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}_{\backslash t}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \\ &\qquad \propto \prod_{d=1}^{D} \prod_{r=1}^{A} \left(\sum_{t'=1}^{T} \frac{\bar{N}^{(t'|d)}}{\bar{N}^{(d)}} p^{(l_{t'})}_{a^{(d)}r} \right)^{y_r^{(d)}} \left(1 - \sum_{t'=1}^{T} \frac{\bar{N}^{(t'|d)}}{\bar{N}^{(d)}} p^{(l_{t'})}_{a^{(d)}r} \right)^{1 - y_r^{(d)}}. \end{aligned} \tag{14}$$

We can resample the values of $b^{(c)}$, $\gamma^{(c)}$, and $\boldsymbol{s}_a^{(c)}$ using Metropolis–Hastings or slice sampling. With either algorithm, we can separately resample the values of the variables or we can jointly resample the values of $\mathcal{B}$, Γ, and $\mathcal{S}$. Let's first consider the latter scenario using Metropolis–Hastings.

As a quick reminder, Metropolis–Hastings uses a proposal density Q, which depends on the current values of the variables to be sampled: $Q(\mathcal{B}', \Gamma', \mathcal{S}' \,|\, \mathcal{B}, \Gamma, \mathcal{S})$. Here, we'll assume that Q is a multivariate Gaussian distribution, with a diagonal covariance matrix, centered on a vector that consists of the current state of the Markov chain (i.e., the current values of $\mathcal{B}$, Γ, and $\mathcal{S}$). Each iteration of the algorithm involves drawing a proposal $\mathcal{B}'$, Γ', and $\mathcal{S}'$ from Q given the current state of the Markov chain $\mathcal{B}$, Γ, and $\mathcal{S}$. This proposal is accepted as the new state of the chain if

$$\frac{Q(\mathcal{B}, \Gamma, \mathcal{S} \,|\, \mathcal{B}', \Gamma', \mathcal{S}')}{Q(\mathcal{B}', \Gamma', \mathcal{S}' \,|\, \mathcal{B}, \Gamma, \mathcal{S})} \frac{P(\mathcal{B}', \Gamma', \mathcal{S}' \,|\, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A})}{P(\mathcal{B}, \Gamma, \mathcal{S} \,|\, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A})} \geq 1. \tag{15}$$

If Q is a Gaussian distribution, then the first term disappears because

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-x')^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x'-x)^2}{2\sigma^2}\right)} = \exp\left(\frac{1}{2\sigma^2}\left((x'-x)^2 - (x-x')^2\right)\right) = \exp\left(0\right) = 1. \tag{16}$$

To simplify the second term, note that

$$\frac{P(\mathcal{B}', \Gamma', \mathcal{S}' \,|\, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A})}{P(\mathcal{B}, \Gamma, \mathcal{S} \,|\, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A})} = \frac{P(\mathcal{B}', \Gamma', \mathcal{S}', \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A})}{P(\mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A})} \tag{17}$$

because the normalization constants that turn the joint distribution into the conditional distribution cancel. Any terms in the joint distribution that don't involve $\mathcal{B}$, Γ, or $\mathcal{S}$ also cancel, so

