General Mixture

written by Khashayar

February 2018

1 Problem formulation

We have N number of binary mother sequences $\mathcal{Y} = \{y_i\}_{i=1}^{\mathrm{N}}$ with length L, each of them have a corresponding probability f_i, where $\sum_{i=1}^{\mathrm{N}} f_i = 1$. In each time step $1 \leq t \leq m$, one of the mother sequences such as y_{r_t} is picked with respect to the probabilities $\mathcal{F} = \{f_i\}$, and is passed through a symmetric noisy channel with flip probability f. In the other end, we observe the noisy samples $\{z_i\}_{i=1}^{m}$. We know the channel's flip rate f and the sequences' length L, and we want to estimate N, the mother sequences themselves, and their frequencies $\{f_i\}_{i=1}^{\mathrm{N}}$.
We will introduce an estimator and bound it's norm-2 error using notions from high dimensional statistics.

2 The Estimator

We denote the space of all binary sequences with length L by $\mathcal{A}_{\mathrm{L}}$, which has size $n = 2^{\mathrm{L}}$. By $\mathcal{A}_{\mathrm{L}}^m$ and $\mathcal{A}_{\mathrm{L}}^s$ we refer to the space of mother sequences and sample sequences respectively. Let $p = \{p_i\}_{i=1}^n$ be the probability distribution induced by the set of mother sequences $\mathcal{Y}$ on $\mathcal{A}_{\mathrm{L}}^s$. Furthermore, for sequences s_1, s_2, let $d(s_1, s_2)$ be their hamming distance, and $p(s_1, s_2) = f^{d(s_1,s_2)}(1-f)^{\mathrm{L}-d(s_1,s_2)}$.

2.1 Preliminaries

We denote the empirical distribution on $\mathcal{A}_{\mathrm{L}}^s$ by $q = \{q_i\}_{i=1}^n$, which is obtained by counting the number of observed sequences for each type in $\mathcal{A}_{\mathrm{L}}^s$, and normalizing them by their sum m. Look at q as a random distribution, thus each q_i is a random variable. Obviously, we see that $\mathbb{E}[q] = p$. In fact, $\{q_i\}_{i=1}^{\mathrm{N}}$ is the normalized version of a multinomial distribution with probabilities $\{p_i\}_{i=1}^n$. Hence

$$Cov[q_i, q_j] = -\frac{p_i p_j}{m}\ ,\ Var[p_i] = \frac{p_i(1-p_i)}{m},\ \forall\ 1 \leq i < j \leq n. \tag{1}$$

Let $\epsilon = \{\epsilon_i\}_{i=1}^{n}$ be $\epsilon = p - q$, for which $\mathbb{E}[\epsilon] = 0$. Equation (1) gives

$$\mathbb{E}[\epsilon_i \epsilon_j] = -\frac{p_i p_j}{m} \ , \ \mathbb{E}[\epsilon_i^2] = \frac{p_i(1-p_i)}{m}, \ \forall\ 1 \le i < j \le n. \tag{2}$$

Now, Let B be a $n \times n$ matrix which has its rows and columns corresponded to sequences in $\mathcal{A}_{\mathrm{L}}$, such that for every sequences $s_1, s_2 \in \mathcal{A}_{\mathrm{L}}$,

$$B(s_1, s_2) = p(s_1, s_2),$$

where $B(s_1, s_2)$ is the matrix element corresponding to row s_1 and column s_2. Let $\mathcal{Q}(\mathcal{A}_{\mathrm{L}})$ be the space of probability distributions on $\mathcal{A}_{\mathrm{L}}$. For desired $v \in \mathcal{Q}(\mathcal{A}_{\mathrm{L}})$ and $s \in \mathcal{A}_{\mathrm{L}}$, the element in vector v which corresponds to sequence s is denoted by v_s. It can be easily checked that B is an invertible probability matrix, hence it defines an onto and one to one linear transformation from $\mathcal{Q}(\mathcal{A}_{\mathrm{L}}^m)$ to $\mathcal{Q}(\mathcal{A}_{\mathrm{L}}^s)$. Now We apply the inverse map B^{-1} to the empirical distribution q in $\mathcal{Q}(\mathcal{A}_{\mathrm{L}}^s)$, to obtain the empirical distribution $\mathrm{Y} = \{\mathrm{Y}_s\}_{s \in \mathcal{A}_{\mathrm{L}}^m}$ in $\mathcal{Q}(\mathcal{A}_{\mathrm{L}}^m)$. Let $\mathcal{F}$ be a vector in $\mathcal{Q}(\mathcal{A}_{\mathrm{L}}^m)$ such that for each $s \in \mathcal{A}_{\mathrm{L}}$

$$\mathcal{F}_s = \begin{cases} f_i & s = y_i \in \mathcal{Y} \\ 0 & o.w. \end{cases}$$

