The Metagenomic Binning Problem: Clustering Markov Sequences

Grant Greenberg
Electrical and Computer Engineering Department
University of Illinois at Urbana-Champaign
Champaign, IL, USA
gcgreen2@illinois.edu

Ilan Shomorony
Electrical and Computer Engineering Department
University of Illinois at Urbana-Champaign
Champaign, IL, USA
ilans@illinois.edu

***Abstract*—The goal of metagenomics is to study the composition of microbial communities, typically using high-throughput shotgun sequencing. In the metagenomic binning problem, we observe random substrings (called contigs) from a mixture of genomes and want to cluster them according to their genome of origin. Based on the empirical observation that genomes of different bacterial species can be distinguished based on their *tetranucleotide frequencies*, we model this task as the problem of clustering N sequences generated by M distinct Markov processes, where $M \ll N$. Utilizing the large-deviation principle for Markov processes, we establish the information-theoretic limit for perfect binning. Specifically, we show that the length of the contigs must scale with the inverse of the Chernoff Information between the two most similar species. Our result also implies that contigs should be binned using the conditional relative entropy as a measure of distance, as opposed to the Euclidean distance often used in practice.**

I. Introduction

In the last decade, advances in high-throughput DNA sequencing technologies have allowed a vast amount of genomic data to be generated. Countless tasks such as genome assembly, RNA quantification, and genome-wide association studies have become a reality, opening up exciting new research directions within biology and medicine [1].

Significant attention has recently been given to the analysis of the human microbiome through *metagenomics* [2]. In metagenomics, a sample is taken from a microbial community, such as the human gut. The genetic material in the sample is then sequenced and analyzed to determine the microbial composition of the community [3]. Recent research, including the Human Microbiome Project [4], has shown that the composition of the microbiome is a "snapshot" into an individual's overall health, providing great potential for personalized medicine.

Full reconstruction of the genomes in a metagenomic sample is generally infeasible due to insufficient coverage and high similarity across species [5]. In the typical analysis pipeline, the millions of reads obtained via high-throughput sequencing are used to create a much smaller number of contiguous sequences, known as *contigs*, by merging reads with large overlaps [6]. The set of resulting contigs typically make up only a small fraction of the full genomes of all species present in the sample and have no significant overlaps with each other.

Metagenomic binning is concerned with the following question: is it possible to group the resulting contigs based on the genome from which they were derived? Somewhat surprisingly, it has been shown that contigs belonging to the same species typically have similar sequence compositions. Specfically, it was empirically verified that the distribution of four-letter strings (e.g. $AGCG$) remains relatively constant across an entire bacterial genome [7], [8]. Hence one can compute for each contig the *tetranucleotide frequency* (TNF) vector, and group together contigs with "similar" TNF vectors. Provided that the underlying TNF distributions are distinct enough, metagenomic binning can thus be performed. Based on this idea[1], many different algorithms and software packages have been proposed to perform metagenomic binning [5], [6], [9]. Other algorithms use a supervised learning approach by comparing the sequence composition of reads to a database of known bacterial genomes [10]–[13], or through direct alignment to said database [14], [15].

The fact that the distribution of four-symbol strings is consistent throughout a given genome motivates the modeling of each genome as a *third*-order Markov process. Hence we assume that a contig is generated by one out of M distinct, *unknown* Markov processes $p_1, \dots, p_M$ with equal probability, where each p_k corresponds to a certain species. In order to study the fundamental limits of this problem, we assume all N contigs have length L, and consider an asymptotic regime where $N \to \infty$ and the contig length grows slowly with the number of contigs (specifically we set $L = \bar{L} \log N$). Our goal is to characterize how large $\bar{L}$ needs to be in order to allow perfect binning with high probability.

To obtain our main result, we establish the equivalent of the Chernoff Information [16, Chapter 11.9] for Markov processes, which gives the error exponent for the Bayesian error probability when testing between two known Markov processes. This result, combined with a scheme to estimate the M Markov distributions, allows us to show that perfect binning is possible if and only if

$$\bar{L} > \frac{1}{\min_{k,\ell} C(p_k, p_\ell)},$$

where $C(p_k, p_\ell)$ is the Chernoff Information between p_k and p_ℓ. To estimate the unknown distributions, we consider build-

[1]Usually, in addition to the TNF vector, the read coverage of each contig is used as another feature to help with the clustering. However, in this paper, we focus on using the TNF alone.

ing a graph where contigs whose empirical distributions are close are connected. We then show that, with high probability, M large cliques can be found, which can be used to find estimates $\tilde{p}_k$, $k = 1, ..., M$, of the Markov distributions. Each contig $\mathbf{x}$ is then then placed in bin k given by

$$\arg\min_{k\in\{1,\dots,M\}} D_c\left(\hat{p}_{\mathbf{x}}\|\tilde{p}_k\right)$$

where $\hat{p}_{\mathbf{x}}$ is the empirical 4-symbol distribution of $\mathbf{x}$ and $D_c(\cdot\|\cdot)$ is the conditional relative entropy [16, Definition 2.65].

Our main result suggests that the optimal way to bin metagenomic contigs is to estimate the underlying TNF distributions and then bin contigs using the conditional relative entropy as a metric, as opposed to the commonly used Euclidean distance. By simulating contigs from real bacterial genomes, we show that this metric can lead to lower binning error probabilities.

The paper is organized as follows. In Section II we describe the problem formulation in detail and state our main result. In Section III we describe our achievability scheme and the main technical ingredients used to prove it, and in Section IV we describe the converse argument. In Section V we provide preliminary simulation results, and we conclude the paper with a discussion in Section VI.

II. Problem Statement

As shown in [7], the distribution of tetranucleotides (four-letter strings), tends to be stationary across an individual bacterial genome. Hence, it is natural to assume that each of the species in our sample corresponds to a distribution over all possible tetranucleotides[2] $\{AAAA, AAAC, ..., TTTT\}$.

Let $\mathcal{P}$ be the $|\mathcal{X}|^4$-dimensional simplex, where $\mathcal{X} = \{A, C, G, T\}$. Notice that not all distributions in $\mathcal{P}$ are valid tetranucleotide distributions, as the tetranucleotides in a sequence overlap with each other. Let $\widetilde{\mathcal{P}}$ be the set of all $p \in \mathcal{P}$ with $p(\mathbf{c}) > 0$, $\forall\, \mathbf{c} \in \mathcal{X}^4$, which satisfy for all $\mathbf{a} \in \mathcal{X}^3$

$$\sum_{b\in\mathcal{X}} p(\mathbf{a}b) = \sum_{b\in\mathcal{X}} p(b\mathbf{a}). \tag{1}$$

Condition (1) ensures that a given $p \in \widetilde{\mathcal{P}}$ corresponds to the tetranucleotide distribution of a specific, *stationary*, third-order Markov chain. More precisely, we can let the induced distribution over 3-letter strings be

$$p(\mathbf{a}) = \sum_{b\in\mathcal{X}} p(\mathbf{a}b). \tag{2}$$

This uniquely determines a stationary Markov process with initial state distributed as (2) and transition probabilities (i.e. conditional distribution)

$$p(b|\mathbf{a}) = \frac{p(\mathbf{a}b)}{p^{(3)}(\mathbf{a})}. \tag{3}$$

Hence we will model each species in the sample using a distribution $p_k \in \widetilde{\mathcal{P}}$. Notice that the constraint that $p_k(\mathbf{c}) > 0$ for any $\mathbf{c} \in \mathcal{X}^4$ guarantees that the resulting Markov processes are irreducible.