$$\begin{aligned}
&\frac{P(\mathcal{B}', \Gamma', \mathcal{S}', \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A})}{P(\mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A})} \\
&\quad = \frac{P(\mathcal{B}')P(\Gamma')P(\mathcal{S}')P(\mathcal{L})P(\mathcal{Z})P(\mathcal{W} \,|\, \mathcal{Z})P(\mathcal{Y} \,|\, \mathcal{B}', \Gamma', \mathcal{S}', \mathcal{L}, \mathcal{Z}, \mathcal{X}, \mathcal{A})}{P(\mathcal{B})P(\Gamma)P(\mathcal{S})P(\mathcal{L})P(\mathcal{Z})P(\mathcal{W} \,|\, \mathcal{Z})P(\mathcal{Y} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{X}, \mathcal{A})} && (18) \\
&\quad = \frac{P(\mathcal{B}')P(\Gamma')P(\mathcal{S}')P(\mathcal{Y} \,|\, \mathcal{B}', \Gamma', \mathcal{S}', \mathcal{L}, \mathcal{Z}, \mathcal{X}, \mathcal{A})}{P(\mathcal{B})P(\Gamma)P(\mathcal{S})P(\mathcal{Y} \,|\, \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{X}, \mathcal{A})} && (19) \\
&\quad = \frac{P(\mathcal{B}')P(\Gamma')P(\mathcal{S}')}{P(\mathcal{B})P(\Gamma)P(\mathcal{S})} \frac{\prod_{d=1}^{D} \prod_{r=1}^{A} \left(\sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} {p'}^{(l_t)}_{a^{(d)}r}\right)^{y_r^{(d)}} \left(1 - \sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} {p'}^{(l_t)}_{a^{(d)}r}\right)^{1-y_r^{(d)}}}{\prod_{d=1}^{D} \prod_{r=1}^{A} \left(\sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} p^{(l_t)}_{a^{(d)}r}\right)^{y_r^{(d)}} \left(1 - \sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} p^{(l_t)}_{a^{(d)}r}\right)^{1-y_r^{(d)}}} && (20) \\
&\quad = \frac{P(\mathcal{B}')P(\Gamma')P(\mathcal{S}')}{P(\mathcal{B})P(\Gamma)P(\mathcal{S})} \prod_{d=1}^{D} \prod_{r=1}^{A} \left(\frac{\sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} {p'}^{(l_t)}_{a^{(d)}r}}{\sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} p^{(l_t)}_{a^{(d)}r}}\right)^{y_r^{(d)}} \left(\frac{1 - \sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} {p'}^{(l_t)}_{a^{(d)}r}}{1 - \sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} p^{(l_t)}_{a^{(d)}r}}\right)^{1-y_r^{(d)}}, && (21)
\end{aligned}$$

where ${p'}^{(c)}_{ar} = \sigma({b'}^{(c)} + {\boldsymbol{\gamma}'}^{(c)\top} \boldsymbol{x}^{(ar)} - ||{\boldsymbol{s}'}^{(c)}_{a} - {\boldsymbol{s}'}^{(c)}_{r}||)$.

Bruce suggests an alternative to step 14 of the generative process for the local variables:

$$y_r^{(d)} \sim \text{Bern}\left(\frac{\prod_{c=1}^{C} \left(p^{(c)}_{a^{(d)}r}\right)^{\bar{N}^{(c|d)}}}{\prod_{c=1}^{C} \left(p^{(c)}_{a^{(d)}r}\right)^{\bar{N}^{(c|d)}} + \prod_{c=1}^{C} \left(1 - p^{(c)}_{a^{(d)}r}\right)^{\bar{N}^{(c|d)}}}\right), \tag{22}$$

where $\bar{N}^{(c|d)} = \sum_{t=1}^{T} \bar{N}^{(t|d)} \delta(l_t = c)$. Currently, however, we have

$$y_r^{(d)} \sim \text{Bern}\left(\sum_{t=1}^{T} \frac{\bar{N}^{(t|d)}}{\bar{N}^{(d)}} p_{a^{(d)} r}^{(l_t)}\right) = \text{Bern}\left(\sum_{c=1}^{C} \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} p_{a^{(d)} r}^{(c)}\right), \tag{23}$$

which implies that $y_r^{(d)}$ is drawn from a mixture model, where each mixture component is an interaction pattern. The (normalized) weight for each interaction pattern c is the proportion of tokens in email d that are associated with (topics associated with) that interaction pattern. In contrast, Bruce's suggestion implies that $y^{(d)}$ is drawn from a product model. Here too, the components are interaction patterns, however the weight for each pattern is now the number of tokens in email d that are associated with that pattern, and the components' contributions are combined differently.

