Secondphase General Mixture

May 2018

1 The New Algorithm

By trying to infer the mother sequences in a restricted subset of columns S, we are actually driving their projection on S. Notice the fact that if two mother sequences y_1, y_2 have the same sequence of bits in the subset S, then the same source sequence will be derived from them, with a frequency equal to the sum of their frequencies.

Instead of applying the matrix B^{-1} directly, we apply it to subsets of columns, in order to draw the mother sequences and their frequencies with respect to that columns. We start with one column, and at the end of step i, we have already obtained a set of projected mother sequences onto the first i columns, namely the set $\{\hat{y}_i^{(j)}\}_{j=1}^{\eta_i}$ with frequencies $\{\hat{f}_i^{(j)}\}_{j=1}^{\eta_i}$. In the meanwhile, we keep track of another set of columns $R_i \subseteq \{1, 2, ..., i\}$ which is a minimal subset of indices with the property that projecting $\hat{y}_i^{(j)}$'s on R_i are pairwise different. Then, In step $i+1$, we drive our estimation of the mother sequences projected on the set $R_i \cup \{i+1\}$, using the estimator that we will introduce in continue, and we name them $\{\hat{y}_i^{(j)}\}_{j=1}^{\eta_{i+1}}$. Note that each of the projected mother sequences $\hat{y}_i^{(j)}$ in the previous step can result in at most two different sequences $\hat{y}_{i+1}^{(j_1)}, \hat{y}_{i+1}^{(j_2)}$ on $R_i \cup \{i+1\}$. Hence we have $\eta_i \leq \eta_{i+1} \leq 2\eta_i$. If there exist any j such that $\hat{y}_i^j$ turn into two sequences $\hat{y}_{i+1}^{(j_1)}, \hat{y}_{i+1}^{(j_2)}$, or equivalently if $\eta_{i+1} > \eta_i$, then we set $R_{i+1} = R_i \cup \{i+1\}$. Otherwise if $\eta_{i+1} = \eta_i$, define $R_{i+1} = R_i$. This way, because we have N number of mother sequences in total, we get $\forall i \ , \ |R_i| \leq \mathrm{N}$.

2 Estimator

The estimator is similar to the past. We apply the matrix B^{-1} to the empirical distribution on the projected columns, and we threshold the obtained frequencies with λ, where λ is a lower bound with respect to the frequencies of the mother sequences.

3 Analysis

As we defined before, ϵ_s is the error we see on the sequence s regarding the empirical distribution. Moreover, α_{s^*} is the error we see on each source sequence, which is

$$\alpha_{s^*} = \sum_{s \in \mathcal{A}_{\mathrm{L}}} p(s^*, s)\epsilon_s.$$

Moreover, define $\acute{\epsilon_s} = m\epsilon_s$, where m is the sample size, $p = \{p_s\}_{s \in \mathcal{A}_{\mathrm{L}}}$ to be the statistical distribution on the samples, and $p(s^*) = (p(s^*, s))_{s \in \mathcal{A}_{\mathrm{L}}}$.

In the following, we are going to drive an exponential bound for the probability of growth of α_{s^*} with respect to our sample size. To this end, we calculate the moment generating function of α_{s^*}. Suppose $\{z_k\}_{k=1}^m$ are vectors with dimension 2^L that are the one-hot representation of our samples. Then,

$$\mathbb{E}[e^{t\alpha_{s^*}}] = \mathbb{E}[e^{t\sum_{s \in \mathcal{A}_{\mathrm{L}}} p(s^*,s)\epsilon_s}] = \mathbb{E}[e^{t\frac{\sum_{s \in \mathcal{A}_{\mathrm{L}}} p(s^*,s)\acute{\epsilon_s}}{m}}] = \Pi_{k=1}^m \mathbb{E}[e^{t\frac{\sum_{s \in \mathcal{A}_{\mathrm{L}}} p(s^*,s)(z_{k_s} - p_s)}{m}}] \tag{1}$$

$$= \big(\mathbb{E}[e^{t\frac{\sum_{s \in \mathcal{A}_{\mathrm{L}}} p(s^*,s) z_s}{m}}]\big)^m e^{-tp(s^*)^\top p} = \big(\sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s e^{\frac{tp(s^*,s)}{m}}\big)^m e^{-tp(s^*)^\top p}. \tag{2}$$

