Fundamental limits and insights: from wireless communication to DNA sequencing

Guy Bresler

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2013-43
http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-43.html

May 1, 2013

Acknowledgement

Advisor: David Tse

Fundamental limits and insights: from wireless communication to DNA sequencing

by

Guy Bresler

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering — Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Tse, Chair
Professor Kannan Ramchandran
Professor David Aldous

Fall 2012

Fundamental limits and insights: from wireless communication to DNA sequencing

Abstract

Fundamental limits and insights: from wireless communication to DNA sequencing

by

Guy Bresler

Doctor of Philosophy in Engineering — Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor David Tse, Chair

Interference is a central phenomenon in wireless networks of all types, occurring whenever multiple users attempt to communicate over a shared medium. Current state-of-the-art systems rely on two basic approaches: orthogonalizing communication links, or treating interference as noise. These approaches both suffer from a swift degradation in performance as the number of users in the system grows large. Recently, interference alignment has emerged as a promising new perspective towards mitigating interference. The extent of the potential benefit of interference alignment was observed by Cadambe and Jafar [1], who showed that for time-varying or frequency selective channels, $\frac{K}{2}$ total degrees of freedom are achievable in a K-user interference channel. In the context of the result, this means that interference causes essentially no degradation at all in performance as the number of users grows. However, a caveat is that the number of independent channel realizations needed over time or frequency, i.e. the channel diversity, is *unbounded*. Actual communication systems have only *finite* channel diversity, and thus the practical implications of interference alignment are uncertain: Just how much channel diversity is required in order to get substantial benefit from interference alignment? The first part of this thesis focuses on this question. Our first result characterizes the degrees of freedom for the three-user interference channel as a function of time or frequency diversity. We next focus on spatial diversity, in the form of multiple antennas at transmitters and receivers. We characterize the degrees of freedom for the symmetric three-user multiple-input multiple-output interference channel. This result is partially generalized to an arbitrary number of users, under a further symmetry assumption.

The second part of this thesis studies DNA sequencing from an information theory point-of-view. DNA sequencing is the basic workhorse of modern day biology and medicine. Shotgun sequencing is the dominant technique used: many randomly located short fragments called reads are extracted from the DNA sequence, and these reads are assembled to reconstruct the original sequence. During the last two decades, many assembly algorithms have been proposed, but comparing and evaluating them is difficult. To clarify this, we ask: Given N reads of length L sampled from an arbitrary DNA sequence, is it possible to achieve some target probability $1-\epsilon$ of successful reconstruction? We show that the answer

depends on the repeat statistics of the DNA sequence to be assembled, and we compute these statistics for a number of reference genomes. We construct lower bounds showing that reconstruction is impossible for certain choices of N and L, and complement this by analytically deriving the performance of several algorithms, both in terms of repeat statistics. In seeking an algorithm that matches the lower bounds on real DNA data, we are able to methodically progress towards an optimal assembly algorithm. The goal of this work is to advocate a new systematic approach to the design of assembly algorithms with an optimality or near-optimality guarantee.

To Ima, Aba, Ma'ayan, and Sarah.

Contents

Acknowledgments

I have been extremely fortunate to have David Tse as my advisor. As a researcher, he inspires me with his deep intellectual curiosity and a level of intuition that lets him go to the very core of difficult research problems. He savors and enjoys the process of doing research, both the challenges and the small discoveries along the way. It is David's qualities as a person that I am most grateful for; his patience, compassion, and generosity.

A huge thanks to Kannan Ramchandran and David Aldous for serving on my dissertation committee. I will forever be grateful to Bruce Hajek, who mentored me with generosity and wisdom during my undergraduate years at Illinois. I thank Bernd Sturmfels for warmly inviting me to interact with him and his students. A heartfelt thank you to Elchanan Mossel for many conversations and opportunities to learn. I would like to thank Abhay Parekh for his perspective on life and his encouragement.

I would like to thank my wonderful collaborators, Ma'ayan Bresler, Dustin Cartwright, Allan Sly, Shankar Bhamidi, Abolfazl Motahari. Thanks to Yun Song, Lior Pachter, Sharon Aviran, Serafim Batzoglou, and Venkat Anantharam for stimulating discussions. I am grateful to my colleagues, Galen Reeves, Hari Palaiyanur, Bobak Nazer, Alex Dimakis, Anand Sarwate, Salman Avestimehr, Lenny Grokop, Gireeja Ranade, Baosen Zhang, Changho Suh, Rahul Tandra, Sudeep Kamath, Hao Zhang, Kristen Woyach, Pulkit Grover, Dapo Omidiran, Sahand Negahban, and all the other members of the Wifo community.

A special thank you to my dear friends, who are like a family to me. I am grateful beyond words to my parents, Liora and Yoram, for their constant love, support, and encouragement. My sister Ma'ayan has been with me since the beginning. I feel incredibly lucky to have her as my closest friend. Finally, I would like to thank my soulmate Sarah Dendy for her infinite love.

Chapter 1

Overview

Wireless communication systems have become an almost indispensable part of our daily lives. This has required extensive technological innovation, which is in turn made possible by theoretical development. Information theory, in particular, has provided a unified and powerful way to think about the design of high-performance wireless systems in terms of *capacity*, the highest data rate that can be supported. The basic information theory approach to understanding a communication problem has three steps: first, come up with a physical model for the scenario; second, derive fundamental bounds on the capacity; and third, devise a communication scheme that is either optimal or at least close to capacity. Within the last 15 years this approach has resulted in a number of new ideas that have been adopted in industry standards, two examples of which include use of multiple antennas (MIMO) and opportunistic communication. Such progress has improved the performance of wireless systems, and cemented the role of information theory in their design.

Moving forward, much of the research on wireless communication is focused on dealing with interference. A prototypical scenario has some number of transmit-receiver pairs (called links), each transmitter (a cellular base-station, for example) wishing to communicate a message to a receiver (e.g. mobile phone). Each receiver attempts to decode the message from the corresponding transmitter, but receives a superposition of the signals from all of the transmitters; the challenge is to determine the desired message amongst the interference. Current state-of-the-art communication systems deal with interference by either orthogonalizing the links over time or frequency (in other words links take turns), or treating interference as noise. Both of these strategies result in performance degradation as the number of users K grows.

A basic information theory model, the so-called *interference channel*, distills the problem to its essence (Fig. 1.1). Finding the capacity of this channel, and near-optimal communication schemes, would presumably give insight into the design of good wireless systems. Unfortunately, the capacity of even the two user interference channel has been an open problem for nearly 40 years. About 6 years ago, Etkin et al. [2] made significant progress by finding the capacity of the Gaussian interference channel with two users to within a single bit per user. Such an approximation gives a lot of insight into the problem; however, for

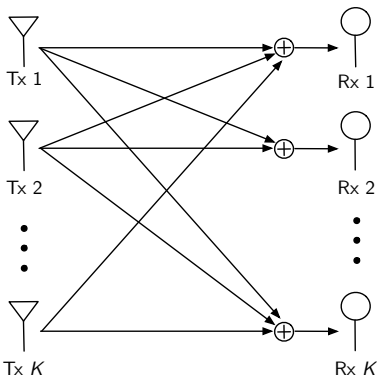


Figure 1.1: Interference channel model with K users. Receiver i, on the right, wishes to decode a message from transmitter i, on the left. All other signals are interference.

more than two users getting such a tight bound seems difficult.

Instead of seeking to bound the capacity within a constant gap, recent work has focused on the leading term in the capacity, captured by the *degrees of freedom*. The number of degrees of freedom in a system is given by the total capacity normalized by the capacity of a single point-to-point link, in the limit of high transmit power:

$$\mathsf{DoF} = \lim_{P \to \infty} \frac{C_\Sigma(P)}{\log P}.$$

A single transmit-receive link has one degree of freedom, and so the degrees of freedom achievable in a multi-user channel measures how efficiently the channel is being used relative to the baseline scheme of orthogonalizing the links. Exceeding the baseline indicates an interesting potential for an improved communication scheme. In the context of a different communication channel, the MIMO X-channel, Maddah-Ali et al. [3] and subsequently Jafar and Shamai [4] introduced the idea of *interference alignment*, and showed that up to 4/3 degrees of freedom were attainable. The basic idea is to align multiple interfering signals at each receiver in order to reduce the effective interference, while still allowing the desired signal to be discerned.

The extent of the potential benefit of interference alignment was discovered by Cadambe and Jafar [1] in application to the K-user interference channel, when they showed that for time-varying or frequency selective channels, $\frac{K}{2}$ total degrees of freedom are achievable using a basic linear precoding scheme. In other words, somewhat amazingly, each user gets half the degrees of freedom they would get if they were the only user in the system, independent of the total number of users K.

The $\frac{K}{2}$ degrees of freedom result has the major caveat that the number of independent channel realizations needed (in the form of parallel channels), i.e. the channel diversity, is *unbounded* and in fact grows as $O(K^{2K^2})$. Actual communication systems have only *finite* channel diversity, and the $O(K^{2K^2})$ requirement is prohibitive even for moderately many

users. This issue stands as a major obstacle to determining whether interference alignment will deliver on its initial promise in practical communication systems with many users.

How much channel diversity, precisely, is required in order to align interference? This question motivates Chapters 2 of this thesis, which focuses on the $K = 3$ user case. We focus on time or frequency diversity and characterize the degrees of freedom as a function of channel diversity. Aside from time or frequency diversity, many systems have *spatial diversity*, in the form of multiple antennas at transmitters and receivers. Chapter 3 studies the degrees of freedom of interference channels as a function of the spatial diversity, obtaining results for $K = 3$ users as well as a partial generalization to an arbitrary number of users.

In the last 65 years, information theory has achieved astounding success in guiding the development of communication systems. Can the success of this way of thinking be applied to other problems as well? The second part of this thesis aims to do just that, and studies the problem of DNA sequencing from an information theory point of view.