In general, a C-component mixture model for a random variable $y \in \{0, 1\}$ implies that y is drawn from a linear combination of C Bernoulli components, parameterized by $p^{(1)}, \ldots, p^{(C)}$—i.e.,

$$\begin{aligned} P(y = 1) = p &= \frac{\sum_{c=1}^{C} \pi^{(c)} p^{(c)}}{\sum_{c=1}^{C} \pi^{(c)} p^{(c)} + \sum_{c=1} \pi^{(c)} \left(1 - p^{(c)}\right)} && (24) \\ &= \frac{\sum_{c=1}^{C} \pi^{(c)} p^{(c)}}{\sum_{c=1}^{C} \pi^{(c)} \left(p^{(c)} + 1 - p^{(c)}\right)} && (25) \\ &= \sum_{c=1}^{C} \frac{\pi^{(c)}}{\sum_{c=1}^{C} \pi^{(c)}} p^{(c)}, && (26) \end{aligned}$$

where $\pi^{(1)} \geq 0, \ldots, \pi^{(C)} \geq 0$ are the component weights. In contrast, a C-component product model for y, formed from the same components and component weights, implies that

$$P(y = 1) = p = \frac{\prod_{c=1}^{C} \left(p^{(c)}\right)^{\pi^{(c)}}}{\prod_{c=1}^{C} \left(p^{(c)}\right)^{\pi^{(c)}} + \prod_{c=1}^{C} \left(1 - p^{(c)}\right)^{\pi^{(c)}}}. \tag{27}$$

There are two main differences between a mixture model and a product model. The first is in the way that the components' contributions are combined. In a mixture model, the component-specific probabilities are combined additively (i.e., via an "or" function); in a product model, the component-specific probabilities are combined multiplicatively (i.e., via an "and" function).

To illustrate this difference, let's consider a simple example in which the value of $y \in \{0, 1\}$ is drawn from either a mixture model with two Bernoulli components, parameterized by $p^{(1)}$ and $p^{(2)}$, respectively, or a product model formed from the same two components. In both models, we'll assume that the components are equally weighted with $\pi^{(1)} = \pi^{(2)} = 1$. The mixture model therefore implies that

$$P(y{=}1) = p = \frac{1}{2}\sum_{c=1}^{2} p^{(c)}, \tag{28}$$

while the product model implies that

$$P(y{=}1) = p = \frac{\prod_{c=1}^{2} p^{(c)}}{\prod_{c=1}^{2} p^{(c)} + \prod_{c=1}^{2}(1 - p^{(c)})}. \tag{29}$$

In the mixture model, the components can compensate for one another. For example, if $p^{(1)} = 0.9$ and $p^{(2)} = 0.2$, then $p = 0.55$. To better understand this example, let's quantify the uncertainty of a distribution as its entropy, so less certain distributions have higher entropy and vice versa. For a two dimensional distribution q, $0 \leq H(q) \leq 1.0$. Here, $H\left(p^{(1)}\right) = 0.469$, $H\left(p^{(2)}\right) = 0.722$, and $H(p) = 0.993$. Because the components disagree, both components are made less certain by their combination. In contrast, if $p^{(1)} = 0.9$ and $p^{(2)} = 0.7$, then $p = 0.8$. Here, $H\left(p^{(1)}\right) = 0.469$, $H\left(p^{(2)}\right) = 0.881$, and $H(p) = 0.722$. Because the components agree, but with differing certainties, the more certain component is made less certain while the less certain component is made more so. This compensatory behavior has the biggest substantive impact on components that are very certain. For example, if $p^{(1)} = 0$ and $p^{(2)} = 0.4$, then $p = 0.2$. Equivalently, $H\left(p^{(1)}\right) = 0$, $H\left(p^{(2)}\right) = 0.971$, and $H(p) = 0.722$. In other words, even though the first component was absolutely certain, the mixture model is not certain (i.e., $p \neq 0$) because of the second component's non-zero probability.

In general, a mixture model cannot be more certain than its most certain component. (It can, however, be more or less certain than the other components, though.) We can use Jensen's inequality to see why this is the case. Since entropy is a concave function, Jensen's inequality implies that

$$H\left(\sum_{c=1}^{C} \frac{\pi^{(c)}}{\sum_{c=1}^{C} \pi^{(c)}} p^{(c)}\right) \geq \sum_{c=1}^{C} \frac{\pi^{(c)}}{\sum_{c=1}^{C} \pi^{(c)}} H\left(p^{(c)}\right). \tag{30}$$

