On the Fundamental Limits of Heterogeneous Distributed Detection: Price of Anonymity

Wei-Ning Chen, Ho-Chun Chen, and I-Hsiang Wang
Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan
Email: {r05942078,r04942046,ihwang}@ntu.edu.tw

***Abstract*—In this paper, we explore the fundamental limits of heterogeneous distributed detection in an anonymous sensor network with n sensors and a single fusion center. The fusion center collects the single observation from each of the n sensors to detect a binary parameter. The sensors are clustered into multiple groups, and different groups follow different discrete distributions under a given hypothesis. The key challenge for the fusion center is the anonymity of sensors – although it knows the exact number of sensors and the distribution of observations in each group, it does not know which group each sensor belongs to. It is hence natural to consider it as a composite hypothesis testing problem. We focus on the Neyman-Pearson setting and give upper and lower bounds of the error exponent of the worst-case type-II probability of error as n tends to infinity, assuming the number of sensors in each group is proportional to n. Our results elucidate the price of anonymity in heterogeneous distributed detection. The results are also applied to distributed detection under Byzantine attacks, which hints that the conventional simple hypothesis testing approach might be too pessimistic.**

A full version of this paper is accessible at:
http://homepage.ntu.edu.tw/~ihwang/Eprint/isit18hd.pdf

I. Introduction

In wireless sensor networks, the cost of identifying individual sensors increases drastically as the number of sensors grows. For distributed detection [1], when the observations follow i.i.d. distributions across all sensors, identifying individual sensors is not very important. When the fusion center can fully access the observations, the empirical distribution (types) of the collected observation is a sufficient statistics. When the communication between each sensor and the fusion center is limited, for binary hypothesis testing it is asymptotically optimal to use the same local decision function at all sensors [2]. Hence, anonymity is not a critical issue for the classical (homogeneous) distributed detection problem.

However, when the distributions of the observations are *heterogeneous*, that is, the distribution of the observation varies across sensors, sensor anonymity may deteriorate the performance of distributed detection, even for binary hypothesis testing. One such example is distributed detection under Byzantine attack [3], where a fixed number of sensors are compromised by malicious attackers and report fake observations following certain distributions. Even if the fusion center is aware of the number of compromised sensors and the attacking strategy that renders worst-case detection performance (the least favorable distribution as considered in [4]–[6]), it is more difficult to detect the hidden parameter when the fusion center does not know which sensors are compromised.

In this paper, we aim to quantify the performance loss due to sensor anonymity in heterogeneous distributed detection, with n sensors and a single fusion center. Each sensor (say sensor i, $i \in \{1, ..., n\}$) has a single random observation X_i. The goal of the fusion center is to estimate the hidden parameter $\theta \in \{0, 1\}$ (that is, binary hypothesis testing) from the collected observations. The distributions of the observations, however, are *heterogeneous* – observations at different sensors may follow different sets of distributions. In particular, we assume that these n sensors are clustered into K groups $\{\mathcal{I}_1, ..., \mathcal{I}_K\}$, and group $\mathcal{I}_k \subseteq \{1, ..., n\}$ comprises $n\alpha_k$ sensors, for $k = 1, ..., K$. Under hypothesis $\mathcal{H}_\theta$, $\theta \in \{0, 1\}$,

$$X_i \sim P_{\theta;k}, \text{ for } i \in \mathcal{I}_k.$$

Moreover, the sensors are *anonymous*, that is, the collected observations at the fusion center is *unordered*. In other words, although the fusion center is fully aware of the *heterogeneity* of it observation, including the set of distributions $\{P_{\theta;k} \mid \theta \in \{0, 1\},\ k = 1, ..., K\}$ and $\{\alpha_k \mid k = 1, ..., K\}$, it does not know what distribution each individual sensor will follow.

To overcome the difficulty of not knowing the exact distributions of the observations, we formulate the detection problem as a *composite hypothesis testing* problem, where the length-n vector observation follows a product distribution within a finite class of n-letter product distributions under a given parameter θ. The class consists of $\binom{n}{n\alpha_1,...,n\alpha_K}$ possible product distributions, each of which follows one of the $\binom{n}{n\alpha_1,...,n\alpha_K}$ possible partitions of the sensors. The fusion center takes all the possible partitions into consideration when detecting the hidden parameter. We focus on a Neyman-Pearson setting, where the goal is to minimize the worst-case type-II probability of error such that the worst-case type-I probability of error is not larger than a constant. As a first step towards understanding the performance loss due to sensor anonymity, our goal is to characterize the error exponent of the minimum worst-case type-II probability of error as $n \to \infty$ with $\{\alpha_k \mid k = 1, ..., K\}$ being fixed.

Our main contribution is a set of upper and lower bounds on the error exponent. For the achievability, we develop a threshold test on the KullbackLeibler (KL) divergence from a chosen distribution to the empirical distribution (type) of the collected observations, similar to the Hoeffding test [7]. The resulting lower bound on the error exponent is the minimization of a linear combination of KL divergences with the k-th term being $D(U_k \| P_{1;k})$ and α_k being the coefficient, for $k = 1, ..., K$. The minimization is over all possible distributions $U_1, ..., U_K$ such that $\sum_{k=1}^{K} \alpha_k U_k = \sum_{k=1}^{K} \alpha_k P_{0;k}$. For the converse, we relax the composite testing problem to a simple one, where the null hypothesis is a tailored mixture of several n-letter heterogeneous product distributions, and the alternative hypothesis is a specific n-letter heterogeneous product distribution. The resulting upper bound on the error exponent takes the same form as the lower bound except that the constraint in the minimization is slightly relaxed.

As a by-product, we apply our results for $K = 2$ to the distributed detection problem under Byzantine attack and further obtain bounds on the worst-case type-II error exponent. Compared with the worst-case exponent in an alternative Bayesian formulation [3] where the observation of sensors are assumed to be i.i.d. according to a mixture distribution, it is shown that the worst-case exponent in the composite testing formulation is strictly larger. This hints that the conventional approach taken in [3] might be too pessimistic.

II. Problem Formulation

Following the description of the setting in Section I, let us formulate the composite hypothesis testing problem. Let $\sigma(i)$ denote the label of the group that sensor i belongs to. This labeling $\sigma(\cdot)$, however, is not revealed to the fusion center. Hence, the fusion center needs to consider all $\binom{n}{n\alpha_1, ..., n\alpha_K}$ possible $\sigma : \{1, ..., n\} \to \{1, ..., K\}$ satisfying

$$|\{i \mid \sigma(i) = k\}| = n\alpha_k,\ \forall\, k = 1, ..., K, \tag{1}$$

and decides whether the hidden θ is 0 or 1. For notational convenience, let $\boldsymbol{\alpha}$ denote the vector $\begin{bmatrix}\alpha_1 & ... & \alpha_K\end{bmatrix}^\mathsf{T}$, and let $\mathcal{S}_{n,\boldsymbol{\alpha}}$ denote the collection of all labelings satisfying (1).

Hence, the fusion center is faced with the following *composite* hypothesis testing problem:

$$\mathcal{H}_\theta : X^n \sim \mathbb{P}_{\theta;\sigma} \triangleq \textstyle\prod_{i=1}^{n} P_{\theta;\sigma(i)},\ \text{for some } \sigma \in \mathcal{S}_{n,\boldsymbol{\alpha}}.$$

The observations take values in a finite alphabet $\mathcal{X}$, and hence $P_{\theta;k} \in \mathcal{P}_\mathcal{X}$ for all $\theta \in \{0, 1\}$ and $k \in \{1, ..., K\}$, where $\mathcal{P}_\mathcal{X}$ denote the collection of all possible distributions over $\mathcal{X}$.

The worst-case type-I and type-II probability of error of a decision rule ϕ are defined as

$$\mathsf{P_F}^{(n)}(\phi) \triangleq \max_{\sigma\in\mathcal{S}_{n,\boldsymbol{\alpha}}} \mathbb{P}_{0;\sigma}\{\phi(X^n) = 1\} \quad \text{(Type I)}$$

$$\mathsf{P_M}^{(n)}(\phi) \triangleq \max_{\sigma\in\mathcal{S}_{n,\boldsymbol{\alpha}}} \mathbb{P}_{1;\sigma}\{\phi(X^n) = 0\} \quad \text{(Type II)}.$$

Our focus is on the Neyman-Pearson setting: find a decision rule ϕ satisfying $\mathsf{P_F}^{(n)}(\phi) \le \epsilon$ such that $\mathsf{P_M}^{(n)}(\phi)$ is minimized, and let us use $\beta^{(n)}(\epsilon, \boldsymbol{\alpha})$ to denote the minimum type-II probability of error. For the asymptotic regime, we aim to explore if $\beta^{(n)}(\epsilon, \boldsymbol{\alpha})$ decays exponentially fast as $n \to \infty$ with $\boldsymbol{\alpha}$ fixed, and characterize the corresponding error exponent. For notational convenience, we define upper and lower bounds on the exponent:

$$\overline{E}^*(\epsilon, \boldsymbol{\alpha}) \triangleq \limsup_{n\to\infty} \left\{-\tfrac{1}{n}\log_2 \beta^{(n)}(\epsilon, \boldsymbol{\alpha})\right\}$$
$$\underline{E}^*(\epsilon, \boldsymbol{\alpha}) \triangleq \liminf_{n\to\infty} \left\{-\tfrac{1}{n}\log_2 \beta^{(n)}(\epsilon, \boldsymbol{\alpha})\right\}.$$

Remark 2.1: The original distributed detection problem [1], [2], [6] involves local decision functions at the sensors to address the limited communication between each sensor and the fusion center. In this work, we neglect this part and assume that the fusion center can collect all unordered observations, in order to focus on the impact of anonymity.

