Fundamental Bounds for Sequence Reconstruction from Nanopore Sequencers

Abram Magner, Jarosław Duda, Wojciech Szpankowski and Ananth Grama

***Abstract*—Nanopore sequencers are emerging as promising new platforms for high-throughput sequencing. As with other technologies, sequencer errors pose a major challenge for their effective use. In this paper, we present a novel information theoretic analysis of the impact of insertion-deletion (indel) errors in nanopore sequencers. In particular, we consider the following problems: (i) for given indel error characteristics and rate, what is the probability of accurate reconstruction as a function of sequence length; (ii) using replicated extrusion (the process of passing a DNA strand through the nanopore), what is the number of replicas needed to accurately reconstruct the true sequence with high probability?**

Our results provide a number of important insights: (i) the maximum length of a sequence that can be accurately reconstructed from a single sample in the presence of indel errors is relatively small; and (ii) replicated extrusion is an effective technique for accurate reconstruction. We show that for typical distributions of indel errors, the required number of replicas is a slow function (polylogarithmic) of sequence length – implying that through replicated extrusion, we can sequence large reads using nanopore sequencers. Moreover, we show that in certain cases, the required number of replicas can be related to information-theoretic parameters of the indel error distributions.

I. Introduction

The past few years have seen significant advances in sequencing technologies. Sequencing platforms from Illumina, Roche, PacBio and other vendors are commonly available in laboratories. Accompanying these hardware advances, significant progress has been made in statistical methods, algorithms, and software for tasks ranging from base calling to complete assembly. Among the key distinguishing features of these sequencing platforms are their read lengths and error rates. Short read lengths pose problems for sequencing high-repeat regions. Higher error rates, on the other hand, require oversampling to either correct, or discard erroneous reads without adversely impacting sequencing/ mapping quality. Significant research efforts have studied tradeoffs of read length, error rates, and sequencing complexity. An excellent survey of these efforts is presented by Quail et al. [21].

More recently, nanopores have been proposed as platforms for sequencing. Nanopores are fabricated either using organic channels (pore-forming proteins in a bilayer) or solid-state material (silicon nitride or graphene). An ionic current is passed through this nanopore by establishing an electrostatic potential. When an analyte simultaneously passes through the nanopore, the current flow is disrupted. This disruption of the current flow is monitored, and used to characterize the analyte. This general principle can be used to characterize nucleotides, small molecules, and proteins. Complete solutions based on this technology are available from Oxford Nanopore Technologies [16]. In this platform, a DNA strand is extruded through a protein channel in a membrane. The rate of extrusion must be slower than current measurement (sampling) for characterizing each base (or groups of small number of bases, up to four, in the nanopore at any point of time).

In principle, nanopores have several attractive features – long reads (beyond 100K bases) and minimal sample preparation. However, there are potential challenges that must be overcome – among them, the associated error rate. The extrusion rate of a DNA strand through a protein channel is controlled using an enzyme [19]. This rate is typically modeled as an exponential distribution. When a number of identical bases pass through the nanopore, the observed signal must be parsed to determine the precise number of bases. This results in one of the dominant error modes for nanopore sequencers. Insertion-deletion errors in such sequencers are reported to range from 4% by O'Donnell et al. [19], [23], approximately 13% in addition to a 5% substitution error by Jain et al. [11],

A. Magner is with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign (e-mail: anmagner@illinois.edu).

J. Duda was with Department of Computer Science and Center for Science of Information, Purdue University. He is now with the Institute of Computer Science, Faculty of Mathematics and Computer Science Jagiellonian University (e-mail: dudajar@gmail.com).

W. Szpankowski and A. Grama are with the Department of Computer Science and Center for Science of Information, Purdue University (e-mail: spa@cs.purdue.edu, ayg@cs.purdue.edu). W. Szpankowski is also with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland.

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01.

and up to 38.2% by Laver et al. [15].

High error rate can be handled using replicated reads for de-novo assembly, or through algorithmic techniques using reference genomes. The Oxford Nanosequencer claims a scalable matrix of pores and associated sensors using which replicated reads can be generated. Alternately, other technologies based on bi-directional extrusion have been proposed. In either case, two fundamental questions arise for de novo assembly: (i) for single reads, what is the bound on read length that can be accurately reconstructed using a nanopore sequencer with known error rates; and (ii) what is the number of replicas needed to accurately reconstruct the sequence with high probability (analytically defined). We provide well characterized bounds for both of these questions in this paper. For replicated reads, we assume that each fragment is read multiple times. Since it is not currently possible to exercise fine-grain control over the nanopore to read the same sequence multiple times, we achieve this using PCR amplification and resulting reads from multiple copies. These reads are aligned to achieve the same effect as reading individual fragments multiple times. Note that the alignment problem is simpler in this case, owing to longer reads.

We present a novel information theoretic analysis of the impact of indel errors in nanopore sequencers. We model the sequencer as a sticky insertion-deletion channel. The DNA sequence is fed into this channel and the output of the channel is used to reconstruct the input sequence. Using this model, we solve the following problems: (i) for given error characteristics and rate, what is the probability of accurate reconstruction as a function of sequence length; and (ii) what is the number of replicas needed to exactly reconstruct the input sequence with high probability?

Our results provide a number of important insights: (i) the maximum length of sequence that can be accurately reconstructed from a single sample in the presence of indel errors is relatively small; and (ii) the number of replicas required for accurate reconstruction is a slow function (polylogarithmic) of the sequence length – implying that through replicated extrusion, we can sequence large reads using nanopore sequencers. The bounds we derive are fundamental in nature – i.e., they hold for any re-sequencing/ processing technique. Please note this this study does not model substitution errors. However, our lower bounds (Theorems 1 and 2) on the number of replicas necessary for accurate reconstruction carry over to models featuring substitution errors in addition to indel errors. Modeling substitution errors poses significant additional challenges within an information theoretic framework, which are topics of ongoing efforts.

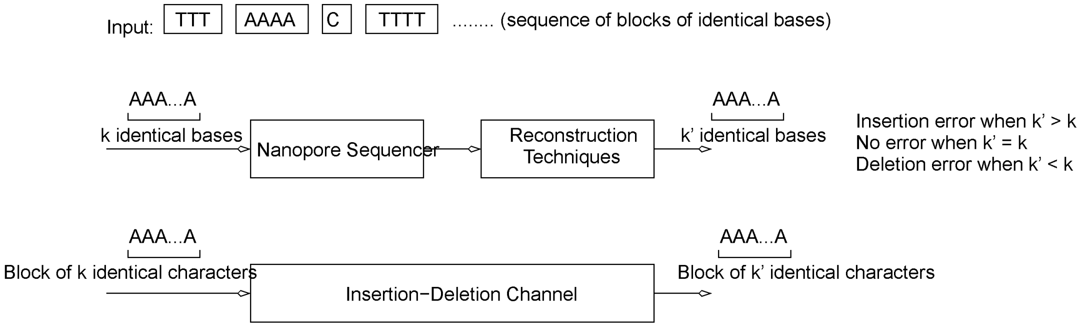
II. Approach

In this section, we present our model and the underlying concepts in information theory that provide the analytical substrates. We define notions of a channel, reconstruction, and an insertion-deletion channel. We then describe how these concepts are mapped to the problem of sequence reconstruction in nanopore sequencers.

Our basic model for a nanopore sequencer is illustrated in Figure 1. A DNA sequence is input to the nanopore sequencer. This sequence is read and suitably processed to produce an output sequence. We view the input sequence as a sequence of blocks. Each block is comprised of a variable number (k) of identical bases. The nanopore sequencer potentially introduces errors into each block by altering the number of repeated bases. If the output block size k' is not the same as the input block size k, an indel error occurs. Specifically, $k' < k$ corresponds to a deletion error, and $k' > k$ to an insertion error (we emphasize that, though several insertions and deletions may occur in the physical processing of a block, our model is meant to capture the net effect of these errors). Note that this model does not account for substitution errors.

We model the sequencing process (both the sequencer and the associated processing) as a channel. A channel in information theory is a model (traditionally for a storage or communication device, but in our case, used more generally) for information transfer with certain error characteristics. The input sequence of blocks is sent into this channel. The error characteristics of the channel transform a block of $k > 0$ characters into a block of k' characters. This transformation is modeled as a probability distribution: $k' {\sim} G(k, P)$, where G is the distribution and P, the associated set of parameters tuned to the sequencing platform. In typical scenarios, the distribution peaks at k and decays rapidly on either side. The distribution may be asymmetric around k depending on relative frequency of insertion and deletion errors. We refer to such a channel as a sticky channel if it maintains the structure of blocks: $k' > 0$.

a) Insertion-Deletion Channels: In ideal communication systems, one often assumes that senders and receivers are perfectly synchronized – i.e., each sent bit is read by the receiver. However, in real systems, such perfect synchronization is often not possible. This leads to sent bits missed by the receiver (a deletion error), or read more than once (an insertion error). Such communication systems are traditionally modeled as insertion-deletion


Input:
TTT
AAAA
C
TTTT
.... (sequence of blocks of identical bases)
AAA...A
k identical bases
Nanopore Sequencer
Reconstruction Techniques
AAA...A
k' identical bases
Insertion error when k' > k
No error when k' = k
Deletion error when k' < k
AAA...A
Block of k identical characters
Insertion−Deletion Channel
AAA...A
Block of k' identical characters

An insertion−deletion channel transforms each block of k identical characters to an output block of k' identical characters. This transformation of k to k' is drawn from a distribution that is derived from machine characteristics.