For a fixed set of components $p^{(1)}, \ldots, p^{(C)}$, making the mixture model as certain as possible is equivalent to minimizing its entropy with respect to the weights $\pi^{(1)}, \ldots, \pi^{(C)}$—i.e.,

$$\min_{\boldsymbol{\pi}} H\left(\sum_{c=1}^{C} \frac{\pi^{(c)}}{\sum_{c=1}^{C} \pi^{(c)}} p^{(c)}\right) \geq \min_{\boldsymbol{\pi}} \sum_{c=1}^{C} \frac{\pi^{(c)}}{\sum_{c=1}^{C} \pi^{(c)}} H\left(p^{(c)}\right). \tag{31}$$

The right-hand side of equation 30 can only be minimized by placing all the weight (i.e., $\sum_{c=1}^{C} \pi^{(c)}$)

on the most certain component (and no weight on the other components). Therefore,

$$\min_{\boldsymbol{\pi}} H\left(\sum_{c=1}^{C} \frac{\pi^{(c)}}{\sum_{c=1}^{C} \pi^{(c)}} p^{(c)}\right) \geq \min_{c} H\left(p^{(c)}\right). \tag{32}$$

In other words, no matter how we choose the weights $\pi^{(1)}, \ldots, \pi^{(C)}$, the mixture model cannot be more certain than its most certain component. Moreover, by examining equation 32, we can see that the mixture model will only be as certain as its most certain component if $p^{(1)} = \ldots = p^{(C)}$ or if we place all the weight on the most certain component and none on the other components.

Returning to our model, let's consider an email whose tokens $\boldsymbol{w}^{(d)}$ are associated (via $\boldsymbol{z}^{(d)}$ and $\mathcal{L}$) with interaction patterns c and c'. Even if interaction pattern c specifies that sender $a^{(d)}$ never interacts with recipient r, then provided interaction pattern c' specifies a non-zero probability of interaction, the probability of $a^{(d)}$ sending the email to r will be non-zero. For example, even if I never email my boss about personal matters (but do email her about work), there is a non-zero probability that I will send an her an email containing both personal and work content. Is this a reasonable assumption? Intuitively, I'm not sure that it is. Realistically, I can't see myself sending my boss an email containing personal content, even if some portion of the email is about work.

In contrast, a product model can be more or less certain than its most certain component. For example, if $p^{(1)} = 0.9$ and $p^{(2)} = 0.2$, then $p = 0.692$. Equivalently, $H\left(p^{(1)}\right) = 0.469$, $H\left(p^{(2)}\right) = 0.722$, and $H(p) = 0.890$. In other words, both components are made less certain by their combination. However, if $p^{(1)} = 0.9$ and $p^{(2)} = 0.7$, then $p = 0.955$. Here, $H\left(p^{(1)}\right) = 0.469$, $H\left(p^{(2)}\right) = 0.881$, and $H(p) = 0.267$. Both components are made more certain by their combination—even the most certain component. In addition to this difference, components can veto one another. For example, if $p^{(1)} = 0$ and $p^{(2)} = 0.4$, then $p = 0$. Equivalently, $H\left(p^{(1)}\right) = 0$, $H\left(p^{(2)}\right) = 0.971$, and $H(p) = 0$. In other words, the product model is certain that $y = 0$ because the first component was certain.

Returning to Bruce's alternative to our model, if interaction pattern c specifies that sender $a^{(d)}$ never interacts with recipient r, then the probability of $a^{(d)}$ sending to r an email that contains some tokens associated with interaction pattern c is zero. For example, if I never email my boss about personal matters (but do email her about work), then I will not send her an email containing both personal and work content. Is this a reasonable assumption? Intuitively, I think it is.

The second difference between a mixture model and a product model is in the components' weights. Because the weights in a mixture model are normalized (see equation 26), their scale is ignored. For example, in a two-component mixture model, $\pi^{(1)} = 1$ and $\pi^{(2)} = 4$ and $\pi^{(1)} = 100$ and $\pi^{(2)} = 400$ will both be normalized to $\pi^{(1)} = 0.2$ and $\pi^{(2)} = 0.8$ and therefore yield the same mixture model.