III. Main Results

A. Bounds on the Exponent

Theorem 3.1 (Lower Bound on the Exponent): $\forall\, \epsilon \in (0, 1)$,

$$\begin{aligned}\underline{E}^*(\epsilon, \boldsymbol{\alpha}) \ge \min_{\boldsymbol{U}\in(\mathcal{P}_\mathcal{X})^K} &\textstyle\sum_{k=1}^{K} \alpha_k D\left(U_k \,\|P_{1;k}\right)\\ \text{subject to} \quad &\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{U} = \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{P}_0\end{aligned} \tag{2}$$

Here $\boldsymbol{U} \triangleq \begin{bmatrix}U_1 & ... & U_K\end{bmatrix}^\mathsf{T}$ denotes a K-tuple of distributions, and similarly $\boldsymbol{P}_0 \triangleq \begin{bmatrix}P_{0;1} & ... & P_{0;K}\end{bmatrix}^\mathsf{T}$. Hence, $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{U} = \sum_{k=1}^{K} \alpha_k U_k$ denotes the mixture of distributions. For notational convenience, denote the specific mixture distribution $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{P}_0$ as $M_0(\boldsymbol{\alpha})$, representing the mixture of distributions in $\mathcal{H}_0$ with respect to mixing parameter $\boldsymbol{\alpha}$.

Proof Sketch: To achieve the desired error exponent, consider the following threshold test: we accept the null hypothesis $\mathcal{H}_0$ when the type (empirical distribution) of the observations, denoted as Π_{x^n} (see Appendix A for the definition) , *close* to $M_0(\boldsymbol{\alpha})$ within a threshold ϵ_n. More precisely, the acceptance region is defined as $\mathcal{A}^{(n)} = \{x^n : D\left(\Pi_{x^n} \| M_0(\boldsymbol{\alpha})\right) \le \epsilon_n\}$. The key observation is that for all samples X^n drawn from $\mathbb{P}_{0;\sigma}$, regardless of which σ, their type converges to $M_0(\boldsymbol{\alpha})$ according to the law of large numbers. Hence if we properly choose a threshold ϵ_n, the acceptance region $\mathcal{A}^{(n)}$ defines a *level-ϵ test* (i.e. a test with $\mathsf{P_F}^{(n)} \le \epsilon$) for any $\epsilon \ge 0$. Note that we can choose $\epsilon_n \to 0$. It can be shown that under $\mathcal{H}_1$, $\mathbb{P}_{1;\sigma'}\left\{\mathcal{A}^{(n)}\right\} \to 2^{-n(\underline{E}^*(\epsilon,\boldsymbol{\alpha})+o(1))}$ for any σ'. Details are left in Appendix A of the full version. ■

Theorem 3.2 (Upper Bound on the Exponent): $\forall\, \epsilon \in (0, 1)$,

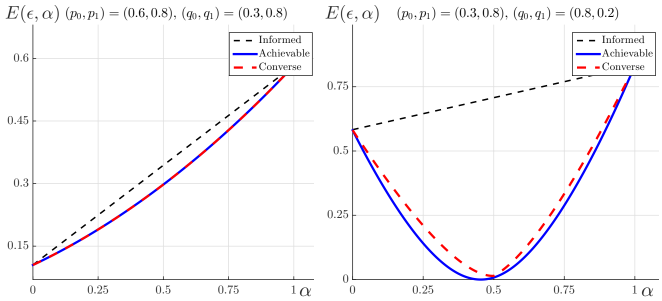
$$\begin{aligned}\overline{E}^*(\epsilon, \boldsymbol{\alpha}) \le \min_{\substack{\boldsymbol{U}\in(\mathcal{P}_\mathcal{X})^K\\ \mathbf{B}\in\mathbb{R}^{K\times K}}} &\textstyle\sum_{k=1}^{K} \alpha_k D\left(U_k \,\|P_{1;k}\right)\\ \text{subject to} \quad &(\mathbf{B})_{i,j} \ge 0\ \forall\, i, j, \quad \mathbf{B}\boldsymbol{\alpha} = \boldsymbol{\alpha}\\ &\boldsymbol{U} = \mathbf{B}^\mathsf{T}\boldsymbol{P}_0\end{aligned} \tag{3}$$

Proof: See Section IV. The proof of all related lemmas can be found in Appendix B of the full version. ■

Remark 3.1: It can be readily seen that the minimization problem in (3) is more restrictive than that in (2), since any K-tuple of distributions $\boldsymbol{U}$ satisfying the constraint in (3), that is, $\boldsymbol{U} = \mathbf{B}^\mathsf{T}\boldsymbol{P}_0$ and $\mathbf{B}\boldsymbol{\alpha} = \boldsymbol{\alpha}$ for some $\mathbf{B}$, will also satisfy the constraint in (2): $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{U} = \boldsymbol{\alpha}^\mathsf{T}\mathbf{B}^\mathsf{T}\boldsymbol{P}_0 = \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{P}_0$.

Though in general, the upper and lower bound do not match, they do in some non-degenerate regimes, and the error exponent can be specified there. Note that the upper and lower bound are characterized by convex programs, due to the convexity of KL divergence and probability simplex, so it is not hard to compute the bounds.

For ease of illustration, in the rest of this section we restrict our discussions and derivations to the special case of binary alphabet, that is, $|\mathcal{X}| = 2$, and $K = 2$ groups. Let $P_{\theta;1} = \mathrm{Ber}(p_\theta)$ and $P_{\theta;2} = \mathrm{Ber}(q_\theta)$, for $\theta = 0, 1$. Since there are only two groups, we set $\boldsymbol{\alpha} \equiv \begin{bmatrix} 1-\alpha & \alpha \end{bmatrix}^\mathsf{T}$. In Figure 1, we give two examples where the bounds match and do not match, respectively.



(a) Upper bound matches lower bound

(b) Upper and lower bound not tight

Fig. 1: Price of anonymity

To quantify the price of anonymity, note that when the sensors are not anonymous (termed the "informed" setting), it becomes a simple hypothesis testing problem, and the error exponent of the type-II probability of error in the Neyman-Pearson setting is straightforward to derive:

$$E^*_{\text{Informed}}(\boldsymbol{\alpha}) = \textstyle\sum_{k=1}^K \alpha_k D\left(P_{0;k} \,\|P_{1;k}\right).$$

Numerical examples are given in Figure 1 to illustrate the price of anonymity versus the mixing parameter α. In general, anonymity may cause significant performance loss. In certain regimes, the type-II error exponent can even be pushed to zero.

B. Distributed Detection with Byzantine Attacks

We further apply the results to distributed detection with Byzantine attacks, where the sensors are partitioned into two groups. One group consists of $n(1-\alpha)$ *honest* sensors reporting true i.i.d. observations, while the other consists of $n\alpha$ *Byzantine* sensors reporting fake i.i.d. observations. Here we again neglect the local decision function and assume that each sensor can report its observation to the fusion center. The true observations follow P_θ i.i.d. across honest sensors, while the fake ones follow Q_θ i.i.d. across Byzantine sensors, for $\theta = 0, 1$. In general, Q_θ is unknown to the fusion center, but in terms of error exponent, one can find the least favorable Q_0, Q_1 which minimize the error exponent. Hence, our results can be applied here and arrive at the upper and lower bounds for the worst-case type-II error exponent as follows:

Upper bound (Converse):

$$\begin{aligned} &\min_{\substack{Q_0,Q_1,U,V\in\mathcal{P}_\mathcal{X}\\ \mathbf{B}\in\mathbb{R}^{2\times 2},(\mathbf{B})_{i,j}\geq 0\ \forall i,j}} && (1-\alpha)D\left(U\,\|P_1\right) + \alpha D\left(V\,\|Q_1\right)\\ &\text{subject to} && \mathbf{B}\begin{bmatrix}1-\alpha & \alpha\end{bmatrix}^\mathsf{T} = \begin{bmatrix}1-\alpha & \alpha\end{bmatrix}^\mathsf{T}\\ &&& \begin{bmatrix}U & V\end{bmatrix}^\mathsf{T} = \mathbf{B}^\mathsf{T}\begin{bmatrix}P_0 & Q_0\end{bmatrix}^\mathsf{T}\end{aligned}$$

Lower bound (Achievability):

$$\begin{aligned} &\min_{Q_0,Q_1,U,V\in\mathcal{P}_\mathcal{X}} (1-\alpha)D\left(U\,\|P_1\right) + \alpha D\left(V\,\|Q_1\right)\\ &\text{subject to } (1-\alpha)U + \alpha V = (1-\alpha)P_0 + \alpha Q_0\end{aligned} \tag{4}$$

In [3], it assumes that each sensor can be Byzantine with probability α, and hence it becomes a homogeneous distributed detection problem, where the observation of each sensor follows a mixture distribution $(1-\alpha)P_\theta + \alpha Q_\theta$ under hypothesis θ, i.i.d. across all sensors. The worst-case exponent of type-II probability of error, as derived in [3], is hence

$$\min_{Q_0,Q_1\in\mathcal{P}_\mathcal{X}} D\left((1-\alpha)P_0 + \alpha Q_0 \,\|(1-\alpha)P_1 + \alpha Q_1\right). \tag{5}$$

We see that the achievable type-II error exponent (4) in our setting is always greater than that in the i.i.d. scenario (5) (and is *strictly* larger for some α) due to the convexity of KL divergence, implying the i.i.d. mixture model [3] might be too pessimistic. Figure 2 shows the numerical evaluation.

