This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3390068

Fundamental Limits of Reference-Based Sequence Reordering

Nir Weinberger[1] and Ilan Shomorony[2]

***Abstract*—We consider the problem of reconstructing a sequence of independent and identically distributed symbols from a set of equal-size, consecutive, fragments, as well as a dependent reference sequence. First, in the regime in which the fragments are relatively long and typically no fragment appears more than once, we determine the scaling of the failure probability of the maximum-likelihood reconstruction algorithm for a perfect reconstruction, and bound it for a partial reconstruction. Second, we characterize the regime in which the fragments are relatively short and repeating fragments abound. We state a trade-off between the fraction of fragments that cannot be adequately reconstructed vs. the distortion level allowed for the reconstruction of each fragment, while still allowing vanishing failure probability.**

***Index Terms*—Fragment reordering, permutation reconstruction, reference sequence, side information, sequence reconstruction, sliced sequences, DNA sequencing, bee-identification problem.**

I. Introduction

In this paper, we consider the problem of reconstructing a sequence $X^N \in \mathcal{X}^N$ from its non-overlapping consecutive fragments and a reference sequence, as illustrated in Fig. 1. A sequence of N independent and identically distributed (IID) symbols is drawn from a finite alphabet source, and is then partitioned into non-overlapping, consecutive *fragments* of length L each. The fragments are then permuted in an arbitrary manner, and a *multiset* of $M = N/L$ fragments is observed, without any specific order. In order to facilitate the correct reordering of the fragments, the observer of the fragments is supplied with a *reference sequence* $Y^N \in \mathcal{Y}^N$ of length N. This reference sequence is similar, yet not identical, to the sequence of interest; for example, it can be its noisy version, or slightly different due to statistical variations in some population. A reconstruction algorithm observes the M fragments of the original sequence as well as the reference sequence Y^N, and is required to recover the original sequence X^N.

This problem is motivated by settings in which data is observed out of order, and ordering is made possible through side information. It arises in various domains: First, genomic DNA sequencing typically produces short fragments, which should be assembled in order to obtain the correct sequence.[1] Due to the high similarity between the individual genomes, the reconstruction algorithm may have access to a reference sequence that can be aid the assembly of the target sequence [24]. Variations between the two individually sequenced genomes, may arise, e.g., either due to sequencing noise, or due to genetic variations such as single nucleotide polymorphisms (SNPs). As such, it can be modeled as a noisy channel operating on one of the sequences to produce the other. While genetic variations and sequencing noise are not independent across the genome locations, we will follow the literature [25] and for simplicity model them as a discrete memoryless source. Second, as described in [19], such problem arises in transmission of information over (noiseless) permutation channels, such as packet networks employing multipath routing as a means for an end-to-end packet transfer [26]. The transmission of X^N over such link may use short packets, each one encoding fragments of size L. Then, similarly to the standard distributed compression problem — the Slepian-Wolf problem [31] — the reconstruction of the sequence from the fragments can be aided by a side-information sequence Y^N. Third, as we discuss below, the problem is related to the identification of unordered entities marked by barcodes, from noisy fragments of those barcodes [33]. The sequence reordering problem also has connections with the problem of database alignment, for which a framework to study the fundamental limits of database alignment was proposed in [10], motivated by the problem of database de-anonymization. Similar to our setting, one wishes to match a set of (anonymized) fragments to a set of similar fragments (with identity information).

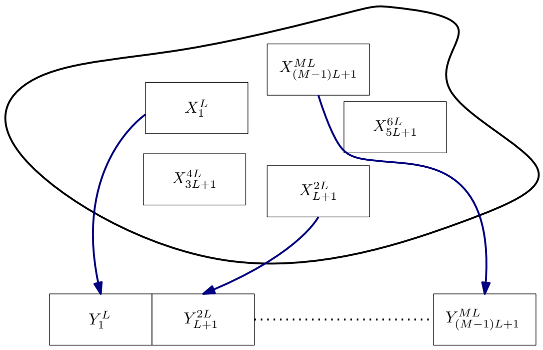

Figure 1. Illustration of the reference-based reordering problem.

[1] Nir Weinberger is with the Viterbi Faculty of Electrical and Computer Engineering, Technion-Israel Institute of Technology, Haifa 3200004, Israel (Email: nirwein@technion.ac.il.). [2] Ilan Shomorony is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (Email: ilans@illinois.edu). This paper was presented in part in the 2023 IEEE International Symposium on Information Theory. The research of N. W. was supported by the Israel Science Foundation (ISF), grant no. 1782/22. The work of I. S. was supported in part by the National Science Foundation under CCF grants 2007597 and 2046991.

[1]More accurately, the fragments in DNA sequencing typically start at random locations along the sequence, and might have overlaps. In this sense, our model is a distilled version of this problem, and extending our results to overlapping fragments is an interesting open problem.

In this paper, we assume that X^N is drawn from a memoryless source P_X, and that Y^N is obtained by passing X^N in a discrete memoryless channel $P_{Y|X}$. Furthermore, we assume that the length of each fragment scales logarithmically with the number of fragments, and specifically set $L = \beta \log M$ for some length-scaling parameter β. As we show, for this choice of scaling there exists a critical value of β that is required for reliable reconstruction. In addition, for this scaling, the problem described above is closely related, and perhaps *prima facie* equivalent, to the *bee identification* (BI) problem, recently introduced in [33] and further studied in [6], [17], [32], [34]. In the BI problem, one assumes that the fragments of Y^N (obtained similarly to the fragments of X^N, as non-overlapping, consecutive segments of length L) are each a barcode used for identification of some objects, via the fragments of X^N, which are noisy unordered observations of the barcodes. A codebook for this problem is comprised of the M fragments of Y^N, where X^N is drawn in a memoryless fashion according to the reverse channel $P_{X|Y}$. A plausible method to generate this codebook is via random coding, and specifically, by drawing the $N = LM$ symbols of the fragments in an IID fashion. In this random coding regime, the *average* error probability over the random ensemble of codebooks is similar to the reconstruction error in the fragments reordering problem we consider, with the inconsequential difference that the channel for the BI problem is the reverse channel $P_{X|Y}$, rather than $P_{Y|X}$.

Nonetheless, there are two subtle, yet significant, differences between the ordering problem using reference sequences considered here and the BI problem. First, in the ordering problem, one is not necessarily interested in recovering the exact permutation of the fragments, but rather just the correctly reordered sequence. Second, the source sequence and the reference sequence are random, and there is no design freedom to optimally choose the source fragments. By contrast, in the BI problem, only a single optimal codebook is sought. As shown in [32], [33], improved bounds can be obtained by considering the average error of the *typical random code* [1], [22], or via expurgation techniques [32]. As said, in the ordering problem considered here, this is impossible, and thus it is the random coding analysis that is of interest. In fact, these two matters are interrelated, as exemplified by the following extreme case: In the event that all fragments of X^N are equal, there is no ambiguity in the reconstruction of the sequence, and *zero reconstruction error* is obtained. By contrast, as a codebook for the BI problem, this has the *maximal* possible error probability. Generalizing upon this observation, the difference between the ordering problem and the BI problem is most pronounced whenever there are *repeated fragments* in the sequence. This typically happens when the fragments are relatively short (small β), and the entropy of the source probability mass function (PMF) P_X is low.

When considering repeated fragments in a sequence, it is also expected that *similar* fragments will also be observed, and in such a scenario, it is unreasonable to expect a perfect reconstruction. Therefore, we consider in this paper a relaxed notion of *imperfect* reconstruction, comprised of two elements. First, we assume that an additive distortion measure is given, and consider a fragment to be successfully reconstructed if its distortion within the source fragment in that location is below a prescribed distortion level $\delta \in \mathbb{R}_+$. Second, we consider the reconstruction to be successful if at most a fraction $\xi \in [0, 1]$ of the fragments were unsuccessfully reconstructed (i.e., their distortion level is larger than δ). We then may analyze the failure probability of the reconstruction algorithm for a pair (δ, ξ), or, the trade-off between δ and ξ . The relaxed definition of failure probability for $\xi > 0$ was also proposed in the conclusion part of [33] for the BI problem, as well as in [32] (although with ξM being replaced by a constant that does not scale with M).

As might be noted, and owing to the above described differences, we describe our setting with a different terminology compared to the way it is formulated in the BI problem. In the BI problem, the fragment length is considered the decoding blocklength, and is expected to be large. The number of bees is then exponential in that blocklength, that is, $M = e^{\frac{L}{\beta}}$.[2] Therefore, $1/\beta$ may be perceived as the *rate* of identification, and reliable identification of the bees is shown to be possible in [32], [33] as long as $\frac{1}{\beta}$ is less than a maximal possible rate, which may be considered the *capacity* of the BI problem. Here, we opt to equivalently refer to the fragment length as a logarithmic function of the sequence length, as common in various other fragmented sequences problems, such as DNA storage [29], [36], [37]. Accordingly, the success of reconstruction will be (equivalently) stated in terms of *lower* bounds on the fragment-length scaling β.

A. Results Overview

We assume throughout that the optimal maximum-likelihood (ML) decoder is used for reconstruction. In the parlance of [33], this is termed *joint decoding* of all the fragments.

1) The No-Repeating-Fragments Regime with Zero Distortion: First, we consider the regime in which no repeated fragments are expected, and assume zero distortion $\delta = 0$, though we allow both perfect and imperfect reconstruction ($\xi = 0$ and $\xi > 0$, respectively). For $\xi = 0$, this revisits the setting of random coding analysis for joint decoding in the BI problem [33]. We show the following in Theorem 1: As long as $\beta > \frac{1}{\psi_2(P_{XY})}$, then the reconstruction algorithm succeeds with high probability, where the threshold $\psi_2(P_{XY})$ is explicitly defined in (10) as a convex optimization problem over joint PMFs $Q_{X_1X_2}$, that is, over $|\mathcal{X}|^2 - 1$ free variables. For example, when X is a uniform Bernoulli random variable (RV), and $P_{Y|X}$ is a binary symmetric channel (BSC) with crossover probability α then $\psi_2(P_{XY}) = \frac{1}{2}[\log 2 - \log(1 + 4\alpha(1 - \alpha))]$. Specifically, if $\xi = 0$ then failure occurs with probability at most $O(M^{2[1-\beta\psi_2(P_{XY})]})$, that is, a polynomial decay in M. If $\xi > 0$ then the failure probability occurs with probability at most $e^{-\xi M \log M \cdot [\beta\psi_2(P_{XY})-1]}$, that is, exponential decay with respect to (w.r.t.) $M \log M$. We then establish in lower bounds in Theorem 3, which make mild, unavoidable assumptions.

[2] In the notation in [33], $L \leftrightarrow n$ and $M \leftrightarrow m$.

For $\xi = 0$, we show that the failure probability rate is in fact tight for $\xi = 0$, and lower bounded as $e^{-\xi M \log M \cdot \beta\psi_2(P_{XY})}$ for $\xi > 0$ (thus, there is a gap of ξ in the exponent w.r.t. $M \log M$).

In the $\xi = 0$ setting, the improvement of Theorem 1 over [33] is twofold. First, [33] only considered a symmetric (uniform) binary source with a BSC, and its analysis heavily utilizes the symmetry properties of this distribution. We obtain this result for a *general* source P_{XY}. In addition, our bound is tighter than the one obtained for the binary symmetric setting considered in [33]. Specifically, the dependence on $\psi_2(P_{XY})$ is related to the error probability of transpositions, i.e., cycles of length 2 (hence the subscript 2), and as we show, this is the dominant error event. The analysis of [33] only showed that the dominant error event may be either a transposition or a cycle of length 3. In terms of proof techniques, as in previous papers, we condition on the source vector, use a union bound of the pairwise error of all possible permutations, and upper bound the pairwise error using the Bhattacharyya bound. As previous analysis also showed, the average of this upper bound over X^N should be computed for permutations which are *cycles*. Such a cycle may have any length from $\{2, 3, ..., M-1, M\}$, and we proved that transpositions (cycles of length 2) dominate the error probability. The key new ingredient is to evaluate this expectation via the *Donsker–Varadhan* variational formula [12] (e.g., [3, Corollary 4.15]). This method is preferred over perhaps the more straightforward *method of types* [8] [9, Sec. 2.1], since the error term in the latter blows up when the cycle length is on the order of M. Interestingly, the argument used to show that transpositions dominate cycles of length 3 and the analogous argument for longer cycles are different. The argument for length-3 cycles is direct, and is based on special symmetry properties along with Han's inequality for the Kullback-Leibler (KL) divergence [3, Theorem 4.9] [16]. The argument for cycles of lengths 4 and larger is based on a *relaxation* of the function $\psi_K(P_{XY})$, characterizing the Bhattacharyya error bound for length-K cycles, which, intuitively speaking, *breaks* the cycle at its end point.

2) The Repeating-Fragments Regime with Positive Distortion: In this regime, we consider both positive fragment-failure rate $\xi > 0$ and positive distortion $\delta > 0$. As in the previous regime and in previous works, the analysis of the reconstruction failure probability is based on a union bound over all possible permutations of the M fragments. In the worst case, in which all fragments are unique, this is a union bound over $M! = e^{M \log M + O(\log M)}$ fragments. However, in the repeating-fragments regime, multiple permutations are in fact equivalent, in the sense that they lead to the same reconstructed sequence (e.g., if all fragments are equal except for one, then there are just M different possible reconstruction vectors). The number of possible distinct permutations is determined by the histogram of the different $|\mathcal{X}|^L$ possible fragments. Clearly, repeated fragments are more prone to occur when fragments are short (small β), or when the entropy of the source $H(P_X)$ is low. The main technical contribution in this regime, and the key ingredient in the analysis, is to show that the possible number of reconstruction vectors is tightly concentrated around $e^{\beta H(P_X) \cdot M \log M + o(M \log M)}$ (see Prop. 7), with high probability of $1 - o(1)$. Thus, if $H(P_X)\beta < 1$ then the effective number of permutations is smaller than the maximal value of order $e^{M \log M}$. In turn, the proof of this property is based on two main ingredients. First, while the histogram vector of the fragments is distributed as a *multinomial* and thus has dependent entries, probabilities defined on events of this random vector are dominated by a Poisson distribution with independent entries (an effect known as *Poissonization*, see Fact 9 and Lemma 10). The logarithm of the number of possible reconstruction vectors is then upper bounded by the entropy of the histogram vector, which, using the Poissonization effect, is the sum of independent terms of the form $-\sum_{i \in [\mathcal{X}^L]} G_i \log G_i$, where G_i follows a Poisson distribution. For $\beta \in (0, 2)$, the proof then uses a concentration inequality on Lipschitz functions of Poisson RVs by Bobkov and Ledoux [2], [18], however in a modified way, since $t \to -t \log t$ is, strictly speaking, not a Lipschitz function. For $\beta > 2$ a standard Bernstein inequality is used. The derivation of the bound on the failure probability then follows a different path compared to previous works. Rather then fixing a permutation and analyzing the probability over a random choice of X^N, we upper bound the probability for a fixed, typical, X^N, in the sense that the number of its possible reconstructions is $e^{\beta H(P_X) \cdot M \log M}$. Per the analysis above, it holds that a-typical X^N occurs with probability at most $o(1)$. Now, if we let $d^*_{P_{Y|X}}(\delta)$ be the minimal Bhattacharyya distance for fragments of distortion larger than δ, it is easily shown that the failure reconstruction for such typical X^N decays as $e^{-\Theta(M \log M)}$ as long as $\xi > H(P_X)/d^*_{P_{Y|X}}(\delta)$. This leads to a trade-off between δ and ξ in the repeating-fragments regime $\beta < 1/H(P_X)$, which is the main result in this regime, stated in Theorem 6.

B. Additional Related Work

An information-theoretic study of sequence reconstruction from short fragments taken at *random* locations was initiated in [25], and its reference-based counterpart was considered in [24]. The analysis in those papers is motivated by DNA sequencing, and thus assumes a uniform source over the DNA alphabet of size 4, and a quaternary symmetric channel $P_{Y|X}$, with the goal of detecting SNPs. The performance metric is the average misdetection and false-alarm probabilities, as defined therein, which is essentially equivalent to the total number of failed fragments, ξM in our notation. In [4], the problem of compressing a non-probabilistic source was considered when the encoder has a possible list of reference vectors. In [14], [15], compression methods were proposed and analyzed for the setting in which fragments are compressed at the encoder side and are reconstructed at the decoder side using a reference sequence. The ordering problem is also tightly related to the DNA storage sampling-shuffling channel [20], [29], [37], in which short unordered fragments store the information. In that setting too, the fragment lengths have the same scaling of $\beta \log M$, where M is the number of fragments, and a positive capacity requires β to be above a critical value. More generally, it is related to *permutation channels* [19], [21], [30], in which the output sequence is a permuted and possibly noisy version of the input sequence.

C. Outline

The outline of the rest of the paper is as follows. In Sec. II we formulate the problem, in Sec. III we state our main results, and in Sec. IV we conclude the paper. Proofs are relegated to the appendices.

II. Problem Formulation

Notation conventions: Let Q_X be a PMF over a finite alphabet $\mathcal{X}$. For $j > i$, the sequence comprised of the components between i and j is denoted by $X_i^j := (X_i, X_{i+1}, \ldots, X_j)$ and is shorthanded as $X^N \equiv X_1^N$ for $i = 1$. Let $\mathcal{P}_L(\mathcal{X})$ denote the set of all types (empirical distributions) of length L, and let $\mathcal{P}(\mathcal{X})$ be the set of all PMFs on $\mathcal{X}$ (i.e., the $(|\mathcal{X}|-1)$-dimensional probability simplex). The type class [9, Ch. 2] of a type $Q_X \in \mathcal{P}_L(\mathcal{X})$ is denoted by $T_L(Q_X)$, that is, the set of all empirical PMFs for length L vectors over $\mathcal{X}$. The Rényi entropy of order $\alpha \geq 0$, $\alpha \neq 1$ is denoted by

$$H_\alpha(Q_X) := \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathcal{X}} Q_X^\alpha(x) \right), \tag{1}$$

and the Shannon entropy is denoted by $H(Q_X) \equiv H_1(Q_X) := \lim_{\alpha \downarrow 1} H_\alpha(Q_X) = -\sum_{x\in\mathcal{X}} Q_X(x) \log Q_X(x)$. Specifically, $H_2(Q_X) = -\log \sum [Q_X(x)]^2$ is the *collision entropy*. The binary entropy function is denoted by $h_{\text{bin}}(t) := -t \log t - (1-t)\log(1-t)$ for $t \in (0,1)$ and $h_{\text{bin}}(0) = h_{\text{bin}}(1) = 0$. For a pair of conditional PMFs $Q_{Y|X}$ and $P_{Y|X}$ and a PMF P_X, the conditional KL divergence is denoted by $D_{\text{KL}}(Q_{Y|X} \;||\; P_{Y|X} \mid P_X)$. The conditioning on P_X is removed when Y is independent of X under both $Q_{Y|X}$ and $P_{Y|X}$. The binary KL divergence function is denoted by $d_{\text{bin}}(p,q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ for $p, q \in (0,1)$, $d_{\text{bin}}(0,0) = d_{\text{bin}}(1,1) = 0$, and $d_{\text{bin}}(1,0) = d_{\text{bin}}(0,1) = \infty$. The total variation distance (ℓ_1 distance) between a pair of PMFs over a countable alphabet is denoted by $d_{\text{TV}}(P,Q) := \sum_{y\in\mathcal{Y}} |P(y) - Q(y)|$. The complement of an event $\mathcal{A}$ is denoted by $\mathcal{A}^c$. For an integer M, let $[M] := \{1, \ldots, M\}$. The maximum (resp. minimum) between a and $b \in \mathbb{R}$ is denoted by $a \vee b$ (resp. $a \wedge b$). The maximum between $t \in \mathbb{R}$ and 0 is denoted by $(t)_+ := t \vee 0$.