In our model, when forming the probability of $y_r^{(d)}\!=\!1$, the weight for each interaction pattern c is the number of tokens in email d that are associated with that interaction pattern. When normalized, each weight becomes a proportion—i.e., $\frac{\pi^{(c)}}{\sum_{c=1}^C \pi^{(c)}} = \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}}$. Normalization means that two emails with the same sender will have identical probabilities of being sent to recipient r, provided their empirical distributions over interaction patterns are the same. For example, let's consider two such emails d and d' and assume that for every interaction pattern, the second email contains ten times as many tokens associated with that pattern as the first—i.e., $\bar{N}^{(c|d')} = 10 \cdot \bar{N}^{(c|d)}$. Since

$$\frac{\bar{N}^{(c|d')}}{\bar{N}^{(d')}} = \frac{\bar{N}^{(c|d')}}{\sum_{c=1}^C \bar{N}^{(c|d')}} = \frac{10 \cdot \bar{N}^{(c|d)}}{\sum_{c=1}^C 10 \cdot \bar{N}^{(c|d)}} = \frac{\bar{N}^{(c|d)}}{\sum_{c=1}^C \bar{N}^{(c|d)}} = \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}}, \tag{33}$$

our model implies that

$$P(y_r^{(d)}\!=\!1) = \sum_{c=1}^C \frac{\bar{N}^{(c|d)}}{\bar{N}^{(d)}} p_{a^{(d)}r}^{(c)} = \sum_{c=1}^C \frac{\bar{N}^{(c|d')}}{\bar{N}^{(d')}} p_{a^{(d')}r}^{(c)} = P(y_r^{(d')}\!=\!1). \tag{34}$$

In other words, our model treats the two emails identically, even though they are different lengths, because their empirical distributions over interaction patterns are the same. Equivalently, our model assumes that email length does not affect the probability of sending an email to a particular recipient; an empirical distribution over interaction patterns formed from two tokens plays exactly the same role as one formed from 200 tokens. Is this a reasonable assumption? Intuitively, I'm not sure it is, given that the latter distribution is formed from much more data than the former.

In the product model, the scale of the components' weights is not ignored. For example, let's consider a two-component mixture model with $p^{(1)}\!=\!0.9$ and $p^{(2)}\!=\!0.2$. If $\pi^{(1)}\!=\!0.2$ and $\pi^{(2)}\!=\!0.8$, then $p = 0.339$; however, if $\pi^{(1)}\!=\!1$ and $\pi^{(2)}\!=\!4$, then $p = 0.033$. If $\pi^{(1)}\!=\!100$ and $\pi^{(2)}\!=\!400$, then $p\!=\!0$. In other words, the larger the scale of the weights, the more certain the product model is.

In Bruce's alternative model, when forming the probability of $y_r^{(d)} = 1$, the weight for interaction pattern c is the number of tokens in email d that are associated with that interaction pattern—i.e., $\pi^{(c)} = \bar{N}^{(c|d)}$. Two emails with the same sender will only have identical probabilities of being sent to recipient r if their empirical distributions over interaction patterns *and* their lengths are the same. Let's again consider two such emails d and d' and assume that for every interaction pattern, the second email contains ten times as many tokens associated with that pattern as the first. Now,

$$
\begin{aligned}
P(y_r^{(d')} \! = \! 1) &= \frac{\prod_{c=1}^C \left(p^{(c)}_{a^{(d')}r}\right)^{\bar{N}^{(c|d')}}}{\prod_{c=1}^C \left(p^{(c)}_{a^{(d')}r}\right)^{\bar{N}^{(c|d')}} + \prod_{c=1}^C \left(1 - p^{(c)}_{a^{(d')}r}\right)^{\bar{N}^{(c|d')}}} && (35)\\
&= \frac{\prod_{c=1}^C \left(p^{(c)}_{a^{(d)}r}\right)^{10 \cdot \bar{N}^{(c|d)}}}{\prod_{c=1}^C \left(p^{(c)}_{a^{(d)}r}\right)^{10 \cdot \bar{N}^{(c|d)}} + \prod_{c=1}^C \left(1 - p^{(c)}_{a^{(d)}r}\right)^{10 \cdot \bar{N}^{(c|d)}}} && (36)\\
&= \frac{\left(\prod_{c=1}^C \left(p^{(c)}_{a^{(d)}r}\right)^{\bar{N}^{(c|d)}}\right)^{10}}{\left(\prod_{c=1}^C \left(p^{(c)}_{a^{(d)}r}\right)^{\bar{N}^{(c|d)}}\right)^{10} + \left(\prod_{c=1}^C \left(1 - p^{(c)}_{a^{(d)}r}\right)^{\bar{N}^{(c|d)}}\right)^{10}} && (37)\\
&\neq P(y_r^{(d)} \! = \! 1). && (38)
\end{aligned}
$$