[2]In practical approaches, *reverse-complementary* tetranuclotides such as $ACAG$ and $CTGT$ are treated as the same tetranucleotide, but we ignore that fact for the sake of simplicity.

A. Metagenomic Binning Problem

We assume that we have M species in our sample (for a known M). Each species is modeled by a stationary third-order Markov process defined by $p_k \in \widetilde{\mathcal{P}}$, for $k = 1, ..., M$. From this genomic mixture, we observe a set of N realizations $\mathcal{C} = \{\mathbf{x}_i\}_{i=1}^N$, which we call *contigs*. Each $\mathbf{x} \in \mathcal{C}$ is generated independently by first choosing a species $k \in \{1, ..., M\}$ with prior probability π_k, and then generating a length-L sequence according to p_k. For each k, let $\mathcal{C}_k$ be the set of contigs generated according to p_k. We wish to determine $\mathcal{C}_k,\ k = 1, \dots, M$, i.e. which contigs originated from the same genome.

We point out that in real metagenomic experiments, the *coverage depth*, that is, the expected number of contigs containing a specific nucleotide from one of the M genomes, is low [17]. Hence, contigs will have no overlap with high probability, allowing us to model them as independent realizations of the different Markov processes in the sample.

B. Perfect Binning

The goal of the metagenomic binning problem is to cluster the N contigs into M "bins", where each bin k corresponds to a unique species with distribution p_k. More precisely, the goal is to find a decision rule $\delta_L : \mathcal{X}^L \to \{1, \dots, M\}$ (using notation from [18]) which correctly maps each contig to its respective genome bin.

Perfect binning would be achieved if for every contig $\mathbf{x}$, $\delta(\mathbf{x})$ chooses the label of the distribution from which it was generated. However, we have the added difficulty that the distributions are unknown. As a result, we can only require the decision rule to be correct up to a consistent relabeling of species indices. Hence the error event for a decision rule δ is

$$\mathcal{E}_\delta = \{\exists \mathbf{x} \in \mathcal{C}_k, \mathbf{y} \in \mathcal{C}_\ell, k \neq \ell : \delta(\mathbf{x}) = \delta(\mathbf{y})\}. \tag{4}$$

We would like to know under what circumstances we can perfectly bin all N contigs. In order to study the information-theoretic limits of this problem, we analyze an asymptotic regime, similar to [19], in which $N \to \infty$ and

$$L = \bar{L} \log N \tag{5}$$

where $\bar{L}$ is the "normalized contig length". This scaling forces the contig length to be small compared to the number of contigs and, as we will show, is a meaningful scaling for the asymptotic problem we consider. Intuitively, a larger value of $\bar{L}$ should allow one to bin a contig with higher accuracy. This asymptotic regime allows us to define when species are resolvable as follows:

Definition 1. *The M species with distributions $\{p_k\}_{k=1}^M$ are **resolvable** if there exists a sequence of decision rules $\{\delta_L\}$ such that $Pr(\mathcal{E}_{\delta_L}) \to 0$ as $N \to \infty$.*

C. Main Result

Interestingly, the fundamental limit of resolvability relies on the Chernoff Information, which we define next.

Definition 2. *For two Markov processes p_k and p_ℓ, the **Chernoff Information** between p_k and p_ℓ is given by*

$$C(p_k, p_\ell) = D_c\left(p^* \| p_k\right) = D_c\left(p^* \| p_\ell\right) \tag{6}$$

where D_c is the conditional relative entropy [16, Def. 2.65], and p^* is the solution to the following minimization problem.

$$\begin{aligned} p^* = &\arg\min_{p \in \widetilde{\mathcal{P}}} D_c\left(p \| p_k\right) \\ &\text{s.t. } D_c\left(p \| p_k\right) = D_c\left(p \| p_\ell\right) \end{aligned} \tag{7}$$

The Chernoff Information can be thought of as a measure of distance to either distribution. Our main result establishes that the minimum normalized contig length, $\bar{L}$, required for *resolvability* depends exclusively on the minimum Chernoff Information between species distributions.

Theorem 1. *Let $C_{\min} = \min_{k,\ell} C(p_k, p_\ell)$. The species' distributions $\{p_k\}_{k=1}^M$ are resolvable if and only if*

$$\boxed{\bar{L} > \frac{1}{C_{\min}}} \tag{8}$$

Intuitively, this means that the contig length must be large enough to distinguish between the two closest distributions.

III. Achievability

The achievability proof of Theorem 1 is first shown in the form of an algorithm so as to highlight the algorithmic nature of metagenomic binning. Given a contig $\mathbf{x} \in \mathcal{C}$, we define the empirical fourth-order distribution of $\mathbf{x}$ as $\hat{p}_{\mathbf{x}}$ and we use d as the ℓ_1 distance between distributions, i.e. $d(p, q) = \sum_{\mathbf{c} \in \mathcal{X}^4} |p(\mathbf{c}) - q(\mathbf{c})|$.

Algorithm 1: Binning Contigs

Result: Decision Rule $\delta(\mathbf{x})$
Input: Contigs $\mathcal{C}$, Parameter $\alpha \in (0, 1)$
begin
  $\mathcal{D} \longleftarrow$ sort in ascending order $\{d(\hat{p}_{\mathbf{x}}, \hat{p}_{\mathbf{y}}) : \mathbf{x}, \mathbf{y} \in \mathcal{C}\}$
  for ϵ *in* $\mathcal{D}$
    $\mathcal{G}_\epsilon \longleftarrow (V = \mathcal{C}, E_\epsilon = \{(\mathbf{x}, \mathbf{y}) : d(\hat{p}_{\mathbf{x}}, \hat{p}_{\mathbf{y}}) \le \epsilon\})$
    if $\mathcal{G}_\epsilon$ has cliques $\{\mathcal{K}_k\}_{k=1}^M, |\mathcal{K}_k| \ge (1-\alpha)\frac{N}{M}$
      for $k \longleftarrow 1$ to M
        $\tilde{p}_k \longleftarrow \frac{1}{|\mathcal{K}_k|} \sum_{\mathbf{x} \in \mathcal{K}_k} \hat{p}_{\mathbf{x}}$
      break
  for $\mathbf{x} \in \mathcal{C}$
    $\delta(\mathbf{x}) \longleftarrow \arg\min_{k \in \{1,\dots,M\}} D_c\left(\hat{p}_{\mathbf{x}} \| \tilde{p}_k\right)$

The algorithm first estimates the species distributions by averaging the empirical distributions of the contigs in each clique, then it bins the contigs based on the estimates. Note that the algorithm as described is not computationally efficient (specifically, finding large cliques), and is used only to establish the achievability of Theorem 1.

