Fundamental Limits of Reference-Based Sequence Reordering

Nir Weinberger
Technion - Israel Institute of Technology
nirwein@technion.ac.il

Ilan Shomorony
University of Illinois Urbana-Champaign
ilans@illinois.edu

***Abstract*—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 is considered. First, in the regime in which the fragments are relatively long, and typically no fragment appears more than once, the exact scaling of the failure probability of maximum likelihood reconstruction algorithm is determined, both for perfect reconstruction as well as partial reconstruction. Second, the regime in which the fragments are relatively short and repeating fragments abound is characterized. A trade-off is stated between the fraction of fragments that fail to be adequately reconstructed vs. the distortion level allowed for the reconstruction of each fragment, while still allowing vanishing failure probability.**

I. Introduction

We consider the problem of reconstructing a sequence $X^N \in \mathcal{X}^N$ from its non-overlapping $M = N/L$ consecutive fragments of length L each, and a reference sequence $Y^N \in \mathcal{Y}^N$. This problem is motivated by settings in which data is observed out of order, and ordering is made possible through side information. The problem can also be seen as a dataset alignment problem, where one needs to find a matching between two pairwise approximately matched data-sets (potentially for subsequent joint compression). For example, DNA sequencing of a genomic sequence based on a reference sequence [17], or the transmission of short encoded packets over packet networks [13], while decoding them with a side-information sequence. More specifically, we assume that X^N is drawn from a memoryless source, and that Y^N is obtained by passing X^N in a discrete memoryless channel $P_{Y|X}$. Furthermore, we assume that the fragment length is logarithmic in M, that is $L = \beta \log M$ for some $\beta > 0$. For this scaling, the problem described above is also closely related to the *bee identification* (BI) problem [5], [12], [22]–[24], for which the M unordered fragments of X^N represent noisy measurements of M ordered fragments of Y^N, and the matching between the corresponding fragments enables us to identify bees from a picture. A codebook for this problem is thus 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 drawing the $N = LM$ symbols of the fragments, in an independently and identically distributed (IID) manner. The average error probability over the random ensemble of codebooks is similar to the reconstruction error in the fragments ordering problem (with the inconsequential notational changes of $P_{Y|X} \leftrightarrow P_{X|Y}$, $L \leftrightarrow n$ and $M \leftrightarrow m$.).

Nonetheless, there are two main differences between the ordering and BI problems. First, in the ordering problem, one is interested in recovering the sequence, and not necessarily the permutation. 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 [22], [23], improved bounds are obtained by considering the average error of the *typical random code* [1], [15], or via expurgation techniques [22]. This is impossible for the ordering problem. In fact, these two matters are interrelated. As an extreme example, in the event that all fragments of X^N are equal, the order reconstruction is trivially *perfect*, whereas maximal error probability is obtained if these identical fragments are chosen as a codebook for the BI problem. More generally, *repeated fragments* in the sequence make the ordering and BI problems different. This typically happens when the fragments are relatively short (small β), or the entropy of X is low.

Our main contributions are twofold: First, we consider the regime of no repeating-fragments, and upper bound the failure probability both for perfect reconstruction, as well as for imperfect reconstruction, i.e., allowing a fraction $\xi \in (0,1)$ of erroneous reconstructed fragments (this was proposed in the future research part of [23], as well as in [22], although with ξM being replaced by a constant). For $\xi = 0$, this revisits the setting of random coding analysis for joint decoding in the BI problem [23]. 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 $\psi_2(P_{XY})$ is given in (6). 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 $e^{-M \log M \cdot \xi(\beta\psi_2(P_{XY})-1)}$, that is, exponential decay with respect to (w.r.t.) $M \log M$. Our Theorem 3 then shows that these rates of failure probability are in fact *tight*, under mild assumptions. In the $\xi = 0$ case, the improvement of Theorem 1 over [23] is by generalizing it to any source P_{XY}, and not just symmetric (uniform) binary source with a binary symmetric channel (BSC), and strictly tightening it, even for the aforementioned binary case. Second, we consider the regime of repeating-fragments. As a technical contribution, we show that this regime is characterized by the condition $\beta < 1/H(P_X)$. Essentially, this assures that the number of distinct sequences