Let $(X^N, Y^N) \sim P_{XY}^{\otimes N}$ be a pair of sequences of length N, drawn IID from P_{XY}, over the finite Cartesian product alphabet $\mathcal{X} \times \mathcal{Y}$. The PMF P_X is assumed without loss of generality (WLOG) to be fully supported on $\mathcal{X}$. Let L denote a fragment length. For simplicity of notation we assume that $M := N/L$ is integer, and ignore in what follows any integer constraints on asymptotically large numbers, as they do no affect the results. The sequence X^N is partitioned into M equal-length and non-overlapping fragments denoted by $\boldsymbol{X}(i) := X_{(i-1)L+1}^{iL}$. A reconstruction algorithm observes the *multiset* of fragments $\{\boldsymbol{X}(i)\}_{i\in[M]}$ and the reference sequence Y^N, and is required to output the original ordered sequence X^N. Let S_M denote the symmetric group of order M, i.e., the group of all bijections from $[M]$ to itself. A permuted sequence of fragments is denoted by

$$\pi[X^N] := (\boldsymbol{X}(\pi(1)), \boldsymbol{X}(\pi(2)), \ldots, \boldsymbol{X}(\pi(M)))\,, \tag{2}$$

and

$$\mathcal{A}_L(X^N) := \left\{\pi[X^N]\right\}_{\pi \in S_M} \tag{3}$$

is then the *set* of all possible reconstructed sequences from fragments of X^N of length L. In essence, conditioned on X^N, the reconstruction problem is a multiple hypothesis testing problem between a *random* number of $|\mathcal{A}_L(X^N)|$ hypotheses. An ML reconstruction algorithm chooses an $\hat{X}^N \in \mathcal{A}_L(X^N)$ that satisfies

$$\hat{X}^N = \operatorname*{arg\,max}_{\tilde{X}^N \in \mathcal{A}_L(X^N)} \mathbb{P}\left[Y^N \,\middle|\, \tilde{X}^N \right], \tag{4}$$

or equivalently, a proper permutation (ordering) of the fragments $\{\boldsymbol{X}(i)\}_{i\in[M]}$. The ML reconstruction can be cast as a max-weight matching problem, and thus can be computed in $O(M^3)$ time [13], or via message passing algorithms [5]. The fragments of $\hat{X}^N$ are similarly denoted by $\hat{\boldsymbol{X}}(i) = \hat{X}_{(i-1)L+1}^{iL}$, and the fragments of Y^N by $\boldsymbol{Y}(i) = Y_{(i-1)L+1}^{iL}$.

Let $\Delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a distortion measure. With a slight abuse of notation, the distortion measure is additively extended to length-L fragments $\tilde{\boldsymbol{X}}, \overline{\boldsymbol{X}} \in \mathcal{X}^L$ as

$$\Delta(\tilde{\boldsymbol{X}}, \overline{\boldsymbol{X}}) = \frac{1}{L} \sum_{j \in [L]} \Delta(\tilde{X}_j, \overline{X}_j). \tag{5}$$

Given a desired distortion level $\delta > 0$, $\hat{\boldsymbol{X}}(i)$ is said to fail to reconstruct $\boldsymbol{X}(i)$ if $\Delta(\boldsymbol{X}(i), \hat{\boldsymbol{X}}(i)) \geq \delta$. Let

$$\Xi_\delta(X^N, \hat{X}^N) := \frac{1}{M} \sum_{i\in[M]} \mathbb{1}\{\Delta(\boldsymbol{X}(i), \hat{\boldsymbol{X}}(i)) > \delta\} \tag{6}$$

be the relative number of fragments that failed to be properly reconstructed at distortion level δ. The reconstruction failure probability at distortion level $\delta \geq 0$ and failure level $\xi \in [0,1)$ is then

$$\mathsf{FP}(\delta, \xi) := \mathbb{P}\left[\Xi_\delta(X^N, \hat{X}^N) \geq \xi \right]. \tag{7}$$

Our goal is to establish conditions under which $\mathsf{FP}(\delta, \xi)$ asymptotically vanishes, as $M \to \infty$. We assume that the length of the fragments scales logarithmically with the number of fragments M, and the scaling is determined by a *fragment length parameter* $\beta > 0$ as

$$L = \beta \cdot \log M. \tag{8}$$

Note that it holds for this parametrization that $|\mathcal{X}|^L = M^\beta$ and $M = e^{\frac{L}{\beta}}$.

In what follows, the probability of a reconstruction failure will be bounded using the *Bhattacharyya distance* and more generally, using the *Chernoff distance*. For a pair of symbols $\overline{x}, \tilde{x} \in \mathcal{X}$, a transition probability kernel $P_{Y|X}$, and a parameter $s \in [0,1]$, the Chernoff distance is denoted by

$$d_{P_{Y|X},s}(\overline{x}, \tilde{x}) := -\log \sum_{y \in \mathcal{Y}} P_{Y|X}^s(y \mid \overline{x}) \cdot P_{Y|X}^{1-s}(y \mid \tilde{x}). \tag{9}$$

In most of this paper, this distance will be used for $s = 1/2$. In this case $d_{P_{Y|X},1/2}(\overline{x}, \tilde{x})$ is symmetric, it will be referred to as the Bhattacharyya distance, and s will be omitted from the notation. The Chernoff distance for a pair of sequences $\overline{\boldsymbol{x}}, \tilde{\boldsymbol{x}} \in \mathcal{X}^L$ is additively defined by $d_{P_{Y|X},s}(\overline{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) := \sum_{i\in[L]} d_{P_{Y|X},s}(\overline{x}_i, \tilde{x}_i)$. This additive distance only depends on the joint type of $(\overline{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$. Accordingly, for a given joint type $Q_{\overline{X}\tilde{X}} \in \mathcal{P}_L(\mathcal{X}^2)$ for some $L \in \mathbb{N}$, we denote (with a slight abuse of notation) $d_{P_{Y|X},s}(Q_{\overline{X}\tilde{X}}) := \frac{1}{L} d_{P_{Y|X},s}(\overline{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ where $(\overline{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) \in \mathcal{T}_L(Q_{\overline{X}\tilde{X}})$ is arbitrary. The definition can then be continuously extended to any joint PMF $Q_{\overline{X}\tilde{X}}$ in the interior of $\mathcal{P}(\mathcal{X}^2)$. Similarly, the distortion $\Delta(\overline{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ between $\overline{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$ only depends on their joint type $Q_{\overline{X}\tilde{X}}$, and so we also denote it by $\Delta(Q_{\overline{X}\tilde{X}})$. The definition is then continuously extended to any $Q_{\overline{X}\tilde{X}}$ in the interior of $\mathcal{P}(\mathcal{X}^2)$.

III. Main Results

We next describe our results for the no-repeating fragments regime (Sec. III-A), and then for the repeating-fragments regime (Sec. III-B).

A. The No-Repeating-Fragments Regime with Zero Distortion

In this section, we address the regime in which typically all fragments of X^N are unique, and no distortion is allowed, i.e., $\delta = 0$. We thus abbreviate to reconstruction failure probability to $\mathsf{FP}(\xi)$. Let

$$\psi_2(P_{XY}) := \min_{Q_{X_1X_2}\in\mathcal{P}(\mathcal{X}^2)} \frac{1}{2} D_{\text{KL}}\left(Q_{X_1X_2} \,||\, P_X^{\otimes 2}\right) + d_{P_{Y|X}}(Q_{X_1X_2}), \tag{10}$$

where, we recall that $d_{P_{Y|X}}(Q_{X_1X_2}) = \mathbb{E}_{Q_{X_1X_2}}[d_{P_{Y|X}}(X_1, X_2)]$, and $d_{P_{Y|X}}(X_1, X_2)$ is the Bhattacharyya distance (9). Essentially, $\psi_2(P_{XY})$ is the *rate function* for the probability of a transposition reconstruction error. Intuitively, $\psi_2(P_{XY})$ can be thought of as follows. For a given pair of fragment sequences $\boldsymbol{x}(1)$ and $\boldsymbol{x}(2)$ that have a joint type $Q_{X_1X_2}$, the term $d_{P_{Y|X}}(Q_{X_1X_2})$ captures how hard it is to confuse them after observing them through the channel $P_{Y|X}$, and the term $D_{\text{KL}}(Q_{X_1X_2} \,||\, P_X^{\otimes 2})$ captures how unlikely it is for us to see fragments $\boldsymbol{X}(1) = \boldsymbol{x}(1)$ and $\boldsymbol{X}(2) = \boldsymbol{x}(2)$, when the two fragments are generated IID according to P_X. Minimizing over $Q_{X_1X_2}$ corresponds to finding the worst-case pair of fragments type, which are most likely to produce a reconstruction error.

The minimization problem in (10) can be easily solved by using Jensen's inequality to obtain a lower bound on the minimized argument and show that it is achievable. The short derivation appears in Appendix A, and the result is

$$\psi_2(P_{XY}) := -\frac{1}{2}\log\left[\sum_{x_1,x_2\in\mathcal{X}^2} P_X(x_1)P_X(x_2)\cdot e^{-2d_{P_{Y|X}}(x_1,x_2)}\right]. \tag{11}$$

Furthermore, since $d_{P_{Y|X}}(x,x) = 0$, an alternative expression is

$$\psi_2(P_{XY}) := -\frac{1}{2}\log\left[e^{-H_2(P_X)} + \sum_{x_1,x_2\in\mathcal{X}^2 : x_1\neq x_2} P_X(x_1)P_X(x_2)\cdot e^{-2d_{P_{Y|X}}(x_1,x_2)}\right], \tag{12}$$

which shows that $\psi_2(P_{XY}) \to \frac{1}{2}H_2(P_X)$ as the channel $P_{Y|X}$ approaches a clean channel.

1) An Upper Bound on the Reconstruction Error:

Theorem 1. *If $\beta > \frac{1}{\psi_2(P_{XY})}$ then for $\xi = 0$*

$$\mathsf{FP}(\xi = 0) = O\left(M^{2[1-\beta\psi_2(P_{XY})]}\right) \tag{13}$$

with a constant that depends on P_{XY}, and for $\xi > 0$

$$\mathsf{FP}(\xi) \leq \exp\left[-M\log M\cdot\xi\left(\beta\psi_2(P_{XY}) - 1 - O\left(\frac{1}{M}\right)\right)\right]. \tag{14}$$

Discussion: The bound of Theorem 1 shows a sharp threshold as a function of ξ. For perfect reconstruction ($\xi = 0$) the failure probability decays polynomially in M, whereas for imperfect reconstruction ($\xi > 0$) it decays exponentially with $M\log M$, which is much faster. The error bound in the $\xi = 0$ case is dominated by transposition errors, that is, an almost perfect reconstruction of the sequence, except for a single pair of fragments that have exchanged their location. The rate function determining the threshold is given by $\psi_2(P_{XY})$. When $\xi > 0$, a wrong placement of less than ξM fragments is not considered to be a failure, and so transpositions and other permutations with $M - K$ fixed points, K fixed, do not lead to a failure. For $\xi > 0$, the error event that dominates this bound is a set of $\frac{\xi M}{2}$ transpositions.

Proof sketch of Theorem 1: The proof of Theorem 1 first addresses a fixed permutation $\pi \in S_M$. For any such π, the error is essentially a pairwise error event between X^N and its permuted version $\tilde{X}^N := \pi[X^N]$. This pairwise error is bounded using the standard Bhattacharyya upper bound (e.g., [35, Sec. 2.3]), and then averaged over X^N. Finally, using the union bound, the reconstruction failure probability is upper bounded by summing over all possible permutations.

Since each permutation is a composition of cycles, as in [33], we upper bound the average pairwise error probability for cycles. In this respect, a main ingredient of the proof is the next lemma, which upper bounds the expected Bhattacharyya upper bound for a cycle of length K.

Lemma 2. *Let $X_1^K \sim P_X^{\otimes K}$ IID over a finite alphabet $\mathcal{X}$. Let $\pi \in S_K$ be a cycle of length K, and let $\tilde{X}_j = X_{\pi(j)}$ for $j \in [K]$. Let $P_{Y|X}$ be a transition probability kernel. Then,*

$$\mathbb{E}\left[\exp\left(-d_{P_{Y|X}}(X_1^K, \tilde{X}_1^K)\right)\right] \leq e^{-K\cdot\psi_2(P_{XY})}, \tag{15}$$

where $\psi_2(P_{XY})$ is defined in (10).

The proof of Lemma 2 is based on first upper bounding the expected Bhattacharyya upper bound (left-hand side of (15)) using the *Donsker–Varadhan variational formula* [12], [3, Corollary 4.15]. The resulting upper bound is given by $e^{-K\cdot\psi_K(P_{XY})}$, where the rate function $\psi_K(P_{XY})$ is a generalized version of $\psi_2(P_{XY})$ for cycles of length K, given as a minimization problem over $\mathcal{P}(\mathcal{X}^K)$ (see (A.8) in Appendix (A)). The proof of the lemma then continues by establishing that transpositions, i.e., cycles of length 2, have the minimal rate function, to wit, $\psi_K(P_{XY}) \geq \psi_2(P_{XY})$ for all $K \geq 2$. The proof of this claim involves two different arguments. First, the special symmetry of the case $K = 3$ is used to show that $\psi_3(P_{XY}) \geq \psi_2(P_{XY})$. Specifically, the Bhattacharyya distance for a length-3 cycle is given by $d_{P_{Y|X}}(Q_{X_1X_2}) + d_{P_{Y|X}}(Q_{X_2X_3}) + d_{P_{Y|X}}(Q_{X_1X_3})$, which is half of the Bhattacharyya distance of 3 length-2 cycles. Favorably, the third-order KL divergence involved in the optimization problem of $\psi_3(P_{XY})$, to wit, $D_{\text{KL}}(Q_{X_1X_2X_3} \mid\mid P_X^{\otimes 3})$, is analogously lower bounded by the KL divergence of the marginal pairs using *Han's inequality for the KL divergence* [3, Theorem 4.9] [16]. For $K \geq 4$, such a symmetry does not seem possible to easily exploit. Instead, we consider a relaxed lower bound $\psi_K(P_{XY}) \geq \varphi_K(P_{XY})$, where $\varphi_K(P_{XY})$ is obtained by a relaxation of the minimization problem involved in the definition of $\psi_K(P_{XY})$, and show that $\varphi_K(P_{XY}) \geq \psi_2(P_{XY})$ for all $K \geq 4$. The relaxation from $\psi_K(P_{XY})$ to $\varphi_K(P_{XY})$, essentially breaks the cycle, by removing the constraint that $\tilde{X}_1 = X_K$. This enables to show that the minimizer of $\varphi_K(P_{XY})$ in $\mathcal{P}(\mathcal{X}^K)$ must satisfy a Markov chain condition $X_1 - X_2 - \cdots - X_K$, and consequently reduces the problem from a K-dimensional joint PMF in $\mathcal{P}(\mathcal{X}^K)$ to a simple pairwise joint PMF in $\mathcal{P}(\mathcal{X}^2)$. This Markov condition clearly cannot be satisfied with the original cyclic constraint of $\tilde{X}_1 = X_K$, and this is why the relaxation from $\psi_K(P_{XY})$ to $\varphi_K(P_{XY})$ is necessary. Substituting the estimate of Lemma 2 to the aforementioned union bound over all permutations, while taking into account the fact that different cycles of a permutation are independent, directly leads to the upper bounds in Theorem 1.

A comparison with [33]: The setting in [33] assumed that P_X is a uniform binary source $P_X(X = 0) = P_X(X = 1) = 1/2$, and that $P_{Y|X}$ is a BSC (as well as $\xi = 0$, although the results therein most likely can be extended to $\xi > 0$ in a simple way). For this setting, it was only established that the worst permutation is either a transposition (length-2 cycle) or a length-3 cycle. As we show here, it in fact holds that the worst case is a transposition, and this holds for a general P_{XY}. The proof of this property leads to the improved bound on the failure probability with polynomial decrease $O(M^{1-\beta\psi_2(P_{XY})})$ compared to $O(M^{1-\beta(\psi_2(P_{XY})\vee\psi_3(P_{XY}))})$ that can be conjectured from [33] for the general case. A similar effect holds for the $\xi > 0$ case. We finally mention that the "break of the cycle" argument that was use here to relax $\psi_K(P_{XY})$ to $\varphi_K(P_{XY})$ is inspired from [33], in which the contribution of the Bhattacharyya distance of the last pair of fragments $d_{P_{Y|X}}(X_K, \tilde{X}_K)$ was ignored, in order to obtain tractable bounds.

2) A Lower Bound on the Reconstruction Error: We next state a lower bound on $\mathsf{FP}(\xi)$:

Theorem 3. *Assume that $d_{P_{Y|X}}(x_1, x_2) < \infty$ and that*

$$\psi_2(P_{XY}) < \frac{1}{2} H_2(P_X). \tag{16}$$

If $\beta > \frac{1}{\psi_2(P_{XY})}$ then it holds for $\xi = 0$ that

$$\mathsf{FP}(\xi = 0) \geq M^{2[1-\beta\psi_2(P_{XY})]+o(1)} \tag{17}$$

and for $\xi > 0$ that

$$\mathsf{FP}(\xi) \geq \exp\left[-\xi M \log M \cdot [\beta\psi_2(P_{XY}) + o(1)]\right]. \tag{18}$$

Theorem 3 establishes the tightness of the upper bound in Theorem 1 for $\xi = 0$, and suffers from a gap of $\xi M \log M$ in the exponent for $\xi > 0$.

The origin of the qualifying assumptions: The condition $d_{P_{Y|X}}(x_1, x_2) < \infty$ is technical, and related to the uniform continuity of $Q_{X_1X_2} \to d_{P_{Y|X}}(Q_{X_1X_2})$ over $\mathcal{P}(\mathcal{X}^2)$ required to modify a maximum over types in $\mathcal{P}_L(\mathcal{X}^2)$ to a maximum over PMFs in the entire probability simplex $\mathcal{P}(\mathcal{X}^2)$. The condition (16) is related to the fact that if $\boldsymbol{X}(1) = \boldsymbol{X}(2)$ has occurred then the probability that the reconstruction algorithm erroneously transposes $\boldsymbol{X}(1)$ and $\boldsymbol{X}(2)$ is zero, simply because they are identical (this is where the design goal in the ordering problem setting defers from that of the BI problem). This is gauged by the second-order Rényi entropy, which is related to the *collision probability* via $\mathbb{P}[\boldsymbol{X}(1) = \boldsymbol{X}(2)] = e^{-H_2(P_X)}$, and the assumption assures that this probability is negligible compared to the probability of erroneous reconstruction exchanging $\boldsymbol{X}(1)$ and $\boldsymbol{X}(2)$, whenever they are different.

Proof sketch of Theorem 3: The proof of Theorem 3 first considers the event in which exchanging the order of $\boldsymbol{X}(i_1)$ and $\boldsymbol{X}(i_2)$ for some $i_1, i_2 \in [M], i_1 < i_2$ is more likely than the correct order (though this does not imply that the ML reconstruction will actually have a transposition error in these locations). The probability of this event can be lower bounded using the technique of Shannon, Gallager and Berlekamp [28, Corollary to Thm. 5]. In turn, this technique is based on Chernoff's bound, and hence, involves an optimized version over $s \in [0, 1]$ of the Chernoff distance, rather than the Bhattacharyya distance. Nonetheless, it is shown in the proof that the optimum is obtained for $s = 1/2$. For $\xi = 0$, the lower bound on the reconstruction failure then considers a union over all possible $\binom{M}{2} = \frac{M(M-1)}{2}$ different transpositions. As is well known, the union bound clipped to 1 is order-tight for independent events (or just pairwise independent events). However, these transpositions are not pairwise independent events, and so it is not obvious that the union bound is actually tight in this case. For $\xi = 0$, we use de Caen's inequality [11] to establish the tightness of the union bound (as was also used in [33]). For $\xi > 0$, we simply lower bound the error via the error occurs for some (arbitrary) $\xi M/2$ transpositions. In principle, de Caen's inequality [11] may be used for the $\xi > 0$ setting too. However, using a seemingly natural extension of the $\xi = 0$ does not lead to an improvement over the simpler bound of a single set of $\xi M/2$ transpositions. It is possible that de Caen's inequality is not sufficiently tight for this setting (at least in the way we have attempted to use it), and the tightness in the $\xi > 0$ remains an open question.