Figure 1: Overview of the proposed channel and its correspondence with a sequencer.

channels. Formally, an independent insertion channel is one in which a single bit transmission is accompanied with the insertion of a random bit with a probability p. An independent deletion channel is one in which a transmitted bit can be deleted (omitted from the output stream) with a probability p'. An insertion-deletion channel contains both insertions and deletions [7]. Note that a number of basic characteristics of insertion-deletion channels, such as their capacity, are as yet unknown in information theory literature as well.

We consider a variant of the independent insertion-deletion model that is more general and better suited to nanopore sequencers. In particular, we recognize the primary source of error in nanopore sequencers is associated with disambiguating the exact number of identical bases in a block passing through the nanopore. We modify the independent insertion-deletion channel to the sticky insertion-deletion channel described above (Figure 1). The key difference is that, whereas the independent insertion-deletion channel operates on individual symbols independently, our model operates on *block lengths* independently.

Naturally, one expects that determining the capacity of this channel should be more difficult than doing the same for the independent insertion-deletion channel (as stated above, this is an open problem). Nonetheless, we are able to provide precise answers to our questions because our goal of analysis of the number of samples necessary for exact recovery is equivalent to analysis of the performance of a particular repetition code (chosen for us by the sequencing technology) for our channel, an easier problem than determining the capacity.

b) Approach to Bounding Sample Complexity: To obtain a bound on the number of samples needed for exact recovery, we must, for a given number of samples, prove a lower bound on the probability of error for *any* estimator of the sequence. That is, we must find the largest number of samples r for which we can prove that, for any sequence estimator that uses r samples from the channel, the probability that the estimator is not equal to the input sequence is asymptotically positive as the sequence length tends to ∞.

To do this, we can lower bound the probability of error uniformly over any estimator via Fano's inequality, which relates the error probability to the conditional entropy of the input sequence, given r output sequences. Then bounds on this quantity yield corresponding sample complexity bounds. The intuition here is that the conditional entropy measures the average amount of uncertainty about the true input sequence if we observe the output samples, so one naturally expects that one needs enough samples to cause this measure of uncertainty to tend to 0 in order to recover the input sequence. The main challenge in lower bounding the sample complexity is to formalize this intuition and to give a tight lower bound on the conditional entropy in as general a setting as possible.

To *upper* bound the sample complexity, we could again tie it to the conditional entropy described above, but this has the disadvantage that it does not immediately imply an algorithmically efficient method for exact recovery. We instead prove an effective upper bound by exhibiting an estimator for the input sequence and upper bounding the number of samples necessary to ensure correctness of this estimator with high probability.

III. Methods

A. Notation and Theoretical Model

The input sequence to the channel/sequencer is drawn from an alphabet $\mathcal{A}$ of fixed, finite size $|\mathcal{A}|$. For DNA sequencing, for example, $\mathcal{A} = \{A, T, C, G\}$.

In the next subsections, we describe the source and noise models.

1) Source Model: We consider an independent, identically distributed sequence (i.e., a memoryless source) of symbols $X_1, X_2, \ldots$, where each X_i takes the value $\alpha \in \mathcal{A}$ with probability denoted by p_α. From this sequence, we take, in particular, the prefix X consisting of the first $N \in \mathbb{N}$ *blocks*: a block is a maximal contiguous substring consisting of repetitions of any given symbol. We denote by $B_i(X)$ and $S_i(X)$, $i \in \mathbb{N}$, the length of the ith block of X and the associated symbol, respectively. Note that $B_i(X)$ is supported on the set of natural numbers $\mathbb{N}$. We denote by $\vec{B}(X)$ the sequence of N block lengths of X. For instance, in the sequence X = "AAACTTCTG", we have $B_1(X) = 3$ and $S_1(X)$ = "A".

An alternate model takes the first n symbols of the source, instead of the first N blocks. The results remain qualitatively the same in either model, as, with high probability, $n = \Theta(N)$ as $N \to \infty$. Moreover, we make our choice because doing so simplifies the presentation, somewhat.

We highlight two features of the distributions of block sizes in our model: first, for any block index j, the jth block size of the input, when conditioned on the block symbol being α, is geometrically distributed with parameter $(1 - p_\alpha)$ (this is *not* an assumption of our model; rather, it is a simple consequence of the assumption that the input sequence is memoryless); that is,

$$\Pr[B_j(X) = k | S_j(X) = \alpha] = p_\alpha^k (1 - p_\alpha).$$

Since two consecutive blocks cannot correspond to the same symbol, there is a Markov dependency between blocks, both between consecutive block symbols and between sizes: for any $\alpha, \beta \in \mathcal{A}$ with $p_\alpha, p_\beta > 0$, $\alpha \neq \beta$, and any $j > 0$,

$$\Pr[S_{j+1}(X) = \beta | S_j(X) = \alpha] = \frac{p_\beta}{1 - p_\alpha}.$$

Throughout, we assume that $0 < p_\alpha < 1$ for all $\alpha \in \mathcal{A}$. This implies that the Markov chain formed by the block symbols is ergodic, which in turn implies that there exists a stationary distribution π on $\mathcal{A}$. We denote by π_α the probability that the stationary distribution assigns to $\alpha \in \mathcal{A}$.

2) Noise Model: Sticky Insertion-Deletion Channel: The noise model that we consider is the *sticky insertion-deletion channel*, which, given an input sequence, increases or decreases the length of each block of symbols according to some distribution, which is parameterized by the original block length. The term *sticky* comes from the fact that we insist that new blocks cannot be inserted, nor can blocks be deleted.

More precisely, the channel is defined by a sequence of *block transformation* distributions $q_1, q_2, \ldots$ (see (1)) supported on the *positive* integers (so that $q_{i,j}$ denotes the probability given to the integer j by the ith distribution).

The channel induced by these q_i operates on the blocks of X independently as follows: for a block in X of any symbol $\alpha \in \mathcal{A}$ of length i, the channel outputs a block in Y of the same symbol, of length distributed according to the distribution q_i. That is, for arbitrary $j \in [N] = \{1, \ldots, N\}$,

$$\Pr[B_j(Y) = k | B_j(X) = m] = q_{m,k}, \quad \sum_{k=1}^{\infty} q_{m,k} = 1. \tag{1}$$

The fact that the sum above starts at $k = 1$ is a symptom of the fact that blocks are neither created nor deleted.

In what follows, instead of a single sample from the channel with input X, symbol Y will denote a sequence of $r = r(N)$ independent samples, conditioned on X. We then index the jth block of the ith sample by (j, i); e.g., we denote the length of the jth block of the ith sample by $B_{j,i}(Y)$, for $j \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, r\}$. Then $\mathbf{B}(Y)$ denotes an $N \times r$ matrix-valued random variable whose (j, i)th entry is $B_{j,i}(Y)$. The vector of samples for a particular block j will be denoted by $\vec{B}_j(Y)$.

We are particularly interested in distributions q_i for which the expected output block lengths are a well behaved function of the input block lengths, and for which the deviation from this mean is sufficiently small.

B. Main Results: Upper and Lower Bounds on Accurate Reconstruction from Nanopore Sequencers

Our main results address the task of *exact recovery* of the input sequence X, given r samples of X corrupted by the sticky insertion-deletion channel (collectively called Y). In particular, we give bounds on the number of samples $r = r(N)$ for which there exists an estimator $F(Y)$ satisfying

$$\Pr[F(Y) \neq X] \xrightarrow{N \to \infty} 0$$

(note that r may tend to ∞ with N).

Our first result gives a fundamental **lower bound** on the number of samples needed by any estimator for exact recovery. We give a complete proof in Section III-C, but we sketch the intuition here (suppressing some details that are dealt with in the full proof). First, since a single sample reveals the symbols associated with the input blocks, the main challenge is in lower bounding the sample complexity for estimators of the true block sizes. For an arbitrary estimator, we express the total probability of error p_e in terms of a product of the error probabilities $\{p_{e_j}\}_{j=1}^N$ of each block (after conditioning under which the block error events are independent):

$$p_e = 1 - \prod_{j=1}^{N}(1 - p_{e_j}).$$

To lower bound the probability of error p_{e_j} for each block j, we condition on the true block size not being too large (i.e., less than or equal to some sufficiently large constant m_*), and then we apply Fano's inequality [5]:

$$p_{e_j} \geq \frac{H(B_j(X)|\vec{B}_j(Y), \mathcal{F}) - h(p_{e_j})}{\log(m_* - 1)},$$

where $\mathcal{F}$ denotes an appropriate σ-field (which captures the conditioning needed to deal with the infinite support of the block lengths), $\vec{B}_j(Y)$ denotes the r samples of the jth block size passed through the channel, and $h(x)$ denotes the binary entropy function: $h(x) = -x\log x - (1-x)\log(1-x)$. Convexity considerations allow us to ignore the binary entropy on the right-hand side.