By the above defnition, our goal is to estimate $\mathcal{F}$. Note that q is a noisy unbiased observation of p, and it can be readily seen that $p = \mathrm{B}(\mathcal{F})$. Noting the fact that B^{-1} is linear, we deduce that

$$\mathbb{E}[\mathrm{Y}] = \mathbb{E}[B^{-1}(q)] = B^{-1}(\mathbb{E}[q]) = B^{-1}(p) = \mathcal{F}. \tag{3}$$

Hence, if we define $\alpha \in \mathcal{Q}(\mathcal{A}_{\mathrm{L}}^m)$ such that for each $s \in \mathcal{A}_{\mathrm{L}}^m$,

$$\alpha_s = \mathrm{Y}_s - \mathcal{F}_s,$$

Then according to equation (3), Y is an unbiased noisy observation of $\mathcal{F}$ with noise α. Thus $\mathbb{E}[\alpha] = 0$. In addition, by definition of ϵ, observe that

$$\alpha = \mathrm{B}^{-1}(\epsilon). \tag{4}$$

2.2 Calculation of Noise Variance

Now, we are interested in calculating the errors $\mathbb{E}[\alpha_s]$ for each $s \in \mathcal{A}_{\mathrm{L}}^m$.

It can bee seen that B is the result of tensoring the matrix $\begin{bmatrix} 1-f & f \\ f & 1-f \end{bmatrix}$ to itself L times. Hence,

$$B^{-1} = \frac{1}{(1-2f)^{\mathrm{L}}} \underbrace{\begin{bmatrix} 1-f & -f \\ -f & 1-f \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1-f & -f \\ -f & 1-f \end{bmatrix}}_{n \text{ times}}. \tag{5}$$

By equation (5) we obtain that

$$\mathrm{B}^{-1}(s_1, s_2) = D_f.(-1)^{d(s_1,s_2)} p(s_1, s_2), \tag{6}$$

where $D_f = \frac{1}{(1-2f)^{\mathrm{L}}}$.
Remark that we can index the elements of p, q, ϵ by sequences in $\mathcal{A}_{\mathrm{L}}$ as well.
According to equations (4) and (6), for each $s^* \in \mathcal{A}_{\mathrm{L}}^m$,

$$\alpha_{s^*} = D_f \sum_{s \in \mathcal{A}_{\mathrm{L}}^s} p(s^*, s)\epsilon_s (-1)^{d(s^*,s)}. \tag{7}$$

Hence,

$$\begin{aligned}
\mathbb{E}[\alpha_{s^*}^2] &= \mathbb{E}[D_f^2 (\sum_{s \in \mathcal{A}_{\mathrm{L}}^s} p(s^*, s)\epsilon_s (-1)^{d(s^*,s)})^2] \\
&= D_f^2 \sum_{s \in \mathcal{A}_{\mathrm{L}}^s} \mathbb{E}[p(s^*, s)^2 \epsilon_s^2] + 2D_f^2 \sum_{s1 \neq s2 \in \mathcal{A}_{\mathrm{L}}^s} \mathbb{E}[p(s^*, s_1)p(s^*, s_2)\epsilon_{s_1}\epsilon_{s_2}(-1)^{d(s_1,s_2)}] \\
&= \frac{D_f^2}{m} \sum_{s \in \mathcal{A}_{\mathrm{L}}^s} p(s^*, s)^2 p_s (1 - p_s) - \frac{2D_f^2}{m} \sum_{s1 \neq s2 \in \mathcal{A}_{\mathrm{L}}^s} p(s^*, s_1)p(s^*, s_2)p_{s_1}p_{s_2}(-1)^{d(s_1,s_2)} \\
&= R_f [\sum_{s \in \mathcal{A}_{\mathrm{L}}^s} p(s^*, s)^2 p_s - (\sum_{s \in \mathcal{A}_{\mathrm{L}}^s} p(s^*, s)p_s(-1)^{d(s^*,s)})^2], \qquad (8)
\end{aligned}$$

where

$$R_f = \frac{1}{(1-2f)^{2\mathrm{L}} m} \; . \tag{9}$$

Note that $p = p_{s s \in \mathcal{A}_{\mathrm{L}}^s}$ is the distribution induced by the mother sequences on $\mathcal{A}_{\mathrm{L}}^s$.