Example 4 ($\psi_2(P_{XY})$ **for binary sources).** Consider a binary source $\mathcal{X} = \{0, 1\}$. The expression in (12) results

$$\psi_2(P_{XY}) = \frac{1}{2}\left[H_2(p) - \log\left(1 + \frac{2p(1-p)}{p^2+(1-p)^2}\cdot \mathsf{BC}^2(P_{Y|X})\right)\right] \tag{19}$$

where $\mathsf{BC}(P_{Y|X}) := e^{-d_{1\leftrightarrow 0}}$ with

$$d_{1\leftrightarrow 0} := -\log \sum_{y\in\mathcal{Y}} \sqrt{P_{Y|X}(y\mid 0)\cdot P_{Y|X}(y\mid 1)} \tag{20}$$

is the *Bhattacharyya coefficient*. More specifically, assume that $\mathcal{Y} = \{0,1\}$, and $P_{Y|X}$ is a BSC with crossover probability $\alpha \in [0, 1/2]$. Then, $\mathsf{BC}(\mathsf{BSC}(\alpha)) = \sqrt{4\alpha(1-\alpha)}$ and then

$$\psi_2(P_{XY}) = \frac{1}{2}\left[H_2(p) - \log\left(1 + \frac{8p(1-p)}{p^2+(1-p)^2}\cdot \alpha(1-\alpha)\right)\right]. \tag{21}$$

It can be seen that as $\alpha \downarrow 0$, it holds that $\psi_2(P_{XY}) \downarrow \frac{1}{2}H_2(p)$. The noiseless case $\alpha = 0$ shows the difference between our problem (sequence recovering) and the BI problem (permutation recovering): In our problem the actual exponent for $\alpha = 0$ is infinite (zero error, since $X = Y$ with probability 1), whereas for the BI problem $\frac{1}{2}H_2(p)$. This agrees with the qualifying condition of Theorem 3, given by $\psi_2(P_{XY}) \leq \frac{1}{2}H_2(p)$.

Example 5 ($\psi_2(P_{XY})$ **for symmetric general sources). Consider** P_X to be uniform over $\mathcal{X} = \mathcal{Y}$, and let the channel $P_{Y|X}$ be symmetric, in the sense that

$$P_{Y|X}(y\mid x;\alpha) := \begin{cases} 1-\alpha, & y = x \\ \frac{\alpha}{|\mathcal{Y}|-1} & \text{otherwise} \end{cases} \tag{22}$$

(this transition kernel generalizes the BSC to larger alphabets). The computed value of $\psi_2(P_{XY})$ for $P_{XY} = P_X \otimes P_{Y|X}(y \mid x;\alpha)$ as a function of α appears in Fig. 2. As might be expected, $\psi_2(P_{XY})$ increases with $|\mathcal{X}|$, and hence the lower bound on β decreases. This agree with intuition since ordering the fragments is easier for larger entropy sources.

B. The Repeating-Fragments Regime with Positive Distortion

In this section, we address the regime in which β is small, or the source PMF P_X has low entropy. This is the setting in which the difference between the BI problem and the ordering problem is most pronounced, since when fragments repeat themselves in the sequence, reconstruction of the sequence is possible without a reconstruction of the permutation. In this regime, multiple identical fragments are typically present in the sequence X^N. Since in that case fragments which are similar according to the distortion measure Δ are also likely to be abound, we tolerate a positive distortion level. Intuitively, in this setting, a successful reconstruction is possible, because if a pair of fragments has distortion larger than the threshold δ, then it also has large Bhattacharyya distance, and so the correct order can be identified using the corresponding fragment in the reference sequence. Concretely, this can be gauged by

$$d^*_{P_{Y|X}}(\delta) := \min_{Q_{X_1X_2}\in\mathcal{P}(\mathcal{X}^2):\Delta(Q_{X_1X_2})\geq\delta} d_{P_{Y|X}}(Q_{X_1X_2}), \tag{23}$$

which is the minimal Bhattacharyya distance possible for any joint PMF of a pair of fragments whose distortion level is above δ. Clearly, there is a trade-off between the distortion level δ and the fraction ξ of failed reconstructed fragments that can be tolerated – increasing the distortion level δ allows to reduce ξ. Our main result in this section characterizes the trade-off between ξ and δ, which still allows for vanishing failure probability, as follows:

Theorem 6. *Assume that $\beta < \frac{1}{H(P_X)}$. Then, if*

$$\xi > \frac{H(P_X)}{d^*_{P_{Y|X}}(\delta)} \tag{24}$$

then $\mathsf{FP}(\delta,\xi) = o(1)$.

Discussion: Theorem 6 states a trade-off between δ and ξ in the repeating-fragments regime $\beta < 1/H(P_X)$. Interestingly, the minimal possible ξ for a given δ does not depend on β (as long as the later is sufficiently small). The resulting reconstruction failure probability then decays to zero, though in an unspecified rate, which is most likely slower rate compared to the no-repeating fragments regime, for which the reconstruction failure probability decays as $e^{-\Theta(\xi M\log M)}$ for $\xi > 0$. Evidently, the lower bound on ξ can be improved by increasing the Bhattacharyya distance, which can be considered as a measure of the signal strength, or *signal-to-noise ratio*. Specifically, given any $\xi > 0$, the "quality" of $P_{Y|X}$ should be such that $d^*_{P_{Y|X}}(\delta) \geq H(P_X)/\xi$. In other words, any arbitrarily small $\xi > 0$ can be compensated by taking $d^*_{P_{Y|X}}(\delta) \to \infty$, that is, making the channel $P_{Y|X}$ "cleaner" (specifically, if $Y = X$ with probability 1 then $d^*_{P_{Y|X}}(\delta) \uparrow \infty$ for any non-trivial distortion measure Δ). Theorem 6 states an achievable trade-off between (ξ,δ) and β, and evaluating the tightness of this trade-off and its possible dependence on β is an interesting open problem.

Proof of Theorem 6 – The typical cardinality of the set $\mathcal{A}_L(X^N)$: As stated in the problem formulation, the reconstruction problem is a hypothesis testing problem between a *random* number of $|\mathcal{A}_L(X^N)|$ hypotheses, or equivalently, all possible *different* reconstructed sequences. Upper bounds on the error probability in multiple hypothesis testing typically involve some sort of a union bound over the alternative hypotheses, and similarly so is our upper bound on the failure probability. Therefore, a main technical part is to establish a tight upper bound on the number of alternative hypotheses. If all fragments $\{\boldsymbol{X}(i)\}_{i\in[M]}$ are unique, then the number of possible reconstruction vectors is $M! = e^{M\log M + O(M)}$. However, if the source PMF P_X is such that some fragments in $\mathcal{X}^L$ are expected to repeat multiple times, then it is expected that $\log|\mathcal{A}_L(X^N)|$ will be significantly smaller than $M\log M + O(M)$. The main ingredient of the analysis of the reconstruction failure in this regime shows that $\log|\mathcal{A}_L(X^N)| \leq \beta H(P_X)\cdot M\log M$ essentially holds with

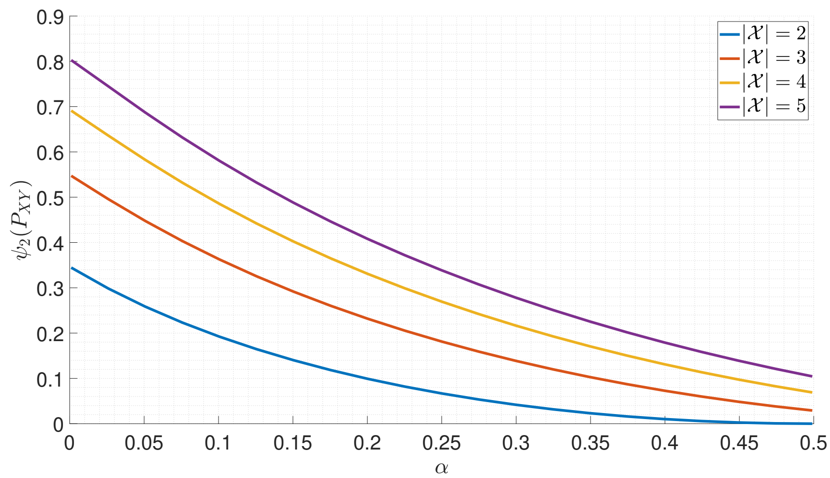

Figure 2. $\psi_2(P_{XY})$ for uniform P_X and symmetric channels parameterized by α.

probability $1 - o(1)$. This cardinality can be much smaller for low β or sources with low entropy.

To accurately present this bound, let us assume for notational simplicity, that the Lth order Cartesian product of $\mathcal{X}$ is arbitrarily ordered as $\mathcal{X}^L \equiv \{a_1 \ldots, a_{M^\beta}\}$, where we recall that $|\mathcal{X}^L| = M^\beta$. Then, for any given vector $x^N \in \mathcal{X}^N$ and any $j \in [M^\beta]$,

$$g_L(j; x^N) := \sum_{i\in[M]} \mathbb{1}\{\boldsymbol{x}(i) = a_j\} \tag{25}$$

is the number of times that the length-L vector $a_j \in \mathcal{X}^L$ appears in the fragments of x^N, and

$$g_L(x^N) := \left(g_L(1; x^N), g_L(2; x^N), \ldots, g_L(M^\beta; x^N)\right) \tag{26}$$
$$\in [M+1]^{M^\beta} \tag{27}$$

is the *histogram* vector of x^N for length-L fragments. It holds that $\sum_{j\in[M^\beta]} g_L(j; x^N) = M$. For brevity, we next denote the random number of appearances of the jth letter of $\mathcal{X}^L$ in the M fragments of X^N as $G(j) := g_L(j; X^N)$. The formal bound is as follows:

Proposition 7. *Assume that $H(P_X) > 0$. There exists a constant $c > 0$ so that for any $\eta \in (0,1)$, the log-cardinality of $\mathcal{A}_L(X^N)$ is concentrated as*

$$\mathbb{P}\left[\frac{1}{M}\log\left|\mathcal{A}_L(X^N)\right| \geq L\cdot H(P_X) + \eta \log M\right] = \begin{cases} \exp\left[-\Omega(\cdot\eta^2 M^{1\vee(2-\beta)})\right], & 0<\beta<2 \\ \frac{2}{M^{\eta/2}}, & \beta\geq 2 \end{cases}, \tag{28}$$

for all $M \geq M_0(P_X, \beta, \eta)$.

The proof of Prop. 7, which is fully presented in Appendix B, is based on the standard entropy bound on the multinomial coefficient, which then leads to the bound

$$\frac{1}{M}\log\left|\mathcal{A}_L(X^N)\right| \leq -\sum_{j\in[M^\beta]} \frac{G(j)}{M}\log\frac{G(j)}{M}. \tag{29}$$

Given the fragments model, the histogram vector $G = (G(1), \ldots, G(M^\beta))$ is distributed as a *multinomial* RV, and thus its components are statistically *dependent*. The upper bound on $\frac{1}{M}\log|\mathcal{A}_L(X^N)|$ is thus a complicated function of this random vector, and so it is difficult to directly analyze its random perturbation around its mean. Nonetheless, as is well known, the probability of an event under the multinomial distribution can be upper bounded by the probability of the same event under a properly defined Poisson distribution that has *independent* components (see Fact 9 and Lemma 10). We thus consider a Poissonized version $\tilde{G}$ of G, and analyze the tail behavior of $f(g)\colon \mathbb{N}_+ \to \mathbb{R}$ for $f(g) := -\frac{g}{M}\log\frac{g}{M}$. For $\beta \in (0,2)$, we show using concentration bounds for Lipschitz functions of Poisson RVs [2] that $f(g)$ is a sub-gamma random variable [3, Ch. 2], and then bound the concentration of $\sum_{j\in[M^\beta]} f(\tilde{G}(j))$ via Bernstein's inequality. A truncation argument is required since, strictly speaking, $f(g)$ involved in the upper bound is *not* Lipschitz continuous on $\mathbb{N}_+$. For $\beta > 2$ we use a standard Bernstein's inequality, after using looser bounding techniques.

Example 8 (A symmetric channel and Hamming distortion measure)**.** Assume that $\mathcal{X} = \mathcal{Y}$ and that $P_{Y|X}$ is a symmetric channel parameterized by α, as in (22). In this case, it holds that

$$d_{P^{(\alpha)}_{Y|X}}(x, \tilde{x}) = d_\alpha \cdot \mathbb{1}[\overline{x} \neq \tilde{x}] \tag{30}$$

where for any $\overline{x}, \tilde{x} \in \mathcal{X}$ with $\overline{x} \neq \tilde{x}$

$$d_\alpha := -\log \sum_{y\in\mathcal{Y}} \sqrt{P_{Y|X}(y \mid \overline{x}) \cdot P_{Y|X}(y \mid \tilde{x})} \tag{31}$$

$$= -\log\left[\sqrt{P_{Y|X}(\tilde{x} \mid \overline{x}) \cdot P_{Y|X}(\tilde{x} \mid \tilde{x})} + \sqrt{P_{Y|X}(\overline{x} \mid \overline{x}) \cdot P_{Y|X}(\overline{x} \mid \tilde{x})} + \sum_{y\in\mathcal{X}\setminus\{\overline{x},\tilde{x}\}} \sqrt{P_{Y|X}(y \mid \overline{x}) \cdot P_{Y|X}(y \mid \tilde{x})}\right] \tag{32}$$

$$= -\log\left[\sqrt{\frac{\alpha}{|\mathcal{Y}|-1}\cdot(1-\alpha)} + \sqrt{(1-\alpha)\cdot\frac{\alpha}{|\mathcal{Y}|-1}} + \frac{(|\mathcal{Y}|-2)\cdot\alpha}{|\mathcal{Y}|-1}\right] \tag{33}$$

$$= -\log\left[\sqrt{\frac{4(1-\alpha)\alpha}{|\mathcal{Y}|-1}} + \frac{(|\mathcal{Y}|-2)\cdot\alpha}{|\mathcal{Y}|-1}\right]. \tag{34}$$

Further assume that the distortion measure is the Hamming distortion measure $\Delta(\overline{x},\tilde{x}) = \mathbb{1}[\overline{x}\neq\tilde{x}]$. Thus, $d_{P^{(\alpha)}_{Y|X}}(x,\tilde{x}) \propto \Delta(\overline{x},\tilde{x})$ and then it is simple to obtain that $d^*(\delta) = \delta\cdot d_\alpha$, and the bound of Theorem 6 results in

$$\xi > \frac{H(P_X)}{\delta\cdot d_\alpha}. \tag{35}$$

The achievable trade-off between ξ and δ is shown in Fig. 3 for $\alpha = 0.1$ and $H(P_X) = 0.1$ [nats], for varying alphabet sizes. As can be seen, the minimal ξ is improving for larger alphabet sizes, though this improvement has diminishing returns. We finally remark that computing $d^*_{P_{Y|X}}(\delta)$ for general channels is a simple linear program (23), and thus can be easily computed for any arbitrary $P_{Y|X}$ and distortion measure Δ.

IV. Conclusion and Future Research

We have considered the problem of ordering the multiset of consecutive fragments of a sequence, based on a reference sequence. We have assumed a general setting, in which each fragment can be reconstructed with a low distortion level, and a constant fraction of the fragments can be reconstructed with a high distortion. First, we considered the regime in which typically all fragments are unique, and so focused on zero distortion. For a general joint source P_{XY}, we have derived an upper bound on the fragment length required for reliable reconstruction, both for a perfect and a partial reconstruction. These results tighten and extend previous results derived for the BI problem in the random coding regime, which were restricted to uniform binary sources P_X and symmetric $P_{Y|X}$. In the perfect reconstruction setting ($\xi = 0$) the bound was proved to be tight, whereas a lower bound was derived for the partial reconstruction ($\xi > 0$) setting. There is a gap between the bounds in the latter setting, which is an interesting and challenging open problem, mainly since using the standard technique based on de Caen's inequality [11] appears to fail. Second, we considered the regime in which repeating fragments abound, and showed its relation to the entropy of the source P_X. In this regime, it is natural to tolerate a positive distortion $\delta > 0$ between the fragments and their reconstruction. We show that as long as β is small enough ($\beta < 1/H(P_X)$) so the reconstruction algorithm operates in the repeating-fragments regime, there is a trade-off (24) between the minimal ξ possible for the given δ, which still assures vanishing failure probability. As said, evaluating the tightness of the trade-off is an interesting open problem, and specifically, whether the optimal trade-off depends on β or not. Furthermore, it is of interest to investigate the optimal decay rate of the reconstruction failure probability, and how it depends on the problem parameters.

Other avenues for future research include: (i) Reconstruction with possibly overlapping fragments and high coverage, taken at random locations (as in [24]) (ii) Reconstruction with a compressed version of either the fragments or the reference sequence (or both), (iii) Reconstruction using fragments that were obtained from Y^N via other channels, e.g., those which include deletions or insertions (or both), (iv) Reconstruction using multiple reference fragments, and more.

Appendix A
Proofs for the No-Repeating-Fragments Regime with Zero Distortion

Proof of (11) and (12): The argument in the optimization problem (10) defining $\psi_2(P_{XY})$ satisfies for any $Q_{X_1X_2} \in \mathcal{P}(\mathcal{X}^2)$

$$\frac{1}{2}D_{\text{KL}}\left(Q_{X_1X_2}\mid\mid P_X^{\otimes 2}\right) + d_{P_{Y|X}}(Q_{X_1X_2})$$
$$= -\frac{1}{2}\sum_{x_1,x_2\in\mathcal{X}^2} Q_{X_1X_2}(x_1,x_2) \times\left[\log\frac{P_X(x_1)P_X(x_2)}{Q_{X_1X_2}(x_1,x_2)} - 2d_{P_{Y|X}}(x_1,x_2)\right] \tag{A.1}$$
$$= -\frac{1}{2}\sum_{x_1,x_2\in\mathcal{X}^2} Q_{X_1X_2}(x_1,x_2) \times\log\frac{P_X(x_1)P_X(x_2)e^{-2d_{P_{Y|X}}(x_1,x_2)}}{Q_{X_1X_2}(x_1,x_2)} \tag{A.2}$$
$$\overset{(a)}{\geq} -\frac{1}{2}\log\sum_{x_1,x_2\in\mathcal{X}^2} P_X(x_1)P_X(x_2)e^{-2d_{P_{Y|X}}(x_1,x_2)}, \tag{A.3}$$

where (a) follows from Jensen's inequality for the convex function $t\to-\log t$, and equality is achieved when the averaged arguments are all equal, that is,

$$Q_{X_1X_2}(x_1,x_2) = \frac{P_X(x_1)P_X(x_2)e^{-2d_{P_{Y|X}}(x_1,x_2)}}{\sum_{x_1',x_2'\in\mathcal{X}^2} P_X(x_1')P_X(x_2')e^{-2d_{P_{Y|X}}(x_1',x_2')}}. \tag{A.4}$$

This proves (11). The expression in (12) follows directly from the definition of the second-order Rényi entropy.

In order to prove Theorem 1, we begin by proving Lemma 2.

Proof of Lemma 2: We denote the length-K cycle, in a two-line notation, as

$$\pi_K^\circ := \begin{pmatrix} 1 & 2 & 3 & \cdots & K-1 & K \\ K & 1 & 2 & \cdots & K-2 & K-1\end{pmatrix}. \tag{A.5}$$

By the variational representation of Donsker–Varadhan [12] (e.g., [3, Corollary 4.15]), for any $Q_{X_1X_2\dots X_K}\in\mathcal{P}(\mathcal{X}^K)$

$$D_{\text{KL}}(Q_{X_1X_2\dots X_K}\mid\mid P_X^{\otimes K}) + \mathbb{E}_{Q_{X_1X_2\dots X_K}}\left[d_{P_{Y|X}}(X_1^K,\tilde{X}_1^K)\right]$$
$$\geq -\log\mathbb{E}_{P_X^{\otimes K}}\left[\exp\left(-d_{P_{Y|X}}(X_1^K,\tilde{X}_1^K)\right)\right]. \tag{A.6}$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3390068

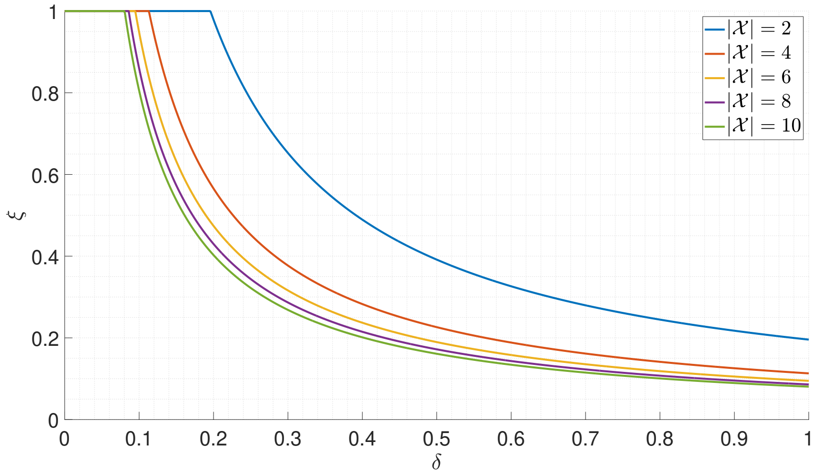

Figure 3. The trade-off between ξ and δ for $H(P_X) = 0.1$[nats] and symmetric channels $P^{(\alpha)}_{Y|X}$ for $\alpha = 0.1$.