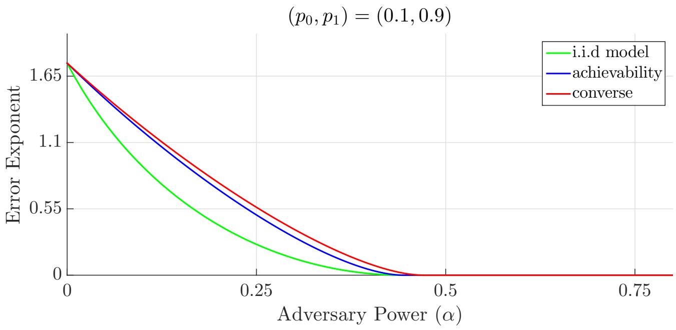


Fig. 2: Comparison between i.i.d. and our setting

IV. Proof of the Converse Upper Bound

In this section, we prove Theorem 3.2. To deal with the composite hypothesis testing problem, we first argue that any relaxed simple testing problem gives us an upper bound. Suppose there exists a test ϕ such that

$$\mathsf{P_F}^{(n)}(\phi) \leq \epsilon, \text{ and } \mathsf{P_M}^{(n)}(\phi) \leq 2^{-nE},$$

then the following simple testing problem must exist a level-ϵ test with type-II error exponent greater than E:

$$\tilde{\mathcal{H}}_0 : X^n \sim \sum_{\sigma'\in\mathcal{S}_{n,\boldsymbol{\alpha}}} w_{\sigma'}\mathbb{P}_{0;\sigma'} \text{ v.s. } \tilde{\mathcal{H}}_1 : X^n \sim \mathbb{P}_{1;\sigma}, \tag{6}$$

where $w_{\sigma'}$ is an arbitrary prior distribution of $\mathbb{P}_{0;\sigma'}$. The reason is as follows: since $\mathsf{P_F}^{(n)}(\phi) \leq \epsilon$, let $\mathcal{A}^{(n)}$ be the $\mathcal{H}_0$-acceptance region of this given test ϕ (i.e. $\phi^{-1}(0)$), we have for any $\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}$, $\mathbb{P}_{0;\sigma'}\left\{\mathcal{A}^{(n)}\right\} \geq 1-\epsilon$, and the same holds for the mixture of all these $\mathbb{P}_{0;\sigma'}$. Therefore, simply choosing $\mathcal{A}^{(n)}$ as the acceptance region gives us a level-ϵ test for the *relaxed* testing problem. Moreover, since $\mathsf{P_M}^{(n)}(\phi) \leq 2^{-nE}$, which means $\mathbb{P}_{1;\sigma}\left\{\mathcal{A}^{(n)}\right\} \leq 2^{-nE}$ for all σ, the type-II error exponent is at least E for any σ.

Our approach to upper bound the type-II error exponent can be summarized as follows. First, pick a specific $\mathbb{P}_{1;\sigma} \in \mathcal{H}_1$, and mix $\mathbb{P}_{0;\sigma'} \in \mathcal{H}_0$ with respect to a prior distribution $w_{\sigma'}$ to obtain a relaxed simple hypothesis testing problem. Then, apply a strong converse lemma [8] for simple hypothesis testing, we obtain an upper bound for the type-II error exponent. Hence, the problem boils down to *choosing a good prior on the hypothesis class* $\mathcal{H}_0$ to minimize the type-II error exponent.

The rest of the proof is organized as follows:

1) First, construct a prior on $\mathcal{H}_0$ with respect to a mixing matrix $\mathbf{B} \in \mathbb{R}^{K\times K}$. This matrix $\mathbf{B}$ determines those $\mathbb{P}_{0;\sigma'} \in \mathcal{H}_0$ upon which we choose to assign uniform prior. For those $\mathbb{P}_{0;\sigma'} \in \mathcal{H}_0$ that are not chosen by $\mathbf{B}$, we assign zero prior. Hence, $\mathbf{B}$ gives a specific relaxed simple testing problem (6).
2) Then, apply the strong converse lemma (Lemma 4.2) to obtain a *multi-letter* upper bound on the type-II error exponent of the relaxed simple testing problem (6).
3) Finally, we single-letterize the multi-letter bound and show that it is upper bounded by (3).

Part 1 (Construction of prior $w_{\sigma'}$): First, let $\mathbb{P}_{1;\sigma}$ be the picked one in $\mathcal{H}_1$. Let $\mathcal{I}_i \triangleq \sigma^{-1}(i)$ denote the collection of indices in group i with respect to the labeling σ.

Let $\mathbf{B} \triangleq \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_K \end{bmatrix} \in \mathbb{R}^{K\times K}$, $(\mathbf{B})_{i,j} \geq 0\ \forall\, i,j$, be the mixing matrix; that is, the i-th column $\boldsymbol{b}_i$ characterizes the component of mixed distribution in $\mathcal{I}_i$. Note that $\mathbf{B}$ must satisfy the constraint $\mathbf{B}\boldsymbol{\alpha} = \boldsymbol{\alpha}$, so that the total number of elements in each group remains the same. For convenience, assume for all $i,j \in \{1,...,K\}$, $n\alpha_i$, $n(\mathbf{B})_{i,j}\alpha_i$ are all non-negative integers, and denote them as n_i and n_{ij} respectively.

Now, consider the following prior distribution on $\mathcal{H}_0$. Put uniform prior on those $\mathbb{P}_{0;\sigma'}$ satisfying the condition $\left|\mathcal{I}'_j \cap \mathcal{I}_i\right| = n_{ij}$ for all j, where $\mathcal{I}'_j = \sigma'^{-1}(j)$, as Figure 3 illustrates. In particular, we set the prior

$$w_{\sigma'} = \left(\binom{n_1}{n_{11},\dots,n_{1K}} \cdot \binom{n_2}{n_{21},\dots,n_{2K}} \cdots \binom{n_K}{n_{K1},\dots,n_{KK}} \right)^{-1} \quad (7)$$

for all σ' such that $\left|\mathcal{I}'_j \cap \mathcal{I}_i\right| = n_{ij}$, and zero otherwise.

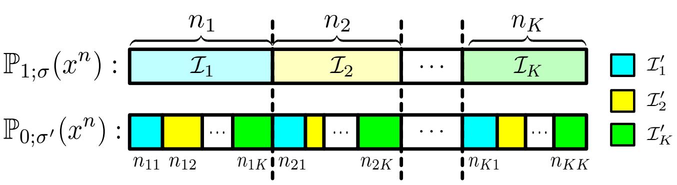


Fig. 3: $\mathbb{P}_{0;\sigma'}$ which satisfying the condition $\left|\mathcal{I}'_j \cap \mathcal{I}_i\right| = n_{ij}$

Then, we claim in the lemma below that the mixed distribution can be decomposed into the product measure with respect to the partition $\mathcal{I}_i$. For notational convenience, we denote $\boldsymbol{x}_{\mathcal{I}_i} \triangleq \{x_j \mid j \in \mathcal{I}_i\}$ as the subsequence.

Lemma 4.1: If we choose the weights $w_{\sigma'}$ as in equation (7), then

$$\sum_{\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}} w_{\sigma'} \left(\prod_{j=1}^n P_{0;\sigma'(j)}(x_j) \right)$$
$$= \prod_{i=1}^K \left(\sum_{\sigma'_i \in \mathcal{S}_{n_i,\boldsymbol{b}_i}} w_{\sigma'_i} \prod_{j\in\mathcal{I}_i} P_{0;\sigma'_i(j)}(x_j) \right) \triangleq \prod_{i=1}^K \boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})$$

where $\sigma'_i : \mathcal{I}_i \to \{1,...,K\}$ is the labeling σ' restricted on $\mathcal{I}_i$. The weights $w_{\sigma'_i}$ satisfy the following equation:

$$w_{\sigma'_i} = \left(\binom{n_i}{n_{i1},\dots,n_{iK}} \right)^{-1}, \text{ if for all } j,\ \left|\sigma_i'^{-1}(j)\right| = n_{ij}$$

and zero otherwise.

In other words, $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})$ is the *uniform* mixture of all distributions $\mathbb{P}_{0;\sigma'_i}$ defined on $\boldsymbol{x}_{\mathcal{I}_i}$, such that $\left|\sigma_i'^{-1}(j)\right| = n_{ij}$, as illustrated in Figure 4. Again, recall that $n_{ij} \triangleq n\alpha_i(\mathbf{B})_{i,j}$, so the matrix $\mathbf{B}$ determines the prior we put on $\mathcal{H}_0$.

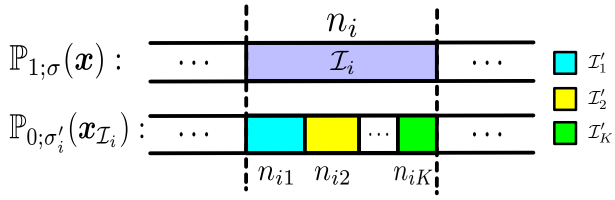


Fig. 4: Example of $\mathbb{P}_{0;\sigma'_i}(\boldsymbol{x}_{\mathcal{I}_i})$ satisfying $\left|\sigma_i'^{-1}(j)\right| = n_{ij}$. $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})$ are the uniform mixture over all such $\mathbb{P}_{0;\sigma'_i}(\boldsymbol{x}_{\mathcal{I}_i})$.

Part 2 (Applying strong converse bound for simple hypothesis testing): So far, we have relaxed the original composite testing problem to the following simple testing problem:

$$\tilde{\mathcal{H}}_0 : X^n \sim \sum_{\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}} w_{\sigma'} \mathbb{P}_{0;\sigma'} = \prod_{i=1}^K \boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i}) \quad (8)$$

$$\tilde{\mathcal{H}}_1 : X^n \sim \mathbb{P}_{1;\sigma} = \prod_{j=1}^n P_{1;\sigma(j)}(x_j) = \prod_{i=1}^K P_{1;i}^{\otimes n_i}(\boldsymbol{x}_{\mathcal{I}_i}), \quad (9)$$

where the superscript '$\otimes n_i$' denotes the i.i.d. extension. Next, the strong converse lemma [8] below will then be used to find upper bounds of the type-II error exponent.