A. Estimating Distributions

Recall that $\mathcal{C}_k$ is the set of contigs generated by p_k. We expect the empirical distribution of the majority of contigs in $\mathcal{C}_k$ to be near p_k. To identify those "good" contigs, let

$$\mathcal{C}_{k,\epsilon} = \{\mathbf{x} \in \mathcal{C}_k : d(\hat{p}_{\mathbf{x}}, p_k) \le \epsilon\}.$$

To prove that the distribution estimates $\{\tilde{p}_k\}_{k=1}^M$ are close to the true distributions $\{p_k\}_{k=1}^M$ (after proper reindexing), we will first show that each clique the algorithm identifies is "pure" in the sense that it only contains "good" contigs from a single species. We will let $d_{\min} \triangleq \min_{k \neq \ell} d(p_k, p_\ell)$ be the minimum ℓ_1 distance between any pair of the M species distributions.

Lemma 1. *If $\mathcal{K}_k$ is a clique in $\mathcal{G}_\epsilon$ for $\epsilon < \frac{d_{\min}}{2}$, then*

$$\sum_{\ell=1}^{M} \mathbb{1}\{\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon/2} \neq \emptyset\} \le 1, \tag{9}$$

Lemma 1 establishes that, if Algorithm 1 finds M cliques of size $(1-\alpha)\frac{N}{M}$ in $\mathcal{G}_\epsilon$ for $\epsilon < \frac{d_{\min}}{2}$, then each clique contains contigs from at most one $\mathcal{C}_{\ell,\epsilon/2}$, $\ell = 1, \dots, M$. In order to establish that these M cliques will exist in $\mathcal{G}_\epsilon$, we use the following lemma, which essentially says that a large fraction of the contigs will be close to their respective generating distributions.

Lemma 2. *For $\epsilon > 0$, $k \in \{1, ..., M\}$, and N large enough,*

$$\Pr\left(|\mathcal{C}_{k,\epsilon/2}| < (1-\alpha)\frac{N}{M}\right) \le \frac{1}{\alpha} e^{-\gamma \alpha^2 L}, \tag{10}$$

where γ is a positive constant.

Fixing $\epsilon < \frac{d_{\min}}{2}$, Lemma 2 guarantees that for a reasonably chosen α, as long as N is large enough, we will have $|\mathcal{C}_{k,\epsilon/2}| \ge (1-\alpha)\frac{N}{M}$ for all $k = 1, ..., M$. Moreover, by the triangle inequality, any two contigs $\mathbf{x}, \mathbf{y} \in \mathcal{C}_{k,\epsilon/2}$ will be at a distance ϵ or less of each other and will thus have an edge between them. Hence, $\mathcal{C}_{k,\epsilon/2}$ forms a clique in $\mathcal{G}_\epsilon$.

Notice that $d_{\min}$ is not known, so the algorithm cannot restrict its search to $\epsilon < \frac{d_{\min}}{2}$. However, since the algorithm considers different values of ϵ in increasing order, for some $\epsilon < \frac{d_{\min}}{2}$, M cliques of size $(1-\alpha)\frac{N}{M}$ will exist with probability $1 - o(1)$. Lemma 1 will then guarantee that any cliques $\mathcal{K}_1, ..., \mathcal{K}_M$ that are found will be pure.

Consider a clique $\mathcal{K}_k$ and let ℓ be such that $\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon/2} \neq \emptyset$. By Lemma 2, the fraction of "good" contigs in $\mathcal{K}_k$ will be

$$\frac{|\mathcal{K}_k \cap \mathcal{C}_{\ell,\epsilon/2}|}{|\mathcal{K}_k|} \ge 1 - \frac{N - M \cdot (1-\alpha)\frac{N}{M}}{(1-\alpha)\frac{N}{M}} = 1 - \frac{\alpha M}{1-\alpha} \tag{11}$$

with probability $1 - o(1)$. The lower bound results from dividing the maximum number of contigs *not* in any clique by the minimum number of contigs in $\mathcal{K}_k$. If we set $\alpha = \frac{1}{\log L}$, (11) converges to 1 and (10) converges to 0 as $N \to \infty$. Thus, with high probability, a vanishing fraction of the contigs in $\mathcal{K}_k$ does not belong to $\mathcal{C}_{\ell,\epsilon/2}$. Since distribution vectors are bounded, their impact on $\tilde{p}_k$ also vanishes, and we conclude that the distribution estimate $\tilde{p}_k = \frac{1}{|\mathcal{K}_k|} \sum_{\mathbf{x} \in \mathcal{K}_k} \hat{p}_{\mathbf{x}} \to p_\ell$ as $N \to \infty$.

B. Binning Contigs

In Subsection III-A, we established that we can construct estimates of the underlying distributions $\{p_k\}_{k=1}^M$ that are arbitrarily accurate as $N \to \infty$. Next we show that, binning the contigs based on the conditional relative entropy using the underlying distributions achieves (8) in the limit.

Consider the hypothesis test between two Markov processes p_k and p_ℓ (i.e. true distributions, not estimates). Given prior probabilities π_k and π_ℓ, the Bayesian probability of error is

$$\pi_k \Pr(\text{choose } \ell | k \text{ true}) + \pi_\ell \Pr(\text{choose } k | \ell \text{ true})$$

for the decision rule on a contig generated by either p_k or p_ℓ.

Theorem 2. *Let $\mathcal{E}_{k,\ell}^{(L)}$ be the error event for the decision rule which minimizes the Bayesian probability of error. Then*

$$\lim_{L\to\infty} \frac{1}{L} \log \Pr(\mathcal{E}_{k,\ell}^{(L)}) = -C(p_k, p_\ell), \tag{12}$$

i.e., $C(p_k, p_\ell)$ is the optimal error exponent.

The proof of Theorems 2 is given in Section VII. For a given contig, the last step of Algorithm 1 can be thought of as $M-1$ binary hypothesis tests between the true distribution and each of the remaining distributions. Thus, we will use Theorem 2 to bound the overall error probability, $\Pr(\mathcal{E}_{\delta_L})$, by considering the two closest distributions.

$$\begin{aligned}
\Pr(\mathcal{E}_{\delta_L}) &\le \sum_{k=1}^{M} \sum_{\mathbf{x}\in\mathcal{C}_k} \pi_k \sum_{\ell\neq k} \Pr(\mathcal{E}_{k,\ell}^{(L)}) && (13)\\
&\le MN(M-1) \max_{k\neq\ell} \Pr(\mathcal{E}_{k,\ell}^{(L)}) && (14)\\
&\le M^2 2^{L\left(1/\bar{L} + \max_{k\neq\ell}(1/L)\log\Pr(\mathcal{E}_{k,\ell}^{(L)})\right)} && (15)
\end{aligned}$$

where (13) follows from the union bound. By Theorem 2,

$$\max_{k\neq\ell} \frac{1}{L} \log \Pr(\mathcal{E}_{k,\ell}^{(L)}) \to -\min_{k\neq\ell} C(p_k, p_\ell) = -C_{\min}$$

as $N \to \infty$. Hence, if $\bar{L} > \frac{1}{C_{\min}}$, $\Pr(\mathcal{E}_{\delta_L}) \to 0$. This concludes the achievability proof of Theorem 1. A continuity argument can be used to show the exponent is identical for the case when instead you have estimates that converge to the true distributions.