Minimizing over $Q_{X_1X_2\ldots X_K}$ while using that $P_X^{\otimes K}$ has full support and thus $P_X^{\otimes K} \gg Q_{X_1X_2\ldots X_K}$ holds for any PMF $Q_{X_1X_2\ldots X_K}$, results

$$\mathbb{E}\left[\exp\left(-d_{P_{Y|X}}(X_1^K, \tilde{X}_1^K)\right)\right] \leq e^{-K\cdot\psi_K(P_{XY})} \tag{A.7}$$

where $\psi_K(P_{XY})$ is given by

$$\begin{aligned}\psi_K&(P_{XY})\\ :=& \min_{Q_{X_1X_2\ldots X_K}\in\mathcal{P}(\mathcal{X}^K)} \frac{1}{K} D_{\mathrm{KL}}\left(Q_{X_1X_2\ldots X_K} \,||\, P_X^{\otimes K}\right)\\ &+ \frac{1}{K}\sum_{i\in[K]} d_{P_{Y|X}}(Q_{X_iX_{\pi_K^\circ(i)}}).\end{aligned} \tag{A.8}$$

We next show that $\psi_K(P_{XY}) \geq \psi_2(P_{XY})$ for all $K \geq 2$. We prove this property separately for $K = 3$ and $K \geq 4$.

We prove that $\psi_3(P_{XY}) \geq \psi_2(P_{XY})$ by utilizing Han's inequality for the KL divergence [3, Thm. 4.9] [16], which states that for any probability measure $Q_{Z_1Z_2\cdots Z_K} \in \mathcal{P}(\mathcal{Z}^K)$ and a product probability measure $P_{Z_1} \otimes P_{Z_2} \cdots \otimes P_{Z_K} \in \mathcal{P}(\mathcal{Z}^K)$ it holds that

$$\begin{aligned}D_{\mathrm{KL}}&(Q_{Z_1Z_2\cdots Z_K} \,||\, P_{Z_1} \otimes P_{Z_2} \cdots \otimes P_{Z_K})\\ &\geq \frac{1}{K-1}\sum_{j\in[K]} D_{\mathrm{KL}}(Q_{Z_1\cdots Z_{j-1}Z_{j+1}\cdots Z_K} \,||\\ &\qquad P_{Z_1} \otimes \cdots P_{Z_{j-1}} \otimes P_{Z_{j+1}} \cdots \otimes P_{Z_K}),\end{aligned} \tag{A.9}$$

where $Q_{Z_1\cdots Z_{j-1}Z_{j+1}\cdots Z_K}$ is understood as the joint PMF of Z_1^K marginalized over Z_j. Indeed, it then holds that

$$\begin{aligned}\psi_3&(P_{XY})\\ =& \min_{Q_{X_1X_2X_3}} \frac{1}{3} D_{\mathrm{KL}}\left(Q_{X_1X_2X_3} \,||\, P_X^{\otimes 3}\right)\\ &+ \frac{1}{3}\sum_{i=1}^{3} d_{P_{Y|X}}(Q_{X_iX_{\pi_K^\circ(i)}})\end{aligned} \tag{A.10}$$

$$\begin{aligned}\stackrel{(a)}{\geq}& \min_{Q_{X_1X_2X_3}} \frac{1}{3}\cdot\frac{1}{2}\Big[D_{\mathrm{KL}}\left(Q_{X_1X_2} \,||\, P_X^{\otimes 2}\right) + 2d_{P_{Y|X}}(Q_{X_1X_2})\\ &+ D_{\mathrm{KL}}\left(Q_{X_1X_3} \,||\, P_X^{\otimes 2}\right) + 2d_{P_{Y|X}}(Q_{X_1X_3})\\ &+ D_{\mathrm{KL}}\left(Q_{X_2X_3} \,||\, P_X^{\otimes 2}\right) + 2d_{P_{Y|X}}(Q_{X_2X_3})\Big]\end{aligned} \tag{A.11}$$

$$\begin{aligned}\geq& \frac{1}{3}\cdot\frac{1}{2}\Big[\min_{Q_{X_1X_2}}\left\{D_{\mathrm{KL}}\left(Q_{X_1X_2} \,||\, P_X^{\otimes 2}\right) + 2d_{P_{Y|X}}(Q_{X_1X_2})\right\}\\ &+ \min_{Q_{X_1X_3}}\left\{D_{\mathrm{KL}}\left(Q_{X_1X_3} \,||\, P_X^{\otimes 2}\right) + 2d_{P_{Y|X}}(Q_{X_1X_3})\right\}\\ &+ \min_{Q_{X_2X_3}}\left\{D_{\mathrm{KL}}\left(Q_{X_2X_3} \,||\, P_X^{\otimes 2}\right) + 2d_{P_{Y|X}}(Q_{X_2X_3})\right\}\Big]\end{aligned} \tag{A.12}$$

$$= \psi_2(P_{XY}), \tag{A.13}$$

where (a) follows from Han's inequality (A.9).

We now turn to prove that $\psi_K(P_{XY}) \geq \psi_2(P_{XY})$ for all $K \geq 4$. To this end, consider the minimization problem involved in the upper bound rate function $\psi_K(P_{XY})$, to wit,

$$\begin{aligned}\min_{Q_{X_1X_2\ldots X_K}} &\frac{1}{K} D_{\mathrm{KL}}\left(Q_{X_1X_2\ldots X_K} \,||\, P_X^{\otimes K}\right)\\ &+ \frac{1}{K}\sum_{i\in[K]} d_{P_{Y|X}}(Q_{X_iX_{\pi_K^\circ(i)}}).\end{aligned} \tag{A.14}$$

Now, suppose that $Q^{(0)}_{X_1\ldots X_K}$ is a solution of the minimization problem in (A.14). Then, due to the circular symmetry of the objective function of (A.14),

$$Q^{(1)}_{X_1\ldots X_K} = Q^{(0)}_{X_{\pi_K^\circ(1)}\ldots X_{\pi_K^\circ(K)}} \tag{A.15}$$

attains the same value for the objective function. We may then recursively define

$$Q^{(j)}_{X_1\ldots X_K} = Q^{(j-1)}_{X_{\pi_K^\circ(1)}\ldots X_{\pi_K^\circ(K)}} \tag{A.16}$$

for all $j \in [K-1]\backslash\{1\}$, and similarly, each $Q^{(j)}_{X_1\ldots X_K}$ also attains the same value for the objective function. Since the KL divergence is convex and the Bhattacharyya distance is linear in $Q_{X_1X_2\ldots X_K}$, the objective function in (A.8) is convex in $Q_{X_1X_2\ldots X_K}$. Thus,

$$\overline{Q}_{X_1\ldots X_K} = \frac{1}{K-1}\sum_{j=0}^{K-1} Q^{(j)}_{X_1\ldots X_K} \tag{A.17}$$

may only attain a lower value for the objective function. Moreover, $\overline{Q}_{X_1 \ldots X_K}$ is such that $\overline{Q}_{X_1 X_2} = \overline{Q}_{X_2 X_3} = \cdots = \overline{Q}_{X_K X_1}$. Thus, the solution of the minimization problem in (A.8) must satisfy that all marginals of consecutive pairs is the same, let say $Q_{\tilde{X}_1 \tilde{X}_2} \in \mathcal{P}(\mathcal{X}^2)$. Let us define this set of PMFs as

$$\mathcal{Q}_{\text{pairs}}(Q_{\tilde{X}_1 \tilde{X}_2}) := \Big\{ Q_{X_1 X_2 \cdots X_K} \in \mathcal{P}(\mathcal{X}^K) \colon Q_{X_1 X_2} = Q_{X_2 X_3} = \cdots Q_{X_{K-1} X_K} = Q_{X_K X_1} = Q_{\tilde{X}_1 \tilde{X}_2} \Big\}. \tag{A.18}$$

Furthermore, let us define a slightly modified version of this set, given as

$$\hat{\mathcal{Q}}_{\text{pairs}}(Q_{\tilde{X}_1 \tilde{X}_2}) := \Big\{ Q_{X_1 X_2 \cdots X_K} \in \mathcal{P}(\mathcal{X}^K) \colon Q_{X_1 X_2} = Q_{X_2 X_3} = \cdots Q_{X_{K-1} X_K} = Q_{\tilde{X}_1 \tilde{X}_2} \Big\}, \tag{A.19}$$

where the only difference between $\mathcal{Q}_{\text{pairs}}(Q_{\tilde{X}_1 \tilde{X}_2})$ and $\hat{\mathcal{Q}}_{\text{pairs}}(Q_{\tilde{X}_1 \tilde{X}_2})$ is the relaxation of the constraint $Q_{X_K X_1} = Q_{\tilde{X}_1 \tilde{X}_2}$. Note that the removal of this constraint effectively "breaks" the cycle. Returning to (A.8), we get from this property that

$$\begin{aligned} &\psi_K(P_{XY}) \\ &:= \min_{Q_{\tilde{X}_1 \tilde{X}_2}} \Bigg\{ \min_{Q_{X_1 X_2 \cdots X_K} : \mathcal{Q}_{\text{pairs}}(Q_{\tilde{X}_1 \tilde{X}_2})} \frac{1}{K} D_{\text{KL}}\left(Q_{X_1 X_2 \ldots X_K} \,||\, P_X^{\otimes K}\right) + d_{P_{Y|X}}(Q_{\tilde{X}_1 \tilde{X}_2}) \Bigg\} &\text{(A.20)} \\ &\geq \min_{Q_{\tilde{X}_1 \tilde{X}_2}} \Bigg\{ \min_{Q_{X_1 X_2 \cdots X_K} : \hat{\mathcal{Q}}_{\text{pairs}}(Q_{\tilde{X}_1 \tilde{X}_2})} \frac{1}{K} D_{\text{KL}}\left(Q_{X_1 X_2 \ldots X_K} \,||\, P_X^{\otimes K}\right) + d_{P_{Y|X}}(Q_{\tilde{X}_1 \tilde{X}_2}) \Bigg\} &\text{(A.21)} \\ &=: \varphi_K(P_{XY}). &\text{(A.22)} \end{aligned}$$

This "cycle-break" of the set $\hat{\mathcal{Q}}_{\text{pairs}}(Q_{\tilde{X}_1 \tilde{X}_2})$ and the relaxation of $\psi_K(P_{XY})$ to $\varphi_K(P_{XY})$ is crucial to establish the following property: The optimal solution $Q^*_{X_1 X_2 \ldots X_K}$ of $\varphi_K(P_{XY})$ must respect the Markov chain $X_1 - X_2 - \cdots X_{K-1} - X_K$. Indeed, assume by contradiction that, under Q, this is not the case for X_K, that is $Q_{X_K|X_{K-1}} \neq Q_{X_K|X_{K-1} \cdots X_1}$. Then,

$$\begin{aligned} &D_{\text{KL}}\left(Q_{X_1 X_2 \cdots X_K} \,||\, P_X^{\otimes K}\right) \\ &= D_{\text{KL}}\left(Q_{X_1 X_2 \cdots X_{K-1}} \,||\, P_X^{\otimes (K-1)}\right) + D_{\text{KL}}\left(Q_{X_K|X_{K-1} X_{K-2} \cdots X_1} \,||\, P_X \mid Q_{X_1 X_2 \cdots X_{K-1}}\right) &\text{(A.23)} \\ &\geq D_{\text{KL}}\left(Q_{X_1 X_2 \cdots X_{K-1}} \,||\, P_X^{\otimes (K-1)}\right) + D_{\text{KL}}\left(Q_{X_K|X_{K-1}} \,||\, P_X \mid Q_{X_{K-1}}\right), &\text{(A.24)} \end{aligned}$$

where the inequality follows since the convexity of the KL divergence implies that

$$\begin{aligned} &D_{\text{KL}}\left(Q_{X_K|X_{K-1} X_{K-2} \cdots X_1} \,||\, P_X \mid Q_{X_1 X_2 \cdots X_{K-1}}\right) \\ &= \mathbb{E}_{Q_{X_1 \cdots X_{K-2}|X_{K-1}}} \left[D_{\text{KL}}\left(Q_{X_K|X_{K-1} X_{K-2} \cdots X_1}(\cdot \mid X_1 X_2 \cdots X_{K-1}) \,||\, P_X\right) \right] &\text{(A.25)} \\ &\geq D_{\text{KL}}\Big(\mathbb{E}_{Q_{X_1 \cdots X_{K-2}|X_{K-1}}} \left[Q_{X_K|X_{K-1} X_{K-2} \cdots X_1} \right] \,||\, P_X \mid Q_{X_1 X_2 \cdots X_{K-1}} \Big) &\text{(A.26)} \\ &\geq D_{\text{KL}}\left(Q_{X_K|X_{K-1}} \,||\, P_X \mid Q_{X_{K-1}}\right). &\text{(A.27)} \end{aligned}$$

Thus, we can replace any $Q^*_{X_1 X_2 \ldots X_K}$ with $Q^*_{X_1 X_2 \ldots X_{K-1}} \otimes Q^*_{X_K|X_{K-1}}$. Next, using a similar argument, the first KL divergence term $D_{\text{KL}}(Q_{X_1 X_2 \cdots X_{K-1}} \,||\, P_X^{\otimes (K-1)})$ in (A.24) can be similarly lower bounded, showing that $Q^*_{X_1 X_2 \ldots X_{K-1}}$ can be replaced by $Q^*_{X_1 X_2 \ldots X_{K-2}} \otimes Q^*_{X_{K-1}|X_{K-2}}$ to obtain a lower objective. Continuing repeating this argument in a recursive fashion results the Markov chain relation any optimal solution must satisfy.

Consequently, by the chain rule for the KL divergence

$$\begin{aligned} &\frac{1}{K} D_{\text{KL}}\left(Q^*_{X_1 X_2 \ldots X_K} \,||\, P_X^{\otimes K}\right) \\ &= \frac{1}{K} \Bigg[D_{\text{KL}}\left(Q^*_{X_1} \,||\, P_X\right) + \sum_{j=2}^{K} D_{\text{KL}}\left(Q^*_{X_j|X_{j-1}} \,||\, P_X \mid Q^*_{X_{j-1}}\right) \Bigg] &\text{(A.28)} \\ &= \frac{1}{K} \Bigg[D_{\text{KL}}\left(Q^*_{\tilde{X}_1} \,||\, P_X\right) + \sum_{j=2}^{K} D_{\text{KL}}\left(Q^*_{\tilde{X}_2|\tilde{X}_1} \,||\, P_X \mid Q^*_{\tilde{X}_1}\right) \Bigg] &\text{(A.29)} \\ &= \frac{1}{K} D_{\text{KL}}\left(Q^*_{\tilde{X}_1} \,||\, P_X\right) + \frac{K-1}{K} D_{\text{KL}}\left(Q^*_{\tilde{X}_2|\tilde{X}_1} \,||\, P_X \mid Q_{\tilde{X}_1}\right), &\text{(A.30)} \end{aligned}$$

where the first equality holds since the optimal solution $Q^*_{X_1 X_2 \ldots X_K}$ must satisfy that Markov chain relation $X_1 - X_2 - \cdots X_{K-1} - X_K$. Moreover, observing (A.21), we may add the constraint $Q_{\tilde{X}_1} = Q_{\tilde{X}_2}$ to the outer minimization, since otherwise the inner constraint $Q_{X_1 X_2} = Q_{X_2 X_3}$, e.g., would make the problem infeasible. Hence, from all the above,

$$\varphi_K(P_{XY}) = \min_{Q_{\tilde{X}_1 \tilde{X}_2} : Q_{\tilde{X}_1} = Q_{\tilde{X}_2}} \Bigg\{ \frac{1}{K} D_{\text{KL}}\left(Q_{\tilde{X}_1} \,||\, P_X\right) + \frac{K-1}{K} D_{\text{KL}}\left(Q_{\tilde{X}_2|\tilde{X}_1} \,||\, P_X \mid Q_{\tilde{X}_1}\right) + d_{P_{Y|X}}(Q_{\tilde{X}_1 \tilde{X}_2}) \Bigg\}. \tag{A.31}$$

Now, by convexity of the KL divergence, it holds that

$$D_{\text{KL}}\left(Q_{\tilde{X}_2|\tilde{X}_1} \,||\, P_X \mid Q_{\tilde{X}_1}\right) \geq D_{\text{KL}}\left(Q_{\tilde{X}_1} \,||\, P_X\right), \tag{A.32}$$

that is, the first KL divergence in (A.31) is smaller than the second one. Thus, the worst bound is obtained for $K = 4$, that is $\varphi_K(P_{XY}) \geq \varphi_4(P_{XY})$ for all $K \geq 4$. Finally,

$\varphi_4(P_{XY})$

$$= \min_{Q_{\tilde{X}_1\tilde{X}_2}: Q_{\tilde{X}_1}=Q_{\tilde{X}_2}} \left\{ \frac{1}{4} D_{\mathrm{KL}}\left(Q_{\tilde{X}_1} \,||\, P_X\right) + \frac{3}{4} D_{\mathrm{KL}}\left(Q_{\tilde{X}_2|\tilde{X}_1} \,||\, P_X \mid Q_{\tilde{X}_1}\right) + d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2}) \right\} \tag{A.33}$$

$$\overset{(a)}{\geq} \min_{Q_{\tilde{X}_1\tilde{X}_2}: Q_{\tilde{X}_1}=Q_{\tilde{X}_2}} \left\{ \frac{1}{4} D_{\mathrm{KL}}\left(Q_{\tilde{X}_1} \,||\, P_X\right) + \frac{1}{4} D_{\mathrm{KL}}\left(Q_{\tilde{X}_2} \,||\, P_X\right) + \frac{1}{2} D_{\mathrm{KL}}\left(Q_{\tilde{X}_2|\tilde{X}_1} \,||\, P_X \mid Q_{\tilde{X}_1}\right) + d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2}) \right\} \tag{A.34}$$

$$\overset{(b)}{=} \min_{Q_{\tilde{X}_1\tilde{X}_2}: Q_{\tilde{X}_1}=Q_{\tilde{X}_2}} \left\{ \frac{1}{2} D_{\mathrm{KL}}\left(Q_{\tilde{X}_1} \,||\, P_X\right) + \frac{1}{2} D_{\mathrm{KL}}\left(Q_{\tilde{X}_2|\tilde{X}_1} \,||\, P_X \mid Q_{\tilde{X}_1}\right) + d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2}) \right\} \tag{A.35}$$

$$\overset{(c)}{=} \min_{Q_{\tilde{X}_1\tilde{X}_2}: Q_{\tilde{X}_1}=Q_{\tilde{X}_2}} \frac{1}{2} D_{\mathrm{KL}}\left(Q_{\tilde{X}_1\tilde{X}_2} \,||\, P_X^{\otimes 2}\right) + d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2}) \tag{A.36}$$

$$= \psi_2(P_{XY}), \tag{A.37}$$

where (a) follows using the convexity of the KL divergence, as in (A.32) (used with a factor of $1/4$), (b) follows from the constraint $Q_{\tilde{X}_1} = Q_{\tilde{X}_2}$, and (c) follows from the chain rule for KL divergence. Thus, $\varphi_K(P_{XY}) \geq \psi_2(P_{XY})$ for all $K \geq 4$. This, combined with the bound $\varphi_3(P_{XY}) \geq \psi_2(P_{XY})$ previously derived completes the proof. ■

The proof of Theorem 1 is then as follows:

Proof of Theorem 1: Let $F(\pi)$ be the number of fixed points of the permutation $\pi \in S_M$, that is, $F(\pi) := |\{i \in [M]\colon \pi(i) = i\}|$. So, if $F(\pi) \geq M(1-\xi)$ then $\pi(X^N)$ is a successful reconstruction of X^N with probability 1. Hence,

$$\mathsf{FP}(\xi) \overset{(a)}{\leq} \sum_{\pi\in S_M: F(\pi)\leq M(1-\xi)} p_e[X^N \to \pi(X^N)] \tag{A.38}$$

$$\overset{(b)}{\leq} \sum_{K=\xi M}^{M} \sum_{\pi\in S_M: F(\pi)=M-K} \mathbb{E}\left[e^{-d_{P_{Y|X}}(X^N,\pi[X^N])}\right] \tag{A.39}$$

$$\overset{(c)}{\leq} \sum_{K=\xi M}^{M} e^{\frac{K}{\beta}L} \cdot \max_{\pi\in S_M: F(\pi)=M-K} \mathbb{E}\left[e^{-d_{P_{Y|X}}(X^N,\pi[X^N])}\right], \tag{A.40}$$

where (a) follows from the union bound, (b) follows from Bhattacharyya's bound [35, Sec. 2.3], and (c) follows since the set of permutations which has exactly $M-K$ fixed points has cardinality of $\binom{M}{K}K! \leq \prod_{j=0}^{K-1}(M-j) \leq M^K = e^{\frac{K}{\beta}L}$.