DNA sequencing is the foundational procedure for modern genomics, with important applications in all of biology and medicine. Since the sequencing of the Human Reference Genome ten years ago, there has been an explosive advance in sequencing technology, resulting in several orders of magnitude increase in throughput and decrease in cost. Multiple "next-generation" sequencing platforms have emerged, all based on shotgun sequencing. First, many short subsequences (called reads) are extracted from a DNA sequence, and then the reads are assembled to reconstruct the original sequence (Fig. 1.2). This is analogous to putting together a gigantic jigsaw puzzle: the difficulty is in deciding which pieces belong next to one another.

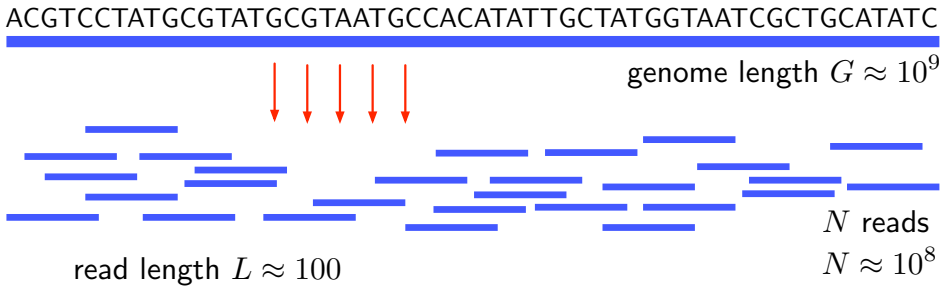


Figure 1.2: Shotgun sequencing: many short subsequences called reads are extracted from the DNA sequence and then assembled to reconstruct the original sequence.

Assembling the reads is a major algorithmic challenge, and in the last two decades dozens of assembly algorithms have been proposed to solve the problem [5]. Still, the assembly problem is far from solved and current DNA sequencing leaves a lot of room for improvement. According to Alkan et al. [6] each genome reconstruction only contains about 85% of the true sequence—far from ideal. This state of affairs hinders scientific progress and limits applications. For example, Baker [7] notes that a recent assembly of the chicken genome found 36 genes to be missing that are all present in such disparate organisms as yeast, plants, and other animals. But further careful analysis showed, as one might expect, that

the genes were missing only from the assembly and not from the chicken itself. Other sequenced organisms are similarly incomplete. Can the reads be assembled more cleverly, resulting in improved genome assemblies? Or is the read data fundamentally insufficient, making better assembly impossible?

The latter question amounts to feasibility: given a set of reads, is it *possible* to reconstruct the original sequence? The feasibility question is a measure of the intrinsic *information* each read provides about the DNA sequence, and for given sequence statistics depends on characteristics of the sequencing technology such as read length and noise statistics. As such, it can provide an algorithm-independent basis for evaluating the efficiency of a sequencing technology.

Equally important, when assembly is possible, we would like to find *algorithms* that can successfully reconstruct. We can compare algorithms based on their feasible regions, and seek an algorithm whose performance approaches the theoretical limit.

Contributions

In general, both of the problems we study highlight the engineering significance of understanding *information* requirements. Finding fundamental limits brings out the salient features, which in turn give engineering insight. For interference channels, our main message is that diversity is a critical resource. We elucidate the role of channel diversity in interference alignment, where diversity plays a new and different role than in previous works. Likewise for DNA sequencing, we are able to extract a few simple statistics of DNA sequences that act as *sufficient statistics*. These statistics, counting various types of repeats in the sequence, determine the performance of various algorithms. The many complicated features of real DNA are captured succinctly in a simple way.

Interference channel

We seek to understand the relationship between channel diversity and the ability to align interference, and obtain several results in this direction.

- **3-user IC with time/frequency diversity:** We derive the degrees of freedom for the three-user interference channel as a function of time or frequency diversity. We derive new converse arguments in this setting.
- **Alignment depth:** As part of this, we introduce the notion of alignment depth, a measure of how much alignment is possible. We quantify the connection between finite diversity and alignment depth.
- **3-user IC with spatial diversity:** We next focus on spatial diversity. For the symmetric three-user MIMO interference channel with M transmit and N receive antennas, we determine the possible alignment depth as a function of parameters M, N, thereby characterizing the degrees of freedom.

- **K-user MIMO IC:** For the many-user MIMO interference channel, we prove a general necessary condition on the parameters to allow alignment. A consequence is that at most *two* degrees of freedom are attainable using spatial diversity, in contrast with the $K/2$ result of [1] for time/frequency diversity. Spatial diversity is thus only mildly useful for interference alignment.

- **K-user MIMO IC:** In the fully symmetric case of N antennas at both the transmitters and receivers, we show that the general necessary conditions are also sufficient, providing a characterization of the degrees of freedom.

DNA sequencing

We next turn to the problem of DNA sequence assembly. We focus on the most basic shotgun sequencing model where N noiseless reads of a fixed length L base pairs are uniformly and independently drawn from a DNA sequence. Our results include the following items.

- **Data-centric view:** A difficulty is that there are no particularly good probabilistic models for DNA sequences. Instead, we propose a *data-centric* view. How do assembly algorithms perform on typical DNA data? What are the features of DNA that actually affect performance? We show that the performance of several algorithms, as well as lower bounds, can be expressed in terms of repeat statistics. These serve as *sufficient statistics* for algorithm performance on DNA sequences.

- **Pipeline to determine feasibility of assembly:** We formulate feasibility as a basic information theoretic question. Our approach results in a pipeline, which takes as input a genome sequence and desired success probability $1 - \epsilon$, computes a few simple repeat statistics, and from these statistics produces a feasibility plot that indicates for which read length L and number of reads N reconstruction is possible.

- **Systematic design and near-optimal algorithm:** Our approach elucidates how various design choices of existing algorithms affect performance. By integrating ideas from existing algorithms, we produce a modification that performs close to optimality on a wide range of genome statistics. In the cases where the algorithm is not optimal, we can bound the gap from optimality.

We contrast two ways to support analytical predictions, each with its own advantages. One approach is to produce rigorous proofs, and the second is to produce experimental results matching analytical predictions. The work on wireless communications, in the tradition of information theory, is supported by rigorous proofs of the various statements. The DNA sequencing work is only partially rigorous. In particular, we prove correctness of the sufficient conditions for the various algorithms, but do not prove rigorous probabilistic bounds. Instead, we produce simulations showing that the predictions given by the analysis are correct. Thus, in the style of physics results, the theoretical predictions are supported by matching experiments.

Chapter 2

Three-user interference channel: degrees of freedom as a function of channel diversity

2.1 Introduction

Interference is a central phenomenon in wireless networks of all types, occurring whenever multiple users attempt to communicate over a shared medium. Cellular networks in densely populated areas, for example, are severely limited by interference. To address this problem, the research community as well as the wireless communications industry have invested a great deal of effort in trying to develop efficient communication schemes to deal with interference. Still, the current state-of-the-art systems rely on two basic approaches: orthogonalizing communication links, or treating interference as noise. These approaches both suffer from a swift degradation in performance as the number of users in the system grows large. This behavior is captured in the seminal work of Gupta and Kumar [8], in which they introduced a scaling law formulation for wireless ad hoc network capacity. They analyzed a multi-hop communication model based on the classical interference-avoidance approaches, with the somewhat pessimistic result that a dense system with K users in a small area can achieve a total throughput of only $O(\sqrt{K})$, i.e. a vanishing throughput per user.

Moving beyond the simple multi-hop communication schemes allowed by [8], a natural question is whether this limitation is *fundamental*, or if better performance is possible by considering more general communication schemes. Indeed, much better performance was shown to be possible when Ozgur et al. [9] invented a hierarchical MIMO scheme that achieves *linear* scaling of total throughput in dense ad hoc wireless networks. This means that such networks are not inherently interference-limited as previous believed. Drawbacks of the hierarchical MIMO scheme include the fact that a lot of cooperation between users is required in order to communicate, the proof of optimality rests on the assumption the users are *randomly* and uniformly located which may not be true in practice, and finally it is not entirely

clear when the (asymptotic) results apply to a finite system. Aside from hierarchical MIMO, *interference alignment* has emerged as a new perspective towards mitigating interference.

Interference alignment offers the potential for considerable increase of performance for interference-limited communication, with little coordination required between users (aside from channel state), and can be applied to any number of users in the system with arbitrary locations (see e.g. [10]). The basic idea is to align multiple interfering signals at each receiver in order to reduce the effective interference, while still allowing the desired signal to be discerned. Interference alignment was introduced by Maddah-Ali et al. [3] and subsequently clarified by Jafar and Shamai [4], both in the context of the MIMO X-channel. But the extent of the potential benefit of interference alignment was observed by Cadambe and Jafar [1] in application to the K-user interference channel, when they showed that for time-varying or frequency selective channels, $\frac{K}{2}$ total degrees of freedom are achievable using a basic linear precoding scheme. The number of degrees of freedom in a system, defined later, is given by the total capacity normalized by the capacity of a single point-to-point link, in the limit of high signal-to-noise ratios (SNR). In other words, somewhat amazingly, each user gets the same degrees of freedom as with only *two* users in the system, independent of the total number of users K.

The $\frac{K}{2}$ degrees of freedom result, achieving linear scaling with number of users, has two caveats. First, it is not clear if such linear scaling is possible at any finite SNR. Second, the number of independent channel realizations needed (in the form of parallel channels), i.e. the channel diversity, is *unbounded* and in fact grows as $O(K^{2K^2})$. Actual communication systems have only *finite* channel diversity, and the $O(K^{2K^2})$ requirement is prohibitive even for moderately many users. These two issues stand as major obstacles to determining whether interference alignment will deliver on its initial promise in practical communication systems with many users.

