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:

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 \( i, \ i \in \{1, \ldots, n\} \)) gives a single random sample \( X_i \). The decision maker estimates the hidden parameter \( \theta \in \Theta \) from the collected samples \( X_1, \ldots, X_n \). 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 \( \{P_{k, \theta} | k = 1, \ldots, K\} \) and the number of sources in each cluster, \( \{n_k | k = 1, \ldots, K\} \), it does not know which distribution each source follows. In words, the sample sequence \( X^n \triangleq (X_1, \ldots, X_n) \) is passed through an unknown permutation \( \pi \) 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 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 \( (\alpha_1, \ldots, \alpha_K) \).

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 $\pi$ 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 $\pi$. 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 $\pi$). 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. $\theta$ 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 $\pi$ is of interest and sharp conditions are set in order to recover $\pi$, which is different from our setting that $\pi$ 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 $i$, $i \in \{1, \ldots, n\}$, gives a single random sample $X_i \in \mathcal{X}$ drawn from one of the $K$ distributions $\{P_{k;\theta} | k = 1, \ldots, K\}$, 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 $\mathcal{X} = \{a_1, \ldots, a_d\}$. Let $n_k$ denote the number of sources that follow distribution $P_{k;\theta}$, for $k = 1, \ldots, K$, and hence $n = \sum_{k=1}^{K} n_k$. The estimator has access to the joint sample $X^n$ and knows $\{n_k | k = 1, \ldots, K\}$ and the parametric family of distributions $\{P_{k;\theta} | k = 1, \ldots, K, \theta \in \Theta\}$. However, it does not know 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 re-order the $n$ samples so that it follows the following product distribution:

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

Meanwhile, the joint observation $X^n$ is perturbed by a $n$-permutation $\pi: \{1, \ldots, n\} \to \{1, \ldots, n\}$ which is not revealed to the estimator. Hence, the estimator needs to consider all $n!$ possible permutation of $\pi$ 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 $\Pi_{X^n}$ as the empirical distribution of $X^n$

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

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

To evaluate an estimator $\phi(X^n)$’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

$$R_{\phi}(\ell) \triangleq \max_{\pi \in \mathcal{S}_n} \mathbb{E}_{X^n \sim \mathbb{P}_\theta}[\ell(\theta, \phi(\pi(X^n)))].$$

When the loss function is the squared error loss, $R_{\phi}(\ell) \equiv \operatorname{MSE}_{\phi}(\theta)$ 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 $\theta \to p_\theta(x)$ is differentiable for all $x \in \mathcal{X}$. Then the Fisher information is defined as

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

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 $\theta$,
2) $\Theta$ is an open interval, and
3) the support of $p_\theta(x)$ does not depend on $\theta$,

so that Fisher information has an alternative form:

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

**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, $\operatorname{MSE}_\theta(\phi) \geq I_\theta(\theta)^{-1}$. 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

$$I_\theta(\theta) = \sum_{k=1}^{K} n_k I_{P_k}(\theta)$$

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 $\hat{\theta}(X^n)$ which depends only on $\Pi_{X^n}$ and

$$R_0(\hat{\theta}) \leq R_0(\hat{\phi}).$$

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 $\Pi_{X^n}$, 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:

$$\hat{\theta}(X^n) \triangleq \frac{1}{n!} \sum_{\sigma \in S_n} \hat{\phi}(\sigma(X^n)).$$

Then we have

$$R_0(\hat{\theta}) = \max_{\pi \in S_n} E_{\pi} \left[ \ell \left( \theta, \frac{1}{n!} \sum_{\sigma \in S_n} \hat{\phi}(\sigma \circ \pi(X^n)) \right) \right] \leq \frac{1}{n!} \sum_{\pi \in S_n} E_{\pi} \left[ \ell \left( \theta, \hat{\phi}(\pi(X^n)) \right) \right] \leq \max_{\pi \in S_n} E_{\pi} \left[ \ell \left( \theta, \hat{\phi}(\pi(X^n)) \right) \right] = R_0(\hat{\phi}),$$

where (a) is due to the fact that

$$\forall \sigma, \pi \circ \sigma \triangleq \{ \sigma \circ \pi | \sigma \in S_n \} = S_n,$$

and (b) is due to the convexity of the loss function $\ell$ 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: $\forall \sigma, \pi \in S_n, \theta(\sigma(X^n)) = \theta(\pi(X^n))$. 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 $\Pi_{X^n}$ as $\bar{P}_\theta$ (note that $X^n \sim \bar{P}_\theta$).