Lemma 4.2 (Strong Converse [8]): For any probability measures $\mathcal{P}, \mathcal{Q}$ on sample space $\mathcal{X}$, measurable set $\mathcal{B} \subset \mathcal{X}$, and any $\gamma > 0$, the following bound holds:

$$\mathcal{P}\{\mathcal{B}\} - 2^{\gamma}\mathcal{Q}\{\mathcal{B}\} \leq \mathcal{P}\left\{ \log \frac{\mathcal{P}(dX)}{\mathcal{Q}(dX)} > \gamma \right\}.$$

Now, let $\mathcal{B}^{(n)} \subseteq \mathcal{X}^n$ be the acceptance region of an *arbitrary* level-ϵ test of $\tilde{\mathcal{H}}_0$ against $\tilde{\mathcal{H}}_1$, with type-II error exponent being E_n. Plugging into (8) and (9), we have

$$\prod_{i=1}^K \boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\mathcal{B}^{(n)}\right\} \geq 1-\epsilon, \text{ and } \prod_{i=1}^K P_{1;i}^{\otimes n_i}\left\{\mathcal{B}^{(n)}\right\} \leq 2^{-nE_n}.$$

For notational convenience, let $\mathcal{P}_n\{dx^n\}$ and $\mathcal{Q}_n\{dx^n\}$ be the measures in (8) and (9) respectively. By Lemma 4.2, we have for any n,

$$nE_n \leq \gamma_n + \log\left(\frac{1}{1-\epsilon-\mathcal{P}_n\left(\log\frac{\mathcal{P}_n(dX^n)}{\mathcal{Q}_n(dX^n)} > \gamma_n\right)} \right). \quad (10)$$

Part 3 (Single-letterization): If γ_n is chosen such that

$$\begin{aligned} &\mathcal{P}_n\left\{ \log\frac{\mathcal{P}_n(dX^n)}{\mathcal{Q}_n(dX^n)} > \gamma_n \right\} \\ &= \mathcal{P}_n\left\{ \sum_{i=1}^K \log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P_{1;i}^{\otimes n_i}(\boldsymbol{x}_{\mathcal{I}_i})} > \gamma_n \right\} \to 0 \end{aligned} \quad (11)$$

as $n \to \infty$, then (10) tells us $\liminf_{n\to\infty} \gamma_n/n$ is an upper bound of the type-II error exponent.

The following proposition is the key to the single-letterization step.

Proposition 4.1: For any $\delta > 0$, choosing γ_n as below satisfies the requirement of (11) (recall that $M_0(\boldsymbol{b}_i) \triangleq \boldsymbol{b}_i^\intercal \boldsymbol{P}_0$):

$$\gamma_n = n\left(\sum_{i=1}^{K} \alpha_i D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right) \tag{12}$$

and hence $\overline{E}^*(\epsilon, \boldsymbol{\alpha}) \le \sum_{i=1}^{K} \alpha_i D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right)$.

Proof of Proposition 4.1: To prove the proposition, we begin with analyzing (11). Observe that

$$\left\{\sum_{i=1}^{K} \log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > n\left(\sum_{i=1}^{K} \alpha_i D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right\}$$
$$\subseteq \left\{\bigcup_{i=1}^{K}\left(\log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > n\alpha_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right)\right\}.$$

Then, applying union bound, we obtain

$$\mathcal{P}_n\left\{\sum_{i=1}^{K} \log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > \sum_{i=1}^{K} n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right\}$$
$$\le \sum_{i=1}^{K} \mathcal{P}_n\left\{\log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right\}.$$

Our goal is to show that for all $\mathcal{I}_i$ and $\delta > 0$,

$$\mathcal{P}_n\left\{\log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right\} \to 0 \tag{13}$$

as $n \to \infty$. Note that $\mathcal{P}_n$ is mixture of probability measures defined on $\mathcal{X}^n$, but the event

$$\left\{\log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right\}$$

only depends on $\boldsymbol{x}_{\mathcal{I}_i} = \{x_j \mid j \in \mathcal{I}_i\}$. According to Lemma 4.1, $\mathcal{P}_n$ can be decomposed into a product measure with respect to $\mathcal{I}_{i'}$, that is,

$$\mathcal{P}_n = \sum_{\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}} w_{\sigma'}\left(\prod_{j=1}^{n} P_{0;\sigma'}(x_j)\right) = \prod_{i=1}^{K} \boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i}),$$

and hence after marginalized over all $\mathcal{I}_{i'}$, $i' \ne i$, (13) can be written as

$$\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right\}.$$

With the above manipulations, to show that (13) holds, we prove the following key lemma. With this lemma, (13) converges to 0, and thus (11) holds.

Lemma 4.3: Let $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})$ be the mixture defined on $\boldsymbol{x}_{\mathcal{I}_i}$ as described before, and $P^{\otimes n_i}_{1;i}$ be the i.i.d. extension of single-letter distribution $P_{1;i}$. Then as $n \to \infty$,

$$\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P^{\otimes n_i}_{1;i}(\boldsymbol{x}_{\mathcal{I}_i})} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right) + \delta\right)\right\} \to 0.$$

The proof of this lemma involves typical sequences and Chebyshev inequality. The details can be found in Appendix B of the full version. ■

Notice that we only require that $\boldsymbol{b}_i$ be a valid mixture vector, that is, $\mathbf{B}\boldsymbol{\alpha} = \boldsymbol{\alpha}$ and $(\mathbf{B})_{i,j} \ge 0$ for all $i, j \in \{1, ..., K\}$. Hence the upper bound can be chosen as the minimum over all feasible $\boldsymbol{B}$, which establishes Theorem 3.2.

V. Conclusion

In this paper, we explore the heterogeneous distributed detection problem with sensor anonymity. To address sensor anonymity, a composite hypothesis testing approach is taken. Focusing on the Neyman-Pearson setting, we prove non-trivial upper and lower bounds of the worst-case type-II error exponent. Unlike the settings considered in robust hypothesis testing literatures [4]–[6], since the hypothesis classes considereds in our framework is discrete, the least favorable distribution might not exist. To circumvent the difficulty, for achievability we propose a type-based test similar to the Hoeffding test [7]. For the converse, we relax the composite testing problem to a tailed simple testing problem and prove the single-letterization of the error exponent lower bound using typical sequences and Chebyshev inequality.

In the follow-up work [9], we further specified the optimal test, termed *mixture likelihood ratio test*, which is a randomized threshold test based on the ratio of the uniform mixture of all the possible distributions under one hypothesis to that under the other hypothesis. Moreover, we showed that the achievability bound (2) in this version is indeed tight, closing the gap between the upper and lower bounds.

Acknowledgement

This work was supported by Ministry of Science and Technology of Taiwan under Grant MOST 107-2634-F-002-009.

References

[1] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, H. V. Poor and J. B. Thomas, Eds. JAI Press Inc., 1990, vol. 2.
[2] ——, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals and Systems*, vol. 1, no. 2, pp. 167–182, 1988.
[3] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 16–29, January 2009.
[4] P. J. Huber, "A robust version of the probability ratio test," *Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.
[5] P. J. Huber and V. Strassen, "Minimax tests and the neyman-pearson lemma for capacities," *Annals of Statistics*, vol. 1, no. 2, pp. 251–263, 1973.
[6] V. V. Veeravalli, T. Başar, and H. V. Poor, "Minimax robust decentralized detection," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 35–40, January 1994.
[7] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 369–401, 1965.
[8] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," August 2017. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf
[9] W.-N. Chen and I.-H. Wang, "Anonymous heterogeneous distributed detection: Optimal decision rules, error exponents, and the price of anonymity," *arXiv: 1805.03554*, May 2018.

Appendix A
Proof of Achievability

Let us introduce some notations.

- First, recall that
$$M_\theta(\boldsymbol{\alpha}) \triangleq \sum_{i=1}^{K} \alpha_k P_{\theta;k}, \text{ where } \boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K]^\intercal$$
$$(n_1, ..., n_K) \triangleq (n\alpha_1, ..., n\alpha_K).$$
- For a sequence $x^n \in \mathcal{X}^n$, where $\mathcal{X} = \{a_1, a_2, ..., a_d\}$, its type (empirical distribution) is defined as
$$\Pi_{x^n} = [\pi(a_1|x^n), \pi(a_2|x^n), ..., \pi(a_d|x^n)],$$
where $\pi(a_i|x^n)$ is the frequency of a_i in the sequence x^n, that is,
$$\pi(a_i|x^n) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{x_j=a_i\}}.$$
- For a given length n, we use $\mathcal{P}_n$ to denote the collection of possible "n-type" of length-n sequences. In other words,
$$\mathcal{P}_n \triangleq \left\{ \left[\frac{i_1}{n}, \frac{i_2}{n}, ..., \frac{i_d}{n}\right] \middle| \forall i_1, ..., i_d \in \mathbb{N} \cup \{0\}, i_1 + i_2 + \cdots + i_d = n \right\}.$$
- Let $U \in \mathcal{P}_n$ be an n-type. The type class $T_n(U)$ is the set of all length-n sequences with type U,
$$T_n(U) \triangleq \{x^n \in \mathcal{X}^n \mid \Pi_{x^n} = U\}.$$

Before proving the achievable bound, let us consider the following lemma.