It thus remains to lower bound the conditional entropy $H(B_j(X)|\vec{B}_j(Y), \mathcal{F})$. Our Proposition 1 below yields

$$H(B_j(X)|\vec{B}_j(Y), \mathcal{F}) \geq e^{-\Theta(r)},$$

uniformly in j. This implies that

$$\begin{aligned} &p_{e_j} \geq e^{-\Theta(r)} \\ &\implies p_e \geq 1 - (1 - e^{-\Theta(r)})^N = 1 - e^{-Ne^{-\Theta(r)}}. \end{aligned}$$

For some small enough positive constant C and a large enough positive constant $D > C$, we then have

$$p_e \geq \begin{cases} \Theta(1) & r < C\log N \\ o(1) & r > D\log N. \end{cases}$$

Note, in particular, that for a single sample ($r = 1$), the probability of error converges exponentially to 1 as the number of blocks increases.

This yields the following theorem.

Theorem 1 (Lower bound on the number of samples necessary for exact reconstruction)**.** *Let X denote a sequence generated by taking the first N runs of an infinite sequence of i.i.d. samples from the distribution $\{p_\alpha\}_{\alpha\in\mathcal{A}}$. Suppose, further, that*

- Nontrivial alphabet distribution: *for all $\alpha \in \mathcal{A}$, $p_\alpha \neq 0$.*
- Nontrivial block transformation distributions: *each block transformation distribution q_ℓ has positive, finite mean and variance.*
- Shared support for block transformation distributions: $\operatorname{supp}(q_\ell) = \operatorname{supp}(q_m)$ *for all $\ell, m \geq 1$. Here,* $\operatorname{supp}(\cdot)$ *denotes the support of a distribution (i.e., for a discrete distribution such as q_ℓ, the set of integers to which it assigns positive probability).*
- Exponential tails for the block transformation distributions: *the block transformation distributions have at least exponentially decaying tails: There are some fixed $\gamma \geq 0$ and $c > 0$ such that, for a random variable Z distributed according to q_m, for all $\tau > 0$,*

$$\Pr[|Z - \mathbb{E}[Z]| \geq \tau m^\gamma] \leq 2e^{-c\tau}.$$

In particular, this means that

$$\Pr[|B_{j,i}(Y) - B_j(X)| \geq \tau m^\gamma \Big| B_j(X) = m] \leq 2e^{-c\tau}. \tag{2}$$

Then the number of samples r needed to recover X exactly with high probability is at least $r = \Omega(\log N)$.

We remark that this lower bound may be tightened in many cases by a more careful analysis of the conditional entropy described above.

It is of practical interest to relax the shared support condition in Theorem 1 (in particular, all of our concrete examples in Section IV require an extension). Our next result does this, at the expense of a looser bound on the constant hidden in the $\Omega(\cdot)$.

Theorem 2. *In the setting of Theorem 1, suppose that we replace the shared support condition with the following weaker condition:*

- Overlapping support for block transformation distributions: *for all $\ell \geq 1$,* $\operatorname{supp}(q_\ell) \cap \operatorname{supp}(q_{\ell+1}) \neq \emptyset$.

Then the number of samples r needed to recover X exactly with high probability is at least $r = \Omega(\log N)$.

The proof is a slight modification of that of Theorem 1, so we relegate it to Appendix VII-B.

Our next main result gives an **upper bound** on the number of samples needed for exact recovery, in terms of the variance and tail behavior of the block transforming distributions. We do this by exhibiting a natural estimator for block sizes, then proving an upper bound on the error probability of this estimator.

In particular, under the hypothesis that the expected value of each distribution q_ℓ is ℓ (to be relaxed in Corollary 1), we choose as our estimator of the true block length $B_j(X)$, the closest integer to the sample average of the observed block lengths:

$$\hat{B}_j(Y) = \left[\!\!\left[\frac{1}{r}\sum_{i=1}^{r} B_{j,i}(Y) \right]\!\!\right], \quad \tilde{B}_j(Y) = \frac{1}{r}\sum_{i=1}^{r} B_{j,i}(Y). \tag{3}$$

Here, for any $x \in \mathbb{R}$, $[\![x]\!]$ denotes the closest integer to x.

Using a union bound over all blocks and the hypothesis that the distributions q_ℓ have at least exponentially decaying tails, we are able to upper bound the probability of error of this estimator in terms of a constant $\gamma \geq 0$ that is linearly increasing as a function of the polynomial growth rate of the variance of q_ℓ as $\ell \to \infty$:

$$\begin{aligned}\Pr[\hat{B}(Y) \neq \vec{B}(X)] &\leq \sum_{j=1}^{N} \Pr[\hat{B}_j(Y) \neq B_j(X)] \\ &\leq N \exp(-\Theta(r^{\frac{1}{2\gamma+1}})).\end{aligned}$$

Here, the second inequality follows from our Proposition 2. This upper bound on the total error probability decays to 0 whenever $r \geq C \log^{2\gamma+1} N$, for some large enough constant C.

Theorem 3 (Upper bound on the number of samples necessary for exact recovery)**.** *Let the block transformation distributions q_ℓ and the alphabet distribution $\{p_\alpha\}_{\alpha\in\mathcal{A}}$ satisfy the nontriviality and exponential tails properties as in Theorem 1. Suppose, also, that for each ℓ, the expected value μ_ℓ of the q_ℓ distribution is $\mu_\ell = \ell$.*

Then there exists a large enough constant $C > 0$ such that, for X generated by a memoryless source with N blocks, given $r \geq C \log^{2\gamma+1} N$ samples (where γ is as in Theorem 1, in particular satisfying (2)), the probability of error for the estimator $\hat{B}(Y)$ of X (see (3)) tends to 0 as $N \to \infty$:

$$p_e = \Pr[\hat{B}(Y) \neq \vec{B}(X)] \xrightarrow{N\to\infty} 0.$$

In particular, we note that if the variance of q_ℓ is $\Theta(1)$ as $\ell \to \infty$ (i.e., $\gamma = 0$), then combining Theorems 1 and 3 shows that $\Theta(\log N)$ samples are necessary and sufficient for exact recovery.

Since the estimator is efficiently computable and the required number of samples is polylogarithmic in N, where we recall that N is the number of blocks in the input, this is an indication of the feasibility of nanopore sequencing under a broad range of noise models.

For some noise models, the stipulation that the expected value of the q_ℓ distribution be ℓ is too restrictive. It is thus worthwhile to have the following extension, which generalizes the upper bound estimator to cases where the expected value under q_ℓ is a nicely behaved invertible function of ℓ:

Corollary 1. *In the setting of Theorem 3, suppose that there is an invertible function $F(\ell)$ for which the expected value of each distribution q_ℓ is given by $F(\ell)$, such that F^{-1} is D-Lipschitz for a constant $D > 0$: that is, for each ℓ and denoting by μ_ℓ the expected value of the distribution q_ℓ,*

$$\mu_\ell = F(\ell),$$

and, for all x, y in the domain of F^{-1},

$$|F^{-1}(x) - F^{-1}(y)| \leq D|x-y|.$$

Then there exists an estimator $\hat{B}'(Y)$ for which the conclusion of Theorem 3 holds: that is, there exists a large enough constant $C > 0$ such that, given $r \geq C \log^{2\gamma+1} N$ samples (where γ is as in Theorem 3), the probability of error for $\hat{B}'(Y)$ tends to 0 as $N \to \infty$:

$$p_e = \Pr[\hat{B}'(Y) \neq \vec{B}(X)] \xrightarrow{N\to\infty} 0.$$

Proof: We propose the following estimator, based on the empirical mean estimator:

$$\hat{B}'_j(Y) = \left[\!\!\left[F^{-1}(\tilde{B}_j(Y)) \right]\!\!\right],$$

where $\tilde{B}_j(Y)$ is as in (3). Then, by the analysis in the proof of Theorem 3, with high probability,

$$|\tilde{B}_j(Y) - \mu_{B_j(X)}| \leq 1/(3D).$$

Using the Lipschitz condition on F^{-1}, this implies

$$\begin{aligned}|F^{-1}(\tilde{B}_j(Y)) - B_j(X)| &\leq D|\tilde{B}_j(Y) - \mu_{B_j(X)}| \\ &\leq D/(3D) = 1/3.\end{aligned}$$

Applying the rounding shows that, with high probability,

$$\hat{B}'(Y) = \vec{B}(X).$$

This completes the proof. ■

a) Extensions: Several extensions to these results are easy modifications of our proofs. For example, the extension of both bounds to Markov sources is trivial, with similar results. One may also wish for approximate results: how many samples are needed to recover the original sequence within a certain error tolerance? The upper bound analysis readily adapts to this, while the lower bound requires more work. Finally, the model may be tweaked by making the block transformation distributions q_ℓ dependent on the symbol associated to the block. This also changes the analysis very little.