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

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

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 $\alpha_k, k = 1, ..., K$. For the case $K = 2$, we use the notation $(n_1/n, n_2/n) \to (\alpha, \alpha)$, with $\alpha \triangleq 1 - \alpha$. 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 $\theta$. Therefore, the asymptotic Fisher information of $\bar{P}_\theta$ characterizes the asymptote of the non-trivial lower bound of the worst-case MSE.

Theorem 3.2 (Asymptotic Fisher information): Suppose $\bar{P}_\theta$, the distribution of $\Pi_{X^n}$, is “good enough” such that the family of functions of $\phi$,

$$\left\{ \frac{1}{n(\theta - \phi)} D \left( \bar{P}_\theta \bigg\| \bar{P}_\phi \right) n \in \mathbb{N} \right\},$$

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

$$n E_{\bar{P}_\theta} \left[ - \frac{\partial^2 \log M_\theta(X)}{\partial \theta^2} \right] + o(n),$$

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 $\phi, \{ g_n(\phi) | n \in \mathbb{N} \}$, 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 $\delta$ is independent of $n$.

Sketch of Proof: One can start with the exact FI of $\bar{P}_\theta$ 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 $(n \bar{I}_{\theta}(\theta))^{-1}$ is a tight bound. However, in our setting, the underlying distribution of $\Pi_{X^n}$ 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 $Ber(\theta)$ and $Ber(1 - \theta)$ respectively. Then the mixture distribution $M_{\theta} = [\alpha \theta + (1 - \alpha)(1 - \theta), \alpha(1 - \theta) + (1 - \alpha)\theta]$. According to Theorem 3.2, the asymptotic Fisher information normalized by $n$ is given by

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

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 $p$ and $q$ be two d-dimensional probability vectors on $\mathcal{X}$, and $D(p || q)$ be the KL-divergence (with base $e$) between them. Then we use $\nabla_{pp} D(p || q)$ to denote the Hessian matrix of $D(\cdot || q)$ (namely, regarding $D(p || q)$ as a function of $p$ with $q$ being fixed). $\nabla_{qq} D(p || q)$. 
\[ \nabla_{pq} D (p \parallel q) \text{, and } \nabla_{qp} D (p \parallel q) \] are defined in a similar manner. For the ease of presentation, we shall omit the dummy variable:

\[ \begin{align*}
\nabla_{pp} D (p_0 \parallel q_0) \big|_{ij} & \triangleq \frac{\partial^2 D(p \parallel p)}{\partial p_i \partial p_j} \\
\nabla_{qq} D (p_0 \parallel q_0) \big|_{ij} & \triangleq \frac{\partial^2 D(p \parallel q)}{\partial q_i \partial q_j} \\
\nabla_{pq} D (p_0 \parallel q_0) \big|_{ij} & \triangleq \frac{\partial^2 D(p \parallel q)}{\partial p_i \partial q_j} \\
\nabla_{qp} D (p_0 \parallel q_0) \big|_{ij} & \triangleq \frac{\partial^2 D(p \parallel q)}{\partial q_i \partial p_j} 
\end{align*} \]  

In words, the subscript \( p \) (or \( q \)) indicates which argument the derivative operator acts on. In addition, denote \( p'_0 \triangleq \partial p_0 / \partial \theta \) (a \( d \)-dimensional vector).

Next, let us introduce some technical lemmas.

**Lemma 4.1:** Let \( p_0(x) \) satisfy the regularity conditions in Section II-B. Then its Fisher information can be expressed by

\[ I_p(\theta) = p'_0 \top \nabla_{qq} D (p_0 \parallel p_0) p'_0. \]

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

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

\[ D \left( \tilde{P}_\theta \parallel \tilde{P}_\phi \right) = n \left( \mathbb{E} \left[ D (V_{0} \parallel P_{\theta}) \right] + o(n) \right), \]

where \( \alpha \triangleq (1 - \alpha) \).