Several works address the first concern, studying interference alignment at finite SNR rather than the asymptotic degrees of freedom setting. Ozgur and Tse [11] showed that the scheme of [1] does in fact achieves linear scaling of total rate as the number of users grows, for a certain random phase channel model. The basic scheme, being the same as [1], still requires exponential channel diversity, namely $O(2^{K^2})$. Nazer et al. [12] introduced a new scheme named ergodic interference alignment, whereby each user can obtain $\frac{1}{2}$ of the rate possible with no interference, i.e. linear scaling at the best possible rate. Ergodic alignment needs a mild symmetry assumption on the channel fading process, but more significantly requires diversity $O(K^{2K^2})$. These finite SNR results show that the degrees of freedom formulation on its own is not necessarily misleading, and establish as the main question: just how much channel diversity is required in order to get substantial benefit from interference alignment?

Grokop et al. [13] examine the basic assumptions of [1] through the lens of a different channel model and arrive again at the conclusion that alignment is viable at finite SNR; however, they are additionally able to make headway on the finite channel diversity question. First they showed that for a line-of-sight channel model, spectral efficiency can increase linearly with number of users K, even for any fixed transmit power, as long as the bandwidth scales at a rate $O(K^{2K^2})$. In their model, bandwidth roughly corresponds to channel diversity.

They next provided a partial converse, showing that if the bandwidth scales sufficiently slowly (approximately at rate $O(K/\log K)$), then linear scaling is impossible. This is, to our knowledge, the only result showing a limitation on alignment in terms of diversity, in a situation where linear scaling of throughput in number of users is possible.

Despite major effort by researchers over the last six or so years, little is known about how diversity affects the ability to align interference. Because [1] uses linear (vector space) precoding, we can attempt to simplify matters by restricting to the class of such vector space schemes (defined carefully in Sec. 2.2). Specifically, we let $\mathsf{DoF}(L, K)$ denote the total degrees of freedom achievable using vector space schemes in the K-user interference channel with diversity L (given by L real parallel channels).

$\mathsf{DoF}(L, K)$	$K = 3$	$K = 4$	general K
$L = 1$	1	1	1
$L = 2$	$6/5$ [CJW10]		
$L = 3$			
$L = 4$			
$\vdots$			
$L = \infty$ [CJ08]	$3/2$	2	$K/2$

Figure 2.1: Previously known values of $\mathsf{DoF}(L, K)$. We fill in the $K = 3$ column.

$\mathsf{DoF}(L, K)$ is known for a few parameter values. One case is easy: for $L = 1$, all channels are proportional to the identity, so all receivers observe essentially the same output. It follows that each receiver can decode *all* the transmitted signals, so the degrees of freedom is limited to that of a multiple-access channel, i.e. at most one. Conversely, one degree of freedom is trivially achievable, hence for any K we get

$$\mathsf{DoF}(1, K) = 1\,.$$

Next, Cadambe et al. [14] showed for the complex scalar interference channel with $K = 3$ users that there are $6/5$ degrees of freedom. The complex scalar channel translates to $L = 2$ real parallel channels, and thus

$$\mathsf{DoF}(2, 3) = 6/5\,.$$

Beyond this nothing is known besides the original result of [1], which amounts to

$$\mathsf{DoF}(\infty, K) = K/2\,.$$

The known results are summarized in Fig. 2.1. In this chapter we make some progress, filling in the entire $K = 3$ column (for all values of channel diversity L). To simplify notation in the remainder of the chapter, since $K = 3$ always, we write DoF in place of $\mathsf{DoF}(L, 3)$.

Main result

The main result of this chapter characterizes the degrees-of-freedom for three users as a function of channel diversity.

Theorem 1. *The three-user parallel real-valued interference channel has the following degrees-of-freedom as a function of the channel diversity L when restricted to vector space strategies:*

$$\mathsf{DoF} = \frac{3D}{2D+1},$$

where $D := 2L - \lfloor L/2 \rfloor - 1$. This holds for generic channel gains.

The key innovation is the concept of *alignment depth*, which describes how intertwined the transmit signal spaces are due to alignment (see Fig. 2.2). If transmitter 1 transmits along a vector $\mathbf{v}_1$, then transmitter 2 can choose a vector $\mathbf{v}_2$ that is *aligned* with $\mathbf{v}_1$ at receiver 3. This constitutes an alignment path of depth 2 and results in a saving, because receiver 3 observes only one interference dimension rather than the two it would normally have if $\mathbf{v}_1$ and $\mathbf{v}_2$ were arbitrary. Transmitter 3 now chooses a vector $\mathbf{v}_3$, and it can choose $\mathbf{v}_3$ to be aligned with $\mathbf{v}_2$ at receiver 1, creating an alignment path of depth 3, and this process can continue arbitrarily. Fig. 2.2 depicts an alignment path of depth 4 starting at transmitter 1.

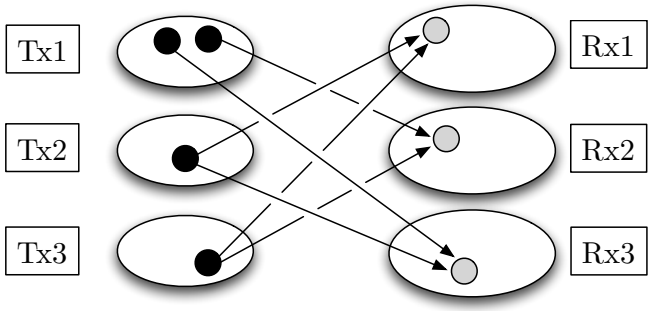


Figure 2.2: Alignment path of depth 4.

The resources saved due to an alignment path of depth D can be heuristically computed as follows. The first vector in the alignment path, say $\mathbf{v}_1$, is not aligned with any other vectors, and occupies 3 total dimensions, one at each receiver. The second vector in the alignment path, $\mathbf{v}_2$ in the previous paragraph, occupies only 2 additional dimensions at the receivers. The same is true for each successive transmit vector in the alignment path. The total number of dimensions occupied at all three receivers is thus $2 \cdot (D-1) + 3 = 2D + 1$, as

compared to the $3D$ dimensions that would ordinarily be occupied if no vectors were aligned. The ratio is the quantity $\mathsf{DoF} = 3D/(2D+1)$ in Theorem 1. If we could take $D \to \infty$ there would be $3/2$ degrees of freedom, in agreement with the $K/2$ result of [1], but for finite channel diversity the alignment depth must be finite as well.

It turns out that for each value L of channel diversity there is a corresponding *maximum* allowed alignment depth $D = 2L - \lfloor L/2 \rfloor - 1$. We show that alignment at depth greater than D forces the desired signal into the interference space at some receiver, thereby preventing decoding. The intuition is that it is actually not difficult to align the signals—for example all users can transmit in the same frequency band—but at some point one loses the ability to distinguish the desired signal from the interference. This occurs because all the transmit vectors on an alignment path are related to one another in a rigid manner through the channel matrices. Once the alignment path is longer than D, the interfering signals (by virtue of the channel matrices all lying in an L-dimensional vector space) are able to *emulate* the direct channel. This highlights the role of channel diversity: the channel between transmitter and intended receiver needs enough diversity, or richness, to transform the desired signal to a part of the received space free of interference.

The achievable strategy for odd values of L is due to Cadambe and Jafar [1], with a slight modification required for even values of L. It essentially consists of creating alignment paths of maximum depth. Our contribution is thus mainly in showing the optimality of this approach, together with a precise understanding of the relationship between diversity and alignment depth.

Overview

The rest of the chapter is organized as follows. In Section 2.2 we present the parallel interference channel model as well as discuss vector space strategies and degrees of freedom. Section 2.3 describes the strategy achieving optimality in Theorem 1. In Section 2.4 we prove the converse to Theorem 1. Finally, Section 2.5 contains a lemma on linear independence used in the proof of both the achievability and converse parts of the main result.

2.2 Formulation

Interference channel model

For $i = 1, 2, 3$, receiver i wishes to obtain a message from the corresponding transmitter i. The remaining signals from transmitters $j \neq i$ are undesired interference. The three-user real Gaussian interference channel is represented by the input-output relationship at each time-step

$$\mathbf{y}_i = \sum_{j=1}^{3} \bar{\mathbf{H}}_{ji} \mathbf{x}_j + \mathbf{z}_i \,. \tag{2.1}$$

Here for each user i we have $\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i \in \mathbb{R}^L$, where $\mathbf{x}_i$ is the transmitted signal, $\mathbf{y}_i$ is the received signal, and $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I}_L)$ is additive isotropic white Gaussian noise. Channel diversity, in the form of L independent channel realizations (for example in a time-varying or frequency-selective fading channel), is modeled via matrices $\bar{\mathbf{H}}_{ij} \in \mathbb{R}^{L\times L}$, assumed to be diagonal with generic entries:

$$\bar{\mathbf{H}}_{ij} = \begin{pmatrix} h_{ij}(1) & & & \\ & h_{ij}(2) & & \\ & & \ddots & \\ & & & h_{ij}(L) \end{pmatrix}.$$

In particular, entries drawn from a non-degenerate continuous distribution will be generic. Each input must satisfy an average power constraint over a length-T block, $\frac{1}{T}E(||\mathbf{x}_i^T||^2) \leq P$.

Since our goal is to understand the effect of finite channel diversity, we will allow T uses of the channel, but the diversity is to remain fixed at L. To illustrate this, a simple scenario is L independently faded frequency bands, but with constant channels for the duration of communication. Using T time-slots amounts to scaling the ambient dimension to $N = T \cdot L$, with a resulting change in vectors $\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i \in \mathbb{R}^N$ and channel matrix

$$\mathbf{H}_{ij} := \mathbf{I}_T \otimes \bar{\mathbf{H}}_{ij} \in \mathbb{R}^{N\times N}. \tag{2.2}$$

The Kronecker product $\mathbf{I}_T \otimes \bar{\mathbf{H}}_{ij}$ places T identical blocks of $\bar{\mathbf{H}}_{ij}$ along the diagonal. We emphasize that throughout this chapter, the nonzero entries $h_{ij}(1), h_{ij}(2), \ldots, h_{ij}(L)$ of the $\bar{\mathbf{H}}_{ij}$ matrices are assumed to be generic; we will refer to generic channel matrices of diversity L, meaning $\mathbf{H}_{ij}$ are of the form $\mathbf{I}_T \otimes \bar{\mathbf{H}}_{ij}$. Obviously, the entries of $\mathbf{H}_{ij}$ are repeated if $T > 1$ and are not generic.

