On the Price of Source Anonymity in Heterogeneous Parametric Point Estimation

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

***Abstract*—Parametric point estimation from anonymous and heterogeneous data is studied. For heterogeneity, we assume n samples are independently drawn, each following one of K possible distributions. For anonymity, we assume the estimator knows the number of samples drawn from each distribution, but which one each sample follows is hidden. In words, samples as a sequence are passed through an unknown permutation prior to being observed. The goal is to find an estimator that minimizes the worst-case statistical risk over all possible permutations. We prove that an optimal estimator depends only on the empirical distribution (type) of samples, and when the risk function is the mean squared error (MSE), it follows a non-trivial Cramer-Rao lower bound. We further characterize its asymptote as $n \to \infty$, assuming the number of samples from each distribution is proportional to n. The lower bound is of the order of $1/n$, and the reciprocal of its prefactor is the Fisher information of the mixture of the K distributions.**

A full version of this paper is accessible at:
http://homepage.ntu.edu.tw/~ihwang/Eprint/isit19ahest.pdf

I. Introduction

Statistical inference is a fundamental task in data science, where a decision maker aims to determine a hidden parameter based on the data it collects, as well as how the data depends on the target parameter statistically. In many modern applications such as crowdsourcing and sensor networks, data is heterogeneous and collected from various sources following different distributions. These sources, however, may be anonymous to the decision maker due to considerations in identification costs and privacy [1]. Since the distribution becomes unknown, it is unclear how to carry out optimal inference, and hence the impact of source anonymity on the performance of statistical inference remains elusive. In our previous work [2], [3], binary hypothesis testing from anonymous and heterogeneous data was investigated. We proved that the empirical distribution (type) is sufficient for hypothesis testing when the sources are anonymous. Under the Neyman-Pearson formulation, we found that anonymity severely decreases the type-II error exponent.

In this work, we study heterogeneous parametric point estimation under source anonymity. There are n independent sources, and each source (say source i, $i \in \{1, ..., n\}$) gives a single random sample X_i. The decision maker estimates the hidden parameter $\theta \in \Theta$ from the collected samples $X_1, ..., X_n$. For heterogeneity of the sources, we assume that they are clustered into K groups, and the k-th group comprises n_k sources. The sample drawn from a source in the k-th group follows distribution $P_{k;\theta}$, $\theta \in \Theta$. For anonymity of the sources, we assume that although the decision maker is fully aware of the *heterogeneity* of samples collected from the data sources, including the set of distributions $\{P_{k;\theta} \mid k = 1, ..., K\}$ and the number of sources in each cluster, $\{n_k \mid k = 1, ..., K\}$, it does not know which distribution each source follows. In words, the sample sequence $X^n \triangleq (X_1, ..., X_n)$ is passed through an unknown permutation π prior to being observed. Our goal is to find an estimator that minimizes the *worst-case* statistical risk over all possible permutations.

Our main contribution is two-fold. In the first part of our contribution, we show that an optimal estimator that minimizes the worst-case statistical risk depends only on the type of the samples X^n, meaning that the *order* of samples, despite that they are drawn from heterogeneous sources, provides no information for inference. Hence, when the loss function is the the squared error loss and the statistical risk is the mean squared error (MSE), the worst-case MSE has a non-trivial Cramer-Rao Lower Bound (CRLB) under suitable regularity conditions, and the reciprocal of this lower bound is the Fisher information (FI) of the distribution of the type of X^n. Due to the data processing inequality, it is not greater than the FI of the product distribution of X^n when the order is known. This motivates us to further investigate how the FI scales asymptotically as the number of samples n grows, so that the asymptotic performance loss due to anonymity can be characterized. In the second part of our contribution, we characterize the *asymptotic Fisher information*. Under the assumption that the proportion of sources in each group, that is, $\frac{n_k}{n} \to \alpha_k$ for each k as $n \to \infty$, we show that the FI of the type of X^n grows linearly with n as $n \to \infty$, and the prefactor turns out to be the FI of the mixture distributions the K distributions with the mixing ratio being $(\alpha_1, ..., \alpha_K)$.

For proving the sufficiency of type in heterogeneous parametric point estimation under source anonymity, we exploit the symmetric structure of the problem and show that for an arbitrary estimator, if one *symmetrizes* it by averaging with respect to all permutations of the samples, the statistical risk cannot increase. For characterizing the asymptotic FI of empirical distribution, we first analyze the *Kullback-Leibler divergence* (KLD) between two empirical distributions, and then by leveraging the fact that KLD behaves locally as FI and taking the second order derivative to KLD, we obtain the asymptotic FI. The rationale behind using such a two-step approach is that, the large deviation analysis on KLD allows us to circumvent the direct computation on FI, the asymptote of which is difficult to analyze since it involves summing the joint distributions over all possible permutations of samples.

Related works

The unknown permutation π in our problem is a *nuisance parameter*, that is, parameters which are not directly of interest but may influence the distribution of samples. There is a rich literature studying how to eliminate nuisance parameters, see, for example, a detailed study in [4]. Among them, our approach aligns with the *marginalization* methods, in which we choose a suitable statistic (the empirical distribution function in our case) so that the resulting distributions no longer depend on π. Interestingly, by leveraging the symmetry of nuisance parameters in our problem, we prove that such marginalization approach achieves optimality under the minimax principle (i.e. worst case π). Inference with an unknown permutation has also been studied in the literature of compressed sensing [5], [6] and linear regression [7], in which the measurements are permuted before being observed. [5], [6] aim to recover the signal (i.e. θ as an d-dim vector) under noiseless assumption and thus the authors mainly focus on designing the measurement matrix. For the linear regression with additive i.i.d. Gaussian noise [7], the unknown permutation π is of interest and sharp conditions are set in order to recover π, which is different from our setting that π is treated as a nuisance parameter.

II. Problem Formulation and Preliminaries

A. Problem formulation

Following the description of the setting in Section I, let us formulate the heterogeneous parametric point estimation problem under source anonymity. Consider n independent heterogeneous sources, numbered from 1 to n, and source i, $i \in \{1, ..., n\}$, gives a single random sample $X_i \in \mathcal{X}$ drawn from one of the K distributions $\{P_{k;\theta} \mid k = 1, ..., K\}$, and the hidden parameter to be estimated is $\theta \in \Theta$. Throughout this paper, we shall assume that the alphabet has finite cardinality $|\mathcal{X}| = d$ and $\mathcal{X} = \{a_1, ..., a_d\}$. Let n_k denote the number of sources that follow distribution $P_{k;\theta}$, for $k = 1, .., K$, and hence $n = \sum_{k=1}^{K} n_k$. The estimator has access to the joint sample X^n and knows $\{n_k \mid k = 1, ..., K\}$ and the parametric family of distributions $\{P_{k;\theta} \mid k = 1, ..., K, \theta \in \Theta\}$. However, it does not know the which distribution each source follows, and we would like to guarantee the worst-case performance over all possibilities of the source distributions.

Hence, without loss of generality, we can equivalently reorder the n samples so that it follows the following product distribution:

$$X^n \sim \mathbb{P}_\theta \triangleq \left(P_{1;\theta}\right)^{\otimes n_1} \left(P_{2;\theta}\right)^{\otimes n_2} \cdots \left(P_{K;\theta}\right)^{\otimes n_K}. \qquad (1)$$

Meanwhile, the joint observation X^n is perturbed by a n-permutation $\pi : \{1, ..., n\} \to \{1, ..., n\}$ which is not revealed to the estimator. Hence, the estimator needs to consider all $n!$ possible permutation of π and estimates the underlying $\theta \in \Theta$.

For notational convenience, we will use $\mathcal{S}_n$ to denote the collection of n-permutation and use Π_{X^n} as the empirical distribution of X^n

$$\Pi_{X^n} \triangleq \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = a_1\}}, ..., \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = a_d\}} \right]^{\mathsf{T}} \in \mathbb{R}^d.$$

With a slight abuse of notation, we sometimes use $\pi \in \mathcal{S}_n$ as a coordinate permutation acting on samples $X^n = (X_1, ..., X_n)$, namely, $\pi\left(X^n\right) \triangleq \left(X_{\pi(1)}, ..., X_{\pi(n)}\right)$.