Now, recall that the Bhattacharyya distance is additive, that is, $d_{P_{Y|X}}(X^N,\pi[X^N]) = \sum_{i\in[M]} d_{P_{Y|X}}(\boldsymbol{X}(i),\boldsymbol{X}(\pi(i)))$. Now, consider a permutation with $F(\pi) = M-2$ fixed points. This is a transposition, and since $d_{P_{Y|X}}(\boldsymbol{X}(i),\boldsymbol{X}(\pi(i))) = 0$ if $\pi(i) = i$, it follows from Lemma 2 that

$$\mathbb{E}\left[e^{-d_{P_{Y|X}}(X^N,\pi[X^N])}\right] \leq e^{-2L\cdot\psi_2(P_{XY})}. \tag{A.41}$$

Similarly, a permutation with $F(\pi) = M-3$ fixed points can only be a cycle of length 3, and so it follows again from Lemma 2 that

$$\mathbb{E}\left[e^{-d_{P_{Y|X}}(X^N,\pi[X^N])}\right] \leq e^{-3L\cdot\psi_3(P_{XY})} \leq e^{-3L\cdot\psi_2(P_{XY})}. \tag{A.42}$$

Next, a permutation with $F(\pi) \leq M-4$ may be either a cycle or comprised of independent cycles. Suppose that the permutation has C cycles of lengths $\{K_j\}_{j\in[C]}$. By permuting the fragments of X^N if necessary, we may assume WLOG that they are consecutive, that is, the first cycle includes indices $\mathcal{I}_1 := \{1,2,\ldots,K_1\}$, the second includes $\mathcal{I}_2 := \{K_1+1,\ldots,K_1+K_2\}$ and so on. Since there are $F(\pi)$ fixed points it holds that $\sum_{j\in[C]} K_j = M - F(\pi)$. Thus we may write

$$d_{P_{Y|X}}(X^N,\pi[X^N]) = \sum_{j=1}^{C}\sum_{i\in\mathcal{I}_j} d_{P_{Y|X}}\left(\boldsymbol{X}(i),\boldsymbol{X}(\pi(i))\right), \tag{A.43}$$

in which the outer summation is over independent RVs. Then,

$$\mathbb{E}\left[e^{-d_{P_{Y|X}}(X^N,\pi[X^N])}\right] \overset{(a)}{=} \prod_{j=1}^{C} \mathbb{E}\left[e^{-\sum_{i\in[M]} d_{P_{Y|X}}(\boldsymbol{X}(i),\boldsymbol{X}(\pi(i)))}\right] \tag{A.44}$$

$$\overset{(b)}{\leq} \prod_{j=1}^{C} e^{-LK_j\psi_2(P_{XY})} \tag{A.45}$$

$$\overset{(c)}{=} e^{-(M-F(\pi))L\cdot\psi_2(P_{XY})}, \tag{A.46}$$

where (a) holds by independence, (b) follows from Lemma 2, and (c) follows since π has $F(\pi)$ fixed points.

Combining all the above,

$$\mathsf{FP}(\xi) \leq \sum_{K=\xi M}^{M} \exp\left[-KL\cdot\left(\psi_2(P_{XY}) - \frac{1}{\beta}\right)\right]. \tag{A.47}$$

Now, suppose that $\xi = 0$. Then if $K = 0$ then the permutation must be the identity permutation, and this is a perfect reconstruction. $K = 1$ is impossible, since a permutation in S_M cannot have $M-1$ fixed points. Hence,

$$\mathsf{FP}(\xi=0) \leq \sum_{K=2}^{\infty} \exp\left[-KL\cdot\left(\psi_2(P_{XY}) - \frac{1}{\beta}\right)\right] \tag{A.48}$$

$$\overset{(a)}{=} \frac{\exp\left[-2L\cdot\left(\psi_2(P_{XY}) - \frac{1}{\beta}\right)\right]}{1-\exp\left[-L\cdot\left(\psi_2(P_{XY}) - \frac{1}{\beta}\right)\right]} \tag{A.49}$$

$$\overset{(b)}{\leq} 2\cdot M^{2(1-\beta\psi_2(P_{XY}))}, \tag{A.50}$$

where (a) is a geometric series, and (b) holds when $1 - \exp[-L\cdot(\psi_2(P_{XY}) - \frac{1}{\beta})] \leq \frac{1}{2}$, which holds for all $M \geq M_0(P_{XY})$. Otherwise, if $\xi > 0$ then we may upper bound the sum by M times its maximal term, and so

$$\mathsf{FP}(\xi) \leq \exp\left[-M\log M\cdot\xi\left(\beta\psi_2(P_{XY})-1-O\left(\frac{1}{M}\right)\right)\right]. \quad \text{(A.51)}$$

This completes the proof of the theorem. ■

We now turn to prove the lower bound in Theorem 3.

Proof of Theorem 3: To lower bound the error probability, we first focus on a single pair of fragments (i_1, i_2) and lower bound the probability that a reconstruction which puts $\boldsymbol{X}(i_1)$ in the i_2th location and $\boldsymbol{X}(i_2)$ in the i_1th location is more likely than the opposite ordering. Concretely, for a pair $(i_1, i_2) \in [M]^2$ such that $i_1 < i_2$, we let

$$\tilde{\mathcal{E}}_{i_1 i_2} := \left\{\frac{P_{Y|X}^{\otimes 2L}\left[(\boldsymbol{Y}(i_2),\boldsymbol{Y}(i_1)) \mid (\boldsymbol{X}(i_1),\boldsymbol{X}(i_2))\right]}{P_{Y|X}^{\otimes 2L}\left[(\boldsymbol{Y}(i_1),\boldsymbol{Y}(i_1)) \mid (\boldsymbol{X}(i_1),\boldsymbol{X}(i_2))\right]} \geq 1\right\}, \quad \text{(A.52)}$$

and the event of interest is defined as the intersection of this event with the event that the two fragments are different, to wit,

$$\mathcal{E}_{i_1 i_2} := \tilde{\mathcal{E}}_{i_1 i_2} \cap \{\boldsymbol{X}(i_1) \neq \boldsymbol{X}(i_2)\}. \quad \text{(A.53)}$$

It should be noted, however, that the event $\mathcal{E}_{i_1 i_2}$ does not necessarily imply that the ML reconstruction transposes the pair (i_1, i_2), since, for example, placing $\boldsymbol{X}(i_1)$ in an index different from i_2 could result larger likelihood. For notational simplicity, we next assume that $i_1 = 1$ and $i_2 = 2$. Then, the probability of $\mathcal{E}_{12}$ equals to the error probability of the hypothesis testing problem between $H_0 : (\boldsymbol{X}(1), \boldsymbol{X}(2)) = (\tilde{\boldsymbol{X}}(1), \tilde{\boldsymbol{X}}(2))$ and $H_1 : (\boldsymbol{X}(1), \boldsymbol{X}(2)) = (\tilde{\boldsymbol{X}}(2), \tilde{\boldsymbol{X}}(1))$ based on the observations $\boldsymbol{Y} \sim P_{Y|X}^{\otimes 2L}(\cdot \mid (\boldsymbol{X}(1), \boldsymbol{X}(2)))$, when $\tilde{\boldsymbol{X}}(1), \tilde{\boldsymbol{X}}(2) \sim P_X^{\otimes L}$, independently, except whenever $\tilde{\boldsymbol{X}}(1) = \tilde{\boldsymbol{X}}(2)$, because then the hypothesis testing problem has large error probability, whereas the reconstruction failure probability is zero.

Consider first a fixed $\tilde{\boldsymbol{X}}(1) = \tilde{\boldsymbol{x}}(1), \tilde{\boldsymbol{X}}(2) = \tilde{\boldsymbol{x}}(2)$, and let the probability of erroneously deciding H_1 when H_0 is true (resp. deciding H_0 when H_1 is true) be $p_{0\to1}(\tilde{\boldsymbol{x}}(1), \tilde{\boldsymbol{x}}(2))$ (resp. $p_{1\to0}(\tilde{\boldsymbol{x}}(1), \tilde{\boldsymbol{x}}(2))$). By the celebrated result of Shannon, Gallager and Berlekamp [28, Corollary to Thm. 5] the error probability is lower bounded using the Chernoff distance (9). Specifically, the version in [9, Problem 10.20(b)] states that for any $\delta > 0$

$$\max\{p_{0\to1}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2)), p_{1\to0}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2))\} \geq e^{-L\cdot\max_{s\in[0,1]} d_{P_{Y|X},s}((\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2)),(\tilde{\boldsymbol{x}}(2),\tilde{\boldsymbol{x}}(1)))-L\delta} \quad \text{(A.54)}$$

$$= e^{-L\max_{s\in[0,1]}\left\{d_{P_{Y|X},s}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2))+d_{P_{Y|X},1-s}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2))\right\}-L\delta} \quad \text{(A.55)}$$

holds for all $L \geq L_0(\delta, P_{Y|X})$. Now, $s \to d_{P_{Y|X},s}(\overline{x}, \tilde{x})$ is a concave function of $s \in [0,1]$ (its second derivative $\frac{\partial^2}{\partial^2 s} d_{P_{Y|X},s}(\overline{x}, \tilde{x})$ is the variance of the tilted distribution $\frac{P_{Y|X}^s[y|\overline{x}]\cdot P_{Y|X}^{1-s}[y|\tilde{x}]}{\exp[-d_{P_{Y|X},s}(\overline{x},\tilde{x})]}$ and hence nonnegative; see [28, Thm. 5] and [35, Proof of Thm. 3.5.1.]). Then, $s \to d_{P_{Y|X},s}(\tilde{\boldsymbol{x}}(1), \tilde{\boldsymbol{x}}(2))$ is an average of concave functions and thus concave. Thus, for any $s \in [0,1]$

$$\frac{1}{2}d_{P_{Y|X},s}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2)) + \frac{1}{2}d_{P_{Y|X},1-s}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2)) \leq d_{P_{Y|X},s/2+(1-s)/2}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2)) = d(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2)), \quad \text{(A.56)}$$

that is, the Bhattacharyya distance. We thus get from the above that

$$\max\{p_{0\to1}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2)), p_{1\to0}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2))\} \geq \exp\left[-L\cdot\left(2d_{P_{Y|X}}(\tilde{\boldsymbol{x}}(1),\tilde{\boldsymbol{x}}(2))+\delta\right)\right]. \quad \text{(A.57)}$$

We next average this bound over the randomness of $\tilde{\boldsymbol{X}}(1), \tilde{\boldsymbol{X}}(2)$, while accounting for the requirement that $\tilde{\boldsymbol{X}}(1) \neq \tilde{\boldsymbol{X}}(2)$. To this end, let $a(\tilde{\boldsymbol{X}}(1), \tilde{\boldsymbol{X}}(2))$ be any integrable function of $\tilde{\boldsymbol{X}}(1), \tilde{\boldsymbol{X}}(2)$ (w.r.t. the probability measure $P_X^{\otimes 2L}$), that is upper bounded as $a(\tilde{\boldsymbol{x}}(1), \tilde{\boldsymbol{x}}(2)) \leq 1$ for all $\tilde{\boldsymbol{x}}(1), \tilde{\boldsymbol{x}}(2) \in \mathcal{X}^L$. Then,

$$\begin{aligned}
&\mathbb{E}\left[a(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\cdot\mathbb{1}\{\tilde{\boldsymbol{X}}(1)\neq\tilde{\boldsymbol{X}}(2)\}\right] \\
&= \mathbb{E}\left[a(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\right] - \mathbb{E}\left[a(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\cdot\mathbb{1}\{\tilde{\boldsymbol{X}}(1)=\tilde{\boldsymbol{X}}(2)\}\right] && \text{(A.58)}\\
&\geq \mathbb{E}\left[a(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\right] - \mathbb{P}\left[\tilde{\boldsymbol{X}}(1)=\tilde{\boldsymbol{X}}(2)\right] && \text{(A.59)}\\
&= \mathbb{E}\left[a(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\right] - e^{-LH_2(P_X)}, && \text{(A.60)}
\end{aligned}$$

where $H_2(P_X)$ is the second-order Rényi entropy (the collision entropy). Hence, using the method of types [9, Sec. 2.1]

$$\begin{aligned}
&\mathbb{E}\Big[\max\left\{p_{0\to1}(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2)), p_{1\to0}(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\right\} \times \mathbb{1}\{\tilde{\boldsymbol{X}}(1)\neq\tilde{\boldsymbol{X}}(2)\}\Big] \\
&\geq \mathbb{E}\left[\exp\left[-L\cdot\left(2d_{P_{Y|X}}\left(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2)\right)+\delta\right)\right]\right] - e^{-LH_2(P_X)} && \text{(A.61)}\\
&= \sum_{Q_{\tilde{X}_1\tilde{X}_2}\in\mathcal{P}_L(\mathcal{X}^2)} \mathbb{P}\left[(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\in T_L(Q_{\tilde{X}_1\tilde{X}_2})\right] \times \exp\left[-L\cdot\left(2d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2})+\delta\right)\right] - e^{-LH_2(P_X)} && \text{(A.62)}\\
&\geq \frac{1}{(L+1)^{|\mathcal{X}|^2}} \times \max_{Q_{\tilde{X}_1\tilde{X}_2}\in\mathcal{P}_L(\mathcal{X}^2)} e^{-2L\cdot\left(\frac{1}{2}D_{\text{KL}}(Q_{\tilde{X}_1\tilde{X}_2}||P_X^{\otimes 2})+d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2})+\frac{\delta}{2}\right)} - e^{-LH_2(P_X)} && \text{(A.63)}\\
&= e^{-2L\cdot\left[\min_{Q_{\tilde{X}_1\tilde{X}_2}\in\mathcal{P}_L(\mathcal{X}^2)} f(Q_{\tilde{X}_1\tilde{X}_2})+\frac{\delta}{2}-O\left(\frac{|\mathcal{X}|^2\log L}{L}\right)\right]} - e^{-LH_2(P_X)}, && \text{(A.64)}
\end{aligned}$$

where

$$f(Q_{\tilde{X}_1\tilde{X}_2}) := \frac{1}{2}D_{\text{KL}}(Q_{\tilde{X}_1\tilde{X}_2} \mid\mid P_X^{\otimes 2}) + d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2}) \quad \text{(A.65)}$$

is the objective function involved in the optimization of the above rate function. Our next goal is to consider the optimization over this function, which in (A.64) is over $\mathcal{P}_L(\mathcal{X}^2)$. In order to replace this optimization set with a simpler optimization over the entire probability simplex $\mathcal{P}(\mathcal{X}^2)$, it suffices to prove that the function $Q_{\tilde{X}_1\tilde{X}_2} \to f(Q_{\tilde{X}_1\tilde{X}_2})$ is equicontinuous w.r.t. to $Q_{\tilde{X}_1\tilde{X}_2}$ over the probability simplex $\mathcal{P}(\mathcal{X}^2)$. First, we decompose the KL divergence as

$$D_{\mathrm{KL}}(Q_{\tilde{X}_1\tilde{X}_2} \,||\, P_X^{\otimes 2}) = -H(Q_{\tilde{X}_1\tilde{X}_2}) - \mathbb{E}_Q\left[\log P_X^{\otimes 2}(\tilde{X}_1,\tilde{X}_2)\right]. \tag{A.66}$$

Now, first, for any $Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2} \in \mathcal{P}(\mathcal{X}^2)$ it holds that [9, Lemma 2.7]

$$\left|H(Q_{\tilde{X}_1\tilde{X}_2}) - H(Q_{\overline{X}_1\overline{X}_2})\right| \le d_{\mathrm{TV}}\left(Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2}\right) \times \log \frac{|\mathcal{X}|^2}{d_{\mathrm{TV}}\left(Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2}\right)}, \tag{A.67}$$

and, furthermore,

$$\begin{aligned}
&\left|\mathbb{E}_Q\left[\log P_X^{\otimes 2}(\tilde{X}_1,\tilde{X}_2)\right] - \mathbb{E}_Q\left[\log P_X^{\otimes 2}(\overline{X}_1,\overline{X}_2)\right]\right| \\
&= \left|\sum_{x_1,x_2\in\mathcal{X}} \left[Q_{\tilde{X}_1\tilde{X}_2}(x_1,x_2) - Q_{\overline{X}_1\overline{X}_2}(x_1,x_2)\right] \times \log\left(P_X^{\otimes 2}(x_1,x_2)\right)\right| && \text{(A.68)}\\
&\le d_{\mathrm{TV}}\left(Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2}\right) \cdot 2\max_{x\in\mathcal{X}}\left[-\log P_X(x)\right]. && \text{(A.69)}
\end{aligned}$$

Second, for the Bhattacharyya distance

$$\begin{aligned}
&\left|d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2}) - d_{P_{Y|X}}(Q_{\overline{X}_1\overline{X}_2})\right| \\
&\le \left|\sum_{x_1,x_2\in\mathcal{X}} \left[Q_{\tilde{X}_1\tilde{X}_2}(x_1,x_2) - Q_{\overline{X}_1\overline{X}_2}(x_1,x_2)\right] \times d_{P_{Y|X}}(x_1,x_2)\right| && \text{(A.70)}\\
&\le d_{\mathrm{TV}}\left(Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2}\right) \cdot \max_{x_1,x_2\in\mathcal{X}} d_{P_{Y|X}}(x_1,x_2). && \text{(A.71)}
\end{aligned}$$

Thus, the triangle inequality implies

$$\left|f(Q_{\tilde{X}_1\tilde{X}_2}) - f(Q_{\overline{X}_1\overline{X}_2})\right| \le d_{\mathrm{TV}}\left(Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2}\right) \times \left[\log\frac{|\mathcal{X}|^2}{d_{\mathrm{TV}}\left(Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2}\right)} + c(P_{XY})\right], \tag{A.72}$$

where

$$c(P_{XY}) = 2\max_{x\in\mathcal{X}}\left[-\log P_X(x)\right] + \max_{x_1,x_2\in\mathcal{X}} d_{P_{Y|X}}(x_1,x_2). \tag{A.73}$$

Now, for any given PMF $Q_{\overline{X}_1\overline{X}_2} \in \mathcal{P}(\mathcal{X}^2)$ there exists a type $Q_{\tilde{X}_1\tilde{X}_2} \in \mathcal{P}_L(\mathcal{X}^2)$ such that

$$\begin{aligned}
&d_{\mathrm{TV}}\left(Q_{\tilde{X}_1\tilde{X}_2}, Q_{\overline{X}_1\overline{X}_2}\right) \\
&= \sum_{x_1,x_2\in\mathcal{X}^2} \left|Q_{\tilde{X}_1\tilde{X}_2}(x_1,x_2) - Q_{\overline{X}_1\overline{X}_2}(x_1,x_2)\right| && \text{(A.74)}\\
&\le \frac{|\mathcal{X}|^2}{L}. && \text{(A.75)}
\end{aligned}$$

It further holds that $t \to t\log(1/t)$ is increasing for $t \in [0, e^{-1}]$. Hence, if L is large enough so that $\frac{|\mathcal{X}|^2}{L} \le e^{-1}$ and $c(P_{XY}) \le \log L$, it holds that

$$\begin{aligned}
&\left|f(Q_{\tilde{X}_1\tilde{X}_2}) - f(Q_{\overline{X}_1\overline{X}_2})\right| \\
&\le \frac{|\mathcal{X}|^2}{L}\left[\log(L) + c(P_{XY})\right] && \text{(A.76)}\\
&\le \frac{2|\mathcal{X}|^2\log L}{L}. && \text{(A.77)}
\end{aligned}$$