Thus, by the Markov inequality

$$\mathbb{P}(\alpha_{s^*} > k) \leq \frac{\mathbb{E}[e^{t\alpha_{s^*}}]}{e^{tk}} = e^{-tk}\big(\sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s e^{\frac{tp(s^*,s)}{m}}\big)^m e^{-tp(s^*)^\top p}. \tag{3}$$

Now by replacing mt with t,

$$\mathbb{P}(\alpha_{s^*} > k) \leq \big(e^{-tk}\big(\sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s e^{tp(s^*,s)}\big)e^{-tp(s^*)^\top p}\big)^m \tag{4}$$

$$= \big(\sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s e^{t(-k+p(s^*,s)-p(s^*)^\top p)}\big)^m. \tag{5}$$

Now define

$$\Phi(t) = \sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s e^{t(-k+p(s^*,s)-p(s^*)^\top p)}. \tag{6}$$

We have

$$\Phi(0) = 1, \tag{7}$$

$$\dot{\Phi}(0) = \sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s(-k + b_s - p(s^*)^\top p) = -k. \tag{8}$$

$$\ddot{\Phi}(t) = \sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s(-k + b_s - p(s^*)^\top p)^2 e^{t(-k+p(s^*,s)-p(s^*)^\top p)} \tag{9}$$

$$\tag{10}$$

But note that

$$p(s^*)^\top p = \begin{cases} 0 & s^* \text{ is not a mother sequence} \\ (1-2f)^{\mathrm{L}} & \exists i \ s^* = y_i \end{cases}.$$

Define

$$\tau = \big(f^3 + (1-f)^3\big)^{\frac{\mathrm{L}}{2}}.$$

for $0 < \beta < \frac{4e\tau}{(1-f)^{\mathrm{L}}}$, put $k = \beta\tau$ and $t_0 = \frac{\beta}{4e\tau}$. Then for $0 < t \leq t_0$

$$e^{t(-k+p(s^*,s)-p(s^*)^\top p)} \leq e^{t_0(-k+p(s^*,s)-p(s^*)^\top p)} \leq e^{t_0 p(s^*,s)} \leq e^{t_0(1-f)^{\mathrm{L}}} \leq e.$$

Therefore

$$\ddot{\Phi}(t) \leq e \sum_{s \in \mathcal{A}_{\mathrm{L}}} p_s(-k + b_s - p(s^*)^\top p)^2 = e\big(k^2 + p(s^*)^{2^\top} p - (p(s^*)^\top p)^2\big) \quad (11)$$

$$\leq e\big(k^2 + p(s^*)^{2^\top} p\big) \quad (12)$$

But note that we have

$$p(s^*)^{2^\top} p \leq \tau^2,$$

Hence, for $0 \leq t \leq t_0$,

$$\ddot{\Phi}(t) \leq e\big(\beta^2\tau^2 + \tau^2\big) = e(1+\beta^2)\tau^2.$$

Therefore, for $0 \leq t \leq t_0$

$$\dot{\Phi}(t) = \dot{\Phi}(0) + \int_0^t \ddot{\Phi}(r)dr \leq -k + \int_0^t e(1+\beta^2)\tau^2 dr = -\beta\tau + te(1+\beta^2)\tau^2 \quad (13)$$

$$\leq -\beta\tau + t_0 e(1+\beta^2)\tau^2 \leq -\beta\tau + \frac{\beta\tau}{2} = -\frac{\beta\tau}{2}. \quad (14)$$

Thus, we conclude

$$\Phi(t_0) = \Phi(0) + \int_0^{t_0} \dot{\Phi}(t)dt \leq 1 - \int_0^{t_0} \frac{\beta\tau}{2} dt = 1 - t_0\frac{\beta\tau}{2} = 1 - \frac{\beta^2}{8e}. \quad (15)$$