To evaluate an estimator $\phi\left(X^n\right)$'s performance, let us consider a general convex loss function $\ell : \Theta \times \Theta \to [0, \infty)$ and the corresponding worst-case statistical risk as

$$\mathsf{R}^*_\theta(\phi) \triangleq \max\nolimits_{\pi \in \mathcal{S}_n} \mathbb{E}_{X^n \sim \mathbb{P}_\theta} \left[\ell\left(\theta, \phi(\pi(X^n))\right) \right].$$

When the loss function is the squared error loss, $\mathsf{R}^*_\theta(\phi) \equiv \mathsf{MSE}^*_\theta(\phi)$ denotes the worst-case mean squared error (MSE).

B. Fisher information and Cramer-Rao lower bound

Let us recap some classical results of point estimation, including Fisher information and Cramer-Rao lower bound.

Definition 2.1 (Fisher information): Suppose $p_\theta(x)$ is a parametric model and $\theta \mapsto p_\theta(x)$ is differentiable for all $x \in \mathcal{X}$. Then the Fisher information is defined as

$$I_p(\theta) \triangleq \mathbb{E}_{p_\theta} \left[\left(\tfrac{\partial}{\partial \theta} \log p_\theta(X) \right)^2 \right].$$

Throughout this paper, we assume some regularity conditions on $p_\theta(x)$ (for more details, see [8, Chapter 5] for example), including

1) $p_\theta(x)$ is *twice* differentiable with respect to θ,
2) Θ is an open interval, and
3) the support of $p_\theta(x)$ does not depend on θ,

so that Fisher information has an alternative form:

$$I_p(\theta) = -\mathbb{E}_{p_\theta} \left[\tfrac{\partial^2}{\partial \theta^2} \log p_\theta(X) \right].$$

Remark 2.1: Usually we need another condition to exchange the order of the derivative and the expectation operator, but in this paper we assume the alphabet is finite, and hence we can always swap them.

If the aforementioned conditions are satisfied, then the MSE of any unbiased estimator is lower bounded by the *Cramer-Rao lower bound (CRLB)*, namely, $\mathsf{MSE}_\theta(\hat{\theta}) \geq I_p(\theta)^{-1}$. There is also an generalized version of CRLB for biased estimator [8, Theorem 5.10], in which FI also plays an essential role. It is noteworthy that if neglecting the unknown permutation in our problem, that is, the estimator *knows* it and there is no anonymity, the Fisher information of $\mathbb{P}_\theta$ (as defined in (1)) becomes

$$I_{\mathbb{P}}(\theta) = \sum\nolimits_{k=1}^{K} n_k I_{P_k}(\theta)$$

due to the independence across samples, where $I_{P_k}(\theta)$ is Fisher information of the k-th source $P_{k;\theta}$.

III. Main Results

Our first result is about the characterization of an optimal estimator that minimizes the worst-case statistical risk.

Theorem 3.1 (Empirical Distribution is Sufficient): For the anonymous estimation problem, the empirical distribution of X^n is sufficient to estimate $\hat{\theta}$. That is, for any estimator $\hat{\phi}(X^n)$, there exists an estimator $\hat{\theta}\left(X^n\right)$ which depends only on Π_{X^n} and

$$\mathsf{R}^*_\theta(\hat{\theta}) \le \mathsf{R}^*_\theta(\hat{\phi}).$$

Proof: We adopt a similar approach as in Rao-Blackwell Theorem [8, Theorem 7.8] by mapping an arbitrary estimator $\hat{\phi}$ to the functional space of Π_{X^n}, and argue that the worst-case statistical risk of mapped estimator is not greater than that of $\hat{\phi}$. Construct the mapping by averaging over all possible permutations as follows:

$$\hat{\theta}\left(X^n\right) \triangleq \frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\hat{\phi}\left(\sigma\left(X^n\right)\right).$$

Then we have

$$\begin{aligned}\mathsf{R}^*_\theta\left(\hat{\theta}\right) &= \max_{\pi\in\mathcal{S}_n}\mathbb{E}_\theta\left[\ell\left(\theta,\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\hat{\phi}\left(\sigma\circ\pi\left(X^n\right)\right)\right)\right]\\ &\overset{\text{(a)}}{=} \mathbb{E}_\theta\left[\ell\left(\theta,\frac{1}{n!}\sum_{\pi\in\mathcal{S}_n}\hat{\phi}\left(\pi\left(X^n\right)\right)\right)\right]\\ &\overset{\text{(b)}}{\le} \frac{1}{n!}\sum_{\pi\in\mathcal{S}_n}\mathbb{E}_\theta\left[\ell\left(\theta,\hat{\phi}\left(\pi\left(X^n\right)\right)\right)\right]\\ &\le \max_{\pi\in\mathcal{S}_n}\mathbb{E}_\theta\left[\ell\left(\theta,\hat{\phi}\left(\pi\left(X^n\right)\right)\right)\right] = \mathsf{R}^*_\theta\left(\hat{\phi}\right),\end{aligned}$$

where (a) is due to the fact that

$$\forall\pi,\ \mathcal{S}_n\circ\pi \triangleq \{\sigma\circ\pi \mid \sigma\in\mathcal{S}_n\} = \mathcal{S}_n,$$

and (b) is due to the convexity of the loss function ℓ and the linearity of expectation. In particular, we notice that after the mapping, $\hat{\theta}$ depends only on the empirical distribution since it outputs the same value for any two realizations differing each other by a permutation: $\forall\sigma,\tau\in\mathcal{S}_n$, $\hat{\theta}\left(\sigma(X^n)\right) = \hat{\theta}\left(\tau(X^n)\right)$. This establishes the theorem. ■

Theorem 3.1 implies we can find an optimal estimator that is a function of the empirical distribution of the samples, and the order does not matter at all. Before we proceed, let us denote the distribution of Π_{X^n} as $\tilde{\mathbb{P}}_\theta$ (note that $X^n \sim \mathbb{P}_\theta$).

For ease of presentation, in the rest of this paper we assume $K = 2$ and set $P_{1;\theta} = P_\theta$, $P_{2;\theta} = Q_\theta$, so

$$X^n \sim \mathbb{P}_\theta = P_\theta^{\otimes n_1}Q_\theta^{\otimes n_2}.$$

The extension to general (finite) K is straightforward.

Our second result is about the asymptote of the minimum worst-case MSE as the number of samples $n\to\infty$ with n_k/n converges to a constant α_k, for $k = 1,...,K$. For the case $K = 2$, we use the notation $(n_1/n, n_2/n)\to(\alpha,\bar{\alpha})$, with $\bar{\alpha}\triangleq 1-\alpha$. As stated in Section II, Fisher information gives a non-trivial lower bound on MSE, and by Theorem 3.1 we know that empirical distribution includes all information we need to estimate θ. Therefore, the asymptotic Fisher information of $\tilde{\mathbb{P}}_\theta$ characterizes the asymptote of the non-trivial lower bound of the worst-case MSE.

Theorem 3.2 (Asymptotic Fisher information): Suppose $\tilde{\mathbb{P}}_\theta$, the distribution of Π_{X^n}, is "good enough" such that the family of functions of ϕ,

$$\left\{\tfrac{1}{n(\theta-\phi)^2}D\left(\tilde{\mathbb{P}}_\theta\middle\|\tilde{\mathbb{P}}_\phi\right)\middle| n\in\mathbb{N}\right\},$$

is *equicontinuous* at $\phi=\theta$. Then the Fisher information of $\tilde{\mathbb{P}}_\theta$ is given by

$$n\mathbb{E}_\theta\left[-\tfrac{\partial^2}{\partial\theta^2}\log M_\theta(X)\right] + o(n), \tag{2}$$

where $M_\theta(x)$ is the *mixture* distribution of $P_\theta(x)$ and $Q_\theta(x)$:

$$M_\theta(x) \triangleq \alpha P_\theta(x) + (1-\alpha)Q_\theta(x).$$

Remark 3.1: A family of functions of ϕ, $\{g_n(\phi)\,|\,n\in\mathbb{N}\}$, is equicontinuous at $\phi=\phi_0$ if $|g_n(\phi_0)-g_n(\phi)|<\epsilon$ for all n as long as $|\phi_0-\phi|<\delta$, where δ is independent of n.