Lemma A.1: $\forall \epsilon_n > 0$ and $\forall \sigma \in \mathcal{S}_{n,\boldsymbol{\alpha}}$,
$$\mathbb{P}_{0;\sigma}\{X^n : D(\Pi_{X^n} \| M_0(\boldsymbol{\alpha})) > \epsilon_n\} < 2^{-n\left(\epsilon_n - \frac{|\mathcal{X}|}{n}(\log(n_1+1)+\cdots+\log(n_K+1))\right)}$$

Lemma A.2: Let $R \in \mathcal{P}_n$ and define the region $\mathcal{A}_R^{(n)} \triangleq \{x^n : D(\Pi_{x^n} \| R) \le \epsilon_n\} \subseteq \mathcal{X}^n$. Then for all $\sigma \in \mathcal{S}_{n,\boldsymbol{\alpha}}$,
$$\mathbb{P}_{1;\sigma}\left\{\mathcal{A}_R^{(n)}\right\} \le \left(\prod_{k=1}^{K} |\mathcal{P}_{n_k}|\right) 2^{-nD_n^*} = 2^{-n(D_n^* + o(1))},$$
$$\text{where } D_n^* = \min_{\substack{\boldsymbol{U} \in \mathcal{P}_{n_1} \times \cdots \times \mathcal{P}_{n_k} \\ D(\boldsymbol{\alpha}^\intercal \boldsymbol{U} \| R) \le \epsilon_n}} \alpha_1 D(U_1 \| P_{1;1}) + \cdots + \alpha_K D(U_K \| P_{1;K}).$$

Note that here we use $\boldsymbol{\alpha}^\intercal \boldsymbol{U}$ to denote $\sum_{k=1}^{K} \alpha_k U_k$.
Moreover, if $\lim_{n\to\infty} \epsilon_n \to 0$, then we have
$$\lim_{n\to\infty} D_n^* = \min_{\substack{\boldsymbol{U} \in (\mathcal{P}_\mathcal{X})^K \\ \boldsymbol{\alpha}^\intercal \boldsymbol{U} = R}} \alpha_1 D(U_1 \| P_{1;1}) + \cdots + \alpha_K D(U_K \| P_{1;K}). \tag{14}$$

Proof of Thoerem 3.1: The acceptance region with respect to test ϕ is $\mathcal{A}_\phi^{(n)} = \{x^n \mid D(\Pi_{x^n} \| M_0(\boldsymbol{\alpha})) \le \epsilon_n\}$. Hence applying lemma A.1 with
$$\epsilon_n = \frac{|\mathcal{X}|}{n}(\log(n_1+1) + ... + \log(n_K+1)) + \log\left(\frac{1}{\epsilon}\right)/n,$$
we see that for any $\sigma \in \mathcal{S}_{n,\boldsymbol{\alpha}}$,
$$\begin{aligned}\mathbb{P}_{0;\sigma}\left\{\left(\mathcal{A}_\phi^{(n)}\right)^c\right\} &= \mathbb{P}_{0;\sigma}\{X^n : D(\Pi_{x^n} \| M_0(\boldsymbol{\alpha})) > \epsilon_n\} \\ &< 2^{-n\left(\epsilon_n - \frac{|\mathcal{X}|}{n}\left(\sum_{k=1}^{K}\log(n_k+1)\right)\right)} \\ &= 2^{-\log\left(\frac{1}{\epsilon}\right)} = \epsilon,\end{aligned}$$
and thus
$$\mathsf{P_F}^{(n)}(\phi) \le \epsilon.$$

Second, observe that $\epsilon_n \to 0$ as $n \to \infty$. Applying lemma A.2 and plugging in $R = M_0(\boldsymbol{\alpha})$, we obtain

$$\begin{aligned} \underline{E}^*(\epsilon, \boldsymbol{\alpha}) \geq & \min_{\boldsymbol{U} \in (\mathcal{P}_{\mathcal{X}})^K} \sum\nolimits_{k=1}^{K} \alpha_k D\left(U_k \,\|P_{1;k}\right) \\ & \text{subject to} \quad \boldsymbol{\alpha}^\intercal \boldsymbol{U} = \boldsymbol{\alpha}^\intercal \boldsymbol{P}_0 \end{aligned} \tag{15}$$

which completes the proof. ■

Appendix B
Proof of Technical Lemmas

Lemma 4.1: If we choose the weights $w_{\sigma'}$ as in equation (7), then

$$\begin{aligned} & \sum_{\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}} w_{\sigma'} \left(\prod\nolimits_{j=1}^{n} P_{0;\sigma'(j)}(x_j) \right) \\ & = \prod_{i=1}^{K} \left(\sum_{\sigma_i' \in \mathcal{S}_{n_i,\boldsymbol{b}_i}} w_{\sigma_i'} \prod_{j \in \mathcal{I}_i} P_{0;\sigma_i'(j)}(x_j) \right) \triangleq \prod_{i=1}^{K} \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}_{\mathcal{I}_i}) \end{aligned}$$

where $\sigma_i' : \mathcal{I}_i \to \{1, ..., K\}$ is the labeling σ' restricted on $\mathcal{I}_i$. The weights $w_{\sigma_i'}$ satisfy the following equation:

$$w_{\sigma_i'} = \left(\binom{n_i}{n_{i1}, ..., n_{iK}} \right)^{-1}, \text{ if for all } j, \left| \sigma_i'^{-1}(j) \right| = n_{ij}$$

and zero otherwise.

proof of Lemma 4.1: Recall that

$$w_{\sigma'} = \begin{cases} \frac{1}{\binom{n_1}{n_{11}, ..., n_{1K}} \cdot \binom{n_2}{n_{21}, ..., n_{2K}} \cdots \binom{n_K}{n_{K1}, ..., n_{KK}}}, \text{ if } \left| \mathcal{I}_i \cap \mathcal{I}_j' \right| = n_{ij} \\ 0, \text{ else} \end{cases}$$

and

$$w_{\sigma_i'} = \begin{cases} \frac{1}{\binom{n_i}{n_{i1}, ..., n_{iK}}}, \text{ if } \left| \left(\sigma_i'\right)^{-1}(j) \right| = n_{ij} \\ 0, \text{ else} \end{cases}.$$

With a slight abuse of notation, let σ_i' be the mapping σ' restrict on $\mathcal{I}_i$, that is,

$$\sigma_i' : \mathcal{I}_i \to \{1, ..., K\}, \, \sigma_i(j) = \sigma(j), \, \forall j \in \mathcal{I}_i.$$

Then we have

$$w_\sigma = \prod_{i=1}^{K} w_{\sigma_i'},$$

and

$$\begin{aligned} & \sum_{\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}} w_{\sigma'} \left(\prod_{j=1}^{n} P_{0;\sigma'(j)}(x_j) \right) \\ & = \sum_{\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}} \left(\prod_{i=1}^{K} w_{\sigma_i'} \right) \left(\prod_{i=1}^{K} \prod_{j \in \mathcal{I}_i} P_{0;\sigma'(j)}(x_j) \right) \\ & = \sum_{\sigma' \in \mathcal{S}_{n,\boldsymbol{\alpha}}} \left(\prod_{i=1}^{K} w_{\sigma_i'} \prod_{j \in \mathcal{I}_i} P_{0;\sigma'(j)}(x_j) \right) \\ & = \left(\sum_{\sigma_1' \in \mathcal{S}_{n_1,\boldsymbol{b}_1}} \cdots \sum_{\sigma_K' \in \mathcal{S}_{n_K,\boldsymbol{b}_K}} \right) \left(\prod_{i=1}^{K} w_{\sigma_i'} \prod_{j \in \mathcal{I}_i} P_{0;\sigma'(j)}(x_j) \right) \\ & = \prod_{i=1}^{K} \left(\sum_{\sigma_i' \in \mathcal{S}_{n_i,\boldsymbol{b}_i}} w_{\sigma_i'} \prod_{j \in \mathcal{I}_i} P_{0;\sigma_i'(j)}(x_j) \right). \end{aligned}$$

■

Lemma 4.3: Let $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})$ be the mixture defined on $\boldsymbol{x}_{\mathcal{I}_i}$ as described before, and $P_{1;i}^{\otimes n_i}$ be the i.i.d. extension of single-letter distribution $P_{1;i}$. Then as $n \to \infty$,

$$\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\log \frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i})}{P_{1;i}^{\otimes n_i}(\boldsymbol{x}_{\mathcal{I}_i})} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right)+\delta\right)\right\} \to 0.$$

proof of Lemma 4.3: Recall that

$$\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i}) \triangleq \left(\sum_{\sigma_i' \in \mathcal{S}_{n_i,\boldsymbol{b}_i}} w_{\sigma_i'} \prod_{j\in\mathcal{I}_i} P_{0;\sigma_i'(j)}(x_j)\right)$$

where

$$w_{\sigma_i'} = \begin{cases} \frac{1}{\binom{n_i}{n_{i1},\dots,n_{iK}}}, \text{ if } \left|\left(\sigma_i'\right)^{-1}(j)\right| = n_{ij} \\ 0, \text{ else.} \end{cases}$$

WLOG, we assume $\mathcal{I}_i = \{1,...,n_i\}$, and omit the subscript $\mathcal{I}_i$ for convenience. Also, we use $\mathbb{P}_{0;\sigma_i'}$ to denote the product measure defined on $\mathcal{I}_i$:

$$\mathbb{P}_{0;\sigma_i'} \triangleq \prod_{j\in\mathcal{I}_i} P_{0;\sigma_i'(j)}(x_j)$$

(therefore $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}(\boldsymbol{x}_{\mathcal{I}_i}) = \sum_{\sigma_i' \in \mathcal{S}_{n_i,\boldsymbol{b}_i}} w_{\sigma_i'}\mathbb{P}_{0;\sigma_i'}$.) Now, (4.3) can be written as

$$\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\log\frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left(X^{n_i}\right)}{\left(P_{1;i}^{\otimes n_i}\right)\left(X^{n_i}\right)} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right)+\delta\right)\right\} \tag{16}$$

$$=\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\log\frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left(X^{n_i}\right)}{\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}\left(X^{n_i}\right)} + \log\frac{\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}\left(X^{n_i}\right)}{\left(P_{1;i}^{\otimes n_i}\right)\left(X^{n_i}\right)} > n_i\left(\frac{\delta}{2}\right) + n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right)+\frac{\delta}{2}\right)\right\} \tag{17}$$

$$\le\underbrace{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\log\frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left(X^{n_i}\right)}{\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}\left(X^{n_i}\right)} > n_i\left(\frac{\delta}{2}\right)\right\}}_{(1)} + \underbrace{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\log\frac{\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}\left(X^{n_i}\right)}{\left(P_{1;i}^{\otimes n_i}\right)\left(X^{n_i}\right)} > n_i\left(D\left(M_0(\boldsymbol{b}_i)\,\|P_{1;i}\right)+\frac{\delta}{2}\right)\right\}}_{(2)}. \tag{18}$$

We claim that both (1) and (2) converge to 0 as $n \to \infty$. Before continuing the rest of the proof, we first sketch our strategies of bounding equation (18).