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

**Lemma 4.3:** If the family of functions of \( \phi \),

\[ \left\{ \frac{1}{n} \frac{\partial}{\partial \phi} D \left( \tilde{P}_\theta \parallel \tilde{P}_\phi \right) \mid n \in \mathbb{N} \right\}, \]

is equicontinuous at \( \phi = \theta \), the following holds at \( \phi = \theta \):

\[ \lim_{n \to \infty} \frac{1}{n} \frac{\partial}{\partial \phi} D \left( \tilde{P}_\theta \parallel \tilde{P}_\phi \right) = \frac{\partial^2}{\partial \theta^2} \left( \lim_{n \to \infty} \frac{1}{n} \frac{\partial}{\partial \phi} D \left( \tilde{P}_\theta \parallel \tilde{P}_\phi \right) \right). \]

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:

\[ \lim_{\Delta \theta \to 0} \frac{1}{\Delta \theta^2} \left( \frac{\partial}{\partial \theta} \left( \frac{1}{n} D \left( \tilde{P}_\theta \parallel \tilde{P}_\phi \right) \right) \right) \]

Part 1 (Change of Variables): Now let us rewrite the constraint by transforming the variables \( V_0, V_1 \) in \( P_X \) to the tangent space

\[ \begin{align*}
V_0 &= P_\theta + f_0 \Delta \theta \\
V_1 &= Q_\theta + f_1 \Delta \theta,
\end{align*} \]

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

\[ \begin{align*}
P_{\theta + \Delta \theta} &= P_\theta + P_\theta' \Delta \theta + o(\Delta \theta) \\
Q_{\theta + \Delta \theta} &= Q_\theta + Q_\theta' \Delta \theta + o(\Delta \theta).
\end{align*} \]

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

\[ \min_{(f_0, f_1) \in F} \alpha \cdot D (P_0 + f_0 \Delta \theta \parallel P_0 + P_0' \Delta \theta + o(\Delta \theta)) / \Delta \theta^2 \]

Part A

\[ + \bar{\alpha} \cdot D (Q_\theta + f_1 \Delta \theta \parallel Q_\theta + Q_\theta' \Delta \theta + o(\Delta \theta)) / \Delta \theta^2, \]

Part B

where for notational simplicity, we define the feasible set \( F \triangleq \left( T_0 \times T_1 \right) \cap T \), with \( T_i \triangleq \{ f_i : \sum_{x \in X} f_i(x) = 0 \} \), for \( i = 0, 1 \), and \( T \triangleq \{ (f_0, f_1) : \alpha f_0 + \bar{\alpha} f_1 = 0 \} \).

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

\[ \left[ f_0, P'_{00} \right] \left[ \nabla_{pp} D (P_0 \parallel P_0) \nabla_{pq} D (P_0 \parallel P_0) \right] \left[ f_0 \parallel P_0' \right] + o(1). \]

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 \( P_0' \) are replaced by \( f_1 \) and \( Q_0' \). Therefore, (6) can be written as a constrained minimization problem of a quadratic function in \( f_0 \) and \( f_1 \) plus some error term:

\[ \min_{(f_0, f_1) \in F} \alpha \cdot \left( \nabla_{pp} D (P_0 \parallel P_0) f_0 + 2 \bar{\alpha} \nabla_{pq} D (P_0 \parallel P_0) \right) f_0 \\
+ \bar{\alpha} \cdot \left( \nabla_{pp} D (Q_\theta \parallel Q_\theta) f_1 + 2 \alpha \nabla_{pq} D (Q_\theta \parallel Q_\theta) \right) f_1 \\
+ \alpha I_{P_0}(\theta) + \bar{\alpha} I_{Q_\theta}(\theta) + o(1), \]

where we use Lemma 4.1 and the fact that

\[ \nabla_{pq} D (P_0 \parallel P_0) = \nabla_{qp} D (P_0 \parallel P_0). \]