Sketch of Proof: One can start with the exact FI of $\tilde{\mathbb{P}}_\theta$ and then analyze its asymptotic behavior. However, the direct analysis involves bounding complicated terms that need to consider all possible permutations. To circumvent the difficulties, we first compute the asymptote of KLD, which is much easier to control than FI if proper large deviation tools are applied, and then exploit the relation between KLD and FI by taking the second order derivative of KLD. The two-step approach requires an additional equicontinuous assumption in order to exchange the order of limit and derivative. Detailed proof of Theorem 3.2 can be found in Section IV. ■

Remark 3.2: For i.i.d. samples, the maximum likelihood estimator (MLE) achieves CRLB asymptotically under suitable regularity conditions and thus $(nI_p(\theta))^{-1}$ is a tight bound. However, in our setting, the underlying distribution of Π_{X^n} is no longer i.i.d. and lack of structure, making the analysis of asymptote of MLE intractable, so the tightness of the bound in Theorem 3.2 remains unsettled.

Let us consider a toy example to illustrate the loss due to anonymity. Suppose two sources follow distributions $\mathrm{Ber}(\theta)$ and $\mathrm{Ber}(1-\theta)$ respectively. Then the mixture distribution $M_\theta = [\alpha\theta+(1-\alpha)(1-\theta),\alpha(1-\theta)+(1-\alpha)\theta]^\intercal$. According to Theorem 3.2, the asymptotic Fisher information normalized by n is given by

$$(1-2\alpha)^2\left(1-(\alpha+\theta-2\alpha\theta)\right)^{-1}\left(\alpha+\theta-2\alpha\theta\right)^{-1},$$

which is strictly smaller than the asymptotic Fisher information normalized by n in the case without anonymity, that is, $(\theta(1-\theta))^{-1}$, for $0<\alpha<1$.

IV. Asymptotic Fisher Information

In this section, let us prove Theorem 3.2. To begin with, let us set up some notations. Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be two d-dimensional probability vectors on $\mathcal{X}$, and $D\left(\boldsymbol{p}\,\|\boldsymbol{q}\right)$ be the KL-divergence (with base e) between them. Then we use $\nabla_{\boldsymbol{pp}}D\left(\boldsymbol{p}\,\|\boldsymbol{q}\right)$ to denote the *Hessian* matrix of $D\left(\cdot\,\|\boldsymbol{q}\right)$ (namely, regarding $D\left(\boldsymbol{p}\,\|\boldsymbol{q}\right)$ as a function of $\boldsymbol{p}$ with $\boldsymbol{q}$ being fixed). $\nabla_{\boldsymbol{qq}}D\left(\boldsymbol{p}\,\|\boldsymbol{q}\right)$,

$\nabla_{\boldsymbol{pq}} D\left(\boldsymbol{p} \,\|\boldsymbol{q}\right)$, and $\nabla_{\boldsymbol{qp}} D\left(\boldsymbol{p} \,\|\boldsymbol{q}\right)$ are defined in a similar manner. For the ease of presentation, we shall omit the dummy variable:

$$\begin{aligned}
\left[\nabla_{\boldsymbol{pp}} D\left(p_\theta \,\|q_\theta\right)\right]_{ij} &\triangleq \left.\tfrac{\partial^2 D(\boldsymbol{p} \,\|p_\theta)}{\partial p_i \partial p_j}\right|_{\boldsymbol{p}=p_\theta} \\
\left[\nabla_{\boldsymbol{qq}} D\left(p_\theta \,\|q_\theta\right)\right]_{ij} &\triangleq \left.\tfrac{\partial^2 D(p_\theta \,\|\boldsymbol{q})}{\partial q_i \partial q_j}\right|_{\boldsymbol{q}=q_\theta} \\
\left[\nabla_{\boldsymbol{pq}} D\left(p_\theta \,\|q_\theta\right)\right]_{ij} &\triangleq \left.\tfrac{\partial^2 D(\boldsymbol{p} \,\|\boldsymbol{q})}{\partial p_i \partial q_j}\right|_{(\boldsymbol{p},\boldsymbol{q})=(p_\theta,q_\theta)} \\
\left[\nabla_{\boldsymbol{qp}} D\left(p_\theta \,\|q_\theta\right)\right]_{ij} &\triangleq \left.\tfrac{\partial^2 D(\boldsymbol{p} \,\|\boldsymbol{q})}{\partial q_i \partial p_j}\right|_{(\boldsymbol{p},\boldsymbol{q})=(p_\theta,q_\theta)}
\end{aligned} \tag{3}$$

In words, the subscript $\boldsymbol{p}$ (or $\boldsymbol{q}$) indicates which argument the derivative operator acts on. In addition, denote $p'_\theta \triangleq \partial p_\theta / \partial \theta$ (a d-dimensional vector).

Next, let us introduce some technical lemmas.

Lemma 4.1: Let $p_\theta(x)$ satisfy the regularity conditions in Section II-B. Then its Fisher information can be expressed by

$$I_p(\theta) = {p'_\theta}^{\mathsf{T}} \nabla_{\boldsymbol{qq}} D\left(p_\theta \,\|p_\theta\right) p'_\theta.$$

The relation between KL divergence and Fisher information motivates us to first estimate $D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right)$, analyze its limit, and then leverage Lemma 4.1 to obtain $I_{\tilde{\mathbb{P}}}(\theta)$. The following lemma gives us the asymptotic behavior of $D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right)$:

Lemma 4.2 (Asymptotic Formula of KL Divergence): Let $X^n \sim \mathbb{P}_\theta \triangleq \left(P_\theta\right)^{\otimes n_1} \left(Q_\theta\right)^{\otimes n_2}$, and $\tilde{\mathbb{P}}_\theta$ be the distribution of Π_{X^n} (and so does $\tilde{\mathbb{P}}_\phi$). Then we have

$$\begin{aligned}
&D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right) \\
&= n \left(\begin{array}{cc} \min\limits_{V_0, V_1 \in \mathcal{P}_{\mathcal{X}}} & \alpha D(V_0 \,\|P_\phi) + \bar{\alpha} D(V_1 \,\|Q_\phi) \\ \text{s.t.} & \alpha V_0 + \bar{\alpha} V_1 = \alpha P_\theta + \bar{\alpha} Q_\theta \end{array} \right) + o(n),
\end{aligned} \tag{4}$$

where $\bar{\alpha} \triangleq (1-\alpha)$.

Next, by imposing an additional equicontinuous condition on $\tilde{\mathbb{P}}_\theta$, we can exchange the order of derivative and the limit.

Lemma 4.3: If the family of functions of ϕ,

$$\left\{ \tfrac{1}{n(\theta-\phi)^2} D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right) \,\middle|\, n \in \mathbb{N} \right\},$$

is *equicontinuous* at $\phi = \theta$, the following holds at $\phi = \theta$:

$$\lim_{n\to\infty} \tfrac{\partial^2}{\partial \phi^2} \left(\tfrac{1}{n} D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right)\right) = \tfrac{\partial^2}{\partial \phi^2} \left(\lim_{n\to\infty} \tfrac{1}{n} D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right)\right).$$

Proof of Lemma 4.2 can be found in Section V, and proofs for other technical lemmas are delegated to Appendix of the full version. In the following, the proof of Theorem 3.2 is presented.