Bounding term (1) : The idea is constructing a high probability set, actually exact the typical set, and show that every element in this high probability set satisfies the inequality. The key observation here is that the typical set with respect to i.i.d. measure $\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}$ is also a high probability set under the measure $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}$.

- First, find a typical set $\mathcal{T}_{\delta'}^{(n_i)}\left(\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}\right)$, such that

$$\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\mathcal{T}_{\delta'}^{(n_i)}\right\} \to 1.$$

- Then we show that with properly choose δ', we have the following fact:

$$\forall x^{n_i} \in \mathcal{T}_{\delta'}^{(n_i)},\ \log\frac{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}}{\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}}\left(x^{n_i}\right) \le \left(n_i\frac{\delta}{2}\right),$$

  and hence concludes the vanishing probability of (1) as $n \to \infty$.

Bounding term (2) :

- Observe that the second part of (18) can be rewritten as

$$\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i} f(X_i) > \mathbb{E}_{\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}}\left[f(X_1)\right] + \frac{\delta}{2}\right\},\ \text{where } f(\cdot) \text{ is the LLR.}$$

  However, since $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}$ is not product measure, we cannot apply law of large number directly.
- Fortunately, leveraging the fact that $\boldsymbol{P}^{\mathrm{m}}_{0,\boldsymbol{b}_i}$ is *close* to i.i.d. distribution $\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}$, we are able to show a variant of weak law of large number, and thus prove (2) is vanishing.

Part 1 (Bounding term (1)): Consider the δ'-typical set under $\left(M_0(\boldsymbol{b_i})\right)^{\otimes n_i}$ defined as below:

$$\mathcal{T}_{\delta'}^{(n_i)} \triangleq \left\{\boldsymbol{x} \in \mathcal{X}^{n_i} \mid \left|\pi(a_\ell|\boldsymbol{x}) - M_0(\boldsymbol{b_i})(a_\ell)\right| \le \delta' M_0(\boldsymbol{b_i})(a_\ell)\right\}.$$

Fact B.1: According to the AEP, we have

$$\forall \boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)},\ 2^{-n_i H(M_0(\boldsymbol{b_i}))(1+\delta')} \leq (M_0(\boldsymbol{b_i}))^{\otimes n_i}(\boldsymbol{x}) \leq 2^{-n_i H(M_0(\boldsymbol{b_i})(1-\delta')}$$

Fact B.2: The cardinality bounds gives us

$$\left|\mathcal{T}_{\delta'}^{(n_i)}\right| < 2^{n_i H(M_0(\boldsymbol{b_i}))(1+\delta')}$$

Then, according to Lemma A.1, $\mathbb{P}_{0;\sigma_i'}\{\mathcal{T}_{\delta'}^{(n_i)}\} \to 1$ for all $\delta' > 0$ and $k \in \{1, ..., K\}$. Thus we have the following fact:
Fact B.3:

$$\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}\{\mathcal{T}_{\delta'}^{(n_i)}\} = \sum_{\sigma_i'} w_{\sigma_i'} \mathbb{P}_{0;\sigma_i'}\{\mathcal{T}_{\delta'}^{(n_i)}\} \to 1,\ \text{as } n_i \to \infty.$$

That is, for any $\gamma > 0$,

$$\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}\{\mathcal{T}_{\delta'}^{(n_i)}\} > 1 - \gamma,$$

as n_i large enough.

Second, we show that under $\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}$, AEP also holds. Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{T}_{\delta'}^{(n_i)}$, then one can always find a permutation $\tau : \{1, ..., n_i\} \to \{1, ..., n_i\}$, such that

$$d_H(\boldsymbol{x}, \boldsymbol{x}_\tau') \leq 2 n_i \delta',\ \text{where } \boldsymbol{x}_\tau' \triangleq (x_{\tau(1)}', ..., x_{\tau(n_i)}').$$

The is because $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{T}_{\delta'}^{(n_i)}$ implies

$$|\pi(a_\ell|\boldsymbol{x}) - \pi(a_\ell|\boldsymbol{x}')| \leq |\pi(a_\ell|\boldsymbol{x}) - M_0(\boldsymbol{b_i})(a_\ell)| + |\pi(a_\ell|\boldsymbol{x}') - M_0(\boldsymbol{b_i})(a_\ell)| \leq 2\delta' M_0(\boldsymbol{b_i})(a_\ell),$$

so there exists a permutation τ such that

$$d_H(\boldsymbol{x}, \boldsymbol{x}_\tau') \leq n_i \sum_{\ell=1}^{d} |\pi(a_\ell|\boldsymbol{x}) - \pi(a_\ell|\boldsymbol{x}')| \leq n_i \sum_{\ell=1}^{d} 2\delta' M_0(\boldsymbol{b_i})(a_\ell) = 2 n_i \delta'.$$

Observe that

$$\frac{\mathbb{P}_{0;\sigma}(\boldsymbol{x})}{\mathbb{P}_{0;\sigma}(\boldsymbol{x}_\tau')} = \frac{P_{0;\sigma(1)}(x_1) \cdot P_{0;\sigma(2)}(x_2) \cdots P_{0;\sigma(K)}(x_K)}{P_{0;\sigma(1)}(x_{\tau(1)}) \cdot P_{0;\sigma(2)}(x_{\tau(2)}) \cdots P_{0;\sigma(K)}(x_{\tau(K)})} \leq \left(\max_{i,j,\ell,m} \frac{P_{0;i}(a_\ell)}{P_{0;j}(a_m)} \right)^{2 n_i \delta'} \triangleq \Delta^{n_i \delta'} \tag{19}$$

Note that $\Delta \geq 1$. Now we claim the same bound also holds on the mixture measure $\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}} = \sum_{\sigma_i'} w_{\sigma_i'} \mathbb{P}_{0;\sigma_i'}$. The reason is that because the weights $w_{\sigma_i'}$ we construct is uniform, it is *permutation invariant*. That is, for all permutation τ,

$$\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}_\tau) = \sum_{\sigma_i'} w_{\sigma_i'} \mathbb{P}_{0;\sigma_i'}(\boldsymbol{x}_\tau) = \sum_{\sigma_i'} w_{\sigma_i' \circ \tau^{-1}} \mathbb{P}_{0;\sigma_i' \circ \tau^{-1}}(\boldsymbol{x}) = \sum_{\sigma_i'} w_{\sigma_i'} \mathbb{P}_{0;\sigma_i'}(\boldsymbol{x}) = \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}), \tag{20}$$

where the third equality holds due to the mixture measure is permutation invariant. Therefore, according to (19) and (20), we have for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{T}_{\delta'}^{(n_i)}$,

$$\frac{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x})}{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}')} = \frac{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x})}{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}_\tau')} = \frac{\sum_{\sigma_i'} w_{\sigma_i'} \mathbb{P}_{0;\sigma_i'}(\boldsymbol{x})}{\sum_{\sigma_i'} w_{\sigma_i'} \mathbb{P}_{0;\sigma_i'}(\boldsymbol{x}_\tau')} \leq \Delta^{n_i \delta'}.$$

Similar results hold for the lower bound $\Delta^{-n_i \delta'}$. Again, we write it as the following fact:
Fact B.4: For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{T}_{\delta'}^{(n_i)}$

$$\Delta^{-n_i \delta'} \leq \frac{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x})}{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}')} \leq \Delta^{n_i \delta'}$$

According to the cardinality bound (fact B.2) and the probability bound (fact B.3), we obtain for any $\gamma > 0$,

$$1 - \gamma < \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}\{\mathcal{T}_{\delta'}^{(n_i)}\} = \sum_{\boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)}} \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}) \leq \left|\mathcal{T}_{\delta'}^{(n_i)}\right| \min_{\boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)}} \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}) \Delta^{n_i \delta'} \leq 2^{n_i H(M_0(\boldsymbol{b_i}))(1+\delta')} \min_{\boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)}} \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}) \Delta^{n_i \delta'}$$

Rearrange the equation, we obtain

$$\min_{\boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)}} \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}) \geq 2^{-n_i H(M_0(\boldsymbol{b_i}))(1+\delta'+\delta' \log \Delta + o(1))}. \tag{21}$$

Similar trick can be used to give an upper bound on $\max_{\boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)}} \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x})$.