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

\[ \min_{f_0 \in F_0} f_0^T \left( \alpha \nabla_{pp} D (P_0 \parallel P_0) + \bar{\alpha} \nabla_{pp} D (Q_\theta \parallel Q_\theta) \right) f_0 \\
+ 2 \bar{\alpha} f_0^T \nabla_{pq} D (P_0 \parallel P_0) - \bar{\alpha} \nabla_{pq} D (Q_\theta \parallel Q_\theta) f_0 \\
+ \alpha I_{P_0}(\theta) + \bar{\alpha} I_{Q_\theta}(\theta) + o(1). \]

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

\[ -\alpha \left( \nabla_{pq} D (P_0 \parallel P_0) P_0' - \nabla_{pq} D (Q_\theta \parallel Q_\theta) Q_\theta' \right)^T f_0^*, \]

where

\[ f_0^* = -\left( \nabla_{pp} D (P_0 \parallel P_0) + \alpha \nabla_{pp} D (Q_\theta \parallel Q_\theta) -1 \nabla_{pq} D (P_0 \parallel P_0) P_0' - \nabla_{pq} D (Q_\theta \parallel Q_\theta) Q_\theta' \right). \]

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 KULLBACK-LEIBLER DIVERGENCE

As the result in [3] suggests, the large deviation exponent of anonymous hypothesis testing (i.e. to test \( \hat{P}_\theta \) vs \( P_\phi \)) is \( 4 \), so it is reasonable to guess that 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

\[
D \left( \frac{\hat{P}_\theta}{P_\phi} \right) = \mathbb{E}_{\hat{P}_\theta} \left[ \log \left( \frac{\hat{P}_\theta(\mu_n)}{P_\phi(\mu_n)} \right) \right], \tag{11}
\]

we notice that the log likelihood ratio (LLR) in (11) is a function of \( \mu_n \), and by the law of large numbers, the density of \( \mu_n \) concentrates. Hence, if the LLR can be approximated by some continuous function (independent of \( n \)), together with the concentration of \( \hat{P}_\theta \), 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 C of the full version).

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

\[
\hat{P}_\theta \triangleq (P_\theta)^{\otimes n_1}(Q_\theta)^{\otimes n_2}, \quad P_\phi \triangleq (P_\phi)^{\otimes n_1}(Q_\phi)^{\otimes n_2},
\]

we have

\[
\log \left( \frac{\hat{P}_\theta(\mu_n)}{P_\phi(\mu_n)} \right) = -nR_0(\mu_n) + nR_1(\mu_n) + o(n),
\]

where

\[
R_0(\mu_n) \triangleq \min_{\mathcal{V}_1, \mathcal{V}_2 \in \mathcal{P}_2} \alpha D(\mathcal{V}_1 \| P_\theta) + \bar{\alpha} D(\mathcal{V}_2 \| Q_\phi),
\]

\[
R_1(\mu_n) \triangleq \min_{\mathcal{U}_1, \mathcal{U}_2 \in \mathcal{P}_2} \alpha D(\mathcal{U}_1 \| P_\phi) + \bar{\alpha} D(\mathcal{U}_2 \| Q_\phi).
\]

**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 \( \ell_1 \) 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 \( \mu_n \):** Next, we argue that \( \mu_n \) does concentrate. In fact, \( \mu_n \) 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. \( P_\theta \) and \( Q_\theta \) respective. By the uniform law of large number (Glivenko-Cantelli theorem), \( \Pi_{X^{n_1}} \rightarrow P_\theta \) and \( \Pi_{X^{n_2}} \rightarrow Q_\theta \) (here for simplicity, we use the \( \ell_1 \) norm as the metric), so \( \mu_n \rightarrow \alpha P_\theta + \bar{\alpha} Q_\theta \) in \( \ell_1 \). 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

\[
\frac{1}{n} D \left( \frac{\hat{P}_\theta}{P_\phi} \right) \leq \frac{1}{n} D (P_\theta \| P_\phi) + \bar{\alpha} D (Q_\theta \| Q_\phi),
\]