Proof of Theorem 3.2: According to Lemma 4.1, Lemma 4.2, and Lemma 4.3, it suffices to evaluate the second order derivative of (4). Equivalently, we aim to analyze the following limit:

$$\lim_{\Delta\theta\to 0} \frac{1}{\Delta\theta^2} \left(\begin{array}{cc} \min\limits_{V_0, V_1 \in \mathcal{P}_{\mathcal{X}}} & \alpha D(V_0 \,\|P_{\theta+\Delta\theta}) + \bar{\alpha} D(V_1 \,\|Q_{\theta+\Delta\theta}) \\ \text{s.t.} & \alpha V_0 + \bar{\alpha} V_1 = \alpha P_\theta + \bar{\alpha} Q_\theta \end{array} \right) \tag{5}$$

Part 1 (Change of Variables): Now let us rewrite the constraint by transforming the variables V_0, V_1 in $\mathcal{P}_{\mathcal{X}}$ to the tangent space

$$\begin{cases} V_0 = P_\theta + f_0 \Delta\theta \\ V_1 = Q_\theta + f_1 \Delta\theta, \end{cases}$$

where $\sum_{x\in\mathcal{X}} f_i(x) = 0$, for $i = 0, 1$. By the Taylor expansion we also have

$$\begin{cases} P_{\theta+\Delta\theta} = P_\theta + P'_\theta \Delta\theta + o(\Delta\theta) \\ Q_{\theta+\Delta\theta} = Q_\theta + Q'_\theta \Delta\theta + o(\Delta\theta). \end{cases}$$

The optimization problem (5) (if temporarily omitting the limit on $\Delta\theta$) thus becomes

$$\begin{aligned}
&\min_{(f_0,f_1)\in\mathcal{F}} \alpha \underbrace{D\left(P_\theta + f_0 \Delta\theta \,\|P_\theta + P'_\theta \Delta\theta + o(\Delta\theta)\right)/\Delta\theta^2}_{\text{Part A}} \\
&\quad + \bar{\alpha} \underbrace{D\left(Q_\theta + f_1 \Delta\theta \,\|Q_\theta + Q'_\theta \Delta\theta + o(\Delta\theta)\right)/\Delta\theta^2}_{\text{Part B}},
\end{aligned} \tag{6}$$

where for notational simplicity, we define the feasible set $\mathcal{F} \triangleq (\mathcal{T}_0 \times \mathcal{T}_1) \cap \mathcal{T}$, with $\mathcal{T}_i \triangleq \{f_i : \sum_{x\in\mathcal{X}} f_i(x) = 0\}$, for $i = 0, 1$, and $\mathcal{T} \triangleq \{(f_0, f_1) : \alpha f_0 + \bar{\alpha} f_1 = \mathbf{0}\}$.

Part 2 (Local Approximation): As Lemma 4.1 suggests that KLD is locally quadratic, we then consider the local expansion on the objective function. That is, Part A of (6) can be approximated by

$$[f_0, P'_\theta] \begin{bmatrix} \nabla_{\boldsymbol{pp}} D\left(P_\theta \,\|P_\theta\right) & \nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right) \\ \nabla_{\boldsymbol{qp}} D\left(P_\theta \,\|P_\theta\right) & \nabla_{\boldsymbol{qq}} D\left(P_\theta \,\|P_\theta\right) \end{bmatrix} \begin{bmatrix} f_0 \\ P'_\theta \end{bmatrix} + o(1).$$

Here (and throughout the rest of this section) the error term $o(1)$ is with respect to $\Delta\theta$, namely, $o(1) \to 0$ as $\Delta\theta \to 0$.

Part B of (6) can be approximated similarly, except that f_0 and P'_θ are replaced by f_1 and Q'_θ. Therefore, (6) can be written as a constrained minimization problem of a quadratic function in f_0 and f_1 plus some error term:

$$\begin{aligned}
\min_{(f_0,f_1)\in\mathcal{F}} &\alpha \left(f_0^{\mathsf{T}} \nabla_{\boldsymbol{pp}} D\left(P_\theta \,\|P_\theta\right) f_0 + 2 f_0^{\mathsf{T}} \nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right) P'_\theta \right) \\
&+ \bar{\alpha} \left(f_1^{\mathsf{T}} \nabla_{\boldsymbol{pp}} D\left(Q_\theta \,\|Q_\theta\right) f_1 + 2 f_1^{\mathsf{T}} \nabla_{\boldsymbol{pq}} D\left(Q_\theta \,\|Q_\theta\right) Q'_\theta \right) \\
&+ \alpha I_P(\theta) + \bar{\alpha} I_Q(\theta) + o(1),
\end{aligned} \tag{7}$$

where we use Lemma 4.1 and the fact that

$$\nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right) = \nabla_{\boldsymbol{qp}} D\left(P_\theta \,\|P_\theta\right).$$

Replacing f_1 with $-\alpha f_0 / \bar{\alpha}$, we see that (7) becomes a constrained quadratic optimization problem in f_0:

$$\begin{aligned}
\min_{f_0\in\mathcal{T}_0} \; & f_0^{\mathsf{T}} \left(\alpha \nabla_{\boldsymbol{pp}} D\left(P_\theta \,\|P_\theta\right) + \frac{\alpha^2}{\bar{\alpha}} \nabla_{\boldsymbol{pp}} D\left(Q_\theta \,\|Q_\theta\right) \right) f_0 \\
&+ 2\alpha f_0^{\mathsf{T}} \left(\nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right) P'_\theta - \nabla_{\boldsymbol{pq}} D\left(Q_\theta \,\|Q_\theta\right) Q'_\theta \right) \\
&+ \alpha I_P(\theta) + \bar{\alpha} I_Q(\theta) + o(1).
\end{aligned} \tag{8}$$

Finally, we claim that the $o(1)$ term in (8) can be ignored and hence (8) can be solved as a quadratic function:

Lemma 4.4: As $\Delta\theta \to 0$, the minimum of (8) converges to

$$-\alpha \left(\nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right) P'_\theta - \nabla_{\boldsymbol{pq}} D\left(Q_\theta \,\|Q_\theta\right) Q'_\theta \right)^{\mathsf{T}} f_0^{**}, \tag{9}$$

where

$$\begin{aligned}
f_0^{**} = &- \left(\nabla_{\boldsymbol{pp}} D\left(P_\theta \,\|P_\theta\right) + \frac{\alpha}{\bar{\alpha}} \nabla_{\boldsymbol{pp}} D\left(Q_\theta \,\|Q_\theta\right) \right)^{-1} \\
&\cdot \left(\nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right) P'_\theta - \nabla_{\boldsymbol{pq}} D\left(Q_\theta \,\|Q_\theta\right) Q'_\theta \right).
\end{aligned} \tag{10}$$

By plugging f_0^{**} into (9), together with some tedious calculations (which can be found in Appendix of the full version, we can see that (8) is indeed (2).

The proof of Theorem 3.2 is now complete. ■

V. Asymptotic KullbackLeibler Divergence

As the result in [3] suggests, the large deviation exponent of anonymous hypothesis testing (i.e. to test $\tilde{\mathbb{P}}_\theta$ vs $\tilde{\mathbb{P}}_\phi$) is (4), so it is reasonable to guess the exact KLD is asymptotically equal to (4). In this section we give a formal proof of Lemma 4.2, which is based on the method of types.

Proof of Lemma 4.2: To evaluate

$$D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right) = \mathbb{E}_{\tilde{\mathbb{P}}_\theta}\left[\log\left(\frac{\tilde{\mathbb{P}}_\theta(\mu_n)}{\tilde{\mathbb{P}}_\phi(\mu_n)}\right)\right], \tag{11}$$

we notice that the log likelihood ratio (LLR) in (11) is a function of μ_n, and by the law of large numbers, the density of μ_n concentrates. Hence, if the LLR can be approximated by some continuous function (independent of n), together with the concentration of $\tilde{\mathbb{P}}_\theta$, we can estimate the asymptotic behavior.

Part 1 (Asymptote of LLR): The asymptote of the LLR is given by the lemma below (the proof can be found in Appendix of the full version.)