$$p_s = \sum_{i=1}^{\mathrm{N}} f_i p(y_i, s). \tag{10}$$

Thus (8) gives

$$\mathbb{E}[\alpha_{s^*}^2] \tag{11}$$

$$=R_f[\sum_{s\in\mathcal{A}_{\mathrm{L}}^s} p(s^*,s)^2(\sum_{i=1}^{\mathrm{N}} f_i p(y_i,s)) - [\sum_{s\in\mathcal{A}_{\mathrm{L}}^s} p(s^*,s)(\sum_{i=1}^{\mathrm{N}} f_i p(y_i,s))(-1)^{d(s^*,s)}]^2] \tag{12}$$

$$=R_f[\sum_{i=1}^{\mathrm{N}} f_i \sum_{s\in\mathcal{A}_{\mathrm{L}}^s} p(s^*,s)^2 p(y_i,s) - [\sum_{s\in\mathcal{A}_{\mathrm{L}}^s} p(s^*,s)(\sum_{i=1}^{\mathrm{N}} f_i p(y_i,s))(-1)^{d(s^*,s)}]^2] \tag{13}$$

$$=R_f[\sum_{i=1}^{\mathrm{N}} f_i \sum_{s\in\mathcal{A}_{\mathrm{L}}^s} p(s^*,s)^2 p(y_i,s) - [\sum_{i=1}^{\mathrm{N}} f_i \sum_{s\in\mathcal{A}_{\mathrm{L}}^s} p(s^*,s)p(y_i,s)(-1)^{d(s^*,s)}]^2] \tag{14}$$

$$=R_f[(\sum_{i=1}^{\mathrm{N}} f_i(f^3+(1-f)^3)^{\mathrm{L}-d(s^*,y_i)}(f(1-f))^{d(s^*,y_i)}) - \psi(s^*)], \tag{15}$$

$$\tag{16}$$

Where $\psi(s^*) = \begin{cases} f_i^2(1-2f)^{2\mathrm{L}} & s^* = y_i \in \mathcal{Y} \\ 0 & o.w. \end{cases}$

We observe that if f is relatively small, then the major mass of noise is on the mother sequences themselves.

2.3 Estimation

For a subset κ of sequences in $\mathcal{A}_{\mathrm{L}}^m$, define $\hat{\mathcal{F}}_\kappa$ such that for each sequence $s \in \mathcal{A}_{\mathrm{L}}^m$,

$$\hat{\mathcal{F}}_{\kappa s} = \begin{cases} \mathrm{Y}_s & s \in \kappa \\ 0 & o.w. \end{cases}$$

Define

$$\hat{\mathcal{F}} = \hat{\mathcal{F}}_{\hat{\kappa}} = \arg\min_{\kappa}\{\| \hat{\mathcal{F}}_\kappa - \mathrm{Y} \|^2 + \mathrm{pen}(\kappa)\}. \tag{17}$$

At the moment, we define

$$\mathrm{pen}(\kappa) = R_f[(f^3+(1-f)^3)^{\mathrm{L}} + (|\kappa|-1)f(1-f)(f^3+(1-f)^3)^{\mathrm{L}-1}]. \tag{18}$$

However, we will show later that we can use a more complicated penalty term which gives a better accuracy, with the cost of a higher computational cost. Also, note that in equation (18) the first term is a constant and does not depend on κ, thus has no effect on our model selection, but for sake of simplicity of the proof, we don't omit it.

Remark that if we sort the elements of vector Y, it is sufficient to evaluate the minimization problem in (17) only on the top r sequences with the largest values in Y. Therefore, the computational cost of our estimator is linear.

2.4 Error Bound

Note that by definition, $\mathcal{F}_\mathcal{Y}$ is a vector equal to Y with respect to the mother sequences, and zero otherwise. Now according to equation (17), we can write

$$\parallel \hat{\mathcal{F}} - \mathrm{Y} \parallel^2 + \mathrm{pen}(\hat{\kappa}) \leq \parallel \hat{\mathcal{F}}_\mathcal{Y} - \mathrm{Y} \parallel^2 + \mathrm{pen}(\mathcal{Y}),$$

which gives

$$\begin{aligned} &\parallel \hat{\mathcal{F}} - \mathcal{F} \parallel^2 + \parallel \mathcal{F} - \mathrm{Y} \parallel^2 + 2 < \hat{\mathcal{F}} - \mathcal{F}, \mathcal{F} - \mathrm{Y} > + \mathrm{pen}(\hat{\kappa}) \leq \\ &\parallel \hat{\mathcal{F}}_\mathcal{Y} - \mathcal{F} \parallel^2 + \parallel \mathcal{F} - \mathrm{Y} \parallel^2 + 2 < \hat{\mathcal{F}}_\mathcal{Y} - \mathcal{F}, \mathcal{F} - \mathrm{Y} > + \mathrm{pen}(\mathcal{Y}). \end{aligned} \tag{19}$$