Fact B.5: for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{T}_{\delta'}^{(n_i)}$,

$$2^{-n_i H(M_0(\boldsymbol{b_i}))(1+\delta'+\delta' \log \Delta+o(1))} \leq \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}(\boldsymbol{x}) \leq 2^{-n_i H(M_0(\boldsymbol{b_i}))(1-\delta'-\delta' \log \Delta+o(1))}$$

Finally, according to Fact B.1 and Fact B.5, we see that

$$\forall \boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)}, \log \frac{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}}{(M_0(\boldsymbol{b_i}))^{\otimes n_i}}(\boldsymbol{x}) \leq n_i \delta'(2+\log \Delta)+o(n).$$

Choosing $\delta' < \frac{\delta}{4+2\log \Delta}$, we obtain that

$$\forall \boldsymbol{x} \in \mathcal{T}_{\delta'}^{(n_i)}, \log \frac{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}}{(M_0(\boldsymbol{b_i}))^{\otimes n_i}}(\boldsymbol{x}) \leq n_i\left(\frac{\delta}{2}\right),$$

for n_i large enough. Also, for n_i large enough, $\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}\{\mathcal{T}_{\delta'}^{(n_i)}\} \to 1$; hence we conclude

$$\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}\left\{\log \frac{\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}}{(M_0(\boldsymbol{b_i}))^{\otimes n_i}}(\boldsymbol{x}) > n_i\left(\frac{\delta}{2}\right)\right\} \to 0.$$

Part 2 (Bounding term (2)): In the second part of the proof, our goal is to give a similar result as law of large number under the mixture measure:

$$\frac{1}{n_i} \log \frac{(M_0(\boldsymbol{b_i}))^{\otimes n_i}}{P_{1;i}^{\otimes n_i}}(X^{n_i}) \xrightarrow{\mathcal{P}} D\left(M_0(\boldsymbol{b}_i) \| P_{1;i}\right).$$

Note that the mixed measure $\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}$ has identical marginal distribution $M_0(\boldsymbol{b}_i)$. Rewriting the above equation, it is equivalent to prove

$$\boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i} f\left(X_j\right) > \mathbb{E}_{X\sim M_0(\boldsymbol{b}_i)}\left[f\left(X\right)\right]+\frac{\delta}{2}\right\} \to 0. \tag{22}$$

where we denote $\log \frac{(M_0(\boldsymbol{b_i}))}{P_{1;i}}(\cdot)$ as $f(\cdot)$ for simplicity. Although $X^{n_i} \sim \boldsymbol{P}_{0,\boldsymbol{b}_i}^{\mathrm{m}}$ is not i.i.d. measure, we can still apply Chebyshev's inequality on (22) and obtain a similar result. Hence it suffices to show that $\mathsf{Var}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} f(X_j)\right) \to 0$. Notice that

$$\mathsf{Var}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} f(X_j)\right) = \frac{1}{n_i}\mathsf{Var}_{X\sim M_0(\boldsymbol{b}_i)}\left(f(X)\right)+\frac{1}{n_i^2}\sum_{\ell,m\in\{1,\dots,K\}}\mathsf{Cov}\left(f(X_\ell), f(X_m)\right),$$

and thus we only need to show that for all $j,k \in \{1,\dots,n_i\}$ such that $j \neq k$, $\mathsf{Cov}\left(f(X_j), f(X_k)\right) \to 0$ as $n \to \infty$. This statement is intuitively true, since the pairwise density of X_j, X_k converges to independent distribution, and hence the covariance must converge to 0. In the below we give a proof. First, note that the single-letter marginal distribution

$$P_{X_i} = P_{X_j} = M_0(\boldsymbol{b_i}).$$

With a slight abuse of notation, we omit the subscript i, denoting as $M_0(\boldsymbol{b}) = (b_1 P_{0;1}+\dots+b_K P_{0;K})$. Now, consider the joint density between X_i, X_j are

$$\begin{aligned}
&\sum_{x_l, l\neq j,k} \boldsymbol{P}_{0,\boldsymbol{b}}^{\mathrm{m}}(x_1,\dots,x_{n_i}) = \sum_{x_l, l\neq j,k}\sum_{\sigma} w_\sigma \mathbb{P}_{0;\sigma}(x_1,\dots,x_{n_i})\\
&= \sum_{\ell,m\in\{1,\dots,K\},\ell\neq m} \frac{\binom{n_i-2}{n_{i1},\dots,n_{i\ell}-1,\dots,n_{im}-1,\dots,n_{ik}}}{\binom{n_i}{n_{i1},\dots,n_{iK}}} P_{0;\ell}(X_j)P_{0;m}(X_k) + \sum_{\ell\in\{1,\dots,K\}} \frac{\binom{n_i-2}{n_{i1},\dots,n_{i\ell}-2,\dots,n_{ik}}}{\binom{n_i}{n_{i1},\dots,n_{iK}}} P_{0;\ell}(X_j)P_{0;\ell}(X_k)
\end{aligned}$$

Since we have $n_{ij} = n_i b_j$, rewritting the above equation, we obtain that the joint density between X_j, X_k is

$$\begin{aligned}
P(X_j, X_k) &= \sum_{\ell,m\in\{1,\dots,K\},\ell\neq m} \frac{n_\ell n_m}{n_i(n_i-1)} P_{0;\ell}(X_j)P_{0;m}(X_k)\\
&= \sum_{\ell,m\in\{1,\dots,K\},\ell\neq m} \frac{n_i}{n_i-1} b_\ell b_m P_{0;\ell}(X_j)P_{0;m}(X_k) + \sum_{\ell\in\{1,\dots,K\}: b_\ell\neq 0} \frac{n_i-1/b_\ell}{n_i-1} b_\ell^2 P_{0;\ell}(X_j)P_{0;m}(X_k).
\end{aligned}$$

Besides, the product of marginal distibution of X_j, X_k is

$$P(X_j)P(X_k) = M_0(\boldsymbol{b})(X_j)M_0(\boldsymbol{b})(X_k) = \sum_{\ell,m} b_\ell b_m P_{0;\ell}(X_j)P_{0;m}(X_k).$$

Bound the covariance as below:

$$\begin{aligned}
|\mathsf{Cov}\,(f(X_j), f(X_k))| &= |\mathbb{E}f(X_j)f(X_k) - \mathbb{E}f(X_j)\mathbb{E}f(X_k)| \\
&= \left| \sum_{X_j, X_k} \left[(P(X_j, X_k) - P(X_j)P(X_k))\, f(X_j) f(X_k) \right] \right| \\
&= \left| \sum_{X_j, X_k} \left(\sum_{\ell \neq m} \frac{1}{n_i - 1} b_\ell b_m P(X_j)P(X_k) + \sum_{\ell : b_\ell \neq 0} \frac{1 - 1/b_\ell}{n_i - 1} b_\ell^2 P(X_j)P(X_k) \right) f(X_j) f(X_k) \right| \\
&\leq 2 \left| \frac{\max_{\ell : b_\ell \neq 0} \frac{1}{b_\ell} + 1}{n_i - 1} \right| \sum_{\ell, m} b_\ell b_m P(X_j)P(X_k) f(X_j) f(X_k) \\
&= 2 \left| \frac{\max_{\ell : b_\ell \neq 0} \frac{1}{b_\ell} + 1}{n_i - 1} \right| \left(\mathbb{E}_{X \sim M_0(\boldsymbol{b})} f(X) \right)^2 = 2 \left| \frac{\max_{\ell : b_\ell \neq 0} \frac{1}{b_\ell} + 1}{n_i - 1} \right| D\left(M_0(\boldsymbol{b}_i) \,\|P_{1;i} \right)^2 \to 0,
\end{aligned}$$

as $n_i \to \infty$. Applying Chebyshev's inequality, we establish (22), and thus concludes the proof of part (2). ∎

Lemma A.1: $\forall \epsilon_n > 0$ and $\forall \sigma \in \mathcal{S}_{n,\boldsymbol{\alpha}}$,

$$\mathbb{P}_{0;\sigma}\left\{ X^n : D\left(\Pi_{X^n} \,\|M_0\left(\boldsymbol{\alpha}\right)\right) > \epsilon_n \right\} < 2^{-n\left(\epsilon_n - \frac{|\mathcal{X}|}{n}(\log(n_1+1) + \cdots + \log(n_K+1))\right)}$$

Proof: Recall that we denote $n\alpha_k$ as n_k.

$$\begin{aligned}
&\mathbb{P}_{0;\sigma}\left\{ X^n : D\left(\Pi_{X^n} \,\|M_0\left(\boldsymbol{\alpha}\right)\right) > \epsilon_n \right\} \\
&= \sum_{V \in \mathcal{P}_n : D(V \,\|M_0(\boldsymbol{\alpha})) > \epsilon_n} \mathbb{P}_{0;\sigma}\left\{ T_n(V) \right\} \\
&= \sum_{\substack{(U_1, \ldots, U_K) \in \mathcal{P}_{n_1} \times \cdots \times \mathcal{P}_{n_K} : \\ D\left(\sum_k \alpha_k U_k \,\middle\|M_0(\boldsymbol{\alpha})\right) > \epsilon_n}} \prod_{k=1}^{K} P_{0;k}^{\otimes n_k}\left\{ T_{n_k}(U_k) \right\} \\
&\leq \sum_{\substack{(U_1, \ldots, U_K) \in \mathcal{P}_{n_1} \times \cdots \times \mathcal{P}_{n_K} : \\ D\left(\sum_k \alpha_k U_k \,\middle\|M_0(\boldsymbol{\alpha})\right) > \epsilon_n}} \prod_{k=1}^{K} 2^{-n_k D(U_k \,\|P_{0;k})} \\
&= \sum_{\substack{(U_1, \ldots, U_K) \in \mathcal{P}_{n_1} \times \cdots \times \mathcal{P}_{n_K} : \\ D\left(\sum_k \alpha_k U_k \,\middle\|M_0(\boldsymbol{\alpha})\right) > \epsilon_n}} 2^{-n \sum_k \alpha_k D(U_k \,\|P_{0;k})}
\end{aligned}$$