Lemma 5.1: For any empirical distribution $\mu_n = \Pi_{X^n}$ and

$$\mathbb{P}_\theta \triangleq (P_\theta)^{\otimes n_1}(Q_\theta)^{\otimes n_2},\ \mathbb{P}_\phi \triangleq (P_\phi)^{\otimes n_1}(Q_\phi)^{\otimes n_2},$$

we have

$$\log\left(\frac{\tilde{\mathbb{P}}_\theta(\mu_n)}{\tilde{\mathbb{P}}_\phi(\mu_n)}\right) = -nR_0(\mu_n) + nR_1(\mu_n) + o(n),$$

where

$$R_0(\mu_n) \triangleq \min_{\substack{V_1,V_2\in\mathcal{P}_\mathcal{X}\\ \alpha V_1+\bar{\alpha}V_2=\mu_n}} \alpha D\left(V_1 \,\|P_\theta\right) + \bar{\alpha} D\left(V_2 \,\|Q_\theta\right),$$

$$R_1(\mu_n) \triangleq \min_{\substack{U_1,U_2\in\mathcal{P}_\mathcal{X}\\ \alpha U_1+\bar{\alpha}U_2=\mu_n}} \alpha D\left(U_1 \,\|P_\phi\right) + \bar{\alpha} D\left(U_2 \,\|Q_\phi\right).$$

Remark 5.1: The rate functions R_0 and R_1 are well-studied in [3], and was proved to satisfy the following:

- They have compact domains: $\{\mu_n : R_i(\mu_n) < \infty\}$ is compact, for $i = 0, 1$.
- They are continuous functions with respect to ℓ_1 norm.

The compactness is essential because continuous in a compact domain implies that R_0 and R_1 are bounded, say, $\forall\, i = 0, 1$, $R_i(\mu) < B$ for some constant $B > 0$. In words, if $R_0(\mu)$ and $R_1(\mu)$ are finite, they must be bounded by some constant.

Part 2 (Concentration of μ_n): Next, we argue that μ_n does concentrate. In fact, μ_n can be written as the weighted sum of two empirical distribution

$$\mu_n = \frac{n_1}{n}\Pi_{X^{n_1}} + \frac{n_2}{n}\Pi_{X^{n_2}},$$

where X^{n_1} and X^{n_2} follow i.i.d. P_θ and Q_θ respective. By the uniform law of large number (Glivenko-Cantelli theorem), $\Pi_{X^{n_1}} \to P_\theta$ and $\Pi_{X^{n_2}} \to Q_\theta$ (here for simplicity, we use the ℓ_1 norm as the metric), so $\mu_n \to \alpha P_\theta + \bar{\alpha} Q_\theta$ in ℓ_1. Thus for arbitrary $\epsilon > 0$ and $\delta > 0$, we have

$$\Pr\left\{\|\mu_n - (\alpha P_\theta + \bar{\alpha}Q_\theta)\|_1 < \epsilon\right\} > 1 - \delta \tag{12}$$

for $n > M_{\epsilon,\delta}$. For notational simplicity, set $\mu \triangleq \alpha P_\theta + \bar{\alpha} Q_\theta$.

Part 3 (Bounding KLD): Now let us bound (11). We first note that the LLR is always finite according to the data-processing inequality

$$\frac{1}{n}D\left(\tilde{\mathbb{P}}_\theta \middle\| \tilde{\mathbb{P}}_\phi\right) \le \frac{1}{n}D\left(\mathbb{P}_\theta \,\|\mathbb{P}_\phi\right) = \alpha D\left(P_\theta \,\|P_\phi\right) + \bar{\alpha} D\left(Q_\theta \,\|Q_\phi\right),$$

so by Remark 5.1 it is bounded by a constant, say, B. Next, (11) can be written as

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\tilde{\mathbb{P}}_\theta}\left[\log\left(\frac{\tilde{\mathbb{P}}_\theta(\mu_n)}{\tilde{\mathbb{P}}_\phi(\mu_n)}\right)\right]\\
=&\frac{1}{n}\Pr\left\{\|\mu_n-\mu\|\le\epsilon\right\}\mathbb{E}_{\tilde{\mathbb{P}}_\theta}\left[\log\left(\frac{\tilde{\mathbb{P}}_\theta(\mu_n)}{\tilde{\mathbb{P}}_\phi(\mu_n)}\right)\middle|\, \|\mu_n-\mu\|\le\epsilon\right]\\
&+\frac{1}{n}\Pr\left\{\|\mu_n-\mu\|>\epsilon\right\}\mathbb{E}_{\tilde{\mathbb{P}}_\theta}\left[\log\left(\frac{\tilde{\mathbb{P}}_\theta(\mu_n)}{\tilde{\mathbb{P}}_\phi(\mu_n)}\right)\middle|\, \|\mu_n-\mu\|>\epsilon\right]\\
\overset{\text{(a)}}{\le}&\,\mathbb{E}_{\tilde{\mathbb{P}}_\theta}\left[\frac{1}{n}\log\left(\frac{\tilde{\mathbb{P}}_\theta(\mu_n)}{\tilde{\mathbb{P}}_\phi(\mu_n)}\right)\middle|\, \|\mu_n-\mu\|\le\epsilon\right]+\delta B\\
\overset{\text{(b)}}{=}&\,\mathbb{E}_{\tilde{\mathbb{P}}_\theta}\left[-R_0(\mu_n)+R_1(\mu_n)+o_n(1)\middle|\, \|\mu_n-\mu\|\le\epsilon\right]+\delta B\\
\overset{\text{(c)}}{\le}&\,-R_0(\mu)+R_1(\mu)+\Delta(\epsilon)+o_n(1)+\delta B,
\end{aligned}$$

where (a) is due to (12), (b) is due to Lemma 5.1, and (c) is due to the continuity of R_0 and R_1 (and thus $\Delta(\epsilon) \to 0$ as $\epsilon \to 0$). Notice that $R_0(\mu) = 0$ since one can easily choose V_1, V_2 to be P_θ, P_ϕ. Also note that both ϵ and δ can be chosen arbitrarily small as n tends to infinity. Therefore we obtain an upper bound on (11), which is $nR_1(\mu) + o(n)$. On the other hand, by replacing B with $-B$, we also obtain a lower bound (since the inequality (c) is due to the continuity). Finally, by definition, $R_1(\mu)$ is exactly the left term in (4).

The proof of Lemma 4.2 is now complete. ■

Acknowledgement

This work was supported by Ministry of Science and Technology of Taiwan and MOST Joint Research Center for AI Technology and All Vista Healthcare under Grant MOST 108-2634-F-002-006.

References

[1] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Tech. Rep., 1998.

[2] W.-N. Chen, H.-C. Chen, and I.-H. Wang, "On the fundamental limits of heterogeneous distributed detection: Price of anonymity," *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[3] W.-N. Chen and I.-H. Wang, "Anonymous heterogeneous distributed detection: Optimal decision rules, error exponents, and the price of anonymity," *arXiv:1805.03554*, 2018.

[4] D. Basu, "On the elimination of nuisance parameters," *Journal of the American Statistical Association*, vol. 72, June 1977.

[5] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval, "Compressed sensing with unknown sensor permutation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[6] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled sensing with random linear measurements," *IEEE Transactions on Information Theory*, 2018.

[7] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with shuffled data: Statistical and computational limits of permutation recovery," *IEEE Transactions on Information Theory*, 2018.

[8] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. Springer, 2003.