But note that for $s \notin \mathcal{Y}$,

$$\mathbb{E}[(\hat{\mathcal{F}}_{\mathcal{Y}_s} - \mathcal{F}_s)(\mathcal{F}_s - \mathrm{Y}_s)] = \mathbb{E}[-\mathcal{F}_s(\mathcal{F}_s - \mathrm{Y}_s)] = 0,$$

and remark that $\mathcal{F} - \mathrm{Y} = \epsilon$, hence

$$< \hat{\mathcal{F}}_\mathcal{Y} - \mathcal{F}, \mathcal{F} - \mathrm{Y} > = < \alpha_\mathcal{Y}, -\alpha > = - \parallel \alpha_\mathcal{Y} \parallel^2 \leq 0,$$

where $\epsilon_\mathcal{Y}$ is the projection of α on the indices with respect to the mother sequences. In the same manner, observe that

$$< \hat{\mathcal{F}} - \mathcal{F}, \mathcal{F} - \mathrm{Y} > = < \alpha_{\hat{\kappa}}, -\alpha > = - \parallel \alpha_{\hat{\kappa}} \parallel^2 . \tag{20}$$

Thus, equation (19) gives

$$\parallel \hat{\mathcal{F}} - \mathcal{F} \parallel^2 \leq \parallel \hat{\mathcal{F}}_\mathcal{Y} - \mathcal{F} \parallel^2 + \mathrm{pen}(\mathcal{Y}) + \parallel \alpha_{\hat{\kappa}} \parallel^2 - \mathrm{pen}(\hat{\kappa}),$$

which gives

$$\mathbb{E} \parallel \hat{\mathcal{F}} - \mathcal{F} \parallel^2 \leq \mathbb{E} \parallel \hat{\mathcal{F}}_\mathcal{Y} - \mathcal{F} \parallel^2 + \mathrm{pen}(\mathcal{Y}) + \mathbb{E} \left[\parallel \alpha_{\hat{\kappa}} \parallel^2 - \mathrm{pen}(\hat{\kappa}) \right] . \tag{21}$$

But due to equation (15), for a fixed κ we have

$$
\begin{aligned}
\mathbb{E} \parallel \alpha_\kappa \parallel^2 =& \sum_{s^* \in \kappa} R_f [(\sum_{i=1}^{\mathrm{N}} f_i (f^3 + (1-f)^3)^{\mathrm{L} - d(s^*, y_i)} (f(1-f))^{d(s^*, y_i)}) - \psi(s^*)] \\
\leq& R_f \sum_{s^* \in \kappa} (\sum_{i=1}^{\mathrm{N}} f_i (f^3 + (1-f)^3)^{\mathrm{L} - d(s^*, y_i)} (f(1-f))^{d(s^*, y_i)}) \\
=& R_f \sum_{s^* \in \kappa \cap \mathcal{Y}} (\sum_{i=1}^{\mathrm{N}} f_i (f^3 + (1-f)^3)^{\mathrm{L} - d(s^*, y_i)} (f(1-f))^{d(s^*, y_i)}) \\
+& R_f \sum_{\substack{s^* \in \kappa \\ s^* \notin \mathcal{Y}}} (\sum_{i=1}^{\mathrm{N}} f_i (f^3 + (1-f)^3)^{\mathrm{L} - d(s^*, y_i)} (f(1-f))^{d(s^*, y_i)}) \\
\leq& R_f \big(\sum_{s^* = y_j \in \kappa \cap \mathcal{Y}} f_j (f^3 + (1-f)^3)^{\mathrm{L}} + \sum_{i \neq j}^{\mathrm{N}} f_i (f^3 + (1-f)^3)^{\mathrm{L}-1} f(1-f) \big) \\
+& R_f \sum_{\substack{s^* \in \kappa \\ s^* \notin \mathcal{Y}}} \sum_{i=1}^{\mathrm{N}} f_i (f^3 + (1-f)^3)^{\mathrm{L}-1} f(1-f) \\
\leq& R_f [(f^3 + (1-f)^3)^{\mathrm{L}} + (|\kappa| - 1)(f^3 + (1-f)^3)^{\mathrm{L}-1} f(1-f)] = \mathrm{pen}(\kappa).
\end{aligned}
\tag{22}
$$