By the convexity of KL divergence,

$$\sum_k \alpha_k D\left(U_k \,\|P_{0;k}\right) \geq D\left(\sum_k \alpha_k U_k \middle\| \sum_k \alpha_k P_{0;k} \right) = D\left(V \,\|M_0\left(\boldsymbol{\alpha}\right)\right) > \epsilon_n,$$

so we obtain

$$\begin{aligned}
&\mathbb{P}_{0;\sigma}\left\{ X^n : D\left(\Pi_{X^n} \,\|M_0\left(\boldsymbol{\alpha}\right)\right) > \epsilon_n \right\} \\
&< \sum_{\substack{(U_1, \ldots, U_K) \in \mathcal{P}_{n_1} \times \cdots \times \mathcal{P}_{n_K} : \\ D\left(\sum_k \alpha_k U_k \,\middle\|M_0(\boldsymbol{\alpha})\right) > \epsilon_n}} 2^{-n\epsilon_n} \\
&\leq \left(\prod_{k=1}^{K} |\mathcal{P}_{n_k}| \right) 2^{-n\epsilon_n} \\
&\leq 2^{-n\left(\epsilon_n - \frac{|\mathcal{X}|}{n}(\log(n_1+1) + \cdots + \log(n_K+1))\right)}.
\end{aligned}$$

∎

Lemma A.2: Let $R \in \mathcal{P}_n$ and define the region $\mathcal{A}_R^{(n)} \triangleq \left\{ x^n : D\left(\Pi_{x^n} \,\|R\right) \leq \epsilon_n \right\} \subseteq \mathcal{X}^n$. Then for all $\sigma \in \mathcal{S}_{n,\boldsymbol{\alpha}}$,

$$\mathbb{P}_{1;\sigma}\left\{\mathcal{A}_R^{(n)}\right\} \leq \left(\prod_{k=1}^{K} |\mathcal{P}_{n_k}|\right) 2^{-nD_n^*} = 2^{-n(D_n^*+o(1))},$$

$$\text{where } D_n^* = \min_{\substack{\boldsymbol{U}\in\mathcal{P}_{n_1}\times\cdots\times\mathcal{P}_{n_k} \\ D(\boldsymbol{\alpha}^\intercal\boldsymbol{U}\,\|R)\leq\epsilon_n}} \alpha_1 D\left(U_1\,\|P_{1;1}\right) + \cdots + \alpha_K D\left(U_K\,\|P_{1;K}\right).$$

Note that here we use $\boldsymbol{\alpha}^\intercal\boldsymbol{U}$ to denote $\sum_{k=1}^{K}\alpha_k U_k$.
Moreover, if $\lim_{n\to\infty}\epsilon_n \to 0$, then we have

$$\lim_{n\to\infty} D_n^* = \min_{\substack{\boldsymbol{U}\in(\mathcal{P}_{\mathcal{X}})^K \\ \boldsymbol{\alpha}^\intercal\boldsymbol{U}=R}} \alpha_1 D\left(U_1\,\|P_{1;1}\right) + \cdots + \alpha_K D\left(U_K\,\|P_{1;K}\right). \tag{14}$$

Proof:
Part 1: Observe that

$$\begin{aligned}
\mathbb{P}_{1;\sigma}\left\{\mathcal{A}_R^{(n)}\right\} &= \mathbb{P}_{1;\sigma}\left\{X^n : D\left(\Pi_{X^n}\,\|R\right)\leq\epsilon_n\right\} \\
&= \sum_{\substack{U_k\in\mathcal{P}_{n_k}: \\ D\left(\sum_k \alpha_k U_k\,\middle\|R\right)\leq\epsilon_n}} \mathbb{P}_{1;\sigma}\left\{T_n(V)\right\} \\
&= \sum_{\substack{U_k\in\mathcal{P}_{n_k}: \\ D\left(\sum_k \alpha_k U_k\,\middle\|R\right)\leq\epsilon_n}} \prod_{k=1}^{K} P_{1;k}^{\otimes n_k}\left\{T_{n_k}(U_k)\right\} \\
&\leq \sum_{\substack{U_k\in\mathcal{P}_{n_k}: \\ D\left(\sum_k \alpha_k U_k\,\middle\|R\right)\leq\epsilon_n}} 2^{-n\sum_k \alpha_k D(U_k\,\|P_{1;k})} \\
&\leq \sum_{\substack{U_k\in\mathcal{P}_{n_k}: \\ D\left(\sum_k \alpha_k U_k\,\middle\|R\right)\leq\epsilon_n}} 2^{-nD_n^*} \\
&\leq \left(\prod_{k=1}^{K}|\mathcal{P}_k|\right) 2^{-nD_n^*} \leq 2^{-n(D_n^*+o(1))},
\end{aligned}$$

where the last inequality holds since the cardinality of type is sub-linear in n.
Part 2: To show that

$$\lim_{n\to\infty} D_n^* = \min_{\substack{\boldsymbol{U}\in(\mathcal{P}_{\mathcal{X}})^K \\ \boldsymbol{\alpha}^\intercal\boldsymbol{U}=R}} \alpha_1 D\left(U_1\,\|P_{1;1}\right) + \cdots + \alpha_K D\left(U_K\,\|P_{1;K}\right), \tag{23}$$

we observe that

$$\left\{U_k\in\mathcal{P}_{n_k}\mid V=\sum_k \alpha_k U_k,\ D\left(V\,\|R\right)\leq\epsilon_n\right\} \to \left\{U_k\in\mathcal{P}_{\mathcal{X}}\mid R=\sum_k \alpha_k U_k\right\},$$

once $\epsilon_n \to 0$, due to the fact that $D\left(V\,\|R\right)=0 \iff V=R$. More rigorously, define

$$\mathcal{A}^{(n)} \triangleq \left\{\boldsymbol{U}\in\mathcal{P}_{n_1}\times\cdots\times\mathcal{P}_{n_K}\mid D\left(\boldsymbol{\alpha}^\intercal\boldsymbol{U}\,\|R\right)\leq\epsilon_n\right\},$$

$$\text{and } \mathcal{A}^* \triangleq \left\{\boldsymbol{U}\in\mathcal{P}_{\mathcal{X}}^K\mid \boldsymbol{\alpha}^\intercal\boldsymbol{U}=R\right\}.$$

Then we claim that

$$\mathcal{A}^* = \bigcap_{i>0}\text{closure}\Big(\bigcup_{n>i}\mathcal{A}^{(n)}\Big).$$

1) "$\subseteq$" part: $\forall \boldsymbol{U}^*\in\mathcal{A}^*$, one can always find a sequence $\boldsymbol{U}_1,\boldsymbol{U}_2,...,\boldsymbol{U}_j,...$, such that $\{\boldsymbol{U}_j\}\to\boldsymbol{U}^*$, and $\boldsymbol{U}_j\in\mathcal{A}^{(j)}$ (for example, rounding $\boldsymbol{U}^*$). Therefore, $\{\boldsymbol{U}_j\}_{j>i}$ guarantees $\boldsymbol{U}^*\in\text{closure}\left(\bigcup_{n>i}\mathcal{A}^{(n)}\right)$.
2) "$\supseteq$" part: $\forall\tilde{\boldsymbol{U}}\notin\mathcal{A}^*$, $D\left(\boldsymbol{\alpha}^\intercal\tilde{\boldsymbol{U}}\,\middle\|R\right)\triangleq d>0$. Hence there exists M large enough, such that $\tilde{\boldsymbol{U}}\notin\text{closure}\left(\bigcup_{n>M}\mathcal{A}^{(n)}\right)$, by the fact that $\epsilon_n\to 0$ and the continuity of KL-divergence.

For notational convenience, define the function

$$D(\boldsymbol{U}) = D(U_1, ..., U_K) \triangleq \sum_k \alpha_k D\left(U_k \,\|P_{1;k}\right).$$

Therefore, we have

$$D_n^* = \min_{\boldsymbol{U}\in\mathcal{A}^{(n)}} D(\boldsymbol{U}) \geq \min_{\text{closure}\left(\bigcup_{k>n}\mathcal{A}^{(k)}\right)} D(\boldsymbol{U}).$$

This implies

$$\liminf_{n\to\infty} D_n^* \geq \min_{\bigcap_{n>0}\text{closure}\left(\bigcup_{k>n}\mathcal{A}^{(k)}\right)} D(\boldsymbol{U}) = \min_{\mathcal{A}^*} D(\boldsymbol{U}).$$

On the other hand, suppose $\min_{\mathcal{A}^*} D(\boldsymbol{U}) = D(\boldsymbol{U}^*)$, $\boldsymbol{U}^* \in \mathcal{A}^*$ (such $\boldsymbol{U}^*$ always exists since $\mathcal{A}^*$ is a compact region in $\mathbb{R}^{K\times d}$). Then there exists a sequence $\{\boldsymbol{U}_i\} \to \boldsymbol{U}^*$, such that $\boldsymbol{U}_i \in \mathcal{A}^{(i)}$. Hence for i, we see that

$$D_i^* = \min_{\mathcal{A}^{(i)}} D(\boldsymbol{U}) \leq D(\boldsymbol{U}_i),$$

and thus

$$\limsup_{i\to\infty} D_i^* \leq \lim_{i\to\infty} D(\boldsymbol{U}_i) = D(\boldsymbol{U}^*) = \min_{\mathcal{A}^*} D(\boldsymbol{U}),$$

which completes the proof.

■