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

\[
\frac{1}{n} \mathbb{E}_{\hat{P}_\theta} \left[ \log \left( \frac{\hat{P}_\theta(\mu_n)}{P_\phi(\mu_n)} \right) \right] = \frac{1}{n} \Pr \left\{ \| \mu_n - \mu \| \leq \epsilon \right\} \mathbb{E}_{\hat{P}_\theta} \left[ \log \left( \frac{\hat{P}_\theta(\mu_n)}{P_\phi(\mu_n)} \right) \| \mu_n - \mu \| \leq \epsilon \right] + \frac{1}{n} \Pr \left\{ \| \mu_n - \mu \| > \epsilon \right\} \mathbb{E}_{\hat{P}_\theta} \left[ \log \left( \frac{\hat{P}_\theta(\mu_n)}{P_\phi(\mu_n)} \right) \| \mu_n - \mu \| > \epsilon \right]
\]

\[
\leq \mathbb{E}_{\hat{P}_\theta} \left[ \frac{1}{n} \log \left( \frac{\hat{P}_\theta(\mu_n)}{P_\phi(\mu_n)} \right) \| \mu_n - \mu \| \leq \epsilon \right] + \delta B
\]

\[
\leq \mathbb{E}_{\hat{P}_\theta} \left[ -R_0(\mu_n) + R_1(\mu_n) + o_n(1) \| \mu_n - \mu \| \leq \epsilon \right] + \delta B
\]

\[
\leq -R_0(\mu) + R_1(\mu) + \Delta(\epsilon) + o_n(1) + \delta B,
\]

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 \( \Delta(\epsilon) \rightarrow 0 \) as \( \epsilon \rightarrow 0 \)). Notice that \( R_0(\mu) = 0 \) since one can easily choose \( V_1, V_2 \) to be \( P_\theta, P_\phi \). Also note that both \( \epsilon \) and \( \delta \) 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


A. Proof of Lemma 5.1

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

$$\mathcal{P}(\mu_n) = \sum_{x^n: P(x^n) = c(\mu_n)} \mathcal{P}(x^n) = c(\mu_n) \sum_{\pi \in \mathcal{S}_n} \mathcal{P}(\pi(x^n))$$

for any $x^n$ with type $\mu_n$,

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

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

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

$$\log \left( \frac{\mathcal{P}_0(\mu_n)}{\mathcal{P}_1(\mu_n)} \right) = \log \left( \frac{\sum_{\pi} \mathcal{P}_0(\pi(x^n))}{\sum_{\pi} \mathcal{P}_1(\pi(x^n))} \right)$$

for any $x^n$ with type $\mu_n$.

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

$$x^n = \left( a_1, \ldots, a_1, a_2, \ldots, a_2, \ldots, a_d, \ldots, a_d \right),$$

assuming $\mu_n = [m_1, m_2, \ldots, m_d]^T / n$. Next, for an (unsorted) $x^n$, the $\mathcal{P}_0(x^n)$ is completely determined by the empirical distribution of first $n_1$ samples and the empirical distribution of the rest $n_2$ samples. We use $\left( m_1^{(1)}, \ldots, m_d^{(1)} \right)$ and $\left( m_1^{(2)}, \ldots, m_d^{(2)} \right)$ to denote these (unnormalized) product type of $x^n$. See Figure 2 for illustration. Therefore, summation

$$\sum_{\pi} \mathcal{P}_0(\pi(x^n)) = \sum_{m_1^{(1)}, m_2^{(1)}, \ldots, m_d^{(1)}}, m_1^{(2)}, m_2^{(2)}, \ldots, m_d^{(2)} \mathcal{P}_0(\pi(x^n))$$

over all $\pi(x^n)$ can be written as first summing over product types and then over all permutations which possess same product type:

$$\sum_{\pi} \mathcal{P}_0(\pi(x^n)) = \sum_{m_1^{(1)}, m_2^{(2)}} \left( \sum_{\pi: \pi(x^n) \in m_1^{(1)}, m_2^{(2)}} \mathcal{P}_0(\pi(x^n)) \right)$$