A more complicated extension may give a more precise upper bound on the necessary number of samples for exact recovery: namely, Theorem 11 of [10] shows that there exists an estimator (possibly difficult to compute) for X using Y whose probability of error is upper bounded by the conditional entropy $H(Y|X)$. Thus, a precise upper bound on the conditional entropy may yield a tighter (in terms of number of samples) version of Theorem 3.

b) Limitations: The main limitation of the present analysis is that we do not consider noise models that create or delete blocks, since we assumed that $q_{\ell,0} = 0$ for all ℓ. Though natural estimators suggest themselves for the upper bound, their analyses are quite complicated. Intuitively, the complication comes from the fact that a single observed sequence may arise from several different patterns of block insertions and deletions. Thus, we do not explore these variations in this paper. Note further that substitutions can be modeled as insertions followed by corresponding deletions (or vice-versa).

C. Proof of Theorem 1

Note, first, that only a single sample from the channel is needed to recover the sequence of block symbols in the input (since blocks cannot be created or erased in our model). For a lower bound on the necessary number of samples needed for exact recovery of the original sequence X, we thus need to lower bound the number of samples needed to recover the sequence of block sizes.

Consider any estimator $\hat{B}(Y)$ of $\vec{B}(X)$. We denote by p_e its probability of error:

$$p_e = \Pr[\hat{B}(Y) \neq \vec{B}(X)].$$

We want a lower bound on this. To do this, we consider the number E of blocks whose sizes the estimator determines incorrectly. This may be written as a sum of indicators:

$$E = \sum_{i=1}^{N} E_i,$$

where E_i is the indicator that the estimator is erroneous on block i. Then

$$p_e = \Pr[E > 0] = 1 - \Pr[E = 0].$$

Now, conditioning on the symbols of the blocks, the E_i are independent random variables (more precisely, we may without loss of generality consider an estimator for which this is true):

$$\Pr[E = 0] \tag{4}$$

$$= \sum_{s \in \mathcal{A}^N} \Pr[\vec{S}(X) = s] \prod_{j=1}^{N} (1 - \Pr[E_j = 1 | S_j(X) = s_j]) \tag{5}$$

It is thus sufficient to lower bound $\Pr[E_j = 1 | S_j(X) = \alpha]$, for arbitrary $\alpha \in \mathcal{A}$. To do this, we first condition on the true block size $B_j(X)$ not being too large: for some $m_* > 0$ that we will fix later,

$$\begin{aligned}
&\Pr[E_j = 1 | S_j(X) = \alpha] \\
&= \Pr[\hat{B}_j(Y) \neq B_j(X) | S_j(X) = \alpha] \\
&= \Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha] \\
&\quad \cdot \Pr[B_j(X) \leq m_* | S_j(X) = \alpha] \\
&\quad + \Pr[\hat{B}_j(Y) \neq B_j(X) \cap B_j(X) > m_* | S_j(X) = \alpha] \\
&\geq \Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha] \\
&\quad \Pr[B_j(X) \leq m_* | S_j(X) = \alpha].
\end{aligned}$$

Here, the inequality is by simply lower bounding $\Pr[\hat{B}_j(Y) \neq B_j(X) \cap B_j(X) > m_* | S_j(X) = \alpha]$ by 0. Since the support of $B_j(X)$ under this conditioning is finite, we may further lower bound by invoking Fano's inequality [5]:

$$\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha] \tag{6}$$

$$\geq \frac{H(B_j(X) | \vec{B}_j(Y), B_j(X) \leq m_*, S_j(X) = \alpha)}{\log(m_* - 1)} \tag{7}$$

$$- \frac{h(\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha])}{\log(m_* - 1)}. \tag{8}$$

Here, $h(x)$ denotes the binary entropy function: $h(x) = -x \log(x) - (1-x)\log(1-x)$ for $x \in [0,1]$. Note that we have replaced the estimator $\hat{B}_j(Y)$ by the samples $\vec{B}_j(Y)$. This is justified by the data processing inequality.

To upper bound the entropy $h(\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha])$ of the error indicator, we need the following lemma:

Lemma 1 (Upper bound on binary entropy)**.** *For any* $x \in [0,1]$,

$$h(x) \leq 2\sqrt{x}.$$

For brevity, we omit the proof, which is by elementary convexity considerations.

Applying Lemma 1, the inequality (8) thus becomes

$$\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha] \tag{9}$$

$$\geq \frac{H(B_j(X) | \vec{B}_j(Y), B_j(X) \leq m_*, S_j(X) = \alpha)}{\log(m_* - 1)} \tag{10}$$

$$- \frac{2\sqrt{\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha]}}{\log(m_* - 1)}. \tag{11}$$

1) Lower bounding the conditional entropy: It remains to show a lower bound on $H(B_j(X)|\vec{B}_j(Y), B_j(X) \leq m_*, S_j(X) = \alpha)$. The goal of this section is to show the following:

Proposition 1 (Lower bound on conditional entropy)**.** *We have, for any $j \in \{1, ..., \mathbb{N}\}$, and for $m_* = \Theta(1)$ and any $\alpha \in \mathcal{A}$,*

$$H(B_j(X)|\vec{B}_j(Y), B_j(X) \leq m_*, S_j(X) = \alpha) \geq e^{-\Theta(r)}.$$

a) Useful notation, types: To proceed, we define some useful notation: for $\ell \in \mathbb{N}$,

$$P_{\alpha,\ell} = \Pr[B_j(X) = \ell | S_j(X) = \alpha, B_j(X) \leq m_*].$$

For $\vec{k}$ a vector of non-negative integers of length r, we denote by $q_{\ell,\vec{k}}$ the probability that r independent samples of q_ℓ yield $\vec{k}$:

$$q_{\ell,\vec{k}} = \prod_{i=1}^{r} q_{\ell,k_i}. \tag{12}$$

Observe that any permutation of the entries of $\vec{k}$ is given the same probability. This motivates the following definition [5]: the *type* of an r-dimensional vector $\vec{k}$ of non-negative integers is the set of all vectors formed by permuting the entries of $\vec{k}$. Associated bijectively with any type is an infinite-length sequence $T = T(\vec{k})$ whose entries are defined as follows: for $i \in \mathbb{N}$,

$$T_i = |\{j \in \{1, ..., r\} \ : \ k_j = i\}|.$$

That is, the ith entry of T is the number of times i appears in $\vec{k}$. Note that the sequence so defined is the same for any element of the corresponding type, so it is well-defined. In what follows, we abuse notation and refer to these sequences as types. We further define the support $\mathrm{supp}(T)$ as the set of i for which $T_i \neq 0$. We denote the set of all possible types of r-sample vectors by $\mathbb{T}$. Finally, we denote by $q_{\ell,T}$ the probability assigned by q_ℓ to any representative element of T:

$$q_{\ell,T} = \prod_{i \in \mathrm{supp}(T)} q_{\ell,i}^{T_i}. \tag{13}$$

b) Proof of Proposition 1: By an elementary derivation, it may be seen that the conditional entropy to be lower bounded has the following equivalent expression:

$$\begin{aligned} &H(B_j(X)|\vec{B}_j(Y), B_j(X) \leq m_*, S_j(X) = \alpha) \\ &= \sum_{\ell=1}^{\infty} P_{\alpha,\ell} \sum_{\vec{k} \geq \vec{1}} q_{\ell,\vec{k}} \log\left(1 + \sum_{m \neq \ell} \frac{P_{\alpha,m} q_{m,\vec{k}}}{P_{\alpha,\ell} q_{\ell,\vec{k}}}\right). \end{aligned}$$

The existence of this conditional entropy is guaranteed by the exponential tail assumption in the theorem.

Now, we observe that all vectors $\vec{k}$ in a given type contribute the same amount in the inner sum. We thus collect them together and replace the inner sum with a sum over all possible types for r-vectors of samples:

$$\begin{aligned} &H(B_j(X)|\vec{B}_j(Y), B_j(X) \leq m_*, S_j(X) = \alpha) \\ &= \sum_{\ell=1}^{\infty} P_{\alpha,\ell} \sum_{T \in \mathbb{T}} \binom{r}{T} q_{\ell,T} \log\left(1 + \sum_{m \neq \ell} \frac{P_{\alpha,m} q_{m,T}}{P_{\alpha,\ell} q_{\ell,T}}\right), \end{aligned} \tag{14}$$

where $\binom{r}{T}$ is the multinomial coefficient:

$$\binom{r}{T} = \binom{r}{T_1, T_2, ...}$$

(note that only finitely many elements of the sequence T are nonzero).