In other words, we can form $P(y_r^{(d')} = 1)$ by raising each of the constituent terms in $P(y_r^{(d)} = 1)$ to a power of ten. This power is effectively an "inverse temperature"; larger values make the distribution more certain. Therefore, in Bruce's alternative to our model, if two emails d and d' have the same sender and empirical distribution over interaction patterns but different lengths, they will not be treated identically; if $N^{(d')} \geq N^{(d)}$, then the distribution over $y_r^{(d')}$ will be more certain than the distribution over $y_r^{(d)}$ for each recipient r. In general, raising (unnormalized) probabilities to a power has a major impact on the resultant distribution; the larger the power, the more certain the distribution will be. (This principle is used in simulated annealing; there, however, the powers are fractional to make distributions less certain.) For example, if an email d contains a few hundred tokens, then its recipient distributions will be close to delta spikes.

There are (at least) two ways to lessen the impact of email length and preserve some uncertainty for longer emails. First, we could use component weights that are proportions—i.e., $\pi^{(c)} = \frac{\bar{N}(c|d)}{\bar{N}(d)}$. However, this would mean that email length does not affect the probability of sending an email to a particular recipient, just as in our model. As I argued earlier, in the context of our model, I'm not sure this is a reasonable assumption. Another, less drastic, option would be to use component weights that are log counts—i.e., $\pi^{(c)} = \log{(\bar{N}^{(c|d)})}$. Intuitively, this approach is appealing as it implies that the length of an email does matter, but only in terms of its order of magnitude.

How does moving to a product model affect our inference equations?

Let's first consider $z_n^{(d)}$. Under Bruce's alternative to our model,

$$\begin{aligned} &P(z_n^{(d)} \!=\! t \mid \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}_{\setminus d,n}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \\ &\quad \propto \left(\bar{N}_{\setminus d,n}^{(t|d)} + \frac{\alpha}{T}\right) \frac{N_{\setminus d,n}^{(w_n^{(d)}|t)} + \frac{\beta}{V}}{N_{\setminus d,n}^{(t)} + \beta} \times \\ &\quad \prod_{r=1}^{A} \left(\frac{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t'=t)}}{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t'=t)} + \prod_{t'=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t'=t)}} \right)^{y_r^{(d)}} \times \\ &\quad \left(\frac{\prod_{t'=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t'=t)}}{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t'=t)} + \prod_{t'=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}_{\setminus d,n}^{(t'|d)} + \delta(t'=t)}} \right)^{1 - y_r^{(d)}} . \end{aligned} \tag{39}$$

If $N^{(d)} = 0$, the first term becomes $\frac{\alpha}{T}$ and disappears because it is a constant. The second term disappears because there are no tokens. Since $\bar{N}_{\setminus d,n}^{(t'|d)} \!=\! 0$ and $\frac{p_{a^{(d)}r}^{(l_t)}}{p_{a^{(d)}r}^{(l_t)} + \left(1 - p_{a^{(d)}r}^{(l_t)}\right)} = p_{a^{(d)}r}^{(l_t)}$, we have

$$P(z_1^{(d)} \!=\! t \mid \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}_{\setminus d,n}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \propto \prod_{r=1}^{A} \left(p_{a^{(d)}r}^{(l_t)}\right)^{y_r^{(d)}} \left(1 - p_{a^{(d)}r}^{(l_t)}\right)^{1 - y_r^{(d)}} . \tag{40}$$