We note that the matrices $\mathbf{H}_{ij}$ are generically invertible. The main consequence of finite channel diversity for the parallel channel model is that the set of products of matrices and inverses

$$\mathcal{H} := \left\{ \prod \mathbf{H}_{ij}^{\alpha_{ij}} : \alpha_{ij} \in \mathbb{Z} \right\}$$

spans an L-dimensional real vector space, $\text{span}(\mathcal{H}) \cong \mathbb{R}^L$. When we talk about linear independence of channels it is in this ambient space.

Vector space schemes and degrees of freedom

We restrict the class of communication schemes to so-called *vector space* schemes. In this context degrees-of-freedom has a simple interpretation in terms of the dimensions of the transmit subspaces, described in the next paragraph. However, note that one can more generally define the degrees-of-freedom region in terms of an appropriate high transmit-power limit $P \to \infty$ of the Shannon capacity region $C(P)$ normalized by $\log P$ ([1], [3]). In

that general framework, it is well-known and easy to show that vector space schemes give a concrete non-optimal achievable strategy with rates

$$R_i(P) = d_i \log(P) + O(1), \quad 1 \le i \le K\,.$$

Here d_i is the dimension of transmitter i's subspace and P is the transmit power.

We now describe what is meant by vector space scheme. Suppose transmitter j wishes to transmit a vector $\hat{\mathbf{x}}_j \in \mathbb{R}^{d_j}$ of d_j data symbols. These data symbols are modulated on the subspace $V_j \subseteq \mathbb{R}^N$ of dimension d_j, producing the input signal $\mathbf{x}_j = \mathbf{V}_j\hat{\mathbf{x}}_j$, where $\mathbf{V}_j$ is a $N \times d_j$ matrix whose column span is V_j. The signal $\mathbf{x}_j$ propagates to receiver i through the channel as $\mathbf{H}_{ji}\mathbf{V}_j\hat{\mathbf{x}}_j$ The dimension of the transmit space, d_j, determines the number of data streams, or degrees-of-freedom, available to transmitter j. With this restriction to vector space schemes, the output is given by

$$\mathbf{y}_i = \mathbf{H}_{ii}\mathbf{V}_i\hat{\mathbf{x}}_i + \sum_{\substack{1\le j\le K\\ j\ne i}} \mathbf{H}_{ji}\mathbf{V}_j\hat{\mathbf{x}}_j + \mathbf{z}_i\,, \quad 1 \le i \le K\,. \tag{2.3}$$

The desired signal space at receiver i is $\mathbf{H}_{ii}V_i$, while the interference space consists of $\sum_{j\ne i}\mathbf{H}_{ji}V_j$, i.e. the span of the undesired subspaces as observed by receiver i.

In the regime of asymptotically high transmit powers, in order that decoding can be accomplished we impose the constraint at each receiver i that the desired signal space $\mathbf{H}_{ii}V_i$ is *complementary* to the interference space $\sum_{j\ne i}\mathbf{H}_{ji}V_j$. Equivalently, each vector $\mathbf{y}_i \in \mathbf{H}_{ii}V_i + \sum_{j\ne i}\mathbf{H}_{ji}V_j$ in the receive subspace has a unique decomposition $\mathbf{y}_i = \mathbf{u} + \mathbf{v}$ with $\mathbf{u} \in \mathbf{H}_{ii}V_i$ and $\mathbf{v} \in \sum_{j\ne i}\mathbf{H}_{ji}V_j$, so the desired signal can be ascertained. For later reference we record an equivalent condition.

$$\text{Decoding condition:} \qquad \mathbf{H}_{ii}V_i \bigcap \left(\sum_{j\ne i}\mathbf{H}_{ji}V_j\right) = \{0\}\,, \quad 1 \le i \le 3\,. \tag{2.4}$$

The total degrees of freedom achieved by a given vector space strategy is the sum of the transmit dimensions normalized by the number of dimensions achievable by one user in isolation (which is just the total ambient dimension N). Thus the maximum total degrees of freedom achievable is

$$\mathsf{DoF} = \max_{\substack{V_1,V_2,V_3\\ \text{satisfying (2.4)}}} \frac{d_1+d_2+d_3}{N}\,.$$

The goal of this chapter is to determine DoF as a function of L.

2.3 Achievable strategy

Achievable strategy description

The achievable strategy has a straightforward objective: it creates maximal possible alignment between vectors from interfering users. Alignment is done by pairwise alignment of *individual vectors* (Fig. 2.3). To see how this works, suppose user 1 transmits a vector $\mathbf{v}_1$. Then

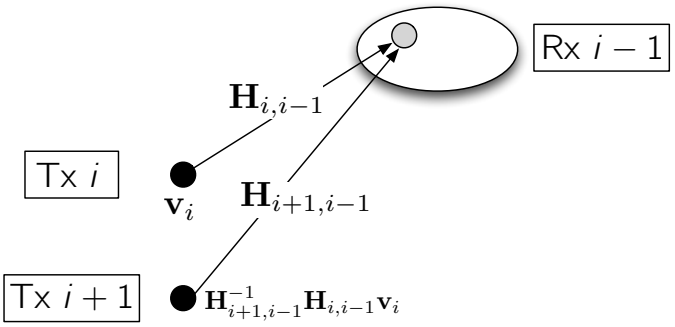


Figure 2.3: The vector at transmitter $i+1$ is aligned with the vector at transmitter i.

user 2 selects the vector $\mathbf{v}_2 = \mathbf{H}_{23}^{-1}\mathbf{H}_{13}\mathbf{v}_1$, which is chosen in order that $\mathbf{H}_{23}\mathbf{v}_2 = \mathbf{H}_{13}\mathbf{v}_1$. These vectors are aligned at receiver 3, and create only one dimension of interference. This process is continued by user 2 transmitting a vector $\mathbf{v}_3 = (\mathbf{H}_{31}^{-1}\mathbf{H}_{21})\mathbf{v}_2 = (\mathbf{H}_{31}^{-1}\mathbf{H}_{21})(\mathbf{H}_{23}^{-1}\mathbf{H}_{13})\mathbf{v}_1$, which aligns $\mathbf{v}_2$ and $\mathbf{v}_3$ at receiver 1: $\mathbf{H}_{21}\mathbf{v}_2 = \mathbf{H}_{31}\mathbf{v}_3$. This process continues iteratively to form *alignment paths*.

To more succinctly describe the vectors obtained by iterating the pairwise alignment construction (c.f. Fig. 2.3) we define the alignment matrices

$$\mathbf{S}_1 = \mathbf{H}_{23}^{-1}\mathbf{H}_{13}\,, \quad \mathbf{S}_2 = \mathbf{H}_{31}^{-1}\mathbf{H}_{21}\,, \quad \mathbf{S}_3 = \mathbf{H}_{12}^{-1}\mathbf{H}_{32}\,. \tag{2.5}$$

Matrix $\mathbf{S}_i$ takes a vector $\mathbf{v}_i \in V_i$ at transmitter i and produces a vector $\mathbf{v}_{i+1} = \mathbf{S}_i\mathbf{v} \in V_{i+1}$ such that $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ are aligned at the interfered receiver $i+1$ (recall that all indices are interpreted modulo 3).

The optimal strategy successively aligns the signals from interfering users up to a total of

$$D = 2L - \lfloor L/2 \rfloor - 1 \tag{2.6}$$

vectors (Fig. 2.2 depicts the case $D = 4$). We call D the *depth* of the alignment. Before giving more detail, we describe the achievable strategies for $L = 3$ and $L = 4$.

Example: $L = 3$

The scheme for odd values of L (and $L = 3$ in particular) was discovered by [1]. It turns out that time extension $T = 1$ suffices, with a resulting ambient dimension $N = T \cdot L = 3$. For $L = 3$, the alignment depth as determined by (2.6) is $D = 2 \cdot 3 - \lfloor 3/2 \rfloor - 1 = 4$. This means we form a sequence of vectors of length 4, starting at (say) user 1:

$$\mathbf{v}_1 = \mathbf{x} \in V_1 \qquad \mathbf{v}_2 = \mathbf{S}_1\mathbf{x} \in V_2 \qquad \mathbf{v}_3 = \mathbf{S}_2\mathbf{S}_1\mathbf{x} \in V_3$$
$$\mathbf{v}_1' = \mathbf{S}_3\mathbf{S}_2\mathbf{S}_1\mathbf{x} \in V_1$$

By construction, the interference $\mathbf{H}_{21}\mathbf{v}_2 = \mathbf{H}_{31}\mathbf{v}_3$ is aligned at receiver 1, which leaves two

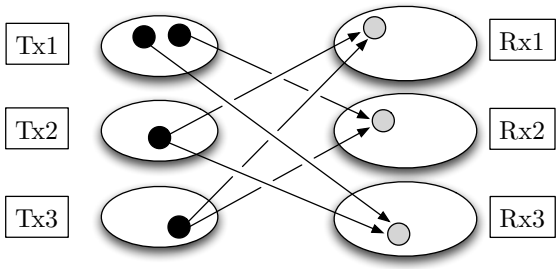


Figure 2.4: Alignment path of depth 4.

free dimensions for the signal vectors $\mathbf{H}_{11}\mathbf{v}_1$ and $\mathbf{H}_{11}\mathbf{v}_1'$. Similarly there is one interference-free dimension at each of receivers 2 and 3, which is occupied by the desired signals $\mathbf{H}_{22}\mathbf{v}_2$ and $\mathbf{H}_{33}\mathbf{v}_3$, respectively. Thus there are enough interference-free dimensions at each receiver to allow the signal to fit. It must be checked that the signal is in fact complementary to the interference, and this is done for the more general case below.