To lower bound this sum, we start by approximating $\binom{r}{T} q_{\ell,T}$ using Stirling's formula [1]. We have

$$\begin{aligned} q_{\ell,T}\binom{r}{T} &= \exp\Big(\sum_{t \in \mathrm{supp}(T)} T_t \log q_{\ell,T_t} + r \log r \\ &\quad - r + \frac{1}{2}\log r + \frac{1}{2}\log(2\pi) + O(1/r) \\ &\quad - \sum_{t \in \mathrm{supp}(T)} (T_t \log T_t - T_t \\ &\quad + \frac{1}{2}\log T_t + \frac{1}{2}\log(2\pi) + O(1/T_t)\Big) \\ &= \exp\Big(-r \sum_{t \in \mathrm{supp}(T)} \frac{T_t}{r} \log \frac{T_t/r}{q_{\ell,T_t}} \\ &\quad + \frac{1}{2}\log \frac{r}{\prod_j T_j} - \frac{|\mathrm{supp}(T)|}{2}\log(2\pi) \\ &\quad + O\Big(\sum_{t \in \mathrm{supp}(T)} \frac{1}{T_t}\Big) + O(1)\Big) \\ &= \exp\Big(-rD(\{T_t/r\}_t \| \{q_\ell\}) + \frac{1}{2}\log\frac{r}{\prod_j T_j} \\ &\quad - \frac{|\mathrm{supp}(T)|}{2}\log(2\pi) + O\Big(\sum_{t \in \mathrm{supp}(T)} \frac{1}{T_t}\Big) + O(1)\Big) \end{aligned}$$

If we restrict our attention to types T with bounded support (i.e., $|\mathrm{supp}(T)| = \Theta(1)$), which yields a lower bound because all terms of (14) are non-negative, then this becomes

$$q_{\ell,T}\binom{r}{T} = e^{-rD(\{T_t/r\}_t\|\{q_\ell\})+O(\log r)}. \tag{15}$$

Here, we note that the entries of each type sum to r (since there are r samples, so $\{T_t/r\}_t$ forms a distribution). Furthermore, $D(\{T_t/r\}_t\|\{q_\ell\})$ denotes the KL-divergence of the two distributions. Terms for which this KL-divergence are smaller contribute more, so we further lower bound by restricting to only those types for which it is small (in a sense to be made precise below).

Thus, we lower bound (14) by restricting as follows:

- In the outer sum, we restrict to $\ell \le \ell_*$, for some fixed positive ℓ_*.
- In the inner sum, we restrict to $T \in \mathbb{T}(\epsilon, t_*)$, where we define $\mathbb{T}(\epsilon, t_*)$, for some small $\epsilon = \epsilon(r)$ and arbitrary fixed positive t_*, to be the set of types $T \in \mathbb{T}$ with $|\mathrm{supp}(T)| \le t_*$ and $D(\{T_t/r\}_t\|\{q_\ell\}) \le \epsilon$. In order for the resulting bound to be nontrivial, we must show the following:

  Claim 1. *The set $\mathbb{T}(\epsilon, t_*)$ is nonempty for any given ϵ, provided that t_* is sufficiently large.*

  This is a consequence of the strong law of large numbers and the fact that we have restricted to $\ell \le \ell_* = O(1)$. We prove it in Appendix VII-A.
- In the innermost sum, inside the logarithm, we restrict m to a single term $m = \ell + 1$ for which $q_{m,T} \neq 0$ (this is nonzero because of the assumption that $\mathrm{supp}(q_\ell) = \mathrm{supp}(q_{\ell+1})$ in the statement of Theorem 1).

We now estimate the expression inside the logarithm in the lower bound:

$$\frac{P_{\alpha,\ell+1}q_{\ell+1,T}}{P_{\alpha,\ell}q_{\ell,T}} = P_{\alpha,\ell+1}/P_{\alpha,\ell} \cdot \exp(-r\sum_{i=1}^{\infty}\frac{T_i}{r}\log\frac{q_{\ell,i}}{q_{\ell+1,i}}), \tag{16}$$

where we recall the definition of $q_{\ell,T}$ given by (13). Now, using the definition of $\mathbb{T}(\epsilon, t_*)$, we may approximate T_i/r by $q_{\ell,i}$ as follows: we know that

$$D(\{T_t/r\}\|q_\ell) < \epsilon.$$

To convert this to a bound on the L_1 distance between $\{T_i/r\}$ and q_ℓ, we apply Pinsker's inequality [5]:

Lemma 2 (Pinsker's inequality)**.** *For two distributions P and Q on the same probability space,*

$$\|P - Q\|_{TV} \le \sqrt{2 \cdot D(P\|Q)}$$

Thus, we have

$$T_i/r = (T_i/r - q_{\ell,i}) + q_{\ell,i} = q_{\ell,i}(1 + O(\epsilon(r))/q_{\ell,i}).$$

Now, since ℓ is bounded, and since $T_i \neq 0 \implies q_{\ell,i} \neq 0$, we have that $q_{\ell,i} = \Theta(1)$, so that we have the asymptotic equivalence

$$T_i/r = q_{\ell,i}(1 + O(\epsilon(r))).$$

(Note that the O is not uniform in ℓ.)

Plugging this approximation into (16), we have

$$\frac{P_{\alpha,\ell+1}q_{\ell+1,T}}{P_{\alpha,\ell}q_{\ell,T}} \\ = P_{\alpha,\ell+1}/P_{\alpha,\ell} \cdot \exp(-rD(q_\ell\|q_{\ell+1})(1 + O(\epsilon(r)))).$$

Plugging this into the logarithm (14), we get

$$\log\left(1 + \sum_{m\neq\ell}\frac{P_{\alpha,m}q_{m,T}}{P_{\alpha,\ell}q_{\ell,T}}\right) \ge \log(1 + \frac{P_{\alpha,m}q_{m,T}}{P_{\alpha,\ell}q_{\ell,T}}) \\ \sim \frac{P_{\alpha,\ell+1}}{P_{\alpha,\ell}} \cdot \exp(-rD(q_\ell\|q_{\ell+1})(1 + O(\epsilon(r)))), \tag{17}$$

provided that $d_{min} = \Omega(1)$, where we define d_{min} to be

$$d_{min} = \min_{\ell,m\in\{1,\ldots,\ell_*\}} D(q_\ell\|q_m).$$

We may, in any case, lower bound further by upper bounding the KL-divergence in the exponent of (17) by some constant. Thus, we have shown that the logarithm in (14) is lower bounded by

$$\log\left(1 + \sum_{m\neq\ell}\frac{P_{\alpha,m}q_{m,T}}{P_{\alpha,\ell}q_{\ell,T}}\right) \ge e^{-\Theta(r)}, \tag{18}$$

after noting that $\frac{P_{\alpha,\ell+1}}{P_{\alpha,\ell}} = \Theta(1)$ if we choose $\ell_* < m_*$.

Combining (18) with (15), we get as a lower bound for (14)

$$H(B_j(X)|\vec{B}_j(Y), B_j(X) \le m_*, S_j(X) = \alpha) \\ \ge \exp(-r(\epsilon + \Theta(1)) + O(\log r)) = \exp(-\Theta(r)).$$

This completes the proof of Proposition 1.

□

2) Finishing the proof of Theorem 1: By Proposition 1, we may further lower bound (11), which yields

$$\Pr[\hat{B}_j(Y) \neq B_j(X)|B_j(X) \le m_*, S_j(X) = \alpha] \\ \ge \frac{e^{-\Theta(r)}}{\log(m_* - 1)} \\ - \frac{2\sqrt{\Pr[\hat{B}_j(Y) \neq B_j(X)|B_j(X) \le m_*, S_j(X) = \alpha]}}{\log(m_* - 1)}.$$

This in particular implies that

$$\begin{aligned}&\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) \leq m_*, S_j(X) = \alpha] \\ &\geq e^{-\Theta(r)}.\end{aligned}$$

Plugging this into (5), we get that

$$\begin{aligned}\Pr[E = 0] &\leq \sum_{s \in \mathcal{A}^N} \Pr[\vec{S}(X) = s] \prod_{j=1}^{N} (1 - e^{-\Theta(r)}) \\ &= \sum_{s \in \mathcal{A}^N} \Pr[\vec{S}(X) = s] (1 - e^{-\Theta(r)})^N \\ &= (1 - e^{-\Theta(r)})^N \\ &\implies p_e \geq 1 - (1 - e^{-\Theta(r)})^N.\end{aligned}$$

Finally, using the fact that

$$(1 - e^{-\Theta(r)})^N = e^{-N e^{-\Theta(r)} (1 + o(1))},$$

we see that this lower bound is $\Theta(1)$ as long as $r < C \log N$, for some small enough positive constant C. This completes the proof of Theorem 1.

□

D. Proof of Theorem 3

For an upper bound on the number of samples needed for exact recovery of the original sequence X from r samples Y, it is sufficient to exhibit an estimator and to upper bound its error probability. Since a single sample reveals all symbols with probability 1, the challenge lies in determining the N original block sizes. We propose the estimator $\hat{B}_j(Y)$ as in (3) for the jth block size, for $j \in \{1, ..., N\}$.

Next, we analyze the error probability of this estimator. By the union bound, we can upper bound the probability of error p_e as a sum over all blocks:

$$\Pr[\hat{B}(Y) \neq \vec{B}(X)] \leq \sum_{j=1}^{N} \Pr[\hat{B}_j(Y) \neq B_j(X)] \quad (19)$$

$$= \sum_{j=1}^{N} \sum_{\alpha \in \mathcal{A}} \Pr[S_j(X) = \alpha] \quad (20)$$

$$\cdot \Pr[\hat{B}_j(Y) \neq B_j(X) | S_j(X) = \alpha] \quad (21)$$

$$= \sum_{j=1}^{N} \sum_{\alpha \in \mathcal{A}} \Pr[S_j(X) = \alpha] \quad (22)$$

$$\sum_{m=2}^{\infty} \Pr[B_j(X) = m | S_j(X) = \alpha] \quad (23)$$

$$\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) = m]. \quad (24)$$

Here, the first equality is by conditioning on the symbol of the jth block of X, and the second equality is by conditioning on the input block size being m and by the fact that $\hat{B}_j(Y)$ is conditionally independent of $S_j(X)$ given the input block size.