IV. Converse

Without loss of generality, let p_1 and p_2 be such that $C_{\min} = C(p_1, p_2)$. Given the decision rule δ_L and a contig $\mathbf{x} \in \mathcal{C}$, let

$$\widetilde{\mathcal{E}}_{1,2,\mathbf{x}} = \{\mathbf{x} \in \mathcal{C}_1, \delta_L(\mathbf{x}) \neq 1\} \cup \{\mathbf{x} \in \mathcal{C}_2, \delta_L(\mathbf{x}) \neq 2\}$$

i.e. the event that $\mathbf{x}$ was generated by either p_1 or p_2 *and* incorrectly binned. Note that $\Pr(\widetilde{\mathcal{E}}_{1,2,\mathbf{x}}) \ge (\pi_1+\pi_2)\Pr(\mathcal{E}_{1,2})$. Then

$$\begin{aligned}
\Pr(\mathcal{E}_{\delta_L}) &\ge \Pr\left(\bigcup_{i=1}^{N} \widetilde{\mathcal{E}}_{1,2,\mathbf{x}_i}\right) = 1 - (1 - \Pr(\widetilde{\mathcal{E}}_{1,2,\mathbf{x}_1}))^N\\
&\ge 1 - \left[(1 - (\pi_1+\pi_2)\Pr(\mathcal{E}_{1,2}))^{1/\Pr(\mathcal{E}_{1,2})}\right]^{N\Pr(\mathcal{E}_{1,2})}\\
&\ge 1 - e^{-(\pi_1+\pi_2)N\Pr(\mathcal{E}_{1,2})} && (16)
\end{aligned}$$

where (16) follows from the bound $(1-ap)^{1/p} \le e^{-a}$ for $p \in [0,1], a \in \mathbb{R}$. We see that, if $N\Pr(\mathcal{E}_{1,2}) \not\to 0$, then $\Pr(\mathcal{E}) \not\to 0$. Since

$$N\Pr(\mathcal{E}_{1,2}) = 2^{L\left(1/\bar{L} + (1/L)\log\Pr(\mathcal{E}_{1,2})\right)},$$

then by Theorem 2, $N\Pr(\mathcal{E}_{1,2}) \not\to 0$ when $\bar{L} \le \frac{1}{C_{\min}}$. This concludes the converse proof for Theorem 1.

V. Experimental Results

From the point of view of practical metagenomic binning algorithms, our main result suggests that:

1) the conditional relative entropy is a good metric for binning contigs,
2) the Chernoff information can be used as a measure of how difficult it is to distinguish two species.

In this section, we provide preliminary empirical evidence of these claims. To this end, we utilized several previously sequenced and assembled bacterial genomes, available at NCBI [20]. For each bacterial species k, we computed its fourth-order distribution p_k (i.e., the overall tetranucleotide frequency vector). We were able to simulate contigs of a desired length L by sampling all possible length-L substrings from the genome with a sliding window. For each experiment, we assume $N = 10^6$ for concreteness (thus $L = \bar{L}\log 10^6$), but the results are not significantly affected by this choice.

In order to verify the usefulness of the conditional relative entropy and compare it to the Euclidean distance (used in state-of-the-art tools such as [5], [9]), we considered the following experiment: we generated random contigs from a species p_1 and then tested whether it was closer to species p_1 or to another species p_2 based on both the Euclidean distance and the conditional relative entropy. In Figure 1a, the conditional relative entropy metric[3] consistently outperforms the Euclidean metric as we vary $\bar{L}$ in the test between the species Alistipes Obesi and Megamonas Funiformi.

We performed this experiment for 45 different choices of pairs of bacterial genomes from NCBI. For each pair (k, ℓ), we considered a fixed normalized contig length given by $\bar{L} = C(p_k, p_\ell)^{-1}$. As shown in Figure 1b, the conditional divergence improves the error compared to the Euclidean distance in almost 90% of cases.

Theorem 1 implies that the inverse of the Chernoff Information characterizes how long the contigs need to be in order for two species to be reliably distinguishable. In order to verify that, we calculated $\bar{L}_{5\%}$, the minimum normalized contig length required to guarantee a 5% error rate in the Bayesian hypothesis test between p_k and p_ℓ with equal priors. In Figure 1c, we plot $\bar{L}_{5\%}$ vs $C^{-1}(p_k, p_\ell)$ for many such pairs and observe a roughly linear relationship between these two quantities. Such a linear relationship agrees with the relationship suggested by Theorem 1. Moreover, it provides support to the claim that $C^{-1}(p_k, p_\ell)$ is a measure of how difficult it is to distinguish contigs from two species based on tetranucleotide frequencies.

[3] $D_c(\cdot||\cdot)$ is not technically a metric as it is not symmetric.

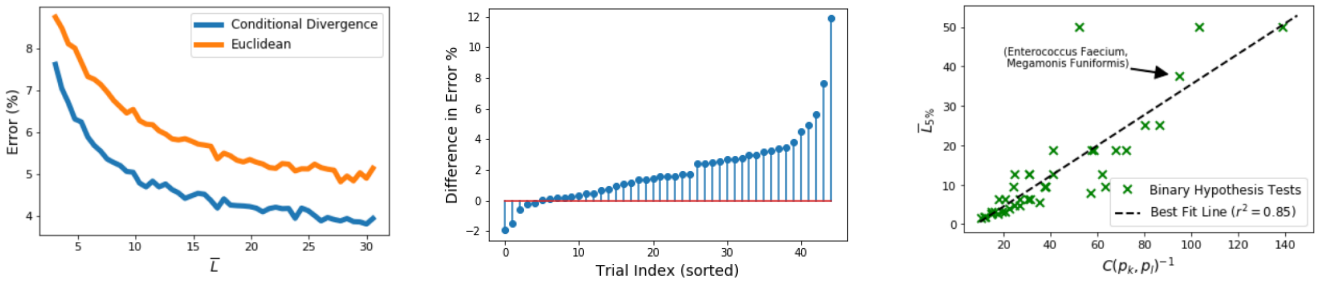

Fig. 1: **(a)** Comparison of conditional divergence and Euclidean distance for a hypothesis test between Alistipes Obesi and Megamonas Funiformis; **(b)** Normalized contig length required for 5% error ($\bar{L}_{5\%}$) vs the inverse of the Chernoff information for several pairs of species; **(c)** The difference between error percentages for Euclidean distance and conditional divergence with $\bar{L} = C(p_k, p_\ell)^{-1}$

.

VI. Discussion

In this paper, we modeled the metagenomic binning problem as the problem of clustering sequences generated by distinct Markov processes. While overly simplistic, this model allowed us to establish the Chernoff Information as a measure of how easy it is to distinguish contigs generated by two species.

The algorithm used to prove the achievability suggests that a good "metric" for binning is the conditional relative entropy between a contig and an estimate of a species TNF. Through experiments, we provided preliminary evidence that this metric often outperforms the Euclidean metric in the problem of assigning a contig to a species bin. However, this assumes knowledge of the overall TNF of a genome, which is not known in practical settings. Therefore a natural direction for future investigation is how to efficiently estimate the TNF distribution for the species present in the sample.