So there exists $L_1(P_{XY})$ such that for all $L \ge L_1(P_{XY})$ it holds that

$$\left|f(Q_{\tilde{X}_1\tilde{X}_2}) - f(Q_{\overline{X}_1\overline{X}_2})\right| \le \frac{2|\mathcal{X}|^2\log L}{L}. \tag{A.78}$$

We may then replace the first exponent in (A.64) with

$$\min_{Q_{\tilde{X}_1\tilde{X}_2}\in\mathcal{P}(\mathcal{X}^2)} \left(\frac{1}{2}D_{\mathrm{KL}}(Q_{\tilde{X}_1\tilde{X}_2} \,||\, P_X^{\otimes 2}) + d_{P_{Y|X}}(Q_{\tilde{X}_1\tilde{X}_2}) + \frac{\delta}{2} - O\left(\frac{|\mathcal{X}|^2\log L}{L}\right)\right), \tag{A.79}$$

where now the outer minimization is over $Q_{\tilde{X}_1\tilde{X}_2}$ that is not necessarily restricted to be a type in $\mathcal{P}_L(\mathcal{X}^2)$, but rather any joint PMF in the probability simplex $\mathcal{P}(\mathcal{X}^2)$. Ignoring the terms of $\delta/2$ and the asymptotically vanishing term, (A.79) is exactly $\psi_2(P_{XY})$ defined in (A.8). After taking $\delta \downarrow 0$, we obtain the lower bound

$$\begin{aligned}
&\mathbb{P}[\mathcal{E}_{12}] \\
&\ge \mathbb{E}\Big[\max\left\{p_{0\to1}(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2)), p_{1\to0}(\tilde{\boldsymbol{X}}(1),\tilde{\boldsymbol{X}}(2))\right\} \times \mathbb{1}\{\tilde{\boldsymbol{X}}(1)\neq\tilde{\boldsymbol{X}}(2)\}\Big] && \text{(A.80)}\\
&\ge e^{-2L\cdot[\psi_2(P_{XY})+o(1)]} - e^{-LH_2(P_X)} && \text{(A.81)}\\
&\overset{(a)}{\ge} e^{-2L\cdot[\psi_2(P_{XY})+o(1)]} && \text{(A.82)}\\
&= M^{-2\beta\psi_2(P_{XY})+o(1)}, && \text{(A.83)}
\end{aligned}$$

where (a) holds under the assumption $2\psi_2(P_{XY}) < H_2(P_X)$ assuming that L is sufficiently large $L \ge L_0(P_{XY}) \vee L_1(P_{XY})$. Note that $\mathbb{P}[\mathcal{E}_{12}]$ is less than the probability of a transposition error of $\boldsymbol{X}(i_1)$ and $\boldsymbol{X}(i_2)$, which was upper bounded in the proof of Theorem 1 as

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_{12}] &\le \mathbb{P}[\tilde{\mathcal{E}}_{12}] && \text{(A.84)}\\
&\le \mathbb{P}\left[\hat{\boldsymbol{X}}(i_1) = \boldsymbol{X}(i_2),\ \hat{\boldsymbol{X}}(i_2) = \boldsymbol{X}(i_1)\right] && \text{(A.85)}\\
&\le e^{-2L\cdot\psi_2(P_{XY})} && \text{(A.86)}\\
&= M^{-2\beta\psi_2(P_{XY})}. && \text{(A.87)}
\end{aligned}$$

So, the bound on $\mathcal{E}_{12}$ is tight in its polynomial decay rate, and we denote its probability as

$$p := \mathbb{P}[\tilde{\mathcal{E}}_{12}] = \mathbb{P}[\mathcal{E}_{12}] = M^{-2\beta\psi_2(P_{XY})+o(1)}. \tag{A.88}$$

We next consider separately the case of $\xi = 0$ and $\xi > 0$, beginning with the former. To this end, we will lower bound the failure probability by the probability of the union $\bigcup_{(i_1,i_2)\in[M]^2:i_1<i_2}\mathcal{E}_{i_1i_2}$, and to lower bound the probability of this union, we will use de Caen's inequality [11]. This inequality requires evaluating the probability of each event, as well as the probability of intersections of events $\mathcal{E}_{i_1i_2}\cap\mathcal{E}_{i_3i_4}$. For single events, it readily holds from the assumption that the fragments $\{\boldsymbol{X}(i)\}_{i\in\mathbb{N}_+}$ are drawn IID and from symmetry that $\mathbb{P}[\mathcal{E}_{i_1i_2}] = p$ for any pair $(i_1,i_2)\in[M]^2$ so that $i_1 < i_2$.

For the probability of an intersection of events, there are two possible cases. If i_1, i_2, i_3, i_4 are all distinct then the events $\mathcal{E}_{i_1 i_2}$ and $\mathcal{E}_{i_3 i_4}$ are independent, and so trivially,

$$\mathbb{P}\left[\mathcal{E}_{i_1 i_2} \cap \mathcal{E}_{i_3 i_4}\right] = \mathbb{P}\left[\mathcal{E}_{i_1 i_2}\right] \cdot \mathbb{P}\left[\mathcal{E}_{i_3 i_4}\right] = p^2. \tag{A.89}$$

Otherwise, if $i_1 = i_3$ and $i_2 \neq i_4$, then the events are dependent, and the probability is larger. We next assume for notational simplicity that $i_1 = 1, i_2 = 2, i_4 = 4$ and upper bound the probability $\mathbb{P}[\mathcal{E}_{12} \cap \mathcal{E}_{14}]$. First, we remove the constraint that the fragments are different and upper bound

$$\begin{aligned}
&\mathbb{P}[\mathcal{E}_{12} \cap \mathcal{E}_{14}] \\
&= \mathbb{P}\left[\tilde{\mathcal{E}}_{12} \cap \{\boldsymbol{X}(1) \neq \boldsymbol{X}(2)\} \cap \tilde{\mathcal{E}}_{14} \cap \{\boldsymbol{X}(1) \neq \boldsymbol{X}(4)\}\right] && \text{(A.90)} \\
&\leq \mathbb{P}[\tilde{\mathcal{E}}_{12} \cap \tilde{\mathcal{E}}_{14}]. && \text{(A.91)}
\end{aligned}$$

We thus bound the probability on the right-hand side, as in the proof of the Bhattacharyya bound [35, Sec. 2.3]. To this end, let $(\boldsymbol{X}(1), \boldsymbol{X}(2), \boldsymbol{X}(4)) \sim P_X^{\otimes 3L}$ and let $(\boldsymbol{Y}(1), \boldsymbol{Y}(2), \boldsymbol{Y}(4)) \mid (\boldsymbol{X}(1), \boldsymbol{X}(2), \boldsymbol{X}(4)) \sim P_{Y|X}^{\otimes 3L}$. Then,

$$\begin{aligned}
&\mathbb{P}\left[\tilde{\mathcal{E}}_{12} \cap \tilde{\mathcal{E}}_{14}\right] \\
&= \mathbb{E}\Bigg[\mathbb{1}\left\{\frac{\mathbb{P}\left[\boldsymbol{Y}(2), \boldsymbol{Y}(1) \mid \boldsymbol{X}(1), \boldsymbol{X}(2)\right]}{\mathbb{P}\left[\boldsymbol{Y}(1), \boldsymbol{Y}(2) \mid \boldsymbol{X}(1), \boldsymbol{X}(2)\right]} \geq 1\right\} \\
&\qquad \times \mathbb{1}\left\{\frac{\mathbb{P}\left[\boldsymbol{Y}(4), \boldsymbol{Y}(1) \mid \boldsymbol{X}(1), \boldsymbol{X}(4)\right]}{\mathbb{P}\left[\boldsymbol{Y}(1), \boldsymbol{Y}(4) \mid \boldsymbol{X}(1), \boldsymbol{X}(4)\right]} \geq 1\right\}\Bigg] && \text{(A.92)} \\
&\overset{(a)}{\leq} \mathbb{E}\Bigg[\sum_{(\boldsymbol{y}(1), \boldsymbol{y}(2), \boldsymbol{y}(4)) \in \mathcal{Y}^{3L}} P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(1) \mid \boldsymbol{X}(1)\right] \\
&\qquad \times P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(2) \mid \boldsymbol{X}(2)\right] \\
&\qquad \times P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(4) \mid \boldsymbol{X}(4)\right] \\
&\qquad \times \sqrt{\frac{P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(2) \mid \boldsymbol{X}(1)\right] P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(1) \mid \boldsymbol{X}(2)\right]}{P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(1) \mid \boldsymbol{X}(1)\right] P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(2) \mid \boldsymbol{X}(2)\right]}} \\
&\qquad \times \sqrt{\frac{P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(4) \mid \boldsymbol{X}(1)\right] P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(1) \mid \boldsymbol{X}(4)\right]}{P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(1) \mid \boldsymbol{X}(1)\right] P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(4) \mid \boldsymbol{X}(4)\right]}}\Bigg] && \text{(A.93)} \\
&= \mathbb{E}\Bigg[\sum_{\boldsymbol{y}(1)} \sqrt{P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(1) \mid \boldsymbol{X}(2)\right] P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(1) \mid \boldsymbol{X}(4)\right]} \\
&\qquad \times \sum_{\boldsymbol{y}(2)} \sqrt{P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(2) \mid \boldsymbol{X}(1)\right] P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(2) \mid \boldsymbol{X}(2)\right]} \\
&\qquad \times \sum_{\boldsymbol{y}(4)} \sqrt{P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(4) \mid \boldsymbol{X}(4)\right] P_{Y|X}^{\otimes L}\left[\boldsymbol{y}(4) \mid \boldsymbol{X}(1)\right]}\Bigg] && \text{(A.94)} \\
&\overset{(b)}{=} \sum_{Q_{X_1 X_2 X_4} \in \mathcal{P}_L(\mathcal{X}^3)} \mathbb{P}\left[(\boldsymbol{X}(1), \boldsymbol{X}(2), \boldsymbol{X}(4)) \in T_L(Q_{X_1 X_2 X_4})\right] \\
&\qquad \times e^{-L \cdot \left(d_{P_{Y|X}}(Q_{X_2 X_4}) + d_{P_{Y|X}}(Q_{X_1 X_2}) + d_{P_{Y|X}}(Q_{X_4 X_1})\right)} && \text{(A.95)} \\
&\overset{(c)}{\leq} \left|\mathcal{P}_L(\mathcal{X}^3)\right| \max_{Q_{X_1 X_2 X_4} \in \mathcal{P}_L(\mathcal{X}^3)} e^{-L \cdot D_{\mathrm{KL}}(Q_{X_1 X_2 X_4} || P_X^{\otimes 3})} \\
&\qquad \times e^{-L \cdot \left(d_{P_{Y|X}}(Q_{X_2 X_4}) + d_{P_{Y|X}}(Q_{X_1 X_2}) + d_{P_{Y|X}}(Q_{X_4 X_1})\right)} && \text{(A.96)} \\
&\overset{(d)}{\leq} \exp\left[-L \cdot \left(3\psi_3(P_{XY}) - \frac{|\mathcal{X}|^3 \log(L+1)}{L}\right)\right] && \text{(A.97)} \\
&= M^{-3\beta\psi_3(P_{XY}) + o(1)} && \text{(A.98)} \\
&\overset{(e)}{\leq} M^{-3\beta\psi_2(P_{XY}) + o(1)} =: q, && \text{(A.99)}
\end{aligned}$$

where (a) follows from the standard Bhattacharyya bound technique of bounding

$$\begin{aligned}
&\mathbb{1}\left\{\frac{\mathbb{P}\left[\boldsymbol{Y}(2), \boldsymbol{Y}(1) \mid \boldsymbol{X}(1), \boldsymbol{X}(2)\right]}{\mathbb{P}\left[\boldsymbol{Y}(1), \boldsymbol{Y}(2) \mid \boldsymbol{X}(1), \boldsymbol{X}(2)\right]} \geq 1\right\} \\
&\leq \sqrt{\frac{\mathbb{P}\left[\boldsymbol{Y}(2), \boldsymbol{Y}(1) \mid \boldsymbol{X}(1), \boldsymbol{X}(2)\right]}{\mathbb{P}\left[\boldsymbol{Y}(1), \boldsymbol{Y}(2) \mid \boldsymbol{X}(1), \boldsymbol{X}(2)\right]}}, && \text{(A.100)}
\end{aligned}$$

(b) follows since the Bhattacharyya distance between $\boldsymbol{X}(1)$ and $\boldsymbol{X}(2)$ only depends on their joint type $Q_{X_1 X_2}$ (and similarly for the other Bhattacharyya distances), (c) follows from the method of types (the upper bound of the probability of a type class [9, Lemma 2.3]), (d) follows from the type counting lemma [9, Lemma 2.2] and the definition of $\psi_K(P_{XY})$ in (A.8), and (e) holds since as was shown in (A.13), it holds that $\psi_3(P_{XY}) \geq \psi_2(P_{XY})$.

Using the above bounds, we may lower bound the failure probability as

$$\begin{aligned}
&\mathsf{FP}(\xi = 0) \\
&\geq \mathbb{P}\left[\bigcup_{(i_1, i_2) \in [M]^2 : i_1 < i_2} \mathcal{E}_{i_1 i_2}\right] && \text{(A.101)} \\
&\overset{(a)}{\geq} \sum_{(i_1, i_2) \in [M]^2 : i_1 < i_2} \frac{\mathbb{P}^2\left[\mathcal{E}_{i_1 i_2}\right]}{\sum_{(i_3, i_4) \in [M]^2 : i_3 < i_4} \mathbb{P}\left[\mathcal{E}_{i_1 i_2} \cap \mathcal{E}_{i_3 i_4}\right]}, && \text{(A.102)}
\end{aligned}$$

where (a) follows from de Caen's inequality [11]. The sum's denominator is bounded for any given (i_1, i_2) as follows. For the single term $(i_3, i_4) = (i_1, i_2)$

$$\mathbb{P}\left[\mathcal{E}_{i_1 i_2} \cap \mathcal{E}_{i_3 i_4}\right] = \mathbb{P}\left[\mathcal{E}_{i_1 i_2}\right] = p. \tag{A.103}$$

For the terms in which either $i_1 = i_3$ or $i_2 = i_4$ it holds that

$$\mathbb{P}\left[\mathcal{E}_{i_1 i_2} \cap \mathcal{E}_{i_3 i_4}\right] \leq q, \tag{A.104}$$

given in (A.99). There are less than $2M$ such pairs of pairs. Finally, for the terms in which i_1, i_2, i_3, i_4 are all distinct it holds that

$$\mathbb{P}\left[\mathcal{E}_{i_1 i_2} \cap \mathcal{E}_{i_3 i_4}\right] = p^2. \tag{A.105}$$

There are less than M^2 such terms. Hence, (A.102) may be further lower bounded as

$$\begin{aligned}
&\mathsf{FP}(\xi = 0) \\
&\geq \sum_{(i_1, i_2) \in [M]^2 : i_1 < i_2} \frac{p^2}{p + qM + M^2 p^2} && \text{(A.106)}
\end{aligned}$$

$$\geq \frac{1}{3} \sum_{(i_1,i_2)\in[M]^2: i_1<i_2} p \wedge \frac{p^2}{2qM} \wedge \frac{p^2}{M^2p^2} \quad \text{(A.107)}$$

$$\stackrel{(a)}{\geq} \frac{1}{24} \cdot \left(M^2 p \wedge M\frac{p^2}{q} \wedge 1\right) \quad \text{(A.108)}$$

$$\stackrel{(b)}{\geq} M^{o(1)} \cdot \left(M^{2(1-\beta\psi_2(P_{XY}))} \wedge M^{1-\beta\psi_2(P_{XY})} \wedge 1\right) \quad \text{(A.109)}$$

$$\stackrel{(c)}{=} M^{2-2\beta\psi_2(P_{XY})+o(1)}, \quad \text{(A.110)}$$

where (a) follows since there are $\frac{M(M-1)}{2} \geq \frac{M^2}{4}$ pairs $(i_1, i_2) \in [M]^2$ such that $i_1 < i_2$ (assuming trivially that $M \geq 2$), (b) follows by the definition of p in (A.83) and the definition of $q = M^{-3\beta\psi_2(P_{XY})+o(1)}$ in (A.99), (c) holds by the assumption of the theorem that $\beta > \frac{1}{\psi_2(P_{XY})}$. This completes the proof of the bound for $\xi = 0$.

We now prove the bound for $\xi > 0$. Let $I := \{\{i_j, i'_j\}\}_{j\in[\xi M/2]}$ be a set of $\xi M/2$ unordered pairs of unique indices in $[M]$, that is $\cap_{j\in[\xi M/2]}\{i_j \cap i'_j\} = \emptyset$.[3] Consider the event of $\frac{\xi M}{2}$ transposition replacements of correct likelihood order between pairs of fragments in I, that is

$$\mathcal{E}_I := \bigcap_{j=1}^{\xi M/2} \left\{\mathcal{E}_{i_j,i'_j}\right\} \quad \text{(A.111)}$$

using the definition of the event $\mathcal{E}_{i_1 i_2}$ in (A.53). Then,

$$\mathbb{P}[\mathcal{E}_I] \stackrel{(a)}{=} \prod_{j=1}^{\xi M/2} \mathbb{P}\left[\mathcal{E}_{i_j,i'_j}\right] \quad \text{(A.112)}$$

$$\stackrel{(b)}{\geq} e^{-\xi ML\cdot[\psi_2(P_{XY})+o(1)]} \quad \text{(A.113)}$$

$$\stackrel{(c)}{=} e^{-\xi M\log M\cdot[\beta\psi_2(P_{XY})+o(1)]}, \quad \text{(A.114)}$$

where (a) follows since disjoint pairwise transpositions are independent events, (b) follows from the lower bound on $\mathbb{P}[\mathcal{E}_{i_1 i_2}]$ in (A.83), and (c) from $L = \beta \log M$. ■

Appendix B
Proofs for the Repeating-Fragments Regime with Nonnegative Distortion

We first prove Prop. 7. Recall that $G \sim \text{Multinomial}(M; (p_1, p_2, \ldots, p_{M^\beta}))$ where M is the fixed number of fragments. Consider the random vector $\tilde{G} = (\tilde{G}(1), \ldots, \tilde{G}(M^\beta))$, which has the same dimension as G, yet each of its components is distributed $\tilde{G}(j) \sim \text{Poisson}(Mp_j)$, where $p_j := \mathbb{P}[X^L = a_j]$ is the probability of $a_j \in \mathcal{X}^L$, the jth letter in $\mathcal{X}^L$, and where the components of $\tilde{G}$ are *independent* (unlike those of G). By construction, the expected value of each coordinate in G and $\tilde{G}$ is equal, and given by $\mathbb{E}[\tilde{G}(j)] = \mathbb{E}[G(j)] = Mp_j$. We recall that "Poissonization of the multinomial" effect (see [23, Sec. 5.4] for the case P_X is uniform and $\{p_j\}_{j\in[M^\beta]}$ are all equal. This has a straightforward extension to non-uniform probabilities, see, e.g., [27, Lecture 11]).

[3] As mentioned we ignore integer constraints, as they do not affect the final result, and thus assume that $\xi M/2$ is integer.

Fact 9 (Poissonization of the multinomial distribution)**.** *Let $\tilde{M} \sim \text{Poisson}(M)$, and let $\tilde{G}$ be a random vector such that $\tilde{G} \sim \text{Multinomial}(\tilde{M}, (p_1, p_2, \ldots p_{M^\beta}))$ conditioned on $\tilde{M}$, where $\sum_{j\in[M^\beta]} p_j = 1$ and $p_j > 0$. Then, $\{\tilde{G}(j)\}_{j\in[M^\beta]}$ are statistically independent and $\tilde{G}(j) \sim \text{Poisson}(Mp_j)$ (unconditioned on $\tilde{M}$).*

Fact 9 can be verified by spelling out the conditional PMF of $\tilde{G}$ conditioned on $\tilde{M}$ [23, Thm. 5.6] in case $\{p_j\}$ are all equal, and easily extended to the non-uniform case (e.g., [27, Lecture 11, Thm. 3.2]). The following then follows from [23, Corollary 5.9]:

Lemma 10. *Let $G \sim \text{Multinomial}(M, (p_1, p_2, \ldots p_{M^\beta}))$, and let $\tilde{G}$ be an independent Poisson vector of the same dimension so that $\mathbb{E}[\tilde{G}(j)] = \mathbb{E}[G(j)] = Mp_j$. Then, for any event $\mathcal{E}$*

$$\mathbb{P}\left[G \in \mathcal{E}\right] \leq \sqrt{eM} \cdot \mathbb{P}\left[\tilde{G} \in \mathcal{E}\right]. \quad \text{(B.1)}$$

We will also need the following results on Poisson RVs. The first one is a standard Chernoff bound for Poisson RVs, and the second one is the aforementioned concentration inequality for Lipschitz functions of Poisson RVs.