$$= \sum_{m_1^{(1)}, m_2^{(2)}} \left( \left( P_0(a_1)^{m_1^{(1)}} \cdots P_0(a_d)^{m_d^{(1)}} \right) \left( Q_0(a_1)^{m_1^{(2)}} \cdots Q_0(a_d)^{m_d^{(2)}} \right) \right)$$

$$\leq \max_{m_1^{(1)}, m_2^{(2)}} \left( \left( P_0(a_1)^{m_1^{(1)}} \cdots P_0(a_d)^{m_d^{(1)}} \right) \left( Q_0(a_1)^{m_1^{(2)}} \cdots Q_0(a_d)^{m_d^{(2)}} \right) \right).$$

Fig. 1: illustration of the representer

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

\[
(14) \leq n^{2d} \max_{m^{(1)}, m^{(2)}} \left\{ \left( \prod_{i=1}^d P_0(a_i)^{m^{(1)}_i} \right) \left( \prod_{i=1}^d Q_0(a_i)^{m^{(2)}_i} \right) \left( (m^{(1)}_1!) \cdots (m^{(1)}_d!) \right) \left( (m^{(2)}_1!) \cdots (m^{(2)}_d!) \right) \right\}.
\]

Hence we have

\[
\log \left( \sum_{\pi} P_0(\pi(x^n)) \right) = \max_{m^{(1)}, m^{(2)}} \left( \sum_{i=1}^d m^{(1)}_i \log P_0(a_i) + \sum_{i=1}^d m^{(2)}_i \log Q_0(a_i) + \sum_{i=1}^d \log (m^{(1)}_i!) + \sum_{i=1}^d \log (m^{(2)}_i!) \right) + o(n). \tag{15}
\]

Notice that the product type must satisfy the constraints

\[
\begin{align*}
\sum_{i=1}^d m^{(1)}_i &= n_1 \\
\sum_{i=1}^d m^{(2)}_i &= n_2 \\
\forall i \in \{1, \ldots, d\}, m^{(1)}_i + m^{(2)}_i &= m_i.
\end{align*}
\tag{16}
\]

Therefore, by letting

\[
V_1 = [m^{(1)}_1/n_1, \ldots, m^{(1)}_d/n_1]^T, \quad V_2 = [m^{(2)}_1/n_2, \ldots, m^{(2)}_d/n_2]^T,
\]

we see that

\[
\begin{align*}
&\sum_{i=1}^d m^{(1)}_i \log P_0(a_i) + \sum_{i=1}^d \log (m^{(1)}_i!)
\begin{align*}
&= n_1 \left( \sum_{i=1}^d \frac{m^{(1)}_i}{n_1} \log P_0(a_i) + \sum_{i=1}^d \frac{m^{(1)}_i}{n_1} \log (m^{(1)}_i!) \right) - \sum_{i=1}^d m^{(1)}_i + o(n_1)
\begin{align*}
&= n_1 \left( \sum_{i=1}^d \frac{m^{(1)}_i}{n_1} \log P_0(a_i) + \sum_{i=1}^d \frac{m^{(1)}_i}{n_1} \log \left( \frac{n_1}{m^{(1)}_i} \right) \right) - n_1 \log n_1 - n_1 + o(n_1)
\begin{align*}
&= -n_1 D(V_1 \parallel P_0) - n_1 \log n_1 - n_1 + o(n_1), \tag{17}
\end{align*}
\end{align*}
\end{align*}
\]

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

\[
\log (m^{(1)}_i!) = m^{(1)}_i \log m^{(1)}_i - m^{(1)}_i + O(\log m^{(1)}_i).
\]

Similarly,

\[
\begin{align*}
&\sum_{i=1}^d m^{(2)}_i \log Q_0(a_i) + \sum_{i=1}^d \log (m^{(2)}_i!)
\begin{align*}
&= -n_2 D(V_2 \parallel Q_0) - n_2 \log n_2 - n_2 + o(n_2). \tag{18}
\end{align*}
\end{align*}
\]