Example: $L = 4$

In the case $L = 4$, the alignment depth is $D = 2 \cdot 4 - 4/2 - 1 = 5$. This means we form a sequence of vectors of length 5, starting (say) at user 1:

$$\mathbf{v}_1 = \mathbf{x} \in V_1 \qquad \mathbf{v}_2 = \mathbf{S}_1\mathbf{x} \in V_2 \qquad \mathbf{v}_3 = \mathbf{S}_2\mathbf{S}_1\mathbf{x} \in V_3$$
$$\mathbf{v}_1' = \mathbf{S}_3\mathbf{S}_2\mathbf{S}_1\mathbf{x} \in V_1 \qquad \mathbf{v}_2' = \mathbf{S}_1\mathbf{S}_3\mathbf{S}_2\mathbf{S}_1\mathbf{x} \in V_2$$

We start by quickly checking the number of interference dimensions at each receiver, as for the $L = 3$ case above. This is made easier by the $L = 3$ computation, since the only additional vector is $\mathbf{v}_2'$, which is aligned to $\mathbf{v}_1'$ at receiver 3, and creates an additional interference dimension at receiver 1. Thus receiver 1 has two signal dimensions and two interference dimensions, receiver 2 likewise has two signal dimensions and two interference dimensions, and receiver 3 has one signal dimension and only two interference dimensions. Receivers 1 and 2 fully utilize their receive spaces, but receiver 3 has a dimension left over. Not all dimensions are utilized, so the scheme is inefficient. How can we make use of the extra dimension at receiver 3?

The converse argument in Section 2.4 shows that alignment at a depth greater than $D = 5$ causes the decoding constraint to be violated. Hence we cannot fill the extra dimension by creating a longer alignment path. Instead, we would like to create more vectors, but still aligned at the maximum depth $D = 5$. This is done by introducing a time-extension $T > 1$, which increases the ambient dimension $N = T \cdot L$. More alignment paths of length 5, completely unrelated to one another, can now be introduced.

If each user initiates an alignment path, we require $4+4+3 = 11$ ambient dimensions at each receiver. But we cannot choose $N = 11$, being indivisible by $L = 4$, so we use a time

extension $T = 11$ and $N = 11 \cdot 4 = 44$. Correspondingly, instead of a single alignment path initiated at each transmitter, 4 alignment paths are initiated.

Use of the time extension parameter T allows left-over dimensions to be occupied by virtue of symmetrizing the alignment path construction, and amounts to a technicality required in order to make the number of incoming vectors to each receiver be equal to N.

We summarize the performance obtained in the $L = 4$ example: $d_i = L(3L/2 - 1) = 20$, $T = 11$, and $N = 44$, giving $\mathsf{DoF} = 3 \cdot 20/44 = 15/11$.

General achievability argument

The construction is a straightforward generalization of the special cases $L = 3$ and $L = 4$ above. The notation introduced here will be used again in the converse argument.

We generalize the alignment matrices $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$ defined in (2.5), where $\mathbf{S}_i$ takes a vector $\mathbf{v}_i \in V_i$ and produces a vector $\mathbf{S}_i\mathbf{v}_i \in V_{i+1}$ aligned with $\mathbf{v}_i$ at receiver $i - 1$. Let

$$\mathbf{S}_{i\uparrow n} := \prod_{j=1}^{n} \mathbf{S}_{i+j-1} \tag{2.7}$$

be the matrix taking a vector $\mathbf{v}_i \in V_i$ at transmitter i to the vector $\mathbf{S}_{i\uparrow n}\mathbf{v}_i \in V_{i+n}$ obtained by iterating the pairwise alignment construction $n - 1$ times. The vector $\mathbf{S}_{i\uparrow n}\mathbf{v}_i \in V_{i+n}$ is the $(n+1)$th vector (n steps away from the start) in the alignment path starting at $\mathbf{v}_i$. For each $n \geq 1$ the vectors $\mathbf{S}_{i\uparrow n-1}\mathbf{v}_i \in V_{i+n-1}$ and $\mathbf{S}_{i\uparrow n}\mathbf{v}_i \in V_{i+n}$ are aligned at receiver $i + n - 2$.

Since every third vector in the alignment path belongs to a given transmitter, an alignment path of length D starting at user 1 (say) with $\mathbf{x} \in V_1$ results in the vectors

$$\begin{aligned}
\mathcal{A}_1 &:= \{\mathbf{S}_{1\uparrow 3\ell}\mathbf{x}\,, 0 \leq \ell \leq n_1\} \subseteq V_1\,,\\
\mathcal{A}_2 &:= \{\mathbf{S}_{1\uparrow 3\ell+1}\mathbf{x}\,, 0 \leq \ell \leq n_2\} \subseteq V_2\,,\\
\mathcal{A}_3 &:= \{\mathbf{S}_{1\uparrow 3\ell+2}\mathbf{x}\,, 0 \leq \ell \leq n_3\} \subseteq V_3\,,
\end{aligned}$$

where

$$n_i := \lfloor (D - i)/3 \rfloor\ .$$

For odd values of L, the scheme consists of making a single alignment path of length D (as discussed in [1]). The initial vector can be arbitrary, but with all entries nonzero. Note that users may have different numbers of vectors. We do not further discuss the case that L is odd, and focus instead on even values of L.

For even values of L, we symmetrize the construction by letting $T = 3L - 1$ and having each transmitter initiate L alignment paths, for a total of $3L$ alignment paths of depth D ($3LD$ total vectors). Writing $\mathbf{v}_i^{1:L} = \{\mathbf{v}_i^{(1)}, \mathbf{v}_i^{(2)}, \ldots, \mathbf{v}_i^{(L)}\}$ for the L initial vectors at user i, we have that user i transmits along the LD vectors

$$\begin{aligned}
\mathcal{A}_{i1} &:= \{\mathbf{S}_{1\uparrow 3\ell}\mathbf{v}_i^{1:L}\,, 0 \leq \ell \leq n_1\}\,,\\
\mathcal{A}_{i2} &:= \{\mathbf{S}_{1\uparrow 3\ell+1}\mathbf{v}_{i-1}^{1:L}\,, 0 \leq \ell \leq n_2\}\,,\\
\mathcal{A}_{i3} &:= \{\mathbf{S}_{1\uparrow 3\ell+2}\mathbf{v}_{i-2}^{1:L}\,, 0 \leq \ell \leq n_3\}\,.
\end{aligned}$$

To check that the decoding constraints are satisfied, we examine the incoming vectors to receiver 1; the other receivers are symmetric. At receiver 1, all vectors from transmitter 3 are aligned to those of transmitter 2, except the L vectors $\mathbf{v}_3^{1:L}$ initiating alignment paths. It follows that there are $2LD + L$ vectors incoming to receiver 1: $\mathbf{H}_{11}(\mathcal{A}_{11} \cup \mathcal{A}_{12} \cup \mathcal{A}_{13})$ from transmitter 1, $\mathbf{H}_{21}(\mathcal{A}_{21} \cup \mathcal{A}_{22} \cup \mathcal{A}_{23})$ from transmitter 2, and $\mathbf{H}_{31}(\mathbf{v}_3^{1:L})$. Now, the number of vectors is $2LD + L = (2D + 1)L = (3L - 1) \cdot L = N$ equal to the ambient dimension, so the interference space is complementary to the signal space if these N vectors are linearly independent. This can be verified by applying Lemma 9.

2.4 Converse

In this section we show that the degrees of freedom of the three-user interference channel is at most $3D/(2D + 1)$, where $D = 2L - \lfloor L/2 \rfloor - 1$. The argument rests largely on Lemma 3 below, which bounds the alignment depth possible without violating the decoding condition at each receiver. Using Lemma 3, the proof essentially shows by the construction in (2.8) that if the sum of transmit dimensions $\sum d_i$ is too large, then there must exist an alignment path of length $D + 1$.

Theorem 2 (Converse)**.** *For generic channel matrices $\mathbf{H}_{ij}$ of diversity L, the* DoF *of the three-user interference channel is bounded as*

$$\mathsf{DoF} \leq \frac{3D}{2D+1},$$

where $D := 2L - \lfloor L/2 \rfloor - 1$.

In order to describe alignment between users we introduce some notation. First, recall the definitions $\mathbf{S}_i = (\mathbf{H}_{i+1,i-1})^{-1}\mathbf{H}_{i,i-1}$ and $\mathbf{S}_{i\uparrow n} = \prod_{j=1}^{n} \mathbf{S}_{i+j-1}$. We let

$$V_{i\uparrow n} = \cap_{j=0}^{n} \mathbf{S}_{i\uparrow j}^{-1} V_{i+j} \tag{2.8}$$

be the part of the transmit space V_i of user i that is aligned to depth at least $n+1$. Informally, starting at user i, $\mathbf{S}_{i\uparrow j}$ goes j steps forward along an alignment path, with increasing user index, and thus $\mathbf{S}_{i\uparrow j}^{-1} V_{i+j}$ brings V_{i+j} *backward* j steps to the beginning of the alignment path at user i. By separating the first term V_i from the intersection in (2.8) we obtain the simple identity

$$V_{i\uparrow n} = V_i \cap (\mathbf{S}_i^{-1} V_{i+1\uparrow n-1}). \tag{2.9}$$

This means that the part of transmit space V_i that is aligned at depth $n + 1$ can obtained by taking the intersection of the transmit space V_i with the portion of transmit space V_{i+1} aligned at depth n and then *pulled back* to transmitter i.

We use the shorthand $d_{i\uparrow n} = \dim(V_{i\uparrow n})$ for the dimension of the depth-$(n + 1)$-aligned part of V_i. An alignment depth of 1 means no alignment at all, and every transmit vector is trivially aligned to a depth at least 1; as a sanity check we observe that $V_{i\uparrow 0} = V_i$ and hence

$d_{i\uparrow 0} = d_i$. This can be seen from (2.8) if we interpret the empty product as the identity, giving $\mathbf{S}_{i\uparrow 0} = \prod_{j=1}^{0} \mathbf{S}_{i+j-1} = \mathbf{I}$.