that can be constructed by ordering the fragments is roughly $e^{\beta H(P_X)\cdot M\log M+o(M\log M)}$, which is strictly less than $M!\leq e^{M\log M+O(\log M)}$. This is shown to hold with probability $1-e^{-\Theta(M)}$. Furthermore, we note that *similar* fragments are also expected to be present in X^N, and so it is unreasonable to require perfect reconstruction. Therefore, we propose to tolerate a distortion level δ between fragments (for some given distortion measure). The reconstruction is then successful if at most a fraction $\xi\in[0,1]$ of the fragments were reconstructed with low distortion. Evidently, this leads to a trade-off between ξ and δ. In Theorem 5 we state an achievable bound on this trade-off, and show that as long as $\beta<1/H(P_X)$, the condition $\xi>H(P_X)/d^*_{P_{Y|X}}(\delta)$ suffices to obtain vanishing failure probability. Here, $d^*_{P_{Y|X}}(\delta)$ is the minimal Bhattacharyya distance for fragments of distortion larger than δ (see (14)).

Other Related Work: An information-theoretic study of DNA sequence reconstruction from short fragments taken at *random* locations was initiated in [18], and its reference-based counterpart was considered in [17]. In [3], the problem of compressing a non-probabilistic source was considered when the encoder has a possible list of reference vectors. In [8], [9], 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 related to the DNA storage sampling-shuffling channel [20], in which short unordered fragments store the information, or, more generally, to *permutation channels* [13], [14], [21], in which the output sequence is a permuted and noisy version of the input sequence.

Outline: In Sec. II we formulate the problem, in Sec. III we state our main results, and in Sec. IV we conclude the paper. Main proofs (and proof sketches for some of the results) are presented in the Appendix.

II. Problem Formulation

For $j>i$, $X_i^j:=(X_i,X_{i+1},\ldots,X_j)$ 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* of length L on $\mathcal{X}$, and let $\mathcal{P}(\mathcal{X})$ be the set of all PMFs on $\mathcal{X}$. The type class [6, Ch. 2] of a type $Q_X\in\mathcal{P}_L(\mathcal{X})$ is denoted by $T_L(Q_X)$. The Rényi entropy of order $\alpha\geq 0$, $\alpha\neq 1$ is denoted by $H_\alpha(Q_X):=\frac{1}{1-\alpha}\log(\sum_{x\in\mathcal{X}}Q_X^\alpha(x))$, and the Shannon entropy is denoted by $H(Q_X):=\lim_{\alpha\downarrow 1}H_\alpha(Q_X)=-\sum_{x\in\mathcal{X}}Q_X(x)\log Q_X(x)$. For a pair of PMFs Q_X and P_X, the Kullback-Leibler (KL) divergence is denoted by $D_{\mathrm{KL}}(Q_X\,||\,P_X)$. For an integer M, $[M]:=\{1,\ldots,M\}$.

Let $(X^N,Y^N)\sim P_{XY}^{\otimes N}$ be a pair of length N IID sequences, over the finite alphabet $\mathcal{X}\times\mathcal{Y}$, and assume without loss of generality (WLOG) that P_X is fully supported on $\mathcal{X}$. Let L denote a fragment length, and assume for notational simplicity that $M:=N/L$ is integer (and ignore in what follows any integer constraints on asymptotically large numbers, as they are inconsequential to 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 (the group of all bijections from $[M]$ to itself). A permuted sequence of fragments is denoted by $\pi[X^N]:=(\boldsymbol{X}(\pi(1)),\boldsymbol{X}(\pi(2)),\ldots,\boldsymbol{X}(\pi(M)))$, and $\mathcal{A}_L(X^N):=\left\{\pi[X^N]\right\}_{\pi\in S_M}$ 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. A maximum likelihood reconstruction algorithm chooses

$$\hat{X}^N=\underset{\tilde{X}^N\in\mathcal{A}_L(X^N)}{\arg\max}\ \mathbb{P}\left[Y^N\mid\tilde{X}^N\right],\tag{1}$$

or equivalently, a proper permutation (ordering) of the fragments $\{\boldsymbol{X}(i)\}_{i\in[M]}$. 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

$$\Delta(\tilde{\boldsymbol{X}},\overline{\boldsymbol{X}})=\frac{1}{L}\sum_{j\in[L]}\Delta(\tilde{X}_j,\overline{X}_j).\tag{2}$$