Appendix

A. Proof of Lemma 5.1

To begin with, we identify the distribution of $\tilde{\mathbb{P}}$. Let μ_n be an empirical distribution, and by definition

$$\tilde{\mathbb{P}}\left(\mu_n\right) = \sum_{x^n : \Pi_{x^n} = \mu_n} \mathbb{P}(x^n) = c(\mu_n) \sum_{\pi \in \mathcal{S}_n} \mathbb{P}\left(\pi\left(x^n\right)\right)\Big|_{\text{for any } x^n \text{ with type } \mu_n}$$

where $c(\mu_n)$ is a constant that normalized the repetition of counting. For example, if

$$\mu_n = \left[\frac{m_1}{n}, \frac{m_2}{n}, ..., \frac{m_d}{n}\right]^{\intercal},$$

then $c(\mu_n) = ((m_1!)(m_2!)\cdots(m_d!))^{-1}$. Therefore,

$$\log\left(\frac{\tilde{\mathbb{P}}_0(\mu_n)}{\tilde{\mathbb{P}}_1(\mu_n)}\right) = \log\left(\frac{\sum_{\pi} \mathbb{P}_0\left(\pi\left(x^n\right)\right)}{\sum_{\pi} \mathbb{P}_1\left(\pi\left(x^n\right)\right)}\right)\Bigg|_{\text{for any } x^n \text{ with type } \mu_n}. \tag{13}$$

Without loss of generality, for each μ_n, we pick the representer x^n as the sorted one, say,

$$x^n = \left(\underbrace{a_1, ..., a_1}_{m_1}, \underbrace{a_2, ..., a_2}_{m_2}, ..., \underbrace{a_d, ..., a_d}_{m_d}\right),$$

assuming $\mu_n = \left[m_1, m_2, ..., m_d\right]^{\intercal}/n$. Next, for an (unsorted) x^n, the $\mathbb{P}_0(x^n)$ is completely determined by the empirical distribution of first n_1 samples and the empirical distribution of the rest n_2 samples. We use $\left(m_1^{(1)}, ..., m_d^{(1)}\right)$ and $\left(m_1^{(2)}, ..., m_d^{(2)}\right)$ to denote these (unnormalized) *product type* of x^n. See Figure 2 for illustration. Therefore, summation over all $\pi(x^n)$ can be written as first summing over product types and then over all permutations which possess same product type:

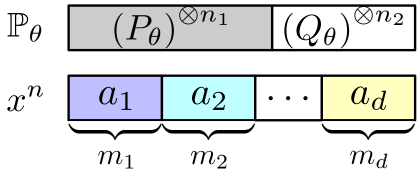

Fig. 1: illustration of the representer

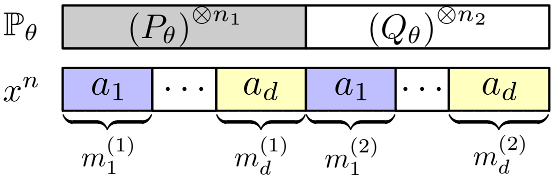

Fig. 2: illustration of the product type

$$\begin{aligned}
\textstyle\sum_{\pi} \mathbb{P}_0\left(\pi\left(x^n\right)\right) &= \sum_{\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}} \left(\sum_{\pi : \pi(x^n) \in \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}} \mathbb{P}_0\left(\pi\left(x^n\right)\right)\right) \\
&= \sum_{\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}} \left(\left(P_0\left(a_1\right)^{m_1^{(1)}} \cdots P_0\left(a_d\right)^{m_d^{(1)}}\right)\left(Q_0\left(a_1\right)^{m_1^{(2)}} \cdots Q_0\left(a_d\right)^{m_d^{(2)}}\right) \cdot \right. \\
&\qquad \left.\left(\left(m_1^{(1)}!\right) \cdots \left(m_d^{(1)}!\right)\right)\left(\left(m_1^{(2)}!\right) \cdots \left(m_d^{(2)}!\right)\right)\right) \\
&\leq \max_{\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}} \left(\left(P_0\left(a_1\right)^{m_1^{(1)}} \cdots P_0\left(a_d\right)^{m_d^{(1)}}\right)\left(Q_0\left(a_1\right)^{m_1^{(2)}} \cdots Q_0\left(a_d\right)^{m_d^{(2)}}\right) \cdot \right. \\
&\qquad \left.\left(\left(m_1^{(1)}!\right) \cdots \left(m_d^{(1)}!\right)\right)\left(\left(m_1^{(2)}!\right) \cdots \left(m_d^{(2)}!\right)\right)\right).
\end{aligned} \tag{14}$$

On the other hand the number of product type is polynomial in n (less than n^{2d}), so it is reasonable to upper bound (14) by the maximum over product types:

$$(14) \leq n^{2d} \max_{\boldsymbol{m}^{(1)},\boldsymbol{m}^{(2)}} \left\{ \left(\prod_{i=1}^{d} P_0\left(a_i\right)^{m_i^{(1)}} \right) \left(\prod_{i=1}^{d} Q_0\left(a_i\right)^{m_i^{(2)}} \right) \left(\left(m_1^{(1)}! \right) ... \left(m_d^{(1)}! \right) \right) \left(\left(m_1^{(2)}! \right) ... \left(m_d^{(2)}! \right) \right) \right\}.$$

Hence we have

$$\begin{aligned}
&\log\left(\sum_{\pi} \mathbb{P}_0\left(\pi\left(x^n\right)\right) \right) \\
&= \max_{\boldsymbol{m}^{(1)},\boldsymbol{m}^{(2)}} \left(\sum_{i=1}^{d} m_i^{(1)} \log P_0(a_i) + \sum_{i=1}^{d} m_i^{(2)} \log Q_0(a_i) + \sum_{i=1}^{d} \log\left(m_i^{(1)}! \right) + \sum_{i=1}^{d} \log\left(m_i^{(2)}! \right) \right) + o(n). \qquad (15)
\end{aligned}$$

Notice that the product type must satisfy the constraints

$$\begin{cases} \sum_{i=1}^{d} m_i^{(1)} = n_1 \\ \sum_{i=1}^{d} m_i^{(2)} = n_2 \\ \forall i \in \{1, ..., d\},\ m_i^{(1)} + m_i^{(2)} = m_i. \end{cases} \qquad (16)$$

Therefore, by letting

$$V_1 = [m_1^{(1)}/n_1, ..., m_d^{(1)}/n_1]^\intercal,\ V_2 = [m_1^{(2)}/n_2, ..., m_d^{(2)}/n_2]^\intercal,$$

we see that

$$\begin{aligned}
&\sum_{i=1}^{d} m_i^{(1)} \log P_0(a_i) + \sum_{i=1}^{d} \log\left(m_i^{(1)}! \right) \\
&\overset{\text{(a)}}{=} n_1 \left(\sum_{i=1}^{d} \frac{m_i^{(1)}}{n_1} \log P_0(a_i) + \sum_{i=1}^{d} \frac{m_i^{(1)}}{n_1} \log\left(m_i^{(1)} \right) \right) - \sum_{i=1}^{d} m_i^{(1)} + o(n_1) \\
&= n_1 \left(\sum_{i=1}^{d} \frac{m_i^{(1)}}{n_1} \log P_0(a_i) + \sum_{i=1}^{d} \frac{m_i^{(1)}}{n_1} \log\left(\frac{n_1}{m_i^{(1)}} \right) \right) - n_1 \log n_1 - n_1 + o(n_1) \\
&= -n_1 D\left(V_1 \,\|P_0\right) - n_1 \log n_1 - n_1 + o(n_1), \qquad (17)
\end{aligned}$$

where (a) is due to the Stirling's formula

$$\log\left(m_i^{(1)}! \right) = m_i^{(1)} \log m_i^{(1)} - m_i^{(1)} + O(\log m_i^{(1)}).$$

Similarly,

$$\sum_{i=1}^{d} m_i^{(2)} \log Q_0(a_i) + \sum_{i=1}^{d} \log\left(m_i^{(2)}! \right) = -n_2 D\left(V_2 \,\|Q_0\right) - n_2 \log n_2 - n_2 + o(n_2). \qquad (18)$$

Recall that the third constraint in (16) requires V_1, V_2 to satisfy

$$n_1 V_1 + n_2 V_2 = n\mu_n,$$

and combine (15), (17) and (18) we get

$$\log\left(\sum_{\pi} \mathbb{P}_0\left(\pi\left(x^n\right)\right) \right) = -\min_{n_1V_1+n_2V_2=-n\mu_n} \left(n_1 D\left(V_1 \,\|P_0\right) + n_2 D\left(V_2 \,\|Q_0\right) \right) - n_1 \log n_1 - n_2 \log n_2 - n_1 - n_2 + o(n).$$