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

\[
\log \left( \sum_{\pi} P_0(\pi(x^n)) \right) = -n_1 V_1 + n_2 V_2 = -n \mu_n
\]

and

\[
\log \left( \sum_{\pi} P_1(\pi(x^n)) \right) = -n_1 U_1 + n_2 U_2 = -n \mu_n
\]

Plugging back into (13), we obtain

\[
\log \left( \frac{\hat{P}_0(\mu_n)}{\hat{P}_1(\mu_n)} \right) = \log \left( \frac{\sum_{\pi} P_0(\pi(x^n))}{\sum_{\pi} P_1(\pi(x^n))} \right)
\begin{align*}
&= - \left( \min_{V_1, V_2 \in \mathbb{R}^d} n_1 D(V_1 \parallel P_0) + n_2 D(V_2 \parallel Q_0) \right) \left( \min_{U_1, U_2 \in \mathbb{R}^d} n_1 D(U_1 \parallel P_1) + n_2 D(U_2 \parallel Q_1) \right) + o(n).
\end{align*}
\]

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

\[ f_{0}^{**} = - \left( \nabla_{\rho \rho} D (P_0 \| P_0) + \frac{\partial}{\partial \alpha} \nabla_{\rho \rho} D (Q_0 \| Q_0) \right)^{-1} \left( \nabla_{\rho q} D (P_0 \| P_0) P'_0 - \nabla_{\rho q} D (Q_0 \| Q_0) Q'_0 \right) . \]

into

\[ -\alpha \left( \nabla_{\rho q} D (P_0 \| P_0) P'_0 - \nabla_{\rho q} D (Q_0 \| Q_0) Q'_0 \right)^{\top} f_{0}^{**} , \tag{19} \]

we obtain the solution of Theorem 3.2.

First by definition, we have

\[ \nabla_{\rho \rho} D (P_0 \| P_0) = \begin{bmatrix} \frac{\partial^{2} D (p \| P_0)}{\partial p_i \partial p_j} \end{bmatrix}_{p \equiv P_0} = \begin{bmatrix} \frac{1}{P_0(a_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{P_0(a_2)} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{P_0(a_d)} \end{bmatrix} , \]

since

\[ \frac{\partial^{2} D (p \| P_0)}{\partial p_i \partial p_j} = \begin{cases} 0, \text{ if } i \neq j, \\ \frac{1}{P_0(a_i)}, \text{ else}. \end{cases} \]

Similarly,

\[ \nabla_{q q} D (P_0 \| P_0) = \begin{bmatrix} \frac{1}{P_0(a_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{P_0(a_2)} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{P_0(a_d)} \end{bmatrix} , \]

and \[ \nabla_{\rho q} D (P_0 \| P_0) = \nabla_{q q} D (P_0 \| P_0) = \begin{bmatrix} -\frac{1}{P_0(a_1)} & 0 & \cdots & 0 \\ 0 & -\frac{1}{P_0(a_2)} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\frac{1}{P_0(a_d)} \end{bmatrix} . \]

Therefore, \( f_{0}^{**} \) becomes

\[ \begin{bmatrix} -\frac{1}{\alpha} \left( \frac{1}{\alpha P_0(a_1)} + \frac{1}{\alpha Q_0(a_1)} \right) & 0 & \cdots & 0 \\ 0 & -\frac{1}{\alpha P_0(a_2)} + \frac{1}{\alpha Q_0(a_2)} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\frac{1}{\alpha P_0(a_d)} + \frac{1}{\alpha Q_0(a_d)} \end{bmatrix} \]

\[ \begin{bmatrix} \frac{1}{P_0(a_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{P_0(a_2)} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{P_0(a_d)} \end{bmatrix} \]

\[ \begin{bmatrix} P'_0(a_1) \\ P'_0(a_2) \\ \vdots \\ P'_0(a_d) \end{bmatrix} \]