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

$$\Xi_\delta(X^N,\hat{X}^N):=\frac{1}{M}\sum_{i\in[M]}\mathbb{1}\{\Delta(\boldsymbol{X}(i),\hat{\boldsymbol{X}}(i))\geq\delta\}\tag{3}$$

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

$$\mathsf{FP}(\delta,\xi):=\mathbb{P}\left[\Xi_\delta(X^N,\hat{X}^N)\geq\xi\right].\tag{4}$$

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 $L=\beta\cdot\log M$.

In what follows, the probability of a reconstruction failure will be bounded using the *Chernoff distance* between $\overline{x},\tilde{x}\in\mathcal{X}$, and $s\in(0,1)$ is denoted by

$$d_{P_{Y|X},s}(\overline{x},\tilde{x}):=-\log\sum_{y\in\mathcal{Y}}P_{Y|X}^s[y\mid\overline{x}]\cdot P_{Y|X}^{1-s}[y\mid\tilde{x}].\tag{5}$$

For brevity, the dependence of the Chernoff distance on $P_{Y|X}$ will often be suppressed henceforth. Moreover, this distance will mostly 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 $d_s(\overline{\boldsymbol{x}},\tilde{\boldsymbol{x}}):=\sum_{i\in[L]}d_s(\overline{x}_i,\tilde{x}_i)$. This additive distance only depends 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) $d_s(Q_{\overline{X}\tilde{X}}):=\frac{1}{L}d_s(\overline{\boldsymbol{x}},\tilde{\boldsymbol{x}})$ 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 $\Delta(\overline{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ only depends on the joint type $Q_{\overline{X}\tilde{X}}$ of $\overline{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$, 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

A. The No Repeating-Fragments Regime with Zero Distortion

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

$$\psi_2(P_{XY}) := \min_{Q_{X_1X_2}} \frac{1}{2} D_{\mathrm{KL}}\left(Q_{X_1X_2} \,||\, P_X^{\otimes 2}\right) + d_{P_{Y|X}}(Q_{X_1X_2}). \tag{6}$$

1) An Upper Bound on the Reconstruction Error:

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

$$\mathsf{FP}(\xi = 0) = O\left(M^{2(1-\beta\psi_2(P_{XY}))}\right) \tag{7}$$

for all $M \geq M_0(P_{XY})$, and for $\xi > 0$

$$\mathsf{FP}(\xi) = \exp\left[-M \log M \cdot \xi \left(\beta\psi_2(P_{XY}) - 1 - O(M^{-1})\right)\right]. \tag{8}$$

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 $\xi > 0$ it decays exponentially with $M \log M$, i.e., much faster. The error bound for $\xi = 0$ is dominated by transposition errors, i.e., an almost perfect reconstruction of the sequence, except for a single pair of fragments that has exchanged location. The rate function determining the threshold is given by $\psi_2(P_{XY})$, which can be computed for any P_{XY}, as a convex optimization problem over $\mathcal{P}(\mathcal{X}^2)$ (6). In addition, the symmetry of the Bhattacharyya distance and convexity of the KL divergence imply that the solution $Q^*_{X_1X_2}$ of the minimization problem in (6) must have equal marginals, i.e., $Q^*_{X_1} = Q^*_{X_2}$. 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 dominant error event in this bound turns out to be a set of $\frac{\xi M}{2}$ transpositions.

Proof sketch of Theorem 1: As in [23], the main technical part of the proof of Theorem 1 is the analysis of the pairwise error probability from the true source vector to the permutation of its fragments, when the permutation is a *cycle*. This is achieved using the Bhattacharyya upper bound (e.g., [25, Sec. 2.3]). Specifically, 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,*

$$\mathbb{E}\left[\exp\left(-d_{P_{Y|X}}(X_1^K, \tilde{X}_1^K)\right)\right] \leq e^{-K\cdot\psi_2(P_{XY})}, \tag{9}$$

where $\psi_2(P_{XY})$ is defined in (6).