It remains to analyze the probability $\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) = m]$. By definition of the closest integer function, we have

$$\begin{aligned}&\Pr[\hat{B}_j(Y) \neq B_j(X) | B_j(X) = m] \\ &= \Pr\left[|\tilde{B}_j(Y) - B_j(X)| \geq 1/2 \,\middle|\, B_j(X) = m\right].\end{aligned}$$

Now, note that $B_{j,i}(Y)$, for $i = 1, ..., r$, under this conditioning, is distributed according to the measure q_m and has mean m. Furthermore, all terms are independent.

At this point in the proof, we use the tail bound assumed in the hypothesis. In particular, this implies [22] that the random variables $B_{j,i}(Y)$ appearing in the definition of the estimator are *weakly sub-Gaussian*, defined as follows:

Definition 1 (Weakly sub-Gaussian)**.** *A random variable M is said to be* weakly sub-Gaussian *(also called* sub-Exponential*) with parameter λ if its centered version $\hat{M} = M - \mathbb{E}[M]$ satisfies*

$$\mathbb{E}[e^{s\hat{M}}] \leq e^{s^2 \lambda^2 / 2}$$

for all $|s| \leq 1/\lambda$. We then write

$$M \sim \mathrm{WeaklySubGaussian}(\lambda).$$

Note that this differs from the definition of a sub-Gaussian random variable in that the inequality on the MGF is not required to hold for all $s \in \mathbb{R}$. Thus, the class of weakly sub-Gaussian random variables includes as a proper subset sub-Gaussian random variables, as well as random variables with only exponential tails.

Sums of i.i.d. weakly sub-Gaussian random variables satisfy the following concentration result [22]:

Lemma 3 (Bernstein's inequality)**.** *Let $M_1, \ldots, M_r$ be independent random variables such that each $M_i \sim \mathrm{WeaklySubGaussian}(\lambda)$ and $\mathbb{E}[M_i] = 0$. Define*

$$\bar{M} = \frac{1}{r} \sum_{i=1}^{r} M_i$$

(i.e., the empirical mean of the random variables M_i). Then for any $t > 0$, we have

$$\Pr[\bar{M} < t] \vee \Pr[\bar{M} > t] \leq \exp\left[-\frac{r}{2}\left(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda}\right)\right].$$

Here, $a \vee b$ denotes $\max\{a, b\}$, and $a \wedge b$ denotes $\min\{a, b\}$.

Applying this with $M_i = B_{j,i}(Y) - \mathbb{E}[B_{j,i}(Y)]$ and $\lambda = \text{const} \cdot m^{\gamma}$, we get that

$$\Pr[|\tilde{B}_j(Y) - \mathbb{E}[\tilde{B}_j(Y)]| \geq 1/2] \leq 2e^{-\Theta(r/m^{2\gamma})}.$$

Plugging this into (24) after recalling that we conditioned on $B_j(X) = m$, this results in

$$\begin{aligned}
&\Pr[\hat{B}(Y) \neq \vec{B}(X)] \\
&\leq 2\sum_{j=1}^{N}\sum_{\alpha\in\mathcal{A}} \Pr[S_j(X) = \alpha] \\
&\qquad \sum_{m=2}^{\infty} \Pr[B_j(X) = m | S_j(X) = \alpha] e^{-\Theta(r/m^{2\gamma})} \\
&= 2\sum_{j=1}^{N}\sum_{\alpha\in\mathcal{A}} \Pr[S_j(X) = \alpha] \sum_{m=2}^{\infty} e^{-\Theta(m)-\Theta(r/m^{2\gamma})}.
\end{aligned}$$

Here, the equality comes from the fact that $\Pr[B_j(X) = m|S_j(X) = \alpha]$ is geometrically distributed with a parameter that is $\Theta(1)$. To evaluate the m sum, we find the largest term by taking the derivative of the exponent with respect to m, setting it equal to 0, and solving. This shows that the largest term comes from

$$m_* = \begin{cases} 2 & \gamma = 0 \\ \Theta(r^{\frac{1}{2\gamma+1}}) & \gamma > 0 \end{cases}$$

and the maximum contribution is $e^{-\Theta(r^{\frac{1}{2\gamma+1}})}$ in either case. Thus, it is easy to see that the m sum is given by

$$\sum_{m=2}^{\infty} \exp(-\Theta(m) - \Theta(r/m^{2\gamma})) = e^{-\Theta(r^{\frac{1}{2\gamma+1}})}.$$

Since the α sum is over a constant number of indices, and since $\Pr[S_j(X) = \alpha] = \Theta(1)$ for all $\alpha \in \mathcal{A}$, we have thus shown the following proposition:

Proposition 2 (Upper bound on block errors)**.** *We have, uniformly for $j \in \{1, ..., N\}$, for the estimator (3),*

$$\Pr[\hat{B}_j(Y) \neq B_j(X)] \leq \exp(-\Theta(r^{\frac{1}{2\gamma+1}})). \tag{25}$$

This implies

$$\Pr[\hat{B}(Y) \neq \vec{B}(X)] \leq \sum_{j=1}^{N} e^{-\Theta(r^{\frac{1}{2\gamma+1}})}.$$

Finally, note that the Θ in the exponent above is uniform in j, which yields the upper bound

$$\Pr[\hat{B}(Y) \neq \vec{B}(X)] \leq N e^{-\Theta(r^{\frac{1}{2\gamma+1}})}.$$

In order for this upper bound to tend to 0 as $N \to \infty$ (so that the estimator is equal to the input sequence), it is sufficient to have a number of samples r satisfying

$$r \geq C \log^{2\gamma+1} N,$$

for a large enough constant C. This completes the proof of the theorem. □

IV. Concrete Examples and Empirical Results

In this section, we illustrate our upper bound results empirically. For two different block transformation distributions, we pass a random sequence X through the corresponding channels r times and use the estimators guaranteed by Theorem 3 and Corollary 1 to calculate an estimate $\hat{Y}$ of X. We plot the resulting number of block errors (averaged over 100 trials of the above experiment) versus the number of samples.

For both examples, X has $N = 10000$ blocks, $r \leq 200$, and the alphabet distribution is given by

$$p_\alpha = 1/4, \qquad \alpha \in \{1, 2, 3, 4\}.$$

A. Bounded Variance Distribution

The first distribution that we consider is defined as follows: fixing any $q \in (0, 1)$,

$$q_{\ell,k} = q^{k-\ell}(1-q),$$

for any $k \geq \ell \geq 1$. After some easy calculation, we find that the expected value of this distribution is given by

$$\mu_\ell = \ell + \frac{q}{1-q}.$$

Moreover, the variance is $\Theta(1)$ (so that the exponential decay condition in Theorem 3 is satisfied with $\gamma = 0$).

Although only insertions are allowed under this distribution, it is easy to tweak it to support deletions, and our theoretical results would still apply.

Thus, Corollary 1 suggests an estimator:

$$\hat{B}'_j(Y) = \left[\!\left[\tilde{B}_j(Y) - \frac{q}{1-q} \right]\!\right].$$

The plot of the performance of this estimator is given in Figure 2. The number of block errors, for a single sample, is of the same order as the number of blocks and decreases exponentially as a function of the number of samples; thus, logarithmically many samples (as a function of N) are required for exact recovery. This is empirically in agreement with the guarantees of Corollary 1 and Theorem 2.

B. Independent Insertion-Deletion Distribution

To illustrate the robustness of our analysis, we consider another distribution with (asymptotically) higher variance: namely, what we call the *independent insertion-deletion* distribution. It captures a transformation model in which, for each block, the first symbol

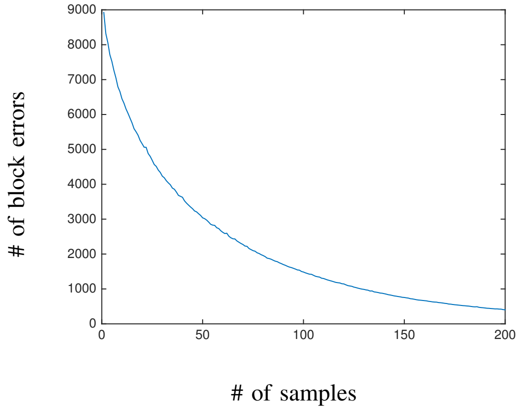


Figure 2: Number of samples versus number of block errors for the bounded variance model.