The main ingredient in the proof of Theorem 2 is a bound on the alignment depth.

Lemma 3 (Bound on alignment depth)**.** *The alignment depth is at most* $D = 2L - \lfloor L/2 \rfloor - 1$. *More precisely,* $V_{i\uparrow D} = \{0\}$*, and hence* $d_{i\uparrow D} = 0$*, for each* i.

Lemmas 4 and 5 record inequalities needed for the proof of Theorem 2.

Lemma 4. *For any nonnegative integers* a, b, n*, we have the inequality* $d_{i\uparrow n} \geq d_{i\uparrow n+a} + d_{i-b\uparrow n+b} - d_{i-b\uparrow n+a+b}$*. A useful special case is* $d_{i\uparrow n} \geq d_{i+1\uparrow n-1} + d_{i\uparrow n-1} - d_{i+1\uparrow n-2}$.

Proof. The proof follows directly from the subadditivity of dimension, i.e. the fact that for two finite-dimensional subspaces W_1, W_2 of some larger vector space, the dimensions satisfy $\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2)$. □

Lemma 5. *We have the inequality* $\frac{(D-1)}{D} \sum d_i \geq \sum d_{i\uparrow 1}$.

Proof. Let $c_n = \sum_{i=1}^{3} d_{i\uparrow n}$. Applying Lemma 4,

$$c_n = \sum d_{i\uparrow n} \geq \sum (d_{i+1\uparrow n-1} + d_{i\uparrow n-1} - d_{i+1\uparrow n-2}) = 2 \sum d_{i\uparrow n-1} - \sum d_{i\uparrow n-2} = 2c_{n-1} - c_{n-2} \,.$$

Using this as a base case, a simple induction argument shows that $c_n \geq i c_{n-i+1} - (i-1) c_{n-i}$ for $1 \leq i \leq n$. Plugging in $i = n = D - 1$ and rearranging, we get

$$(D-2) c_0 \geq (D-1) c_1 - c_{D-1} \,. \tag{2.10}$$

Now, Lemma 4 with parameters $n = 0, a = D - 1, b = 1$ gives

$$d_i \geq d_{i\uparrow D-1} + d_{i-1\uparrow 1} - d_{i-1\uparrow D} = d_{i-1\uparrow 1} + d_{i\uparrow D-1} \,, \tag{2.11}$$

where we used the fact that $d_{i-1\uparrow D} = 0$ due to Lemma 3. Summing the inequality (2.11) over the index i gives

$$c_0 = \sum d_i \geq \sum (d_{i-1\uparrow 1} + d_{i\uparrow D-1}) = c_1 + c_{D-1} \,, \tag{2.12}$$

and adding (2.10) to (2.12) yields $(D-1) \sum d_i \geq D \sum d_{i\uparrow 1}$, completing the proof. □

We now prove Theorem 2 using Lemma 5.

Proof of Theorem 2. Suppose a block-length T is used for a total ambient dimension given by $N = TL$, and each user i, $1 \leq i \leq 3$, uses a transmit subspace $V_i \in \mathbb{R}^N$ with $d_i = \dim(V_i)$. We seek to bound the total degrees of freedom,

$$\frac{d_1 + d_2 + d_3}{N} \,.$$

We examine the signals at receiver 1. The decoding constraint (2.4), requiring that the span of interfering signals $\mathbf{H}_{21}V_2 + \mathbf{H}_{31}V_3$ be complementary to the desired signal space $\mathbf{H}_{11}V_1$, implies that

$$\dim(\mathbf{H}_{11}V_1 + \mathbf{H}_{21}V_2 + \mathbf{H}_{31}V_3) = \dim(\mathbf{H}_{11}V_1) + \dim(\mathbf{H}_{21}V_2 + \mathbf{H}_{31}V_3) = d_1 + d_2 + d_3 - d_{2\uparrow 1}\,.$$

This number is bounded by N, the dimension of the ambient space. Permuting the indices in the argument produces the three inequalities

$$N \geq \sum d_i - d_{i\uparrow 1}\,, \quad 1 \leq i \leq 3\,.$$

Summing over i gives $3N \geq 3\sum d_i - \sum d_{i\uparrow 1}$, and applying Lemma 5 results in

$$3N \geq 3\sum d_i - \frac{(D-1)}{D}\sum d_i = \frac{2D+1}{D}\sum d_i\,.$$

Rearranging the inequality completes the proof.

□

The rest of the section is devoted to the proof of Lemma 3. We begin by identifying a scenario that prevents decoding, and will later show that it occurs if the alignment depth is too large. To this end, a basic observation is that all the channel matrices $\mathbf{H}_{ij}$ have the same eigenspaces, which implies that two transmit spaces V_i, V_{i+1} cannot both contain the same eigenvector $\mathbf{v}$, as otherwise the desired signal $\mathbf{H}_{ii}\mathbf{v} = \mathbf{v}$ at receiver i overlaps the interference $\mathbf{H}_{i+1,i}\mathbf{v} = \mathbf{v}$ from transmitter $i+1$ (i.e. the decoding condition (2.4) is violated). The next two lemmas find conditions implying the occurrence of this scenario.

Let $\mathcal{H}_0$ be the set of nondegenerate products of channel matrices and inverses,

$$\mathcal{H}_0 = \left\{ \prod \mathbf{H}_{ij}^{\alpha_{ij}} : \alpha_{ij} \in \mathbb{Z} \right\} \setminus \mathbf{I}\,.$$

Let $\mathbf{A} \in \mathcal{H}_0$ be such a matrix. The first lemma shows that if V_i contains an $\mathbf{A}$-invariant subspace, then it contains an eigenvector $\mathbf{v}$, and the second lemma gives a simple condition for this eigenvector to be contained in two transmit spaces.

Lemma 6. *Let $\mathbf{A} \in \mathcal{H}_0$. If $W \subseteq V_i$ is an $\mathbf{A}$-invariant subspace, i.e. $W = \mathbf{A}W$, then W contains an eigenvector of $\mathbf{A}$.*

Proof. The proof is slightly encumbered due to $\mathbb{R}$ not being algebraically closed (otherwise it would be immediate). Let us denote by $A : \mathbb{R}^N \to \mathbb{R}^N$ the operator represented by $\mathbf{A}$ in the standard basis and $A|_W$ the restriction of A to W. By definition the characteristic polynomial $c(x)$ vanishes on A, and hence also $c(A|_W) = 0$. But due to $\mathbf{A}$ being diagonal, $c(x) = \prod(\lambda_i - x)^{m_i}$ is a product of linear factors with real λ_i, implying that $(\lambda_i - A|_W)$ is rank-deficient for some λ_i. It follows that $A|_W$, and thus A, has an eigenvector in W. □

Lemma 7. *Let $\mathbf{A}, \mathbf{B} \in \mathcal{H}_0$. If $W \subseteq V_i$ is an $\mathbf{A}$-invariant subspace and $\mathbf{B}W \subseteq V_j$ for some $j \neq i$, then the decoding condition (2.4) is not satisfied at receiver i.*

Proof. By Lemma 6, W contains an eigenvector $\mathbf{v}$ of A. Generically all matrices in $\mathcal{H}_0$ have the same eigenspaces, so $\mathbf{B}\mathbf{v} = \mathbf{v}$ and thus V_j also contains $\mathbf{v}$. As noted immediately before Lemma 6, this implies that the decoding condition (2.4) is not satisfied. □

We now prove Lemma 3, bounding the alignment depth D.

Proof of Lemma 3. The goal is to show that $V_{i\uparrow D} = \{0\}$, which we do for $i = 3$ using an argument that applies also to $i = 1$ and 2 by permuting the indices. We will suppose that $\dim V_{3\uparrow D} \geq 1$ and make use of Lemma 7 to show that the decoding constraint (2.4) is violated at some receiver.

Separating out the first term V_3 from $V_{3\uparrow n}$ as in (2.9), we have that

$$V_{3\uparrow D} = V_3 \cap \mathbf{S}_3^{-1} V_{1\uparrow D-1}\,,$$

and since we assumed $\dim V_{3\uparrow D} \geq 1$, there exists a nonzero vector

$$\mathbf{x} \in \mathbf{S}_3 V_3 \cap V_{1\uparrow D-1} \subseteq V_1\,.$$

The vector $\mathbf{v}_3 := \mathbf{S}_3^{-1}\mathbf{x} \in V_3$ initiates an alignment path of length $D+1$,

$$\mathbf{S}_3^{-1}\mathbf{x},\ \mathbf{x},\ \mathbf{S}_1\mathbf{x},\ \mathbf{S}_2\mathbf{S}_1\mathbf{x},\ \mathbf{S}_3\mathbf{S}_2\mathbf{S}_1\mathbf{x},\ \ldots$$

The point is that now all these vectors are just transformed versions of a *single* vector $\mathbf{x}$, and the finite dimensionality L of the channel space comes into play.