Hence, by replacing $t = t_0$ in (5), we conclude

$$\mathbb{P}(\alpha_{s^*} > \beta\tau) \leq \big(1 - \frac{\beta^2}{8e}\big)^m. \quad (16)$$

For $0 \leq \varepsilon \leq 1$, Assign

$$\beta = \frac{(1-2f)^{\mathrm{L}}}{(f^3 + (1-f)^3)^{\frac{\mathrm{L}}{2}}} \frac{\lambda\varepsilon}{2}$$

Note that we have

$$f^3 + (1-f)^3 \geq (1-f)(1-2f).$$

Hence, the condition $\beta \leq \frac{4e\tau}{(1-f)^{\mathrm{L}}}$ is satisfied by this choice of β. Now assign

$$m = (\frac{8e}{\beta^2}) \ln(2)(\mathrm{N} + 1 + \acute{m}).$$

According to (16)

$$\mathbb{P}(\alpha_{s^*} > \beta\tau) \leq \big(1 - \frac{\beta^2}{8e}\big)^{\frac{8e}{\beta^2} \ln(2)(\mathrm{N}+1+\acute{m})} \leq 2^{-\mathrm{N}-1-\acute{m}} \tag{17}$$

Now consider the algorithm in a desired phase i. We threshold the frequencies obtained from the projected sequences on the set of columns $R_i \cup \{i+1\}$. Suppose that E_i is the event of making a wrong decision in phase i. For each sequence s^* with length $|R_i|+1$, $\vartheta_{s^*} = \frac{\alpha_{s^*}}{(1-2f)^{\mathrm{L}}}$ is the noise of the frequency that we obtain in phase i on sequence s^*. If this noise is less than $\frac{\lambda}{2}$, then due to the fact that the minimum frequency is at least λ, we threshold the frequency of s^* correctly, meaning that if it was a mother sequence, we recover it, and if it was not a mother sequence, we omit it. On the other hand, according to (17) we obtain

$$\mathbb{P}(\vartheta_{s^*} \geq \frac{\lambda}{2}) \leq \mathbb{P}(\vartheta_{s^*} \geq \frac{\lambda\varepsilon}{2}) = \mathbb{P}(\alpha_{s^*} \geq \beta\tau) \leq 2^{-\mathrm{N}-1-\acute{m}}$$

Hence, by the union bound we have

$$\mathbb{P}(E_i) \leq \mathbb{P}(\exists s^* \in \mathcal{A}_{\mathrm{L}};\ \vartheta_{s^*} > \frac{\lambda}{2}) \leq \sum_{s^* \in \mathcal{A}_{\mathrm{L}}} \mathbb{P}(\vartheta_{s^*} > \frac{\lambda}{2}) \leq \sum_{s^* \in \mathcal{A}_{\mathrm{L}}} 2^{-\mathrm{N}-1-\acute{m}} \leq 2^{-\acute{m}} \tag{18}$$

The last inequality is due to the fact that $|R_i \cup \{i+1\}| \leq \mathrm{N}+1$.
Now if we assign $\acute{m} = \log(\mathrm{L}) + \dot{m}$, we get

$$\mathbb{P}(E_i) \leq 2^{-\log(\mathrm{L})-\dot{m}} = \frac{2^{-\dot{m}}}{\mathrm{L}}.$$

Let E be the event of happening an error in recognizing the mother sequences through the hole algorithm. By union bound we get,

$$\mathbb{P}(E) \leq \sum_{i=1}^{\mathrm{L}} \mathbb{P}(E_i) \leq \sum_{i=1}^{\mathrm{L}} \frac{2^{-\dot{m}}}{\mathrm{L}} = 2^{-\dot{m}}. \tag{19}$$

Hence, with probability at least $1 - 2^{\dot{m}}$ we are sure that we have obtained the mother sequences correctly. In addition, we have drawn their frequencies with precision $\frac{\lambda\varepsilon}{2}$. Our sample complexity becomes

$$\frac{32e\ln(2)}{\lambda^2\varepsilon^2}\big(\frac{f^3 + (1-f)^3}{(1-2f)^2}\big)^{\mathrm{N}+1}(\mathrm{N} + 1 + \log(\mathrm{L}) + \dot{m}) \tag{20}$$