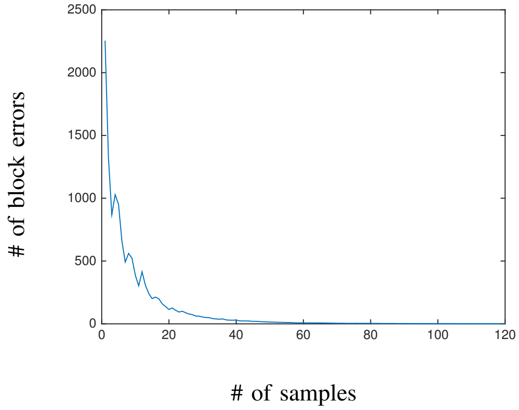


Figure 3: Number of samples versus number of block errors for the independent insertion-deletion model.

passes through untouched, while each subsequent symbol is deleted or duplicated with probability $1/2$ each. That is, for any block $j \leq N$ and sample $i \leq r$,

$$B_{j,i}(Y) = 1 + 2 \cdot \text{Binomial}(B_j(X) - 1, 1/2).$$

The expected value of the q_ℓ distribution is then

$$\mu_\ell = \ell,$$

while the variance is

$$\sigma_\ell^2 = (\ell - 1) = \Theta(\ell).$$

That is, Theorem 3 applies with $\gamma = 1$, guaranteeing an upper bound of $\Theta(\log^3 N)$ samples. Since $\mu_\ell = \ell$, the estimator prescribed by Theorem 3 suffices. By the lower bound theorem, we see that at least $\Theta(\log N)$ samples are needed for exact recovery, as in the previous example.

The plot of the performance of this estimator is given in Figure 3.

V. Related Research

Technologies underlying nanopore sequencers have been investigated for over a decade [6], [2]. Commercial platforms based on these technologies have only recently been announced – with Oxford Nano being the leading platform. An excellent introduction to this platform is available at: https://www.nanoporetech.com/technology/analytes-and-applications-dna-rna-proteins/dna-an-introduction-to-nanopore-sequencing. There have been preliminary efforts aimed at characterizing the performance of nanopore sequencing platforms in terms of error rate, error classification, and run lengths [16], [19], [9], [23].

Churchill and Waterman [4] address similar questions to ours for substitution channels. A key differentiating aspect of their work is that they reconstruct a sequence of read values for a given position as a consensus; i.e., the most frequent value across replicates. In contrast, significantly different estimators are needed for reconstruction in our model. Beyond this, a major difference between our results and those of Churchill and Waterman is that they provide bounds for a specific estimation technique, whereas our bounds are fundamental – they hold irrespective of the chosen estimator. Moreover, for certain choices of model parameters, our lower bounds contain Kullback-Leibler divergences between different block transformation distributions corresponding to different input block lengths, which quantitatively describes the difficulty of distinguishing them.

A historically important paper related to limits of sequencing technologies by Lander and Waterman [14] poses statistical questions about the performance of shotgun sequencing and heuristically answers them. In particular, they consider the following model: fix a DNA string G of length n, which we would like to recover. Samples from this string take the form of substrings of a given fixed small size, taken uniformly at random from G. That is, the randomness in this model is in the *location* of the sampled substrings, and there are no substitutions, insertions, or deletions. The authors define the *coverage depth* of a set of samples as the average number, over each position i in G, of samples that include i. They then relate the expected coverage depth to the expected number of "contigs" (i.e., contiguous substrings of G assembled from overlapping samples), in particular showing that the latter decays exponentially as the expected coverage depth increases. They use this to conclude that, in order to recover G, the expected coverage depth must be at least logarithmic in n. This result differs from ours in several key ways: first, the noise in our model takes the form of erroneous reads of the entire sequence, whereas theirs is in the form of noiseless reads of small portions of the sequence. Second, the logarithmic lower bound in their case is in terms of expected coverage depth, instead of samples. Finally, we give an algorithmically efficient method for reconstruction in our model, whereas they do not.

A more recent paper, by Motahari et al. [18], further (and rigorously) studies DNA shotgun sequencing. In particular, the authors give bounds on the number of reads (which take the same form as in [14]) necessary to reconstruct a given string, this time modeled probabilistically. Moreover, they consider the case of independent substitution errors, and they give algorithms for reconstruction. The major contrast with our work is again that the noise and read models are quite different, leading to different challenges in analysis.

Error characteristics and models for nanopore sequencers have been recently studied by O'Donnell et al [19]. In this study, the authors investigate error characteristics, and build a statistical model for errors. They use this model to show, through a simulation study, that replicated extrusion can be used to improve error characteristics. In particular, they show that using their model, it is possible to achieve 99.99% accuracy by replicating the read 140 times. This empirical study provides excellent context for our analytical study, which provides rigorous bounds and required replication rates.

There have been a number of efforts aimed at analyzing the error characteristics of current generation of sequencing technologies, including the 454 and PacBio sequencers [13], [3], [20]. These efforts are primarily aimed at the problem of alignment of relatively short fragments - of localizing their positions in a longer sequence. In contrast, due to longer reads obtained by nanopore sequencing, in this paper, we do not deal with the issue of localizing positions of the fragments and focus on the problem of correcting indel errors. These considerations can be also used for improving accuracy from multiple (aligned) reads of any sequencing technique. Our approach takes an information theoretic view to the problem. In doing so, we are able to establish fundamental bounds on the performance envelope of the modeled nanopore sequencer. To the best of our knowledge, this paper represents the first information theoretic formulation of its kind.

There has been significant work on different channels, their capacities, and error characteristics since the work of Shannon. Of particular relevance to our results is the work in deletion channels [12] (see also [17] for a survey). As mentioned, the capacity of independent insertion/deletion channels is as yet unknown, though lower bounds have been proven in the form of explicit distributional constructions and coding schemes [24]. There have been efforts aimed at error correction in insertion-deletion channels in the context of communication, storage, and RFID systems [26]. A variant of independent insertion-deletion channels appears in the literature under the name "sticky channels", with the key contrast with our model being that individual symbols are replicated (potentially several times) or deleted independently, whereas our sticky channel operates at the block level. Moreover, the emphasis in the information theory literature has been on determining bounds on the capacity, whereas we have the qualitatively different task of bounding the sample complexity.

Finally, we mention that in practical applications, if one wishes to numerically compute estimates of the bounds on the number of samples necessary for exact recovery, one must at least know estimates of the distribution on the alphabet $\{p_\alpha\}_{\alpha\in\mathcal{A}}$. Since the channel is sticky, it is not utterly obvious how to do this. The paper [8] gives heuristics for this problem in the per-symbol sticky channel model.

VI. Discussion and Conclusion

In this paper, we present a novel modeling methodology based on the abstraction of a nanopore sequencer as an information theoretic channel. We use our methodology to show a number of important results: (i) the indel error rate of the nanopore sequencer limits the sequence length that can be accurately reconstructed from a single sample; (ii) replicated extrusion through the nanopore is an effective technique for increasing the accurate reconstruction length; (iii) the necessary number of replicas is a slowly growing function of the sequence length (polylogarithmic in sequence length), enabling nanopore sequencers to accurately reconstruct long sequences. We demonstrate our results for a wide class of error models and show that our analyses are robust.

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] T. Butler, M. Pavlenok, I. Derrington, M. Niederweis, and J. Gundlach. Single-molecule dna detection with an engineered mspa protein nanopore. *Proceedings of the National Academy of Science*, 105(52):20647–20652, 2008.

[3] Mauricio O Carneiro, Carsten Russ, Michael G Ross, Stacey B Gabriel, Chad Nusbaum, and Mark A DePristo. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics*, 13(1):375, 2012.

[4] Gary A Churchill and Michael S Waterman. The accuracy of dna sequences: estimating sequence quality. *Genomics*, 14(1):89–98, 1992.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[6] D. Deamer and D. Branton. Characterization of nucleic acids by nanopore analysis. *Acc Chem Res*, 35(10):817–825, 2002.

[7] E. Drinea and M. Mitzenmacher. Improved lower bounds for the capacity of i.i.d. deletion and duplication channels. *IEEE Transactions on Information Theory*, 53:8:2693–2714, 2007.

[8] Farzad Farnoud, Olgica Milenkovic, and Narayana Prasad Santhanam. Small-sample distribution over sticky channels. *Proceedings of the International Symposium on Information Theory*, pages 1125–1129, 2009.

[9] E. Hayden. Nanopore genome sequencer makes its debut. *Nature News*, Feb. 2012.

[10] Siu-Wai Ho and Sergio Verdú. On the interplay between conditional entropy and error probability. *IEEE Trans. on Inf. Theory*, 56(12), 2010.

[11] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the minion nanopore sequencer. *Nature methods*, 2015.

[12] Ian A. Kash, Michael Mitzenmacher, Justin Thaler, and Jonathan Ullman. On the zero-error capacity threshold for deletion channels. *CoRR*, abs/1102.0040, 2011.

[13] Sergey Koren, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693–700, 2012.

[14] Eric S. Lander and Michael S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, 1988.

[15] T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D.J. Studholme. Assessing the performance of the oxford nanopore technologies minion. *Biomolecular Detection and Quantification*, 3:1 – 8, 2015.

[16] A. Mikheyev and M. Tin. A first look at the oxford nanopore minion sequencer. *Molecular Ecology Resources*, 14(6):1097–1102, Nov. 2014.