By the definition of $V_{1\uparrow D-1} = \cap_{j=0}^{D-1} \mathbf{S}_{1\uparrow j}^{-1} V_{1+j}$, for each $0 \leq j \leq D-1$ the vector $\mathbf{S}_{1\uparrow j}\mathbf{x}$ is contained in V_{1+j}, and grouping these D vectors according to transmitter gives the lists

$$\begin{aligned}
\mathcal{A}_1 &:= \{\mathbf{S}_{1\uparrow 3\ell}\mathbf{x}\,, 0 \leq \ell \leq n_1\} \subseteq V_1\,,\\
\mathcal{A}_2 &:= \{\mathbf{S}_{1\uparrow 3\ell+1}\mathbf{x}\,, 0 \leq \ell \leq n_2\} \subseteq V_2\,,\\
\mathcal{A}_3 &:= \{\mathbf{S}_{1\uparrow 3\ell+2}\mathbf{x}\,, 0 \leq \ell \leq n_3\} \subseteq V_3\,,
\end{aligned}$$

where

$$n_i := \lfloor (D-i)/3 \rfloor\ .$$

Aside from the vector $\mathbf{v}_3 = \mathbf{S}_3^{-1}\mathbf{x} \in V_3$ at transmitter 3, all the other vectors in $\mathcal{A}_3$ are (by definition of the $\mathbf{S}_{i\uparrow j}$ operators) aligned at receiver 1 with the vectors $\mathcal{A}_2$ from transmitter 2, i.e.

$$\mathbf{H}_{31}\,\mathrm{span}\{\mathcal{A}_3\} \subseteq \mathbf{H}_{21}\,\mathrm{span}\{\mathcal{A}_2\}\,.$$

The result of the alignment at receiver 1 is that we need only concern ourselves with the vectors in $\mathcal{A}_1, \mathcal{A}_2$, and $\mathbf{v}_3 \in V_3$: the vectors in $\mathcal{A}_3$ are redundant. Since we are interested in the view from receiver 1, we define $\mathcal{H}_1$ and $\mathcal{H}_2$ to be the operators in $\mathcal{A}_1, \mathcal{A}_2$ premultiplied by $\mathbf{H}_{11}, \mathbf{H}_{21}$, respectively,

$$\begin{aligned}
\mathcal{H}_1 &:= \{\mathbf{H}_{11}\mathbf{S}_{1\uparrow 3\ell}\,, 0 \leq \ell \leq n_1\}\\
\mathcal{H}_2 &:= \{\mathbf{H}_{21}\mathbf{S}_{1\uparrow 3\ell+1}\,, 0 \leq \ell \leq n_2\}\,.
\end{aligned}$$

The argument rests on the following claim, which shows that the *space of channels* is spanned by the operators in $\mathcal{H}_1$ and $\mathcal{H}_2$, thereby implying linear dependence with $\mathbf{H}_{31}\mathbf{S}_3^{-1}$.

Claim 8. *The operators $\mathcal{H}_1$ and $\mathcal{H}_2$ together generically span the L-dimensional space of channels.*

Proof. The claim is trivially true for $L = 1$ since $\mathcal{H}_1$ is nonempty, so we assume that $L \geq 2$ and hence $D = 2L - \lfloor L/2 \rfloor - 1 \geq 2$ and $n_1, n_2 \geq 0$. A slightly tedious calculation shows that $2 + n_1 + n_2 = L$, so the set $\mathcal{H}_1 \cup \mathcal{H}_2$ consists of L operators. We show that these operators are linearly independent by restricting attention to the upper left $L \times L$ corner of each matrix and applying the linear independence lemma of Section 2.5. Clearly, linear independence of these submatrices implies the claim. Let $[\,\cdot\,]_{L\times L}$ denote the upper left $L \times L$ corner of a square matrix of size at least $L \times L$.

For each $0 \leq \ell \leq n_1$, let $\mathbf{B}_\ell = [\mathbf{S}_{1\uparrow 3\ell} = \mathbf{S}^\ell]_{L\times L}$, and similarly for $0 \leq \ell \leq n_2$ let $\mathbf{A}_\ell = [\mathbf{H}_{21}\mathbf{S}_{1\uparrow 3\ell+1}]_{L\times L} = [\mathbf{H}_{21}\mathbf{S}_1\mathbf{S}^\ell]_{L\times L}$.

Now, define the partitions $P_1 = \{n_1 + 2, \ldots, L\}$ and $P_2 = \{1, \ldots, n_1 + 1\}$. The vectors $\mathbf{B}_\ell|_{P_2}$ (here restricted to the first $|P_2| = n_1 + 1$ entries) are linearly independent because they form the columns of a Vandermonde matrix. Similarly, the $\mathbf{A}_\ell|_{P_1}$ (restricted to the last $|P_1| = n_2 + 1$ entries) arise from an invertible transformation applied to a Vandermonde matrix and are therefore also linearly independent.

The claim follows by applying Lemma 9, with $\mathbf{v}_1 = (1, 1, \ldots, 1)^t$ and $\mathbf{v}_2 = ([\mathbf{H}_{11}]_{L\times L})\mathbf{v}_1$, first $|P_1|$ operators given by $\mathbf{A}_\ell$ and next $|P_2|$ operators given by $\mathbf{B}_\ell$. □

Claim 8 shows that the operators in $\mathcal{H}_1 \cup \mathcal{H}_2$ span the channel space, so there exist real numbers λ_j, μ_j, not all zero, such that

$$\mathbf{H}_{31}\mathbf{S}_3^{-1} = \sum_{j=0}^{n_1} \lambda_j \mathbf{H}_{11}\mathbf{S}_{1\uparrow 3j} + \sum_{j=0}^{n_2} \mu_j \mathbf{H}_{21}\mathbf{S}_{1\uparrow 3j+1}\,. \tag{2.13}$$

The linear dependence between operators will be used to show linear dependence between interfering and desired signal vectors.

We first rule out the case that all the λ_j coefficients are zero. If this were the case, then inserting $\mathbf{x}$ into equation (2.13) on the right reads

$$\mathbf{H}_{31}\mathbf{S}_3^{-1}\mathbf{x} = \mathbf{H}_{31}\mathbf{v}_3 = \sum_{j=0}^{n_2} \mu_j \mathbf{H}_{12}\mathbf{S}_{1\uparrow 3j+1}\mathbf{x}\,,$$

and multiplying through by $\mathbf{H}_{31}^{-1}$ yields

$$\mathbf{v}_3 = \sum_{j=0}^{n_2} \mu_j \mathbf{S}_2\mathbf{S}_{1\uparrow 3j+1}\mathbf{x} = \sum_{j=0}^{n_2} \mu_j \mathbf{S}_{1\uparrow 3j+2}\mathbf{x} = \sum_{j=1}^{n_2+1} \mu_j \mathbf{S}^j(\mathbf{S}_3^{-1}\mathbf{x}) = \sum_{j=1}^{n_2+1} \mu_j \mathbf{S}^j\mathbf{v}_3\,, \tag{2.14}$$

where

$$\mathbf{S} := \mathbf{S}_3\mathbf{S}_2\mathbf{S}_1 = \mathbf{H}_{12}^{-1}\mathbf{H}_{32}\mathbf{H}_{31}^{-1}\mathbf{H}_{21}\mathbf{H}_{23}^{-1}\mathbf{H}_{13}\,.$$

Now, consider the subspace

$$W_3 = \operatorname{span}\{\mathbf{v}_3, \mathbf{S}\mathbf{v}_3, \ldots, \mathbf{S}^{n_2+1}\mathbf{v}_3\} \subseteq V_3\,.$$

By (2.14), $W_3 \subseteq V_3$ is $\mathbf{S}$-invariant, and $\mathbf{S}_2^{-1} W_3 \subseteq V_2$ so we are in the setting of Lemma 7 with $\mathbf{A} = \mathbf{S}$ and $\mathbf{B} = \mathbf{S}_2^{-1}$, which shows that the decoding constraint at receiver 3 is violated. Hence we may assume in (2.13) that $\lambda_j \neq 0$ for some j.

Letting $\mathbf{v}_2 := -\sum_{j=0}^{n_1} \mu_j \mathbf{S}_{1\uparrow 3j+1}\mathbf{x} \in V_2$ and $\mathbf{v}_1 := \left(\sum_{j=0}^{n_1} \lambda_j \mathbf{S}_{1\uparrow 3j}\right)\mathbf{x} \in V_1$, we can rearrange (2.13) to give

$$\mathbf{H}_{11}\mathbf{v}_1 = \mathbf{H}_{12}\mathbf{v}_2 + \mathbf{H}_{13}\mathbf{v}_3\,. \tag{2.15}$$

If $\mathbf{v}_1$ were zero, this would as before imply the existence of an $\mathbf{S}$-invariant subspace

$$W_1 = \mathrm{span}\{\mathbf{x}, \mathbf{S}\mathbf{x}, \ldots, \mathbf{S}^{n_1}\mathbf{x}\} \subseteq V_1$$

with $\mathbf{S}_3^{-1} W_1 \subseteq V_3$, a situation ruled out by Lemma 7. But (2.15) precisely means that the decoding constraint is violated at receiver 1, which was what we set out to prove. □

2.5 Linear independence lemma

In this section we prove a lemma that allows to translate linear independence among *operators* to linear independence among a collection of operators applied to a collection of vectors. Since all matrices in this chapter are diagonal, the lemma also applies when the "vectors" are actually matrices. The proof itself is not particularly insightful and entails manipulating a determinant expression.

Lemma 9 (Linear independence)**.** *Let N be a positive integer and $T^{(i)} : \mathbb{R}^N \to \mathbb{R}^N, 1 \leq i \leq N$ be linear operators each given by a diagonal matrix $\mathbf{T}^{(i)} = \mathrm{diag}(t_1^{(i)}, t_2^{(i)}, \ldots, t_N^{(i)})$. Let $\ell_1, \ldots, \ell_q$ denote an integer partition of N into q parts (some may be empty), i.e. the $\{\ell_i\}$ are nonnegative and $\sum \ell_i = N$.*

The vectors

$$\mathbf{T}^{(1)}\mathbf{v}_1\,, \ldots\,, \mathbf{T}^{(\ell_1)}\mathbf{v}_1\,, \mathbf{T}^{(\ell_1+1)}\mathbf{v}_2, \ldots, \mathbf{T}^{(\ell_1+\ell_2)}\mathbf{v}_2, \mathbf{T}^{(\ell_1+\ell_2+1)}\mathbf{v}_3, \ldots, \mathbf{T}^{(N)}\mathbf{v}_q \tag{2.16}$$

are linearly independent for $\mathbf{v}_1 = (1, 1, \ldots, 1)^t$ and generic $(\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_q) \in \mathbb{R}^{N\times(q-1)}$ if and only if there exists a set partition $P = \{P_1, P_2, \ldots, P_q\}$ of $\{1, \ldots, N\}$ with $|P_j| = \ell_j$ such that for each j, $1 \leq j \leq q$, the set of vectors

$$\{\mathbf{T}^{(i)}|_{P_j} : \ell_1 + \cdots + \ell_{j-1} + 1 \leq i \leq \ell_1 + \cdots + \ell_j\}$$

is linearly independent. Here $\mathbf{T}^{(i)}|_{P_j}$ denotes the length-$|P_j|$ vector with entries $\{t_r^{(i)}, r \in P_j\}$.