The proof of Lemma 2 is based on first upper bounding the expected Bhattacharyya upper bound (left-hand side of (9)) using the *Donsker-Vardhan variational formula*. 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 $\mathcal{P}(\mathcal{X}^K)$. The proof of the lemma then continues by establishing that transpositions, i.e., cycles of length 2, are the worst case, that is, $\psi_K(P_{XY}) \geq \psi_2(P_{XY})$ 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 $d(Q_{X_1X_2}) + d(Q_{X_2X_3}) + d(Q_{X_1X_3})$, 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, $D_{\mathrm{KL}}(Q_{X_1X_2X_3}||P_X^{\otimes 3})$, is analogously lower bounded by the KL divergence of the marginal pairs using *Han's inequality for the KL divergence* [2, Theorem 4.9] [11]. 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 $\tilde{X}_1 = X_K$. 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, then leads to the upper bounds in Theorem 1.

A comparison with [23]: The setting in [23] 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 for a general P_{XY}, that the worst case is a transposition. 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 [23]. A similar effect holds for the $\xi > 0$ case. In [23] the "break of the cycle" that was obtained here by the relaxation to $\varphi_K(P_{XY})$ was obtained by ignoring the contribution of the Bhattacharyya distance of the last pair of fragments $d(X_K, \tilde{X}_K)$.

2) A Lower Bound on the Reconstruction Error: The next lower bound on $\mathsf{FP}(\xi)$ establishes the tightness of Theorem 1.

Theorem 3. *Assume that $d(x_1, x_2) < \infty$ and that $2\psi_2(P_{XY}) < H_2(P_X)$. If $\beta > \frac{1}{\psi_2(P_{XY})}$ then it holds that*

$$\mathsf{FP}(\xi = 0) \geq M^{2[1-\beta\psi_2(P_{XY})]+o(1)}, \tag{10}$$

and for $\xi > 0$ it holds that

$$\mathsf{FP}(\xi) \geq \exp\left[-\xi M \log M \cdot \left[\beta\psi_2(P_{XY}) - 1 + o(1)\right]\right]. \quad (11)$$

The qualifying assumptions: The condition $d_s(x_1, x_2) < \infty$ is somewhat technical, and is related to a continuity requirement. The condition $2\psi_2(P_{XY}) < H_2(P_X)$ 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 error probability of a transposition, i.e., exchanging the order of $\boldsymbol{X}(i_1)$ and $\boldsymbol{X}(i_2)$ for some $i_1, i_2 \in [M], i_1 < i_2$. The probability of this event can be lower bounded using the technique of Shannon, Gallager and Berlekamp [19, 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. 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. We use two techniques to lower bound this probability of a union of events. For $\xi = 0$, we use de Caen's inequality [7] (as was also used in [23]). For $\xi > 0$, we identify that the sequence of error indicators for such transpositions is a sequence of *infinitely exchangeable binary random variables* (or *interchangeable*). We then use de-Finetti's theorem [4, Ch. 7.3] to show that the union bound is tight for this case too. This technique may be of independent interest for other lower bounds.

Example 4 ($\psi_2(P_{XY})$ 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

$$P_{Y|X}(y \mid x; \alpha) := \begin{cases} 1 - \alpha, & y = x \\ \frac{\alpha}{|\mathcal{Y}|-1} & \text{otherwise} \end{cases}. \quad (12)$$

For the case $|\mathcal{X}| = 2$ a simple closed-form solution is

$$\psi_2(P_{XY}) = \frac{1}{2}\left[\log 2 - \log\left(1 + 4\alpha(1-\alpha)\right)\right]. \quad (13)$$

The value of $\psi_2(P_{XY})$ as a function of α (solved using the CVX solver [10] for the case $|\mathcal{X}| > 2$) appears in Fig. 1. As might be expected when $\psi_2(P_{XY})$ increases with $|\mathcal{X}|$, and hence the lower bound on β decreases – ordering the fragments is easier for larger entropy sources.