Next, let's consider l_t. Under Bruce's alternative to our model,

$$\begin{aligned} &P(l_t \!=\! c \mid \mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}_{\setminus t}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}, \mathcal{X}, \mathcal{A}) \\ &\quad \propto \prod_{d=1}^{D} \prod_{r=1}^{A} \left(\frac{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}}}{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}} + \prod_{t'=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}}} \right)^{y_r^{(d)}} \times \\ &\quad \left(\frac{\prod_{t'=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}}}{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}} + \prod_{t'=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}}} \right)^{1 - y_r^{(d)}} , \end{aligned} \tag{41}$$

with $l_t \!=\! c$ throughout. Note that in the product over d, I think we need only consider those emails that actually use topic t; the others will have no terms involving l_t. Also, if $N^{(d)} = 0$ then

$$\frac{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}}}{\prod_{t'=1}^{T} \left(p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}} + \prod_{t'=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_{t'})}\right)^{\bar{N}^{(t'|d)}}} = p_{y_r^{(d)}}^{(l_{z_1^{(d)}})}. \tag{42}$$

We can resample the values of $b^{(c)}$, $\boldsymbol{\gamma}^{(c)}$, and $\boldsymbol{s}_a^{(c)}$ using Metropolis–Hastings or slice sampling. Let's consider jointly resampling the values of $\mathcal{B}$, Γ, and $\mathcal{S}$ using Metropolis–Hastings.

Under Bruce's alternative to our model, the Metropolis–Hastings acceptance ratio is

$$\begin{aligned}
&\frac{P(\mathcal{B}', \Gamma', \mathcal{S}', \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A})}{P(\mathcal{B}, \Gamma, \mathcal{S}, \mathcal{L}, \mathcal{Z}, \mathcal{W}, \mathcal{Y} \,|\, \mathcal{X}, \mathcal{A})} \\
&\quad = \frac{P(\mathcal{B}')P(\Gamma')P(\mathcal{S}')}{P(\mathcal{B})P(\Gamma)P(\mathcal{S}')} \times \\
&\quad\quad \prod_{d=1}^{D} \prod_{r=1}^{A} \left(\frac{\prod_{t=1}^{T} \left(p_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}} + \prod_{t=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}}}{\prod_{t=1}^{T} \left({p'}_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}} + \prod_{t=1}^{T} \left(1 - {p'}_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}}} \prod_{t=1}^{T} \left(\frac{{p'}_{a^{(d)}r}^{(l_t)}}{p_{a^{(d)}r}^{(l_t)}} \right)^{\bar{N}^{(t|d)}} \right)^{y_r^{(d)}} \times \\
&\quad\quad \left(\frac{\prod_{t=1}^{T} \left(p_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}} + \prod_{t=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}}}{\prod_{t=1}^{T} \left({p'}_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}} + \prod_{t=1}^{T} \left(1 - {p'}_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}}} \prod_{t=1}^{T} \left(\frac{1 - {p'}_{a^{(d)}r}^{(l_t)}}{1 - p_{a^{(d)}r}^{(l_t)}} \right)^{\bar{N}^{(t|d)}} \right)^{1 - y_r^{(d)}} \qquad (43) \\
&\quad = \frac{P(\mathcal{B}')P(\Gamma')P(\mathcal{S}')}{P(\mathcal{B})P(\Gamma)P(\mathcal{S}')} \times \\
&\quad\quad \prod_{d=1}^{D} \prod_{r=1}^{A} \frac{\prod_{t=1}^{T} \left(p_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}} + \prod_{t=1}^{T} \left(1 - p_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}}}{\prod_{t=1}^{T} \left({p'}_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}} + \prod_{t=1}^{T} \left(1 - {p'}_{a^{(d)}r}^{(l_t)}\right)^{\bar{N}^{(t|d)}}} \times \\
&\quad\quad \left(\prod_{t=1}^{T} \left(\frac{{p'}_{a^{(d)}r}^{(l_t)}}{p_{a^{(d)}r}^{(l_t)}} \right)^{\bar{N}^{(t|d)}} \right)^{y_r^{(d)}} \left(\prod_{t=1}^{T} \left(\frac{1 - {p'}_{a^{(d)}r}^{(l_t)}}{1 - p_{a^{(d)}r}^{(l_t)}} \right)^{\bar{N}^{(t|d)}} \right)^{1 - y_r^{(d)}}, \qquad (44)
\end{aligned}$$

where ${p'}_{ar}^{(c)} = \sigma({b'}^{(c)} + {\boldsymbol{\gamma}'^{(c)}}^{\top} \boldsymbol{x}^{(ar)} - ||{\boldsymbol{s}'}_a^{(c)} - {\boldsymbol{s}'}_r^{(c)}||)$.