\[ = \begin{bmatrix} \frac{\alpha P_0(a_1) Q_0(a_1)}{\alpha P_0(a_1) + \alpha Q_0(a_1)} & \frac{\alpha P_0(a_2) Q_0(a_2)}{\alpha P_0(a_2) + \alpha Q_0(a_2)} & \cdots & \frac{\alpha P_0(a_d) Q_0(a_d)}{\alpha P_0(a_d) + \alpha Q_0(a_d)} \\ \alpha P_0(a_1) + \alpha Q_0(a_1) & \alpha P_0(a_2) + \alpha Q_0(a_2) & \cdots & \alpha P_0(a_d) + \alpha Q_0(a_d) \end{bmatrix} \]

\[ \begin{bmatrix} \frac{Q'_0(a_1)}{P_0(a_1)} \\ \frac{Q'_0(a_2)}{P_0(a_2)} \\ \vdots \\ \frac{Q'_0(a_d)}{P_0(a_d)} \end{bmatrix} \]

Plugging back to (19), we see that

\[ (19) = - \left( \begin{bmatrix} P'_0(a_1) \\ P'_0(a_2) \\ \vdots \\ P'_0(a_d) \end{bmatrix} - \begin{bmatrix} Q'_0(a_1) \\ Q'_0(a_2) \\ \vdots \\ Q'_0(a_d) \end{bmatrix} \right)^{\top} \begin{bmatrix} \alpha P_0(a_1) Q_0(a_1) \\ \alpha P_0(a_2) Q_0(a_2) \\ \vdots \\ \alpha P_0(a_d) Q_0(a_d) \end{bmatrix} \]

\[ = - \sum_{i=1}^{d} \frac{\alpha \bar{P}_0(a_i) Q_0(a_i)}{\alpha P_0(a_i) + \alpha Q_0(a_i)} \left( \frac{P'_0(a_i)}{P_0(a_i)} - \frac{Q'_0(a_i)}{Q_0(a_i)} \right)^{2} \]

\[ = - \sum_{i=1}^{d} \frac{\alpha \bar{P}_0(a_i) Q_0(a_i) - Q'_0(a_i) P_0(a_i))^{2}}{\alpha P_0(a_i) + \alpha Q_0(a_i)} P_0(a_i) Q_0(a_i) \]
which establishes Lemma 4.3

D. Proof of Lemma 4.3

where (a) is due to the fact that the score function

\[
\frac{\partial^2 D(p_\theta || p_\phi)}{\partial \theta^2} = \frac{\partial^2 D(p_\theta || p_\phi)}{\partial \phi^2}
\]

\[
= \frac{\partial}{\partial \phi} \left( \nabla_q D (p_\theta || q) \right)_{q=p_\phi}
\]

(20)

where (a) is due to the fact that the score function \( \nabla_q D (p_\theta || q) \) is a zero vector.

D. Proof of Lemma 4.3

First, by data process inequality, we have

\[
\frac{1}{n} D \left( \tilde{p}_n \parallel \tilde{p}_n \right) \leq \frac{1}{n} D \left( p_\theta \parallel p_\phi \right) = \frac{1}{n} D \left( p_\theta \parallel p_\phi \right) + \frac{1}{n} D (Q_\theta || Q_\phi),
\]

so

\[
\frac{1}{n} D \left( \tilde{p}_n \parallel \tilde{p}_n \right) \leq \frac{1}{n} D \left( p_\theta \parallel p_\phi \right) + \frac{1}{n} D (Q_\theta || Q_\phi),
\]

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

\[
\lim_{n \to \infty} \lim_{\phi \to \theta} \frac{1}{n} D \left( \tilde{p}_n \parallel \tilde{p}_n \right) = \lim_{\phi \to \theta} \lim_{n \to \infty} \frac{1}{n} D \left( \tilde{p}_n \parallel \tilde{p}_n \right),
\]

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 \( \arg\min_{f_0 \in T_0} G(f_0) \), and $\tilde{G}(f_0)$ denote the objective function \textit{without} $o(1)$ in (8). We aim to show that

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

(21)

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

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

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

$$\tilde{G}(f_0^*) \leq \tilde{G}(f_0^*) \leq G(f_0^*) + o(1),$$

$$G(f_0^*) \leq G(f_0^*) \leq \tilde{G}(f_0^*) + o(1).$$