Similarly we also have

$$\log\left(\sum_{\pi} \mathbb{P}_1\left(\pi\left(x^n\right)\right) \right) = -\min_{n_1U_1+n_2U_2=-n\mu_n} \left(n_1 D\left(U_1 \,\|P_1\right) + n_2 D\left(U_2 \,\|Q_1\right) \right) - n_1 \log n_1 - n_2 \log n_2 - n_1 - n_2 + o(n).$$

Plugging back into (13), we obtain

$$\begin{aligned}
\log\left(\frac{\tilde{\mathbb{P}}_0(\mu_n)}{\tilde{\mathbb{P}}_1(\mu_n)} \right) &= \log\left(\frac{\sum_{\pi} \mathbb{P}_0\left(\pi\left(x^n\right)\right)}{\sum_{\pi} \mathbb{P}_1\left(\pi\left(x^n\right)\right)} \right) \\
&= -\left(\begin{matrix} \min\limits_{V_1, V_2 \in \mathcal{P}_\mathcal{X}} & n_1 D(V_1 \,\|P_0) + n_2 D(V_2 \,\|Q_0) \\ \text{s.t.} & \frac{n_1}{n} V_1 + \frac{n_2}{n} V_2 = \mu_n \end{matrix} \right) + \left(\begin{matrix} \min\limits_{U_1, U_2 \in \mathcal{P}_\mathcal{X}} & n_1 D(U_1 \,\|P_1) + n_2 D(U_2 \,\|Q_1) \\ \text{s.t.} & \frac{n_1}{n} U_1 + \frac{n_2}{n} U_2 = \mu_n \end{matrix} \right) + o(n).
\end{aligned}$$

Finally, by dividing both sides by n and letting n tends to infinity with $n_1/n \to \alpha$ as well as $n_2/n \to 1 - \alpha$, the proof is complete.

B. Calculation of (9)

We prove that by plugging

$$f_0^{**} = -\Big(\nabla_{\boldsymbol{pp}} D\left(P_\theta \,\|P_\theta\right) + \frac{\alpha}{\bar{\alpha}}\nabla_{\boldsymbol{pp}} D\left(Q_\theta \,\|Q_\theta\right)\Big)^{-1}\left(\nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right)P'_\theta - \nabla_{\boldsymbol{pq}} D\left(Q_\theta \,\|Q_\theta\right)Q'_\theta\right).$$

into

$$-\alpha\left(\nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right)P'_\theta - \nabla_{\boldsymbol{pq}} D\left(Q_\theta \,\|Q_\theta\right)Q'_\theta\right)^{\mathsf{T}} f_0^{**}, \tag{19}$$

we obtain the solution of Theorem 3.2.

First by definition, we have

$$\nabla_{\boldsymbol{pp}} D\left(P_\theta \,\|P_\theta\right) \triangleq \left[\frac{\partial^2 D\left(\boldsymbol{p}\,\|P_\theta\right)}{\partial p_i \partial p_j}\right]\Bigg|_{\boldsymbol{p}=P_\theta} = \begin{bmatrix} \frac{1}{P_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{P_\theta(a_d)}, \end{bmatrix}$$

since

$$\frac{\partial^2 D\left(\boldsymbol{p}\,\|P_\theta\right)}{\partial p_i \partial p_j} = \begin{cases} 0, & \text{if } i \neq j, \\ \frac{1}{P_\theta(a_i)}, & \text{else.} \end{cases}$$

Similarly,

$$\nabla_{\boldsymbol{qq}} D\left(P_\theta \,\|P_\theta\right) = \begin{bmatrix} \frac{1}{P_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{P_\theta(a_d)}, \end{bmatrix}, \text{ and } \nabla_{\boldsymbol{pq}} D\left(P_\theta \,\|P_\theta\right) = \nabla_{\boldsymbol{qp}} D\left(P_\theta \,\|P_\theta\right) = \begin{bmatrix} -\frac{1}{P_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & -\frac{1}{P_\theta(a_d)}. \end{bmatrix}.$$

Therefore, f_0^{**} becomes

$$\begin{aligned}
&-\frac{1}{\alpha}\begin{bmatrix} \frac{1}{\alpha P_\theta(a_1)} + \frac{1}{\bar{\alpha} Q_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{\alpha P_\theta(a_d)} + \frac{1}{\bar{\alpha} Q_\theta(a_d)} \end{bmatrix}^{-1} \cdot \\
&\left(\begin{bmatrix} -\frac{1}{P_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & -\frac{1}{P_\theta(a_d)}. \end{bmatrix}\begin{bmatrix} P'_\theta(a_1) \\ \vdots \\ P'_\theta(a_d) \end{bmatrix} - \begin{bmatrix} -\frac{1}{Q_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & -\frac{1}{Q_\theta(a_d)}. \end{bmatrix}\begin{bmatrix} Q'_\theta(a_1) \\ \vdots \\ Q'_\theta(a_d) \end{bmatrix}\right) \\
&= \begin{bmatrix} \frac{\bar{\alpha} P_\theta(a_1) Q_\theta(a_1)}{\alpha P_\theta(a_1) + \bar{\alpha} Q_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{\bar{\alpha} P_\theta(a_d) Q_\theta(a_d)}{\alpha P_\theta(a_d) + \bar{\alpha} Q_\theta(a_d)} \end{bmatrix}^{-1}\left(\begin{bmatrix} \frac{P'_\theta(a_1)}{P_\theta(a_1)} \\ \vdots \\ \frac{P'_\theta(a_d)}{P_\theta(a_d)} \end{bmatrix} - \begin{bmatrix} \frac{Q'_\theta(a_1)}{Q_\theta(a_1)} \\ \vdots \\ \frac{Q'_\theta(a_d)}{Q_\theta(a_d)} \end{bmatrix}\right).
\end{aligned}$$

Plugging back to (19), we see that

$$\begin{aligned}
(19) &= -\left(\begin{bmatrix} \frac{P'_\theta(a_1)}{P_\theta(a_1)} \\ \vdots \\ \frac{P'_\theta(a_d)}{P_\theta(a_d)} \end{bmatrix} - \begin{bmatrix} \frac{Q'_\theta(a_1)}{Q_\theta(a_1)} \\ \vdots \\ \frac{Q'_\theta(a_d)}{Q_\theta(a_d)} \end{bmatrix}\right)^{\mathsf{T}} \begin{bmatrix} \frac{\alpha\bar{\alpha} P_\theta(a_1) Q_\theta(a_1)}{\alpha P_\theta(a_1) + \bar{\alpha} Q_\theta(a_1)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{\alpha\bar{\alpha} P_\theta(a_d) Q_\theta(a_d)}{\alpha P_\theta(a_d) + \bar{\alpha} Q_\theta(a_d)} \end{bmatrix}\left(\begin{bmatrix} \frac{P'_\theta(a_1)}{P_\theta(a_1)} \\ \vdots \\ \frac{P'_\theta(a_d)}{P_\theta(a_d)} \end{bmatrix} - \begin{bmatrix} \frac{Q'_\theta(a_1)}{Q_\theta(a_1)} \\ \vdots \\ \frac{Q'_\theta(a_d)}{Q_\theta(a_d)} \end{bmatrix}\right) \\
&= -\sum_{i=1}^{d} \frac{\alpha\bar{\alpha} P_\theta\left(a_i\right) Q_\theta\left(a_i\right)}{\alpha P_\theta\left(a_i\right) + \bar{\alpha} Q_\theta\left(a_i\right)}\left(\frac{P'_\theta(a_i)}{P_\theta(a_i)} - \frac{Q'_\theta(a_i)}{Q_\theta(a_i)}\right)^2 \\
&= -\sum_{i=1}^{d} \frac{\alpha\bar{\alpha}\left(P'_\theta(a_i) Q_\theta(a_i) - Q'_\theta(a_i) P_\theta(a_i)\right)^2}{\left(\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)\right) P_\theta(a_i) Q_\theta(a_i)}
\end{aligned}$$