Furthermore, it is unclear whether estimating the underlying TNF distributions is necessary to achieve the fundamental limit. Alternatively, one could consider an approach that directly clusters the contigs based on their pairwise distances or based on a graph obtained by thresholding the distances (similar to our $\mathcal{G}_\epsilon$). We point out that, for such a graph, the problem becomes a community detection problem, and bears similarities with the stochastic block model [21], since for each species there is a given probability that an edge is placed among two of its contigs, and for each pair of distinct species, there is another probability that an edge is placed between their contigs. These probabilities would in general depend on the species TNF distributions (or the Markov processes generating the contigs). Notice that, unlike in the standard stochastic block model, here the placing of the edges would not be independent events.

Finally, we point out that in most approaches to metagenomic binning, the read coverage, or *abundance*, is used to compare contigs in addition to the TNF. The read coverage of a contig is essentially the average number of reads that cover any given base in the contig. Intuitively, this number is proportional to the abundance of the corresponding species in the mixture. Hence, one expects contigs from the same species to have similar read coverages, which can be used to improve metagenomic binning. Another direction for future work is thus to consider the metagenomic binning problem where the different species have different abundances and, for each contig, one observes a read coverage value that is related to the species abundance.

Acknowledgements

The idea of modeling the species in a metagenomic sample as distinct Markov processes and utilizing the large deviation principle to characterize their separability came from discussions with Xingyu Jin, Vasilis Ntranos, and David Tse.

References

[1] M. Land *et al.*, "Insights from 20years of bacterial genome sequencing," *Functional & Integrative Genomics*, vol. 15, no. 2, pp. 141–161, Mar 2015. [Online]. Available: https://doi.org/10.1007/s10142-015-0433-4

[2] F. Breitwieser, J. Lu, and S. Salzberg, "A review of methods and databases for metagenomic classification and assembly," *Briefings in bioinformatics*, 2017.

[3] K. Chen and L. Pachter, "Bioinformatics for whole-genome shotgun sequencing of microbial communities," *PLOS Computational Biology*, vol. 1, no. 2, 07 2005. [Online]. Available: https://doi.org/10.1371/journal.pcbi.0010024

[4] Le Chatelier *et al.*, "Richness of human gut microbiome correlates with metabolic markers," *Nature*, vol. 500, pp. 541–546, 2013.

[5] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," *PeerJ*, vol. 3, p. e1165, 2015.

[6] J. N. Nissen *et al.*, "Binning microbial genomes using deep learning," *bioRxiv*, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/12/10/490078

[7] P. A. Noble *et al.*, "Tetranucleotide frequencies in microbial genomes," *Electrophoresis*, vol. 19, no. 4, pp. 528–535, Apr 1998.

[8] J. Mrzek, "Phylogenetic signals in dna composition: Limitations and prospects," *Molecular Biology and Evolution*, vol. 26, no. 5, pp. 1163–1169, 2009. [Online]. Available: http://dx.doi.org/10.1093/molbev/msp032

[9] Y. Y. Lu *et al.*, "COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge," *Bioinformatics*, vol. 33, no. 6, pp. 791–798, 2017.

[10] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome Biology*, vol. 15, no. 3, p. R46, Mar 2014. [Online]. Available: https://doi.org/10.1186/gb-2014-15-3-r46

[11] Y. Luo, Y. W. Yu, J. Zeng, B. Berger, and J. Peng, "Metagenomic binning through low-density hashing," *Bioinformatics*, vol. 35, no. 2, pp. 219–226, 07 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty611

[12] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, "Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers," *BMC Genomics*, vol. 16, no. 1, p. 236, 2015. [Online]. Available: https://doi.org/10.1186/s12864-015-1419-2
[13] L. Schaeffer *et al.*, "Pseudoalignment for metagenomic read assignment," *Bioinformatics*, vol. 33, no. 14, pp. 2082–2088, 02 2017. [Online]. Available: https://doi.org/10.1093/bioinformatics/btx106
[14] A. Brady and S. Salzberg, "Phymmbl expanded: confidence scores, custom databases, parallelization and more," *Nature methods*, vol. 8, no. 5, pp. 367–367, May 2011, 21527926[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21527926
[15] D. H. Huson *et al.*, "Integrative analysis of environmental sequences using megan4," *Genome Research*, vol. 21, pp. 1552–1560, 2011.
[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006.
[17] S. Nurk, . , A. Korobeynikov, and P. Pevzner, "Metaspades: A new versatile metagenomic assembler," *Genome Research*, vol. 27, no. 5, pp. 824–834, 5 2017.
[18] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*. Cambridge University Press, 2018.
[19] A. S. Motahari, G. Bresler, and D. Tse, "Information theory of DNA sequencing," *CoRR*, vol. abs/1203.6233, 2012. [Online]. Available: http://arxiv.org/abs/1203.6233
[20] (1988) National center for biotechnology information. [Online]. Available: https://www.ncbi.nlm.nih.gov/
[21] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2015.
[22] M. Vidyasagar, "An elementary derivation of the large deviation rate function for finite state markov chains," in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC)*, Dec 2009, pp. 1599–1606.
[23] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theor.*, vol. 35, no. 2, pp. 401–408, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/18.32134

VII. Appendix

A. Proof of Lemma 1

Suppose by contradiction that $\mathbf{x}, \mathbf{y} \in \mathcal{K}_j$, $\mathbf{x} \in \mathcal{C}_{k,\epsilon/2}$ and $\mathbf{y} \in \mathcal{C}_{\ell,\epsilon/2}$, for $k \neq \ell$. Then $d(\mathbf{x}, \mathbf{y}) < \epsilon$, and we have

$$\begin{aligned} d(p_k, p_\ell) &\leq d(p_k, \hat{p}_\mathbf{x}) + d(\hat{p}_\mathbf{x}, p_\ell) \\ &\leq d(p_k, \hat{p}_\mathbf{x}) + d(\hat{p}_\mathbf{x}, \hat{p}_\mathbf{y}) + d(\hat{p}_\mathbf{y}, p_l) \\ &< \epsilon/2 + \epsilon + \epsilon/2 < d_{\min}, \end{aligned}$$

which is a contradiction to the definition of $d_{\min}$.

B. Proof of Lemma 2

Let $\mathcal{E}_k$ be the event of interest, $\left\{|\mathcal{C}_{k,\epsilon/2}| < (1-\alpha)\frac{N}{M}\right\}$, and let $\mathcal{A}_k = \left\{|C_k| < (1-\frac{\alpha}{2})\frac{N}{M}\right\}$. Note that we use $\frac{\alpha}{2}$ for $\mathcal{A}_k$ as opposed to α because we need $|C_k|$ to be larger than $|C_{k,\epsilon/2}|$. By Hoeffding's inequality, $\Pr(\mathcal{A}_k) \leq e^{-N\frac{\alpha^2}{2M^2}}$. This means that, with high probability, p_k will generate enough contigs.