Lemma 11 (Chernoff's bound for Poisson RVs [23, Theorem 5.4])**.** *For $W \sim \text{Poisson}(\lambda)$ it holds that*

$$\mathbb{P}\left[W \geq \alpha\mathbb{E}[W]\right] \leq e^{-\lambda}\left(\frac{e}{\alpha}\right)^{\alpha\lambda} \quad \text{(B.2)}$$

$$\leq \left(\frac{e}{\alpha}\right)^{\alpha\lambda} = e^{-\alpha\lambda\log(\alpha/e)} \leq e^{-\alpha\lambda}, \quad \text{(B.3)}$$

where the rightmost inequality holds for any $\alpha > 3e \approx 8.15$.

Lemma 12 (Poisson concentration of Lipschitz functions, a variant of [2], [18])**.** *Let $W \sim \text{Poisson}(\lambda)$, and assume that f is 1-Lipschitz, that is, $|Df(w)| := |f(w+1) - f(w)| \leq 1$ for all $w \in \mathbb{N}$. Then, for any $t > 0$*

$$\mathbb{P}\left[f(W) - \mathbb{E}[f(W)] > t\right] \leq \exp\left[-\frac{t^2}{16\lambda + 3t}\right]. \quad \text{(B.4)}$$

Proof: Under the conditions of the lemma

$$\mathbb{P}\left[f(W) - \mathbb{E}[f(W)] > t\right]$$

$$\stackrel{(a)}{\leq} \exp\left[-\frac{t}{4}\log\left(1 + \frac{t}{2\lambda}\right)\right] \quad \text{(B.5)}$$

$$\stackrel{(b)}{=} \exp\left[-\frac{\lambda}{2}u\log(1+u)\right] \quad \text{(B.6)}$$

$$\stackrel{(c)}{\leq} \exp\left[-\frac{\lambda u^2}{4(1+u/3)}\right], \quad \text{(B.7)}$$

where (a) is the Bobkov and Ledoux's bound [2, Prop. 10] [18], (b) follows by setting $u := \frac{t}{2\lambda}$, (c) follows from

$$u\log(1+u) = (1+u)\log(1+u) - u + u - \log(1+u) \quad \text{(B.8)}$$

$$\stackrel{(*)}{\geq} \frac{u^2}{2(1+u/3)} + u - \log(1+u) \quad \text{(B.9)}$$

$$\stackrel{(**)}{\geq} \frac{u^2}{2(1+u/3)}, \quad \text{(B.10)}$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3390068

where $(*)$ was stated in [3, Exercise 2.8], and $(**)$ follows from $u \geq \log(1+u)$ for $u \geq 0$. The result follows by re-substituting $u = \frac{t}{2\lambda}$, and performing a (minor) numerical relaxation. ■

We may now prove Prop. 7.

Proof of Prop. 7: The entropy upper bound on the multinomial coefficient implies that it surely holds that[4]

$$C(X^N) := \frac{1}{M}\log\left|\mathcal{A}_L(X^N)\right| \tag{B.11}$$
$$= \frac{1}{M}\log\binom{M}{G(1),G(2),\ldots,G(M^\beta)} \tag{B.12}$$
$$\overset{(a)}{\leq} H\left(\frac{G(1)}{M},\frac{G(2)}{M},\ldots,\frac{G(M^\beta)}{M}\right) \tag{B.13}$$
$$= -\sum_{j\in[M^\beta]}\frac{G(j)}{M}\log\frac{G(j)}{M}. \tag{B.14}$$

Let us denote

$$\tilde{C} := \sum_{j\in[M^\beta]} -\frac{\tilde{G}(j)}{M}\log\frac{\tilde{G}(j)}{M}, \tag{B.15}$$

where $\tilde{G}(j) \sim \text{Poisson}(M\cdot\mathbb{P}[X^L = a_j])$, for which it holds that $\mathbb{E}[\tilde{G}(j)] = \mathbb{E}[G(j)]$. Let $\boldsymbol{X} \sim P_X^{\otimes L}$ be a random length-L fragment. By Jensen's inequality for the function $f(t) := -t\log t$ in $\mathbb{R}_+$,

$$\mathbb{E}[\tilde{C}] = \sum_{j\in[M^\beta]}\mathbb{E}\left[-\frac{\tilde{G}(j)}{M}\log\frac{\tilde{G}(j)}{M}\right] \tag{B.16}$$
$$\leq \sum_{j\in[M^\beta]} -\frac{\mathbb{E}[\tilde{G}(j)]}{M}\log\frac{\mathbb{E}[\tilde{G}(j)]}{M} \tag{B.17}$$
$$= -\sum_{j\in[M^\beta]}\mathbb{P}[\boldsymbol{X}=a_j]\cdot\log\mathbb{P}[\boldsymbol{X}=a_j] \tag{B.18}$$
$$= H(\boldsymbol{X}) \tag{B.19}$$
$$= L\cdot H(P_X). \tag{B.20}$$

Hence, for any $\eta > 0$, the Poissonization effect of Lemma 10 implies that

$$\mathbb{P}\left[C(X^N) \geq L\cdot H(P_X) + \eta\log M\right]$$
$$\leq \mathbb{P}\left[C(X^N) \geq \mathbb{E}[\tilde{C}] + \eta\log M\right] \tag{B.21}$$
$$\leq e\sqrt{M}\cdot\mathbb{P}\left[\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta\log M\right]. \tag{B.22}$$

We next bound the concentration of $\tilde{C}$ above its expected value $\mathbb{E}[\tilde{C}]$. We first derive a bound which is effective in the regime $\beta \in (0,2)$. To this end, we would like to invoke the concentration inequality of Lipschitz functions of Poisson RVs by Bobkov and Ledoux [2], [18], see Lemma 12. However, writing $\tilde{C} := \sum_{j\in[M^\beta]} f(\tilde{G}(j))$ for $f(g) := -\frac{g}{M}\log\frac{g}{M}$, it is apparent that this is not a Lipschitz function on $\mathbb{N}_+$ since $f(g)$ has unbounded derivative for $g\uparrow\infty$. To address this, we first consider a Lipschitz approximation to $f(g)$ given by $f^+(g) = (f(g))_+$, and establish the concentration of $\tilde{C}^+ = \sum_{j\in[M^\beta]} f^+(\tilde{G}(j))$ to its expected value $\mathbb{E}[\tilde{C}^+]$.

[4] See e.g., [7, Lemma 17.5.1] for the binomial coefficient; the extension to multinomial is straightforward and well known.

Afterwards, we show that $\mathbb{E}[\tilde{C}^+]$ is close to $\mathbb{E}[\tilde{C}]$. It can be easily verified that the discrete derivative satisfies

$$\left|f^+(g+1) - f^+(g)\right| \leq \frac{\log M}{M}, \tag{B.23}$$

that is, $f^+(g)$ is a $(\frac{\log M}{M})$-Lipschitz continuous function. We first consider the tail behavior of each of the terms defining $\tilde{C}^+$. Let $j\in[M^\beta]$ be given. Then, for any $\eta > 0$

$$\mathbb{P}\left[f^+\left(\tilde{G}(j)\right) - \mathbb{E}\left[f^+\left(\tilde{G}(j)\right)\right] > \eta\log M\right]$$
$$\overset{(a)}{=} \mathbb{P}\left[\hat{f}^+\left(\tilde{G}(j)\right) - \mathbb{E}\left[\hat{f}^+\left(\tilde{G}(j)\right)\right] > \eta M\right] \tag{B.24}$$
$$\overset{(b)}{\leq} \exp\left[-\frac{M^2\eta^2}{16\cdot\mathbb{E}[\tilde{G}(j)] + 3M\eta}\right], \tag{B.25}$$

where (a) is obtained by setting $\hat{f}^+(g) := \frac{Mf^+(g)}{\log M}$ and noting that $\hat{f}^+(g)$ is a 1-Lipschitz continuous function, (b) is obtained by Poisson concentration of Lipschitz functions, as stated in Lemma 12. Since $-\hat{f}^+(g)$ is also a 1-Lipschitz continuous function, an analogous bound holds for the left tail. Denoting for brevity

$$Z(j) := \hat{f}^+\left(\tilde{G}(j)\right) - \mathbb{E}\left[\hat{f}^+\left(\tilde{G}(j)\right)\right], \tag{B.26}$$

we note that if

$$M\eta = \sqrt{32\cdot\mathbb{E}[\tilde{G}(j)]t} + 6t \tag{B.27}$$

then

$$\frac{M^2\eta^2}{16\cdot\mathbb{E}[\tilde{G}(j)] + 3M\eta} \geq \frac{M^2\eta^2}{2\left(16\cdot\mathbb{E}[\tilde{G}(j)] \vee 3M\eta\right)} \tag{B.28}$$
$$= \frac{M^2\eta^2}{32\cdot\mathbb{E}[\tilde{G}(j)]} \wedge \frac{M\eta}{6} \tag{B.29}$$
$$\geq \frac{32\cdot\mathbb{E}[\tilde{G}(j)]t}{32\cdot\mathbb{E}[\tilde{G}(j)]} \wedge \frac{6t}{6} \tag{B.30}$$
$$\geq t, \tag{B.31}$$

and so (B.25) implies that

$$\mathbb{P}\left[|Z(j)| \leq \sqrt{32\cdot\mathbb{E}[\tilde{G}(j)]t} + 6t\right] \leq 2e^{-t} \tag{B.32}$$

holds for any $t \geq 0$. Hence, the two statements of [3, Thm. 2.3] together imply that $Z(j)$ is a sub-gamma random variable with a variance proxy $4(128\cdot\mathbb{E}[\tilde{G}(j)] + 576)$ and a scale parameter 48. Since $\mathbb{E}[\tilde{G}(j)] = M\cdot\mathbb{P}[X^L = a_j]$, there exists a numerical constant $c_1 > 0$ so that

$$\sum_{j\in[M^\beta]}\mathbb{E}[Z^2(j)] \leq 512\sum_{j\in[M^\beta]}\left(\mathbb{E}[\tilde{G}(j)] + 2\cdot(24)^2\right) \tag{B.33}$$
$$\leq c_1(M + M^\beta). \tag{B.34}$$

Furthermore, $|Z(j)| \leq 2\max_{t\in[0,1]} -t\log t \leq 3c_2$ for some numerical constant $c_2 > 0$ (depending on the choice of base for the logarithm), and so $Z(j)$ satisfies Bernstein's condition with a sum of second moments $\sum_{j\in[M^\beta]}\mathbb{E}[Z(j)^2] \leq c_1(M+M^\beta)$ and a scale constant c_2. Since $\{Z(j)\}_{j\in[M^\beta]}$ are independent and centered, Bernstein's inequality [3, Corollary 2.11] then results

$$\mathbb{P}\left[\sum_{j\in[M^\beta]} Z(j) \geq r\right] \leq \exp\left[-\frac{r^2}{2(c_1 M + c_1 M^\beta + c_2 r)}\right]. \tag{B.35}$$

Setting $r = M\eta$, while restricting now that $\eta \in (0,1)$, then results

$$\mathbb{P}\left[\tilde{C}^+ \geq \mathbb{E}[\tilde{C}^+] + \eta \log M\right]$$
$$= \mathbb{P}\left[\sum_{j\in[M^\beta]} f^+\left(\tilde{G}(j)\right) - \mathbb{E}\left[f^+\left(\tilde{G}(j)\right)\right] > \eta \log M\right] \tag{B.36}$$
$$= \mathbb{P}\left[\sum_{j\in[M^\beta]} \hat{f}^+\left(\tilde{G}(j)\right) - \mathbb{E}\left[\hat{f}^+\left(\tilde{G}(j)\right)\right] > M\eta\right] \tag{B.37}$$
$$\leq \exp\left[-\frac{M^2\eta^2}{2(c_1 M + c_1 M^\beta + c_2 M\eta)}\right] \tag{B.38}$$
$$\leq \exp\left[-M \cdot \min\left\{\frac{\eta^2}{4c_1}, \frac{M^{1-\beta}\eta}{c_1}, \frac{\eta}{c_2}\right\}\right] \tag{B.39}$$
$$\leq \exp\left[-c_3 \cdot \eta^2 M^{1\vee(2-\beta)}\right], \tag{B.40}$$

for some numerical constant $c_3 > 0$.

Next, we bound the absolute difference $\mathbb{E}[\tilde{C}^+]$ and $\mathbb{E}[\tilde{C}]$. Note that each of them is comprised of M^β terms, and as before, we first focus on a single term $j \in [M^\beta]$. For brevity, we let $\tilde{G} \sim \text{Poisson}(\lambda)$, where $\lambda \leq \frac{M}{10} \leq M-1$ is assumed. Then,

$$\mathbb{E}\left[f^+(\tilde{G})\right] - \mathbb{E}\left[f(\tilde{G})\right]$$
$$= \mathbb{E}\left[\frac{\tilde{G}}{M}\log\frac{\tilde{G}}{M} \cdot \mathbb{1}\{\tilde{G} \geq M\}\right] \tag{B.41}$$
$$= \sum_{\tilde{g}=M+1}^{\infty} \mathbb{P}[\tilde{G} = \tilde{g}]\frac{\tilde{g}}{M}\log\frac{\tilde{g}}{M} \tag{B.42}$$
$$= \sum_{\tilde{g}=M+1}^{\infty} \frac{\lambda^{\tilde{g}}e^{-\lambda}}{\tilde{g}!}\frac{\tilde{g}}{M}\log\frac{\tilde{g}}{M} \tag{B.43}$$
$$= \frac{\lambda}{M}\sum_{\tilde{g}=M+1}^{\infty} \frac{\lambda^{\tilde{g}-1}e^{-\lambda}}{(\tilde{g}-1)!}\log\frac{\tilde{g}}{M} \tag{B.44}$$
$$\stackrel{(b)}{=} \frac{\lambda}{M}\sum_{\overline{g}=M}^{\infty} \frac{\lambda^{\overline{g}}e^{-\lambda}}{\overline{g}!}\log\frac{\overline{g}+1}{M} \tag{B.45}$$
$$= \frac{\lambda}{M}\mathbb{E}\left[\log\frac{\tilde{G}+1}{M}\mathbb{1}\{\tilde{G} \geq M\}\right] \tag{B.46}$$
$$\stackrel{(c)}{\leq} \frac{\lambda}{M}\mathbb{E}\left[\frac{(\tilde{G}+1-M)}{M} \cdot \mathbb{1}\{\tilde{G} \geq M\}\right] \tag{B.47}$$
$$\stackrel{(d)}{\leq} \frac{\lambda}{M}\mathbb{E}\left[\frac{(\tilde{G}-\lambda)}{M} \cdot \mathbb{1}\{\tilde{G} \geq M\}\right] \tag{B.48}$$
$$\stackrel{(e)}{\leq} \frac{\lambda}{M}\sqrt{\mathbb{E}\frac{(\tilde{G}-\lambda)^2}{M^2}} \cdot \sqrt{\mathbb{P}\left(\tilde{G} \geq M\right)} \tag{B.49}$$
$$= \frac{\lambda^{3/2}}{M^2}\sqrt{\mathbb{P}\left(\tilde{G} \geq M\right)} \tag{B.50}$$
$$\stackrel{(f)}{\leq} 1 \cdot e^{-M/2}, \tag{B.51}$$

where (a) follows since $f^+(g) = (f(g))_+$, (b) is using the change of variables $\overline{g} = \tilde{g} - 1$, (c) follows from $\log(t) \leq t - 1$ for $t \geq 1$, (d) holds since $\lambda \leq M - 1 \leq M$ was assumed, (e) follows from the Cauchy-Schwarz inequality, and (f) follows from Lemma 11 and the assumption that $\lambda \leq \frac{M}{10}$.

Now, under the assumption that $H(P_X) > 0$ it must hold that $\max_{x\in\mathcal{X}} P_X(x) < 1$. Hence,

$$\mathbb{E}[\tilde{G}(j)] = M \cdot \mathbb{P}[X^L = a_j] = M(\max P_X(x))^L = o(M) \tag{B.52}$$

and we may use the approximation of (B.51) assuming that $M \geq M_0(P_X, \beta)$ is sufficiently large. Then,

$$\left|\mathbb{E}[\tilde{C}] - \mathbb{E}[\tilde{C}^+]\right|$$
$$\leq \sum_{j\in[M^\beta]} \left|\mathbb{E}\left[f^+\left(\tilde{G}(j)\right)\right] - \mathbb{E}\left[f\left(\tilde{G}(j)\right)\right]\right| \tag{B.53}$$
$$\leq M^\beta e^{-M/2} \leq e^{-M/4} \tag{B.54}$$

for all $M \geq M_1(P_X, \beta)$ sufficiently large. Returning to (B.22) we obtain

$$\mathbb{P}\left[C(X^N) \geq L \cdot H(P_X) + \eta \log M\right]$$
$$\leq e\sqrt{M} \cdot \mathbb{P}\left[\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta \log M\right] \tag{B.55}$$
$$\stackrel{(a)}{\leq} e\sqrt{M} \cdot \mathbb{P}\left[\tilde{C}^+ \geq \mathbb{E}[\tilde{C}] + \eta \log M\right] \tag{B.56}$$
$$\stackrel{(b)}{\leq} e\sqrt{M} \cdot \mathbb{P}\left[\tilde{C}^+ \geq \mathbb{E}[\tilde{C}^+] + \eta \log M - e^{-M/4}\right] \tag{B.57}$$
$$\stackrel{(c)}{\leq} e\sqrt{M} \cdot \mathbb{P}\left[\tilde{C}^+ \geq \mathbb{E}[\tilde{C}^+] + \frac{\eta}{2}\log M\right] \tag{B.58}$$
$$\stackrel{(d)}{\leq} e\sqrt{M} \cdot \exp\left[-c_4 \cdot \eta^2 M^{1\vee(2-\beta)}\right] \tag{B.59}$$
$$\stackrel{(e)}{\leq} \exp[-c_5 \cdot \eta^2 M^{1\vee(2-\beta)}], \tag{B.60}$$

where (a) follows since $\tilde{C}^+ \geq \tilde{C}$, (b) follows from (B.54), (c) holds for all $M \geq M_2(\eta)$ sufficiently large, (d) holds from (B.40), and (e) holds for some numerical constant $c_5 > 0$ and all $M \geq M_3(P_X, \beta, \eta)$ sufficiently large. The result then follows for all $M \geq M_0 \vee M_1 \vee M_2 \vee M_3$.