On the other hand,

$$
\mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2 = \mathbb{E} \parallel \alpha_{\mathcal{Y}} \parallel^2 \tag{23}
$$

$$
\begin{aligned}
&= \sum_{y_j \in \mathcal{Y}} R_f [(\sum_{i=1}^{\mathrm{N}} f_i (f^3 + (1-f)^3)^{\mathrm{L} - d(y_j, y_i)} (f(1-f))^{d(y_j, y_i)}) - \psi(y_j)] \\
&\geq R_f \sum_{y_j \in \mathcal{Y}} [(f_j (f^3 + (1-f)^3)^{\mathrm{L} - d(y_j, y_j)} (f(1-f))^{d(y_j, y_j)}) - \psi(y_j)] \\
&= R_f [(\sum_{i=1}^{\mathrm{N}} f_i)(f^3 + (1-f)^3)^{\mathrm{L}} - (\sum_{i=1}^{\mathrm{N}} f_i^2)(1-2f)^{2\mathrm{L}}].
\end{aligned}
\tag{24}
$$

But it can easily be checked that

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

thus, we obtain that

$$
\mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2 \geq R_f [(f^3 + (1-f)^3)^{\mathrm{L}} - (\sum_{i=1}^{\mathrm{N}} f_i^2)(f^3 + (1-f)^3)^{\mathrm{L}}]. \tag{25}
$$

Now suppose that we have a constraint for $0 \leq \gamma \leq 1$ as

$$
\min_i f_i \geq \frac{\gamma}{\mathrm{N}},
$$

which is essential so that the problem becomes identifiable. Then, we get

$$\sum_i f_i^2 \geq \frac{(\mathrm{N}-1)\gamma^2}{\mathrm{N}^2} + (1 - \frac{(\mathrm{N}-1)\gamma}{\mathrm{N}})^2 = 1 - \frac{\mathrm{N}-1}{\mathrm{N}}\gamma(2-\gamma). \tag{26}$$

inequalities (25) , (26) give

$$\begin{aligned}\mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2 \geq & \frac{\mathrm{N}-1}{\mathrm{N}}\gamma(2-\gamma) R_f (f^3 + (1-f)^3)^{\mathrm{L}} \\ = & \frac{\mathrm{N}-1}{\mathrm{N}}\gamma(2-\gamma) \frac{1}{1 + \frac{f(1-f)}{f^3+(1-f)^3}(\mathrm{N}-1)} \mathrm{pen}(\mathcal{Y}). \end{aligned} \tag{27}$$

Define the ratio

$$K_{f,\gamma,\mathrm{N}} = \frac{N}{N-1} \frac{\left(1 + \frac{f(1-f)}{f^3+(1-f)^3}(\mathrm{N}-1)\right)}{\gamma(2-\gamma)}.$$

Then, According to inequality (27),

$$\mathrm{pen}(\mathcal{Y}) \leq K_{f,\gamma,\mathrm{N}} \mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2 . \tag{28}$$

Now observe that if $\hat{\kappa}$ was fixed, then inequality (22) gives

$$\mathbb{E}\left[\parallel \alpha_{\hat{\kappa}} \parallel^2 - \mathrm{pen}(\hat{\kappa})\right] \leq 0.$$

Consequently, due to inequality (21),

$$\mathbb{E} \parallel \hat{\mathcal{F}} - \mathcal{F} \parallel^2 \leq \mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2 + \mathrm{pen}(\mathcal{Y}), \tag{29}$$

which combined by (28) reveals

$$\mathbb{E} \parallel \hat{\mathcal{F}} - \mathcal{F} \parallel^2 \leq (1 + K_{f,\gamma,\mathrm{N}}) \mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2 . \tag{30}$$

Moreover, we can find an explicit bound for the expected norm-2 error $\mathbb{E} \parallel \hat{\mathcal{F}} - \mathcal{F} \parallel^2$ by driving a upper bound for $\mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2$. According to the inequality (22), we have

$$\mathrm{pen}(\mathcal{Y}) \geq \parallel \alpha_{\mathcal{Y}} \parallel^2 = \mathbb{E} \parallel \hat{\mathcal{F}}_{\mathcal{Y}} - \mathcal{F} \parallel^2 . \tag{31}$$

Inequalities (31) , (29) reveal

$$\begin{aligned}&\mathbb{E} \parallel \hat{\mathcal{F}} - \mathcal{F} \parallel^2 \leq \\ &\quad 2\mathrm{pen}(\mathcal{Y}) = 2R_f[(f^3 + (1-f)^3)^{\mathrm{L}} + (\mathrm{N}-1)f(1-f)(f^3+(1-f)^3)^{\mathrm{L}-1}].\end{aligned} \tag{32}$$