Let $\mathcal{F}_k$ be the set of distributions "far" from p_k:

$$\mathcal{F}_k = \left\{p \in \widetilde{\mathcal{P}} : d\left(p_k, p\right) \geq \frac{\epsilon}{2}\right\}. \tag{17}$$

By a version of Sanov's theorem for Markov chains, given in Section VII, Theorem 4, for any $\mathbf{x} \in \mathcal{C}_k$,

$$\Pr(\hat{p}_\mathbf{x} \in \mathcal{F}_k) \leq (L+1)^4 2^{-LD_c(p^*\|p_k)} \tag{18}$$

where $p^* = \arg\inf_{p\in\mathcal{F}_k} D_c\left(p\|p_k\right)$; i.e., p^* is the distribution in $\mathcal{F}_k$ closest to p_k in conditional relative entropy. Notice that $\mathcal{E}_k$ occurs when more than $|C_k| - (1-\alpha)\frac{N}{M} + 1$ contigs lie in $\mathcal{F}_k$, leaving an insufficient number of "good" contigs. Letting $\mathbf{x}_0 \in \mathcal{C}_k$ be some contig generated by p_k,

$$\begin{aligned} &\Pr(\mathcal{E}_k|\mathcal{A}_k^c) \\ =&\Pr\left(\sum_{\mathbf{x}\in\mathcal{C}_k} \mathbb{1}\{\hat{p}_\mathbf{x} \in \mathcal{F}_k\} \geq |\mathcal{C}_k| - (1-\alpha)\frac{N}{M} + 1 \middle| \mathcal{A}_k^c\right) \\ \leq&\Pr\left(\sum_{\mathbf{x}\in\mathcal{C}_k} \mathbb{1}\{\hat{p}_\mathbf{x} \in \mathcal{F}_k\} \geq \frac{\alpha}{2}\frac{N}{M} \middle| \mathcal{A}_k^c\right) \qquad (19) \\ \leq&\frac{2M}{\alpha} \cdot \Pr(\hat{p}_{\mathbf{x}_0} \in \mathcal{F}_k|\mathcal{A}_k^c) \qquad (20) \end{aligned}$$

where (19) follows the definition of $\mathcal{A}_k$, and (20) from Markov's inequality and symmetry across contigs. Combining the probabilities,

$$\begin{aligned} \Pr(\mathcal{E}_k) &= \Pr(\mathcal{E}_k|\mathcal{A}_k^c)\Pr(\mathcal{A}_k^c) + \Pr(\mathcal{E}_k|\mathcal{A}_k)\Pr(\mathcal{A}_k) \qquad (21) \\ &\leq \frac{2M}{\alpha} \cdot \Pr(\hat{p}_{\mathbf{x}_0} \in \mathcal{F}_k|\mathcal{A}_k^c)\Pr(\mathcal{A}_k^c) + \Pr(\mathcal{A}_k) \qquad (22) \\ &\leq \frac{2M}{\alpha}(L+1)^4 2^{-LD_c(p^*\|p_k)} + e^{-N\frac{\alpha^2}{2M^2}} \qquad (23) \\ &\leq \frac{1}{\alpha}e^{-\gamma\alpha^2 L} \qquad (24) \end{aligned}$$

where $\gamma > 0$ is a constant that does not depend on α or L, guaranteed to exist such that (24) holds for large N.

C. Proof of Theorem 2

We define the *type* of a contig $\mathbf{x}$ to be its empirical fourth-order distribution, denoted $\hat{p}_\mathbf{x}$. Let the set of all possible types of length-L, stationary, third-order Markov sequences be $\mathcal{P}_L$. The cardinality of $\mathcal{P}_L$ is upper-bounded by $(L+1)^4$ as shown in [22]. The *type class*, T_L, of a given type, $p \in \mathcal{P}_L$, is then defined as the set of all length-L sequences whose types are equal to p:

$$T_L(p) = \left\{\mathbf{x} \in \mathcal{X}^L : \ \hat{p}_\mathbf{x} = p\right\} \tag{25}$$

To facilitate analysis, we use a *cyclical* Markov model, where three artificial transitions are added from the end of the sequence to the beginning. This model ensures that $\hat{p}_\mathbf{x}$ is *consistent*. More precisely, for $\mathbf{a} \in \mathcal{X}^3$,

$$\sum_{b\in\mathcal{X}} \hat{p}_\mathbf{x}(\mathbf{a}b) = \sum_{b\in\mathcal{X}} \hat{p}_\mathbf{x}(b\mathbf{a}) \tag{26}$$

Note that this implies $\hat{p}_\mathbf{x} \in \widetilde{P}$ as defined in Section II. Furthermore, the third-order and conditional empirical distributions can be derived from $\hat{p}_\mathbf{x}$ as follows, for any $b \in \mathcal{X}$.

$$\hat{p}_\mathbf{x}(\mathbf{a}) = \sum_{b\in\mathcal{X}} \hat{p}_\mathbf{x}(\mathbf{a}b) \tag{27}$$

and

$$\hat{p}_\mathbf{x}(b|\mathbf{a}) = \frac{\hat{p}_\mathbf{x}(\mathbf{a}b)}{\hat{p}_\mathbf{x}(\mathbf{a})} \tag{28}$$

We now use some Large Deviations theory to make an argument about the probability of error in the hypothesis test.

1) Large Deviations Principle: Vidyasagar [22] provides an extensive analysis of large deviations theory for Markov processes. Theorems 3 and 4, shown below, utilize this analysis along with [23, Lemma 1], which allows us to make an argument about the probability of error for the subsequent hypothesis test. For the proofs of Theorems 3 and 4, the reader is referred to [22, Theorem 7] and [16, Chapter 11].

The results in [22] show that a Markov process $\mathbf{X} = (X_1, \ldots, X_L)$ with type p and generated by q satisfies the large deviations property with rate function

$$I(p) = D_c\left(p \| q\right) \tag{29}$$

Here, D_c is the conditional relative entropy defined as the KL divergences averaged over p:

$$\begin{aligned} D_c\left(p\|q\right) &= \sum_{\mathbf{a}\in\mathcal{X}^3} p(\mathbf{a}) \sum_{b\in\mathcal{X}} p(b|\mathbf{a}) \log\left(\frac{p(b|\mathbf{a})}{q(b|\mathbf{a})}\right) \\ &= \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p(\mathbf{a}b) \log \frac{p(\mathbf{a}b)}{q(\mathbf{a}b)} \\ &\quad - \sum_{\mathbf{a}\in\mathcal{X}^3} p(\mathbf{a}) \log \frac{p(\mathbf{a})}{q(\mathbf{a})} \\ &= D^{(4)}(p\|q) - D^{(3)}(p\|q) \end{aligned}$$

i.e. the divergence between the fourth-order distributions minus the divergence between the third-order distributions. Similarly, the conditional entropy can be written as

$$\begin{aligned} H_c\left(p\right) &= \sum_{\mathbf{a}\in\mathcal{X}^3} p(\mathbf{a}) \sum_{b\in\mathcal{X}} p(b|\mathbf{a}) \log p(b|\mathbf{a}) \\ &= H^{(4)}(p) - H^{(3)}(p) \end{aligned}$$

We use D_c and H_c so as to distinguish between the normal divergence and entropy.