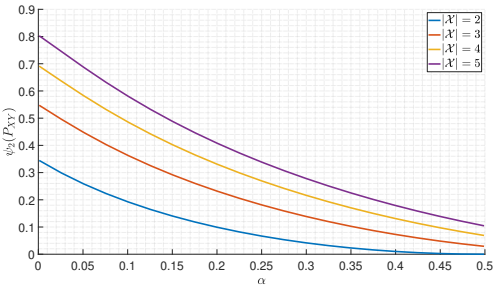

Figure 1. $\psi_2(P_{XY})$ for Example 4.

B. The Repeating-Fragments Regime with Positive Distortion

In this section, we address the small β regime, or low entropy P_X. This is where 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, and so fragments that are close according to the distortion measure Δ also abound. Thus, we propose to 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, let

$$d^*_{P_{Y|X}}(\delta) := \min_{Q_{X_1X_2} \in \mathcal{P}(\mathcal{X}^2): \Delta(Q_{X_1X_2}) \geq \delta} d_{P_{Y|X}}(Q_{X_1X_2}), \quad (14)$$

be 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 is tolerated by a successful reconstruction – increasing the distortion level δ allows to obtain a reduced 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 5. *Assume that $\beta < \frac{1}{H(P_X)}$. If $\xi > \frac{H(P_X)}{d^*_{P_{Y|X}}(\delta)}$ then $\mathsf{FP}(\delta, \xi) = e^{-\Omega(M)}$.*

Discussion: Theorem 5 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 as fast as exponential with M, though in an unspecified rate. This is a 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". Theorem 5 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 5: 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 $\frac{1}{M}\log|\mathcal{A}_L(X^N)|$ 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 probability $1-e^{-\Theta(M)}$. This cardinality can be much smaller for low β or sources with low entropy.

Proposition 6. *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 for all $M \geq M_0(P_X, \beta, \eta)$ as*

$$\mathbb{P}\left[\frac{1}{M}\log\left|\mathcal{A}_L(X^N)\right| \geq L\cdot H(P_X) + \eta\log M\right] \leq e^{-c\cdot\eta^2 M}. \tag{15}$$

To outline the proof of Prop. 6, let us assume, for notational simplicity, that $\mathcal{X}^L \equiv \{a_1\ldots, a_{M^\beta}\}$, where $|\mathcal{X}^L| = M^\beta$. Then, for any $x^N \in \mathcal{X}^N$ and any $j \in [M^\beta]$, $G(j) := \sum_{i\in[M]}\mathbb{1}\{\boldsymbol{X}(i) = a_j\}$ is the number of times that $a_j \in \mathcal{X}^L$ appears in the fragments of x^N, and $G := (G(1),\ldots,G(M^\beta))$ is the *histogram* vector of x^N for length-L fragments. The proof is based on the standard entropy bound on the multinomial coefficient, which leads to

$$\frac{1}{M}\log\left|\mathcal{A}_L(X^N)\right| \leq -\sum_{j\in[M^\beta]}\frac{G(j)}{M}\log\frac{G(j)}{M}. \tag{16}$$

Given the fragments model, the histogram vector $G = (G(1),\ldots,G(M^\beta))$ is distributed as a *multinomial* random variable (RV), whose components are statistically *dependent*. The upper bound in (16) is thus a complicated function of G, and it is difficult to directly analyze its concentration properties. Nonetheless, the probability of an event under the multinomial distribution can be upper bounded by the probability of the same event under a Poisson distribution that has *independent* components [16, Sec. 5.4]. We thus consider a *Poissonized* version $\tilde{G}$ of G, and analyze the tail behavior of $f(\frac{\tilde{G}(j)}{M})$ for $f(t) := -t\log t$. We show using concentration bounds for Lipschitz functions of Poisson random variables that $f(\frac{\tilde{G}(j)}{M})$ is a sub-gamma random variable [2, Ch. 2], and then bound the concentration of $\sum_{j\in[M^\beta]} f(\frac{\tilde{G}(j)}{M})$ via Bernstein's inequality. Additional approximation and truncation arguments are required since, strictly speaking, the function $f(t)$ involved in the upper bound is not Lipschitz continuous on $\mathbb{R}_+$.