[17] Michael Mitzenmacher. *Algorithm Theory – SWAT 2008: 11th Scandinavian Workshop on Algorithm Theory, Gothenburg, Sweden, July 2-4, 2008. Proceedings*, chapter A Survey of Results for Deletion Channels and Related Synchronization Channels, pages 1–3. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[18] Abolfazi Motahari, Guy Bresler, and David Tse. Information theory of DNA shotgun sequencing. *IEEE Trans. on Inf. Theory*, 59(10):6273–6289, 2013.

[19] C. O'Donnell, H. Wang, and W. Dunbar. Error analysis of idealized nanopore sequencing. *Electrophoresis*, 34(15):2137-44, 2013.

[20] Yukiteru Ono, Kiyoshi Asai, and Michiaki Hamada. Pbsim: Pacbio reads simulator—toward accurate genome assembly. *Bioinformatics*, 29(1):119–121, 2013.

[21] M. Quail, M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC Genomics*, 13:341, 2012.

[22] Philippe Rigollet. 18.s997 high-dimensional statistics, chapter 1, spring 2015. Lecture notes, 2015.

[23] J. Schreiber, Z. Wescoe, R. Abu-Shumays, J. Vivian, B. Baatar, K. Karplus, and Mark Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual dna strands. *Proceedings of the National Academy of Science*, 110(47), Nov. 19, 2013.

[24] Ramji Venkataramanan, Sekhar Tatikonda, and Kannan Ramchandran. Achievable rates for channels with deletions and insertions. *Proceedings of the International Symposium on Information Theory*, pages 346–350, 2011.

[25] Sergio Verdú. Total variation distance and the distribution of relative information. *Information Theory and Applications*, 2014.

[26] Guang Yang, Angela I. Barbero, Eirik Rosnes, and Yvind Ytrehus. Error correction on an insertion/deletion channel applying codes from rfid standards. *ITA*, 2012.

VII. Appendix

A. Proof of Claim 1

We prove this by constructing a random type τ and showing that it has the desired properties with positive probability. Consider r independent random samples $\{K_1, ..., K_r\} = \vec{K}$ distributed according to $q_\ell^{\le t_*}$ (the distribution of a random variable distributed according to q_ℓ under the conditioning that it is at most t_*), for arbitrary $\ell \le \ell_*$. Denote by τ the type $T(\vec{K})$ of these samples. Then, trivially, because of the conditioning, $|supp(\tau)| \le t_*$, and $\sum_{i=1}^{\infty} \tau_i = r$ because there are exactly r samples. To show that $D(\{\tau_t\}_t \| q_\ell) \le \epsilon$ for sufficiently large t_*, the plan is to first bound the L_1 distance between the two distributions, then use a reverse Pinkser-type inequality to transfer this to a bound on the KL-divergence.

By the strong law of large numbers, with probability 1, for any $\epsilon' > 0$ and t_*, r can be made large enough so that, for all $i \in \operatorname{supp}(\tau)$,

$$|\tau_i/r - q_{\ell,i}^{\le t_*}| \le \epsilon'/2.$$

In particular, we choose, for an arbitrary fixed $\epsilon'' > 0$,

$$\epsilon' = \epsilon''/t_*. \tag{26}$$

Furthermore, we choose t_* so that, for arbitrary i,

$$|q_{\ell,i}^{\le t_*} - q_{\ell,i}| \le \epsilon'/2. \tag{27}$$

This can be done because, letting W be a random variable distributed according to q_ℓ,

$$\begin{aligned} |q_{\ell,i}^{\le t_*} - q_{\ell,i}| &\le q_{\ell,i} |\frac{1}{1 - \Pr[W > t_*]} - 1| \\ &= q_{\ell,i} \frac{\Pr[W > t_*]}{1 - \Pr[W > t_*]}, \end{aligned}$$

which can be made arbitrarily small by taking t_* large enough.

Next, by the triangle inequality,

$$\begin{aligned} |\tau_i/r - q_{\ell,i}| &\le |\tau_i/r - q_{\ell,i}^{\le t_*}| + |q_{\ell,i}^{\le t_*} - q_{\ell,i}| \\ &\le \epsilon'/2 + \epsilon'/2 = \epsilon''/t_*. \end{aligned}$$

This implies that the L_1 distance between the two distributions is

$$\|\{\tau_i/r\}_i - \{q_{\ell,i}\}_i\|_1 \le t_* \epsilon' = \epsilon''. \tag{28}$$

We have thus shown that, for any fixed $\epsilon'' > 0$, we may choose t_* sufficiently large so that the L_1 distance between the empirical and true distributions is at most ϵ''.

To transfer this upper bound on the L_1 distance to an upper bound on the KL-divergence between the two distributions, we need the following *reverse* Pinsker-type inequality:

Lemma 4 ([25], Theorem 7)**.** *Let P and Q be distributions on a common probability space, taking values in some set S, such that P is absolutely continuous with respect to Q. Define β_1 by $\beta_1^{-1} = \sup_{a \in S} \frac{dP}{dQ}(a)$, where $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P with respect to Q. Then*

$$D(P\|Q) \le \frac{\mathrm{const}}{\sqrt{\beta_1}} \cdot \|P - Q\|_1.$$

Here, P in the lemma is the distribution $\{\tau_i/r\}_{i \in \mathrm{supp}(\tau)}$, and Q is q_ℓ. The absolute continuity of the former with respect to the latter follows from the definition of the conditional distribution $q_\ell^{\le t_*}$. The Radon-Nikodym derivative $\frac{dP}{dQ}$, here, is simply the ratio of the two probability mass functions.

Provided that we can show that the maximum of this ratio over all i is at most a constant, we can then infer that the KL-divergence is bounded. This is a simple consequence of the Hoeffding bound, using the fact that τ_i/r is a sum of r i.i.d. terms bounded by t_*/r in absolute value: for any positive $\delta > 0$,

$$\begin{aligned}\Pr[|\tau_i/r - q_{\ell,i}^{\le t_*}| \ge \delta q_{\ell,i}^{\le t_*}] &\le \exp(-\delta^2 (q_{\ell,i}^{\le t_*})^2/(r \cdot r^{-2})) \\ &= \exp(-\Theta(\delta^2 r/t_*^2)).\end{aligned}$$

Union bounding with this probability over all $i \le t_*$, we have that, with probability at least $1 - t_* \exp(-\Theta(\delta^2 r/t_*^2)) = 1 - \exp(-\Theta(\delta^2 r/t_*^2))$,

$$\max_{i \in \mathrm{supp}(\tau)} \frac{|\tau_i/r - q_{\ell,i}^{\le t_*}|}{q_{\ell,i}^{\le t_*}} \le \delta.$$

This, in particular, implies that with high probability as $r \to \infty$, we have

$$\tau_i/r \le q_{\ell,i}^{\le t_*}(1 + \delta).$$

for all $i \in \mathrm{supp}(\tau)$. Thus,

$$\beta_1^{-1} = \sup_{i \in \mathrm{supp}(\tau)} \frac{\tau_i/r}{q_{\ell,i}} = O(1). \tag{29}$$

Applying Lemma 4, (29), and (28), we thus have

$$D(\{\tau_i/r\}_i \| \{q_{\ell,i}\}_i) \le \mathrm{const} \cdot \epsilon''.$$

We can then set ϵ'' so that this upper bound becomes ϵ.

Since we have proven that there is a positive probability that the random τ so chosen is in the set $\mathbb{T}(\epsilon, t_*)$ provided that we choose t_* large enough, this implies that $\mathbb{T}(\epsilon, t_*)$ is nonempty, as desired. □

B. *Proof of Theorem 2*

The proof is very similar to the proof of Theorem 1, so we only highlight what needs to be changed to derive the desired result. In particular, we need to prove Proposition 1 under the weaker assumption of overlapping distribution supports.

To do this, the challenge is again to lower bound (14) by restricting to appropriate terms. In particular, we restrict as follows:

- We restrict ℓ as before.
- We restrict to the following type T: for an arbitrary $i \in \mathrm{supp}(q_\ell) \cap \mathrm{supp}(q_{\ell+1})$,

  $$T_j = \begin{cases} r & i = j \\ 0 & i \ne j \end{cases}$$

  We then easily have $\mathrm{supp}(T) \subseteq \mathrm{supp}(q_\ell) \cap \mathrm{supp}(q_{\ell+1})$, $|\mathrm{supp}(T)| = 1$, and $\sum_i T_i = r$.
- We restrict to $m = \ell + 1$ as before.

Then the contribution to (14) of $q_{\ell,T}\binom{r}{T}$ becomes

$$q_{\ell,T}\binom{r}{T} = e^{-rD(\{T_t/r\}_t \| \{q_\ell\}) + O(\log r)} \ge e^{-r\Theta(1)}.$$

This is a consequence of the fact that

$$D(\{T_t/r\}_t \| \{q_\ell\}) = T_i/r \log \frac{T_i/r}{q_{\ell,i}} = \log(1/q_{\ell,i}) = \Theta(1).$$

Applying Pinsker's inequality, we see that this implies that the L_1 distance between the two distributions is $O(1)$.

To lower bound the expression inside the logarithm in (14), we start with (16), which becomes

$$P_{\alpha,m}/P_{\alpha,\ell} \cdot \exp(-rD(q_\ell \| q_m)\Theta(1)).$$

The rest of the proof is as in that of Theorem 1, and we omit it. □