Theorem 3. *The probability of* $\mathbf{x}$ *under* q *depends only on its type* $\hat{p}_{\mathbf{x}}$ *and is given by*

$$q^{(L)}(\mathbf{x}) = 2^{-L[D_c(\hat{p}_{\mathbf{x}}\|q) + H_c(\hat{p}_{\mathbf{x}})] + \log\alpha} \tag{30}$$

where $\alpha = q(x_1x_2x_3)$*, i.e. the probability of the initial state of* $\mathbf{x}$*.*

Theorem 4 (Sanov's Theorem for Markov Processes)**.** *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_L)$ *be a Markov process* q*, and let* $\mathcal{F} \subseteq \widetilde{\mathcal{P}}$*. The probability that the empirical distribution of* $\mathbf{X}$ *is contained in* $\mathcal{F}$*, denoted* $q^{(L)}(\mathcal{F})$*, is upper-bounded as*

$$q^{(L)}(\mathcal{F}) \leq |\mathcal{P}_L|\, 2^{-LD_c(p^*\|q) + \log\alpha} \tag{31}$$

where p^* *is the information projection of* q *onto* $\mathcal{F}$*:*

$$p^* = \arg\inf_{p\in\mathcal{F}} D_c\left(p\|q\right) \tag{32}$$

If, in addition, the closure of $\mathcal{F}$ *is equal to the closure of its interior (*$\bar{\mathcal{F}} = \bar{\mathcal{F}^o}$*), then*

$$\lim_{L\to\infty} \frac{1}{L} \log q^{(L)}(\mathcal{F}) = -D_c\left(p^*\|q\right) = -I(p^*) \tag{33}$$

2) Hypothesis Test: In the binary hypothesis test, there are two candidate models for q: p_1 and p_2, where $p_1 \neq p_2$. We decide between the two hypotheses:

- $\mathrm{H}_1: \ q = p_1$
- $\mathrm{H}_2: \ q = p_2$

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the decision regions for H_1 and H_2, respectively. The sets $\mathcal{P}_1$ and $\mathcal{P}_2$ form a disjoint union of $\widetilde{\mathcal{P}}$ ($\mathcal{P}_1 \cup \mathcal{P}_2 = \widetilde{\mathcal{P}}$). As a result, given any $\mathbf{x} \in \mathcal{X}^L$, $\delta_L(\mathbf{x})$ decides H_1 if $\hat{p}_{\mathbf{x}} \in \mathcal{P}_1$ and H_2 if $\hat{p}_{\mathbf{x}} \in \mathcal{P}_2$. The Bayesian probability of error, P_e, for the binary hypothesis test with priors π_1 and π_2 is given by

$$P_e = \pi_1 p_1^{(L)}(\mathcal{P}_2) + \pi_2 p_2^{(L)}(\mathcal{P}_1) \tag{34}$$

To minimize the error, the decision rule[4] uses a Neyman-Pearson test

$$\delta_L(\hat{p}_{\mathbf{x}}) = \begin{cases} \mathrm{H}_1 & \text{if } \mathcal{L}(\mathbf{x}) \geq \frac{\pi_2}{\pi_1} \\ \mathrm{H}_2 & \text{if } \mathcal{L}(\mathbf{x}) < \frac{\pi_2}{\pi_1} \end{cases} \tag{35}$$

where the likelihood ratio, $\mathcal{L}$, is defined as:

$$\mathcal{L}(\mathbf{x}) = \frac{\Pr(\mathbf{x}|\mathrm{H}_1 \text{ true})}{\Pr(\mathbf{x}|\mathrm{H}_2 \text{ true})} = \frac{p_1^{(L)}(\mathbf{x})}{p_2^{(L)}(\mathbf{x})} \tag{36}$$

Using Theorem 3, the normalized log-likelihood ratio is

$$\begin{aligned} \frac{1}{L}\log\mathcal{L}(\mathbf{x}) &= -\left[D_c\left(\hat{p}_{\mathbf{x}}\|p_1\right) + H_c\left(\hat{p}_{\mathbf{x}}\right)\right] + \frac{\log\alpha_1}{L} \\ &\quad + \left[D_c\left(\hat{p}_{\mathbf{x}}\|p_2\right) + H_c\left(\hat{p}_{\mathbf{x}}\right)\right] - \frac{\log\alpha_2}{L} \\ &= D_c\left(\hat{p}_{\mathbf{x}}\|p_2\right) - D_c\left(\hat{p}_{\mathbf{x}}\|p_1\right) + \frac{1}{L}\log\frac{\alpha_1}{\alpha_2} \end{aligned}$$

Again, α_1 and α_2 represent the probabilities of the initial states of $\mathbf{x}$ under p_1 and p_2, respectively. Notice that as $L \to \infty$, the optimal decision rule simply chooses $\arg\min_{k\in\{1,2\}} D_c\left(p\|p_k\right)$ because the effect of the priors washes out with L, along with the probability of the initial states. We will show that, by using the decision regions, $\mathcal{P}_1$ and $\mathcal{P}_2$, given by

$$\mathcal{P}_1 = \{p \in \widetilde{\mathcal{P}} : D_c\left(p\|p_2\right) - D_c\left(p\|p_1\right) \geq 0\} \tag{37}$$

$$\mathcal{P}_2 = \{p \in \widetilde{\mathcal{P}} : D_c\left(p\|p_2\right) - D_c\left(p\|p_1\right) < 0\} \tag{38}$$

the optimal error exponent is achieved in the limit. First we will prove Lemmas 3 and 4, which allow for the use of (33) in Theorem 4.

Lemma 3. *$\mathcal{P}_1$ and $\mathcal{P}_2$ are convex.*

Proof. Let $p_a, p_b \in \mathcal{P}_1$ and let

$$p_{ab} = \lambda p_a + (1-\lambda)p_b, \lambda \ \in \ (0,1)$$

[4]The decision rule δ_L uses overloaded notation with the decision rule for the main problem.

be a convex combination of p_a and p_b. Then

$$\begin{aligned}
&D_c\left(p_{ab}\|p_2\right) - D_c\left(p_{ab}\|p_1\right) \\
&= \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p_{ab}(\mathbf{a}b)\log\frac{p_{ab}(b|\mathbf{a})}{p_2(b|\mathbf{a})} \\
&\quad - \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p_{ab}(\mathbf{a}b)\log\frac{p_{ab}(b|\mathbf{a})}{p_1(b|\mathbf{a})} \\
&= \sum_{\mathbf{a}\in\mathcal{X}^3, b\in\mathcal{X}} p_{ab}(\mathbf{a}b)\log\frac{p_1(b|\mathbf{a})}{p_2(b|\mathbf{a})} \\
&= \lambda\left[D_c\left(p_a\|p_2\right) - D_c\left(p_a\|p_1\right)\right] \\
&\quad + (1-\lambda)\left[D_c\left(p_b\|p_2\right) - D_c\left(p_b\|p_1\right)\right] \geq 0
\end{aligned}$$

so $p_{ab} \in \mathcal{P}_1$ and therefore $\mathcal{P}_1$ is a convex set. A similar argument can be made for the set $\mathcal{P}_2$. Note that since $\mathcal{P}_1$ and $\mathcal{P}_2$ are both convex, this implies that the boundary linearly divides the set of stationary fourth-order distributions, $\widetilde{\mathcal{P}}$. □