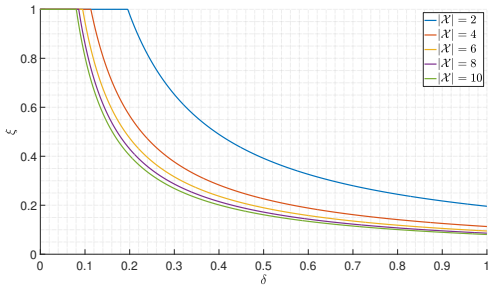

Figure 2. The trade-off between ξ and δ for Example 7.

Example 7 (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 (12). In this case, it holds that $d_{P^{(\alpha)}_{Y|X}}(x,\tilde{x}) = d_\alpha\cdot\mathbb{1}[\overline{x}\neq\tilde{x}]$ where

$$d_\alpha := -\log\left[\sqrt{\frac{4(1-\alpha)\alpha}{|\mathcal{Y}|-1}} + \frac{\alpha(|\mathcal{Y}|-2)}{|\mathcal{Y}|-1}\right]. \tag{17}$$

Further assume that the distortion measure is the Hamming distortion measure $\Delta(\overline{x},\tilde{x}) = \mathbb{1}[\overline{x}\neq\tilde{x}]$. Then, it is simple to obtain that $d^*(\delta) = \delta\cdot d_\alpha$, and the bound of Theorem 5 results $\xi > \frac{H(P_X)}{\delta\cdot d_\alpha}$. The achievable trade-off between ξ and δ is shown in Fig. 2 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 remark that computing $d^*_{P_{Y|X}}(\delta)$ for general channels is a linear program (14) that 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 the consecutive fragments of a sequence based on a reference sequence. We considered the regime of no-repeating fragments, and sharply characterized the failure probability for both perfect and imperfect reconstruction ($\xi = 0$ and $\xi > 0$), thus improving and generalizing [23]. We then characterized the repeating fragments regime as $\beta < 1/H(P_X)$, and obtained an achievable trade-off between the distortion level $\delta > 0$ and failure level ξ for 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 whether the optimal decay rate of the reconstruction failure probability is $e^{-\Theta(M)}$ or faster, and how it depends on the problem parameters.

Acknowledgment

The work of N.W. was partly 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.

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

[3] 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.

[4] Yuan Shih Chow and Henry Teicher. *Probability theory: Independence, interchangeability, martingales*. Springer Science & Business Media, 2003.

[5] 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.

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

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

[8] 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.

[9] Yotam Gershon and Yuval Cassuto. Genomic compression with read alignment at the decoder. *arXiv preprint arXiv:2205.07947*, 2022.

[10] Michael Grant and Stephen Boyd. Cvx: Matlab software for disciplined convex programming, version 2.1, 2014.

[11] Te Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36:133–156, 1978.

[12] Han Mao Kiah, Alexander Vardy, and Hanwen Yao. Efficient algorithms for the bee-identification problem. *arXiv preprint arXiv:2212.09952*, 2022.

[13] 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.

[14] Anuran Makur. Coding theorems for noisy permutation channels. *IEEE Transactions on Information Theory*, 66(11):6723–6748, 2020.

[15] Neri Merhav. Error exponents of typical random codes. *IEEE Transactions on Information Theory*, 64(9):6223–6235, 2018.

[16] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.

[17] 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.

[18] 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.

[19] 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.

[20] I. Shomorony and R. Heckel. DNA-based storage: Models and fundamental limits. *IEEE Transactions on Information Theory*, 67(6):3675–3689, 2021.

[21] J. Sima, N. Raviv, and J. Bruck. On coding over sliced information. *IEEE Transactions on Information Theory*, 67(5):2793–2807, 2021.

[22] Ran Tamir and Neri Merhav. Error exponents in the bee identification problem. *IEEE Transactions on Information Theory*, 67(10):6564–6582, 2021.

[23] 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.

[24] 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.

[25] A.J. Viterbi and J.K. Omura. *Principles of Digital Communication and Coding*. Dover Publications, 2009.