$$
\begin{aligned}
&= - \alpha\bar{\alpha} \sum_{i=1}^{d} \left(\frac{\left(P'_\theta(a_i)\right)^2 Q_\theta(a_i)}{\left(\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)\right) P_\theta(a_i)} + \frac{\left(Q'_\theta(a_i)\right)^2 P_\theta(a_i)}{\left(\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)\right) Q_\theta(a_i)} - 2 \frac{P'_\theta(a_i) Q'_\theta(a_i)}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} \right) \\
&= - \alpha\bar{\alpha} \sum_{i=1}^{d} \left(\frac{\left(P'_\theta(a_i)\right)^2 Q_\theta(a_i)}{\left(\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)\right) P_\theta(a_i)} + \frac{\left(Q'_\theta(a_i)\right)^2 P_\theta(a_i)}{\left(\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)\right) Q_\theta(a_i)} - 2 \frac{P'_\theta(a_i) Q'_\theta(a_i)}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} \right) \\
&= - \sum_{i=1}^{d} \left(\left(\alpha \frac{P'_\theta(a_i)^2}{P_\theta(a_i)} - \frac{\alpha^2 \left(P'_\theta(a_i)\right)^2}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} \right) + \left(\bar{\alpha} \frac{Q'_\theta(a_i)^2}{Q_\theta(a_i)} - \frac{\bar{\alpha}^2 \left(Q'_\theta(a_i)\right)^2}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} \right) - 2 \frac{\alpha\bar{\alpha} P'_\theta(a_i) Q'_\theta(a_i)}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} \right) \\
&= - \sum_{i=1}^{d} \left(\alpha \frac{P'_\theta(a_i)^2}{P_\theta(a_i)} + \bar{\alpha} \frac{Q'_\theta(a_i)^2}{Q_\theta(a_i)} \right) + \sum_{i=1}^{d} \left(\frac{\alpha^2 \left(P'_\theta(a_i)\right)^2}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} + \frac{\bar{\alpha}^2 \left(Q'_\theta(a_i)\right)^2}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} + 2 \frac{\alpha\bar{\alpha} P'_\theta(a_i) Q'_\theta(a_i)}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} \right) \\
&= - \sum_{i=1}^{d} \left(\alpha \frac{\left(P'_\theta(a_i)\right)^2}{P_\theta(a_i)} + \bar{\alpha} \frac{\left(Q'_\theta(a_i)\right)^2}{Q_\theta(a_i)} \right) + \sum_{i=1}^{d} \left(\frac{\left(\alpha P'_\theta(a_i) + \bar{\alpha} Q'_\theta(a_i)\right)^2}{\alpha P_\theta(a_i) + \bar{\alpha} Q_\theta(a_i)} \right) \\
&= - \left(\alpha I_P(\theta) + \bar{\alpha} I_Q(\theta)\right) + I_M(\theta),
\end{aligned}
$$

where $M_\theta(x) \triangleq \alpha P_\theta(x) + \bar{\alpha} Q_\theta(x)$ is the mixture distribution of two sources P_θ and Q_θ.

C. Proof of Lemma 4.1

We prove that the following expressions of Fisher information are equivalent

(1) $\mathbb{E}_{p_\theta}\left[-\frac{\partial^2}{\partial\theta^2} \log p_\theta(X)\right]$

(2) $\frac{\partial^2 D(P_\theta \,\|P_\phi)}{\partial\phi^2}\big|_{\phi=\theta}$

(3) $P_\theta'^{\intercal} \nabla_{\boldsymbol{qq}} D\left(P_\theta \,\|P_\theta\right) P'_\theta$.

"(2) = (1)":

$$
\frac{\partial^2 D\left(P_\theta \,\|P_\phi\right)}{\partial\phi^2}\big|_{\phi=\theta} = \frac{\partial^2}{\partial\phi^2} \mathbb{E}_\theta \left[\log \frac{P_\theta(X)}{P_\phi(X)}\right]\Bigg|_{\phi=\theta} = \frac{\partial^2}{\partial\phi^2} \mathbb{E}_\theta \left[-\log P_\phi(X)\right]\Bigg|_{\phi=\theta}.
$$

"(3) = (2)":

$$
\begin{aligned}
\frac{\partial^2 D\left(P_\theta \,\|P_\phi\right)}{\partial\phi^2}\Bigg|_{\phi=\theta} &= \frac{\partial}{\partial\phi} \left(\frac{\partial D\left(P_\theta \,\|P_\phi\right)}{\partial\phi} \right)\Bigg|_{\phi=\theta} \\
&= \frac{\partial}{\partial\phi} \left(P_\phi'^{\intercal} \nabla_{\boldsymbol{q}} D\left(P_\theta \,\|\boldsymbol{q}\right)\big|_{\boldsymbol{q}=P_\phi} \right)\big|_{\phi=\theta} \\
&= P_\theta''^{\intercal} \nabla_{\boldsymbol{q}} D\left(P_\theta \,\|\boldsymbol{q}\right)\big|_{\boldsymbol{q}=P_\theta} + P_\theta'^{\intercal} \nabla_{\boldsymbol{qq}} D\left(P_\theta \,\|P_\theta\right) P'_\theta \\
&\overset{(a)}{=} P_\theta'^{\intercal} \nabla_{\boldsymbol{qq}} D\left(P_\theta \,\|P_\theta\right) P'_\theta,
\end{aligned}
$$

where (a) is due to the fact that the score function $\nabla_{\boldsymbol{q}} D\left(P_\theta \,\|\boldsymbol{q}\right)\big|_{\boldsymbol{q}=P_\theta}$ is a zero vector.

D. Proof of Lemma 4.3

First, by data process inequality, we have

$$
\frac{1}{(\theta-\phi)^2} \left(\frac{1}{n} D\left(\tilde{\mathbb{P}}_\theta \,\middle\|\tilde{\mathbb{P}}_\phi\right) \right) \le \frac{1}{(\theta-\phi)^2} \left(\frac{1}{n} D\left(\mathbb{P}_\theta \,\|\mathbb{P}_\phi\right) \right) = \frac{1}{(\theta-\phi)^2} \left(\alpha D\left(P_\theta \,\|P_\phi\right) + \bar{\alpha} D\left(Q_\theta \,\|Q_\phi\right) \right),
$$

so

$$
\frac{1}{(\theta-\phi)^2} \left(\frac{1}{n} D\left(\tilde{\mathbb{P}}_\theta \,\middle\|\tilde{\mathbb{P}}_\phi\right) \right) \tag{20}
$$

is uniformly bounded. Together with the equicontinuous assumption, Arzelà-Ascoli theorem applies and hence (20) converges uniformly. Therefore, we have

$$
\lim_{n\to\infty} \lim_{\phi\to\theta} \frac{1}{(\theta-\phi)^2} \left(\frac{1}{n} D\left(\tilde{\mathbb{P}}_\theta \,\middle\|\tilde{\mathbb{P}}_\phi\right) \right) = \lim_{\phi\to\theta} \lim_{n\to\infty} \frac{1}{(\theta-\phi)^2} \left(\frac{1}{n} D\left(\tilde{\mathbb{P}}_\theta \,\middle\|\tilde{\mathbb{P}}_\phi\right) \right),
$$

which establishes Lemma 4.3

E. *Proof of Lemma 4.4*

For notational simplicity, let $G(f_0)$ denote the objective function of (8), f_0^* denote $\operatorname{argmin}_{f_0 \in \mathcal{T}_0} G(f_0)$, and $\tilde{G}(f_0)$ denote the objective function *without* $o(1)$ in (8). We aim to show that

$$\left| G(f_0^*) - \tilde{G}(f_0^{**}) \right| = o(1). \tag{21}$$

To see this, observe that by the local approximation (Taylor expansion), we have for any f_0,

$$\left| G(f_0) - \tilde{G}(f_0) \right| = o(\Delta\theta).$$

Therefore, (21) is proved by combining the following two together

$$\begin{aligned} \tilde{G}(f_0^{**}) &\leq \tilde{G}(f_0^*) \leq G(f_0^*) + o(1), \\ G(f_0^*) &\leq G(f_0^{**}) \leq \tilde{G}(f_0^{**}) + o(1). \end{aligned}$$