Lemma 4.

$$\overline{\mathcal{P}_1} = \overline{\mathcal{P}_1^o} \quad \textit{and} \quad \overline{\mathcal{P}_2} = \overline{\mathcal{P}_2^o} \tag{39}$$

Proof. The boundary between $\mathcal{P}_1$ and $\mathcal{P}_2$ consists of the set of distributions $p \in \widetilde{\mathcal{P}}$ for which $D_c\left(p\|p_2\right) - D_c\left(p\|p_1\right) = 0$. We see that p_1 does not lie on this boundary because

$$D_c\left(p_1\|p_2\right) - D_c\left(p_1\|p_1\right) = D_c\left(p_1\|p_2\right) > 0 \tag{40}$$

by the non-negativity of the KL-divergence. Furthermore, p_1 cannot lie on any other boundary of $\mathcal{P}_1$ because all of the elements of p_1 are nonzero for L large enough (with probability $1 - o(1)$) due to the fact that all the entries of p_1 are nonzero. Thus p_1 is an interior point of $\mathcal{P}_1$ as it does not lie on any of the boundaries.

Finally, we need to show that convexity and a non-empty interior imply (39). Take a point $p \in \overline{\mathcal{P}_1}$. Then either $p \in \mathcal{P}_1^o$ or $p \in \partial\mathcal{P}_1$, the boundary of $\mathcal{P}_1$. If $p \in \mathcal{P}_1^o$, then $p \in \overline{\mathcal{P}_1^o}$, trivially. If $p \in \partial\mathcal{P}_1$, we must prove that p is a limit point of $\mathcal{P}_1^o$. Since p_1 is an interior point of $\mathcal{P}_1$, then there exists an open ball U_1 centered at p_1 which is completely contained in $\mathcal{P}_1$. We define V_1 as the set of distributions that result from a convex combination of p and U_1

$$V_1 = \{\alpha U_1 + (1-\alpha)p : 0 < \alpha \leq 1\} \tag{41}$$

using Minkowski addition. The set V_1 clearly has non-zero volume (by Lebesgue measure) and all of its points are interior points of $\mathcal{P}_1$ due to Lemma 3. Therefore there exists a sequence of interior points $\{p_t\}, p_t \in V_1$ such that $p_t \to p$. Thus $p \in \overline{\mathcal{P}_1^o}$.

A similar argument can be made for $\mathcal{P}_2$. Hence, the proof of Lemma 4 is complete. □

Now, by Theorem 4 and Lemmas 3 and 4, the error exponents are as such.

$$\lim_{L\to\infty}\frac{1}{L}\log p_1^{(L)}(\mathcal{P}_2) = -D_c\left(p_1^*\|p_1\right) \tag{42}$$

$$\lim_{L\to\infty}\frac{1}{L}\log p_2^{(L)}(\mathcal{P}_1) = -D_c\left(p_2^*\|p_2\right). \tag{43}$$

Distribution p_1^* is found by minimizing $D_c\left(p_1^*\|p_1\right)$, subject to the decision boundary constraint,

$$D_c\left(p_1^*\|p_1\right) - D_c\left(p_1^*\|p_2\right) \geq 0, \tag{44}$$

the consistency constraints for all $a \in \mathcal{X}^3$,

$$\sum_{b\in\mathcal{X}} p_1^*(\mathbf{a}b) = \sum_{b\in\mathcal{X}} p_1^*(b\mathbf{a}), \tag{45}$$

and the sum-to-one constraint,

$$\sum_{\mathbf{c}\in\mathcal{X}^4} p_1^*(\mathbf{c}) = 1 \tag{46}$$

This will yield the distribution $p_1^* \in \mathcal{P}_2$ that is closest to p_1. Notice that (7) in Definition 2 implies that p^* actually lies on the boundary, i.e. (44) holds with equality. This can be proven by contradiction: suppose p' is the optimal solution to the minimization problem and suppose

$$D_c(p'\|p_1) - D_c(p'\|p_2) > 0.$$

For $0 \leq \lambda \leq 1$, let $p_\lambda = \lambda\, p_1 + (1-\lambda)\, p'$ be a convex combination of p' and p_1. We know from Lemma 3 that $p_\lambda \in \widetilde{\mathcal{P}}$ for any value of λ and furthermore, there exists a $\lambda = \lambda^*$ such that

$$D_c(p_{\lambda^*}\|p_1) - D_c(p_{\lambda^*}\|p_2) = 0$$

since the boundary linearly divides $\widetilde{\mathcal{P}}$. Now, to show by condtradiction that

$$D_c(p_{\lambda^*}\|p_1) < D_c(p'\|p_1),$$

we will show that conditional relative entropy is convex in its first argument. For some distribution $q \in \widetilde{\mathcal{P}}$,

$$\begin{aligned}
D_c(p_\lambda\|q) &= \sum_{\mathbf{a}b\in\mathcal{X}^4} p_\lambda(\mathbf{a}b)\log\frac{p_\lambda(b|\mathbf{a})}{q(b|\mathbf{a})} \\
&= \lambda\sum_{\mathbf{a}b\in\mathcal{X}^4} p_1(\mathbf{a}b)\log\frac{p_\lambda(b|\mathbf{a})}{q(b|\mathbf{a})} \\
&\quad + (1-\lambda)\sum_{\mathbf{a}b\in\mathcal{X}^4} p'(\mathbf{a}b)\log\frac{p_\lambda(b|\mathbf{a})}{q(b|\mathbf{a})} \\
&= \lambda\sum_{\mathbf{a}b\in\mathcal{X}^4} p_1(\mathbf{a}b)\left(\log\frac{p_1(b|\mathbf{a})}{q(b|\mathbf{a})} - \log\frac{p_1(b|\mathbf{a})}{p_\lambda(b|\mathbf{a})}\right) \\
&\quad + (1-\lambda)\sum_{\mathbf{a}b\in\mathcal{X}^4} p'(\mathbf{a}b)\left(\log\frac{p'(b|\mathbf{a})}{q(b|\mathbf{a})} - \log\frac{p'(b|\mathbf{a})}{p_\lambda(b|\mathbf{a})}\right) \\
&= \lambda D_c(p_1\|q) + (1-\lambda)D_c(p'\|q) \\
&\quad - \lambda D_c(p_1\|p_\lambda) - (1-\lambda)D_c(p'\|p_\lambda) \\
&< \lambda D_c(p_1\|q) + (1-\lambda)D_c(p'\|q)
\end{aligned}$$

where the last step follows from the non-negativity of conditional relative entropy. Finally, setting $q = p_1$, we have

$$D_c(p_\lambda\|p_1) < \lambda D_c(p_1\|p_1) + (1-\lambda)D_c(p'\|p_1) < D_c(p'\|p_1).$$

Therefore, p_{λ^*} must be a better solution, which is a contradiction. Clearly $p_1^* = p_2^*$ since they both lie on the boundary where the minimads, $D_c(\cdot\|p_1)$ and $D_c(\cdot\|p_2)$, are equal.