The bound derived above is non-trivial only for $\beta \in (0, 2)$ Next, we derive a different bound on the one-sided concentration of $\tilde{C}$ above its expected value $\mathbb{E}[\tilde{C}]$ in (B.22), which is valid for any $\beta > 0$. Consider the events

$$\mathcal{F}(j) := \left\{\tilde{G}(j) \geq M\right\}. \tag{B.61}$$

As before, under the assumption that $H(P_X) > 0$ it must hold that $\mathbb{E}[\tilde{G}(j)] \leq \frac{M}{10}$ for all $M \geq M_0(P_X)$. Hence, Lemma 11 implies that

$$\mathbb{P}\left[\mathcal{F}(j)\right] = \mathbb{P}\left[\tilde{G}(j) \geq \frac{\mathbb{E}\left[\tilde{G}(j)\right]}{\mathbb{P}[X^L = a_j]}\right] \leq e^{-M} \tag{B.62}$$

(using $\alpha = 1/\mathbb{P}[X^L = a_j] \geq 3e$). Letting $\mathcal{F} = \bigcap_{j\in[M^\beta]} \mathcal{F}(j)$, the union bound implies that

$$\mathbb{P}[\mathcal{F}] \leq M^\beta e^{-M} = e^{-M\left(1-\frac{\beta \log M}{M}\right)}, \tag{B.63}$$

and the probability in (B.22) is then bounded as

$$\begin{aligned}
&\mathbb{P}\left[\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta \log M\right] \\
&\leq \mathbb{P}\left[\left\{\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta \log M\right\} \cap \mathcal{F}^c\right] + \mathbb{P}\left[\mathcal{F}\right] && \text{(B.64)}\\
&= \mathbb{P}\left[\left\{\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta \log M\right\} \cap \mathcal{F}^c\right] + e^{-M\left(1-\frac{\beta \log M}{M}\right)}. && \text{(B.65)}
\end{aligned}$$

We further upper bound the first probability. Under $\mathcal{F}^c$, it holds that $\tilde{G}(j) \leq M$, the argument of $f(t)$ is less than 1, and $f(t) \geq 0$. Thus,

$$\begin{aligned}
&\mathbb{P}\left[\left\{\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta \log M\right\} \cap \mathcal{F}^c\right] \\
&= \mathbb{P}\left[\left\{\sum_{j\in[M^\beta]} f\left(\tilde{G}(j)\right) \geq \mathbb{E}\left[\sum_{j\in[M^\beta]} f\left(\tilde{G}(j)\right)\right] + \eta \log M\right\} \cap \mathcal{F}^c\right] && \text{(B.66)}\\
&= \mathbb{P}\left[\left\{\sum_{j\in[M^\beta]} f^+\left(\tilde{G}(j)\right) \geq \mathbb{E}\left[\sum_{j\in[M^\beta]} f\left(\tilde{G}(j)\right)\right] + \eta \log M\right\} \cap \mathcal{F}^c\right] && \text{(B.67)}\\
&\leq \mathbb{P}\left[\sum_{j\in[M^\beta]} f^+\left(\tilde{G}(j)\right) \geq \mathbb{E}\left[\sum_{j\in[M^\beta]} f\left(\tilde{G}(j)\right)\right] + \eta \log M\right]. && \text{(B.68)}
\end{aligned}$$

The difference in expectation when switching from f to f^+ is bounded, as in (B.51), as

$$\begin{aligned}
&\sum_{j\in[M^\beta]} \mathbb{E}\left[f\left(\tilde{G}(j)\right)\right] - \mathbb{E}\left[f^+\left(\tilde{G}(j)\right)\right] \\
&\leq M^\beta e^{-M/2} \leq e^{-M/4} && \text{(B.69)}
\end{aligned}$$

for all $M \geq M_1(P_X, \beta)$. Hence,

$$\begin{aligned}
&\mathbb{P}\left[\left\{\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta \log M\right\} \cap \mathcal{F}^c\right] \\
&\leq \mathbb{P}\left[\sum_{j\in[M^\beta]} f^+\left(\frac{\tilde{G}(j)}{M}\right) \geq \mathbb{E}\left[\sum_{j\in[M^\beta]} f^+\left(\frac{\tilde{G}(j)}{M}\right)\right] - e^{-M/4} + \eta \log M\right] && \text{(B.70)}\\
&\leq \mathbb{P}\left[\sum_{j\in[M^\beta]} f^+\left(\frac{\tilde{G}(j)}{M}\right) \geq \mathbb{E}\left[\sum_{j\in[M^\beta]} f^+\left(\frac{\tilde{G}(j)}{M}\right)\right] + \frac{\eta}{2}\eta \log M\right] && \text{(B.71)}
\end{aligned}$$

for all $M \geq M_2(\eta)$. To further bound this probability, we note that $\{f^+(\tilde{G}(j))\}_{j\in[M^\beta]}$ are IID RVs which are bounded from above as

$$f^+\left(\tilde{G}(j)\right) \leq \max_{t\geq 0}(-t\log t) = \frac{1}{e}. \tag{B.72}$$

We thus may use the regular Bernstein's inequality to bound the deviation of their sum from its mean. To this end, we bound their second moment, by noting that $\tilde{G}(j) \in \mathbb{N}_+$, and that for any $g \in \mathbb{N}_+$ it holds that

$$0 \leq f^+(g) \leq \frac{g}{M}\log M \tag{B.73}$$

(as the intersection of the concave function $f(t) = -t\log t$ and $t \log M$ occurs at $t = 1/M$). Hence,

$$\begin{aligned}
&\mathbb{E}\left[\left(f^+\left(\tilde{G}(j)\right)\right)^2\right] \\
&\leq \mathbb{E}\left[\left(\frac{\tilde{G}(j)}{M}\log M\right)^2\right] && \text{(B.74)}\\
&= \frac{\log^2 M}{M^2}\cdot \mathbb{E}\left[\tilde{G}^2(j)\right] && \text{(B.75)}\\
&= \frac{\log^2 M}{M}\cdot \mathbb{P}[X^L = a_j], && \text{(B.76)}
\end{aligned}$$

since $\tilde{G}(j)$ is Poisson with parameter $\mathbb{E}[\tilde{G}(j)] = M\cdot\mathbb{P}[X^L = a_j]$ for all $M \geq M_0$. So, Bernstein's inequality [3, Corollary 2.11 and the discussion that follows it] implies that for any $r \geq 0$

$$\begin{aligned}
&\mathbb{P}\left[\sum_{j\in[M^\beta]} f^+\left(\tilde{G}(j)\right) \geq \mathbb{E}\left[\sum_{j\in[M^\beta]} f^+\left(\tilde{G}(j)\right)\right] + r\right] \\
&\leq \exp\left[-\frac{r^2}{2\left(\frac{\log^2 M}{M} + \frac{r}{3e}\right)}\right]. && \text{(B.77)}
\end{aligned}$$

Setting $r = \frac{\eta}{2}\log M$ in (B.77) we obtain

$$\begin{aligned}
\frac{r^2}{2\left(\frac{\log^2 M}{10M} + \frac{r}{3e}\right)} &\geq \frac{10Mr^2}{4\log^2 M} \vee \frac{3e}{4}r && \text{(B.78)}\\
&\geq \frac{1}{2}\cdot\left(\eta^2 M \vee \eta \log M\right) && \text{(B.79)}\\
&\geq \frac{1}{2}\cdot \eta \log M, && \text{(B.80)}
\end{aligned}$$

where the inequalities hold for all $M \geq M_3(\beta, \eta)$ large enough. This, together with (B.71) and (B.65), implies that

$$\begin{aligned}
&\mathbb{P}\left[\tilde{C} \geq \mathbb{E}[\tilde{C}] + \eta \log M\right] \\
&\leq \exp\left[-\frac{1}{2}\eta \log M\right] + e^{-M\left(1-\frac{\beta \log M}{M}\right)} && \text{(B.81)}\\
&\leq 2\exp\left[-\frac{1}{2}\eta \log M\right] && \text{(B.82)}\\
&= \frac{2}{M^{\eta/2}}, && \text{(B.83)}
\end{aligned}$$

for all $M \geq M_0 \vee M_1 \vee M_2 \vee M_3$. ■

We may now prove Theorem 6.

Proof of Theorem 6: If $H(P_X) = 0$ then trivially $\mathsf{FP}(\delta,\xi) = 0$ for any $\delta \geq 0$ and $\xi \in [0,1]$. We thus assume henceforth that $H(P_X) > 0$. In what follows, we will upper bound the pairwise error probability between two sequences using the Bhattacharyya bound. To this end, let $\tilde{x}^N, \overline{x}^N \in \mathcal{X}^N$ be a pair of sequences, where $\tilde{x}^N \neq \overline{x}^N$. Let $p_e(\tilde{x}^N \to \overline{x}^N)$ denote the error probability of a pairwise test between $\tilde{x}^N$ and $\overline{x}^N$ from the observations $Y^N \sim P_{Y^N|X^N}(\cdot \mid \tilde{x}^N)$. Then, the Bhattacharyya bound on the probability of erroneously deciding in favor of $\overline{x}^N$ in a pairwise test is given by (e.g., [35, Sec. 2.3])

$$\begin{aligned} &p_e(\tilde{x}^N \to \overline{x}^N) \\ &\leq \sum_{y^N \in \mathcal{Y}^N} \sqrt{P_{XY}^{\otimes N}[y^N \mid \tilde{x}^N] \cdot P_{XY}^{\otimes N}[y^N \mid \overline{x}^N]} && \text{(B.84)} \\ &\overset{(a)}{=} \sum_{y^N \in \mathcal{Y}^N} \prod_{i \in [M]} \sqrt{P_{XY}^{\otimes L}[\boldsymbol{y}(i) \mid \tilde{\boldsymbol{x}}(i)] \cdot P_{XY}^{\otimes L}[\boldsymbol{y}(i) \mid \overline{\boldsymbol{x}}(i)]} && \text{(B.85)} \\ &= \prod_{i \in [M]} \sum_{\boldsymbol{y} \in \mathcal{Y}^L} \sqrt{P_{XY}^{\otimes L}[\boldsymbol{y} \mid \tilde{\boldsymbol{x}}(i)] \cdot P_{XY}^{\otimes L}[\boldsymbol{y} \mid \overline{\boldsymbol{x}}(i)]} && \text{(B.86)} \\ &= e^{-\sum_{i \in [M]} d_{P_{Y|X}}(\tilde{\boldsymbol{x}}(i), \overline{\boldsymbol{x}}(i))}, && \text{(B.87)} \end{aligned}$$

where (a) holds since the fragments are independent. Assume that $\delta > 0$ and $\xi \in (0,1)$ are given, set $\eta \in (0,1)$, and define the event

$$\mathcal{F}_\eta := \left\{ x^N \in \mathcal{X}^N : \frac{1}{M} \log \left|\mathcal{A}_L(x^N)\right| \leq L \cdot H(P_X) + \eta \log M \right\}. \tag{B.88}$$

Let us denote the failure error probability conditioned on $X^N = x^N$ by $\mathsf{FP}(\delta, \xi \mid x^N)$. Then,

$$\begin{aligned} &\mathsf{FP}(\delta,\xi) \\ &= \mathbb{E}\left[\mathsf{FP}(\delta,\xi \mid X^N)\right] && \text{(B.89)} \\ &\overset{(a)}{\leq} \mathbb{E}\left[\mathsf{FP}(\delta,\xi \mid X^N) \cdot \mathbb{1}\{X^N \in \mathcal{F}_\eta\}\right] + \mathbb{P}\left[X^N \notin \mathcal{F}_\eta\right] && \text{(B.90)} \\ &\overset{(b)}{\leq} \mathbb{E}\left[\mathsf{FP}(\delta,\xi \mid X^N) \cdot \mathbb{1}\{X^N \in \mathcal{F}_\eta\}\right] + o_\eta(1), && \text{(B.91)} \end{aligned}$$

where (a) follows from the union bound, and (b) follows from Prop. 7. We next focus on the first term. Let $\mathcal{S}_L(x^N) \subset S_M$ be a set of permutations that generates $\mathcal{A}_L(x^N)$, that is, $|\mathcal{S}_L(x^N)| = |\mathcal{A}_L(x^N)|$ and for each $\tilde{x}^N \in \mathcal{A}_L(x^N)$ there exists $\pi \in \mathcal{S}_L(x^N)$ such that $\tilde{x}^n = \pi[x^N]$. Let $\mathsf{FP}(\delta,\xi,\pi[x^N] \mid X^N = x^N)$ be the probability of the event in which the reconstruction failed and the ML output was the erroneous $\pi[x^N]$. For any $x^N \in \mathcal{F}_\eta$ it then holds that

$$\begin{aligned} &\mathsf{FP}(\delta,\xi \mid X^N = x^N) \\ &\overset{(a)}{\leq} \sum_{\pi \in \mathcal{S}_L(x^N)} \mathsf{FP}(\delta,\xi,\pi[x^N] \mid X^N = x^N) && \text{(B.92)} \\ &\leq \left|\mathcal{S}_L(x^N)\right| \cdot \max_{\pi \in \mathcal{S}_L(x^N)} \mathsf{FP}(\delta,\xi,\pi[x^N] \mid X^N = x^N) && \text{(B.93)} \\ &\overset{(b)}{\leq} e^{(\beta \cdot H(P_X) + \eta) M \log M} \\ &\quad \times \max_{\pi \in \mathcal{S}_L(x^N)} \mathsf{FP}(\delta,\xi,\pi[x^N] \mid X^N = x^N) && \text{(B.94)} \\ &\overset{(c)}{\leq} e^{(\beta \cdot H(P_X) + \eta) M \log M} e^{-\xi M \cdot L d^*_{P_{Y|X}}(\delta)}, && \text{(B.95)} \end{aligned}$$

where (a) follows from the union bound, (b) follows from Prop. 7 and the assumption that $x^N \in \mathcal{F}_\eta$, and (c) follows from the following consideration: Consider an arbitrary permutation $\pi \in \mathcal{S}_L(x^N)$, and denote $\tilde{x}^n = \pi[x^N]$. If

$$\sum_{i \in [M]} \mathbb{1}\{\Delta(\boldsymbol{x}(i), \tilde{\boldsymbol{x}}(i)) \geq \delta\} \geq \xi M \tag{B.96}$$

then the definition of $d^*_{P_{Y|X}}(\delta)$ in (23) implies that

$$\sum_{i \in [M]} d_{P_{Y|X}}(\boldsymbol{x}(i), \tilde{\boldsymbol{x}}(i)) \geq \xi M L \cdot d^*_{P_{Y|X}}(\delta). \tag{B.97}$$

In this case, (B.87) implies that

$$\mathsf{FP}(\delta,\xi,\pi[x^N] \mid X^N = x^N) \leq \exp\left[-\xi M L \cdot d^*_{P_{Y|X}}(\delta)\right]. \tag{B.98}$$

Alternatively, if (B.96) does not hold, we have that $\mathsf{FP}(\delta,\xi,\pi[x^N] \mid X^N = x^N) = 0$ (by the definition of reconstruction success at failure level ξ). Inserting (B.95) back to (B.91), using $L = \beta \log M$, shows that if

$$\xi > \frac{H(P_X) - \eta}{d^*_{P_{Y|X}}(\delta)} \tag{B.99}$$

then $\mathsf{FP}(\delta,\xi) = o_\eta(1)$ for all M large enough. The result then follows by taking $\eta \downarrow 0$. ■

REFERENCES

[1] Alexander Barg and G. David Forney. Random codes: Minimum distances and error exponents. *IEEE Transactions on Information Theory*, 48(9):2568–2573, 2002.

[2] Sergey G. Bobkov and Michel Ledoux. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *Journal of functional analysis*, 156(2):347–365, 1998.

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[4] Yuval Cassuto and Jacob Ziv. Efficient compression of long arbitrary sequences with no reference at the encoder. *IEEE Transactions on Information Theory*, 67(1):1–9, 2020.

[5] Yuan-sheng Cheng, Michael Neely, and Keith M. Chugg. Iterative message passing algorithm for bipartite maximum weighted matching. In *2006 IEEE International Symposium on Information Theory*, pages 1934–1938. IEEE, 2006.

[6] Johan Chrisnata, Han Mao Kiah, Alexander Vardy, and Eitan Yaakobi. Bee identification problem for DNA strands. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 969–974. IEEE, 2022.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, USA, 2006.

[8] I. Csiszár. The method of types. *IEEE Transactions on Information Theory*, 44(6):2505–2523, 1998.

[9] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, Cambridge, U.K., 2011.

[10] Daniel Cullina, Prateek Mittal, and Negar Kiyavash. Fundamental limits of database alignment. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 651–655. IEEE, 2018.

[11] D. De Caen. A lower bound on the probability of a union. *Discrete mathematics*, 169(1-3):217–220, 1997.

[12] Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

[13] Jack Edmonds and Richard M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)*, 19(2):248–264, 1972.
[14] Yotam Gershon and Yuval Cassuto. Efficient distributed source coding of fragmented genomic sequencing data. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 3302–3307. IEEE, 2021.
[15] Yotam Gershon and Yuval Cassuto. Genomic compression with read alignment at the decoder. *arXiv preprint arXiv:2205.07947*, 2022.
[16] Te Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36:133–156, 1978.
[17] Han Mao Kiah, Alexander Vardy, and Hanwen Yao. Efficient algorithms for the bee-identification problem. *arXiv preprint arXiv:2212.09952*, 2022.
[18] Ioannis Kontoyiannis and Mokshay Madiman. Measure concentration for compound Poisson distributions. *Electronic Communications in Probability*, 11:45–57, 2006.
[19] M. Kovačević and V. Y. F. Tan. Codes in the space of multisets – Coding for permutation channels with impairments. *IEEE Transactions on Information Theory*, 64(7):5156–5169, 2018.
[20] A. Lenz, L. Welter, and S. Puchinger. Achievable rates of concatenated codes in DNA storage under substitution errors. In *International Symposium on Information Theory and Its Applications*, pages 269–273. IEEE, 2020.
[21] Anuran Makur. Coding theorems for noisy permutation channels. *IEEE Transactions on Information Theory*, 66(11):6723–6748, 2020.
[22] Neri Merhav. Error exponents of typical random codes. *IEEE Transactions on Information Theory*, 64(9):6223–6235, 2018.
[23] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.
[24] Soheil Mohajer, Abolfazl S. Motahari, and David N. C. Tse. Reference-based DNA shotgun sequencing: Information theoretic limits. In *2013 IEEE International Symposium on Information Theory*, pages 1635–1639. IEEE, 2013.
[25] Abolfazl S. Motahari, Guy Bresler, and David N. C. Tse. Information theory of DNA shotgun sequencing. *IEEE Transactions on Information Theory*, 59(10):6273–6289, 2013.
[26] Keith W. Ross and James F. Kurose. Computer networking, 2012.
[27] Comandur Seshadhri. Lecture notes for cse 290a: Randomized algorithms, 2020. Available at https://users.soe.ucsc.edu/~sesh/Teaching/2020/CSE290A/Slides/Lecture11.pdf.
[28] Claude E. Shannon, Robert G. Gallager, and Elwyn R. Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels. I. *Information and Control*, 10(1):65–103, 1967.
[29] I. Shomorony and R. Heckel. DNA-based storage: Models and fundamental limits. *IEEE Transactions on Information Theory*, 67(6):3675–3689, 2021.
[30] J. Sima, N. Raviv, and J. Bruck. On coding over sliced information. *IEEE Transactions on Information Theory*, 67(5):2793–2807, 2021.
[31] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *Information Theory, IEEE Transactions on*, 19(4):471–480, July 1973.
[32] Ran Tamir and Neri Merhav. Error exponents in the bee identification problem. *IEEE Transactions on Information Theory*, 67(10):6564–6582, 2021.
[33] Anshoo Tandon, Vincent Y. F. Tan, and Lav R. Varshney. The bee-identification problem: Bounds on the error exponent. *IEEE Transactions on Communications*, 67(11):7405–7416, 2019.
[34] Anshoo Tandon, Vincent Y. F. Tan, and Lav R. Varshney. The bee-identification error exponent with absentee bees. *IEEE Transactions on Information Theory*, 66(12):7602–7614, 2020.
[35] A. J. Viterbi and J. K. Omura. *Principles of Digital Communication and Coding*. Dover Publications, 2009.
[36] Nir Weinberger. Error probability bounds for coded-index DNA storage channels. *IEEE Transactions on Information Theory*, 68(11):7005–7022, 2022.
[37] Nir Weinberger and Neri Merhav. The DNA storage channel: Capacity and error probability bounds. *IEEE Transactions on Information Theory*, 68(9):5657–5700, 2022.

Nir Weinberger (S'14-M'18-SM'23) is an assistant Professor at the The Viterbi Faculty of Electrical and Computer Engineering, Technion - Israel Institute of Technology. Previously, from 2017 to 2018 he was a post-doctoral fellow at Tel Aviv University, and from 2018-2020 he was a Technion-MIT post-doctoral fellow at the Massachusetts Institute of Technology, Cambridge, MA, USA. He has received the B.Sc. and M.Sc. degrees (both *summa cum laude)* from Tel-Aviv University, Tel-Aviv, Israel, in 2006 and 2009, respectively, and his Ph.D. degree in 2017, from the Technion, Israel Institute of Technology.

Ilan Shomorony (S'09-M'14) received the Ph.D. degree in electrical and computer engineering from Cornell University in 2014. He was a post-doctoral scholar at UC Berkeley through the NSF Center for Science of Information (CSoI) until 2017. After that, he spent a year working as a Researcher and Data Scientist at Human Longevity Inc., a personal genomics company. He is currently an Assistant Professor in electrical and computer engineering at the University of Illinois at Urbana-Champaign (UIUC) where he is a member of the Coordinated Science Laboratory. His research interests include information theory, communications, and computational biology. He received the NSF CAREER Award in 2021.