Proof. We will perform manipulations on an expression for the determinant of the $N \times N$ matrix $\mathbf{M}$ with columns given by the vectors in (2.16). The determinant can be written as

$$\det(\mathbf{M}) = \sum_{\pi} \mathrm{sign}(\pi) \prod_{i=1}^{N} \mathbf{M}_{\pi(i),i}\,, \tag{2.17}$$

where the sum is over the permutations on $\{1,\dots,N\}$. It will be convenient to let $Q = \{Q_1,\dots,Q_q\}$ be the partition of $\{1,\dots,N\}$ with

$$Q_j = \{\ell_1+\cdots+\ell_{j-1}+1,\dots,\ell_1+\cdots+\ell_j\}\,,\quad 1\le j\le q\,.$$

Now we may break up each product in (2.17) according to the index of $\mathbf{v}_i$ to get

$$\det(\mathbf{M}) = \sum_\pi \operatorname{sign}(\pi) \prod_{i\in Q_1} \mathbf{M}_{\pi(i),i} \prod_{i\in Q_2} \mathbf{M}_{\pi(i),i} \;\cdots \prod_{i\in Q_q} \mathbf{M}_{\pi(i),i}$$

$$= \sum_\pi \operatorname{sign}(\pi) \prod_{i\in Q_1} t^{(i)}_{\pi(i)} v_{1,\pi(i)} \prod_{i\in Q_2} t^{(i)}_{\pi(i)} v_{2,\pi(i)} \;\cdots \prod_{i\in Q_q} t^{(i)}_{\pi(i)} v_{q,\pi(i)}\,.$$

Denote by $P_\pi = \{P_1,P_2,\dots,P_q\}$ the partition induced by π on Q, i.e. the blocks are given by $P_i = \pi(Q_i)$. Each permutation π that induces the same partition P contributes to the coefficient of the monomial

$$v_P := \prod_{j=1}^q \prod_{i\in Q_j} v_{j,\pi(i)} = \prod_{j=1}^q \prod_{i\in P_j} v_{j,i}\,,$$

and different partitions P give unique monomials. Summing according to partitions P,

$$\begin{aligned}\det(\mathbf{M}) &= \sum_{P\in\mathcal{P}} v_P \sum_{\pi:\pi(Q)=P} \operatorname{sign}(\pi) \prod_{i\in Q_1} t^{(i)}_{\pi(i)} \prod_{i\in Q_2} t^{(i)}_{\pi(i)} \;\cdots \prod_{i\in Q_q} t^{(i)}_{\pi(i)} \\ &= \sum_{P\in\mathcal{P}} v_P \sum_{\pi:\pi(Q)=P} \operatorname{sign}(\pi) \prod_{j=1}^q \prod_{i\in Q_j} t^{(i)}_{\pi(i)}\,.\end{aligned}$$

Fix a partition P and permutation $\bar\pi$ with $\bar\pi(Q) = P$. Then each permutation π with $\pi(Q) = P$ can be written uniquely as $\pi = \pi_1\circ\pi_2\circ\cdots\circ\pi_q\circ\bar\pi$, where $\pi_j|_{P_j} : P_j \to P_j$ is a permutation on P_j and π_j is identity on elements not in P_j. Note that for $i\in Q_j$, $\pi(i) = (\pi_j\circ\bar\pi)(i)$, and so for fixed $\bar\pi$ the map $\pi\mapsto(\pi_1,\pi_2,\dots,\pi_q)$ is a bijection. We fix a choice $\bar\pi_P$ for each partition P, and thus

$$\begin{aligned}\det(\mathbf{M}) &= \sum_{P\in\mathcal{P}} v_P\cdot\operatorname{sign}(\bar\pi_P) \prod_{j=1}^q \left(\sum_{\substack{\pi_j \text{ perm.}\\ \text{on } P_j}} \operatorname{sign}(\pi_j) \prod_{i\in Q_j} t^{(i)}_{(\pi_j\circ\bar\pi)(i)} \right) \\ &= \sum_{P\in\mathcal{P}} v_P\cdot\operatorname{sign}(\bar\pi_P) \prod_{j=1}^q \det(\mathbf{W}_j(P_j))\,, \qquad (2.18)\end{aligned}$$

where

$$\mathbf{W}_j(P_j) = [\mathbf{T}^{(i)}|_{P_j}]_{i\in Q_j}$$

is a matrix with column i given by the vector $\mathbf{T}^{(i)}|_{P_j}$ (and $\mathbf{T}^{(i)}|_{P_j}$ is defined in the statement of the lemma).

The polynomial (2.18) in variables $v_{i,j}$ is not identically zero if and only if at least one monomial is nonzero, which is true if there exists a partition P such that $\det(\mathbf{W}(P_j))\neq 0$ for $1\le j\le q$. This precisely matches the lemma statement. □

Chapter 3

Feasibility of interference alignment for the multiple-user MIMO interference channel

3.1 Introduction

In this chapter we continue our study of interference alignment and the role of channel diversity. As discussed in Chapter 2, interference alignment is a promising approach to mitigating interference in wireless networks, but the ability to align depends on channel diversity. Cadambe and Jafar [1], in their surprising result, showed that $\frac{K}{2}$ degrees of freedom are achievable for the K-user Gaussian interference channel (IC), assuming the channels had *unbounded* diversity in the form of time or frequency variation. A vital question is therefore: how much diversity is required in order to align interference?

The main result of Chapter 2 shows that for the three-user IC, channel diversity is indeed a precious resource that determines the ability to align interference. Alignment is impossible with no channel diversity, and performance gradually increases as a function of the diversity. Chapter 2 focuses on time and frequency diversity, but most practical systems are also equipped with multiple antennas and thus have *spatial diversity*. Multiple antennas are known to greatly increase the degrees of freedom of point-to-point systems.

In this chapter we focus on how spatial diversity helps to deal with interference by studying the MIMO IC, where each of K transmitters and K receivers has multiple antennas, and each transmitter wishes to communicate with the corresponding receiver. We let M_i and N_i denote the number of antennas of the ith transmitter and the ith receiver respectively. In order to focus on the effect of spatial diversity, we assume there is no time or frequency diversity, i.e. the channel is constant over time and frequency. Similar in flavor to the situation with finite time or frequency diversity in Chapter 2, here we have a fixed amount of spatial diversity and the goal is to design the best communication scheme—achieving the most degrees of freedom—for the system at hand.

To simplify matters, we again restrict attention to vector space schemes. With this restriction, as shown in Section 3.2, the alignment problem reduces to finding vector spaces $U_i \subset \mathbb{C}^{M_i}$ and $V_i \subset \mathbb{C}^{N_i}$ where $\dim U_i = \dim V_i$ is denoted by d_i, and such that

$$\mathbf{H}_{ij} U_i \perp V_j \,, \quad 1 \leq i, j \leq K, \quad i \neq j \,, \tag{3.1}$$

where the matrix $\mathbf{H}_{ij} \in \mathcal{C}^{N_i \times M_j}$ signifies the channel between transmitter j and receiver i. Each entry of $\mathbf{H}_{ij}$ is the gain between one of transmitter i's antennas and one of receiver j's antennas and is in general nonzero. For the rest of the paper we assume that the $\mathbf{H}_{ij}$ are generic, meaning that their entries lie outside of some algebraic hypersurface. If the entries are randomly chosen from some non-singular probability distribution, this will be true with probability 1.

The existence of a receive space V_j satisfying the orthogonality condition (3.1) amounts to a requirement that the interference spaces $\mathbf{H}_{ij} U_j$ are sufficiently aligned that there is room left over for the desired signal space $\mathbf{H}_{jj} U_j$.

Our goal is to maximize the signal dimensions d_i subject to the constraint that there exist vector spaces satisfying (3.1). This is, of course, equivalent to fixing a choice of dimensions d_i and answer the *feasibility* question of whether there exist subspaces satisfying (3.1). This problem was posed by [15], who then proposed a heuristic iterative algorithm but left the question open.

We obtain results for the three user ($K = 3$) MIMO IC, and then partially generalize to an arbitrary number of users. For the symmetric three-user channel, we focus on the symmetric case where $d_i = d$, $M_i = M$, and $N_i = N$ for all i. Theorem 10 below determines the region of M and N for which there exists a valid linear encoding and decoding strategy (as defined in Subsection 3.2); the region is depicted in Fig. 3.1.

Theorem 10 (3-user MIMO IC)**.** *Fix the number of desired transmit dimensions $d_i = d$, transmit antennas $M_i = M$, and receive antennas $N_i = N$. Assume without loss of generality that $N \geq M$. Then alignment is feasible if and only if*

$$(2r+1)d \leq \max(rN, (r+1)M), \quad \textit{for all integers } r \geq 0 \,. \tag{3.2}$$

Just as in the three-user parallel IC of Chapter 2, the concept of alignment depth plays a central role. The maximum alignment depth corresponds to $r + 1$ in Theorem 10 and depends on the ratio of M and N.

Interestingly, the reason for the limitation on alignment depth in the MIMO IC is completely different from in the parallel channel. In the parallel channel, recall that aligning the signals is easy; the challenge is ensuring that the desired signal is distinguishable at each receiver. Accordingly, if the diversity is too low, the direct channel is not sufficiently rich to send the desired signal to a part of the space without interference, even if there are plenty of free dimensions.

In the MIMO case, it is the *alignment* that is difficult. Once the interference space is small enough, the direct channels (as given by full generic matrices) are such that the

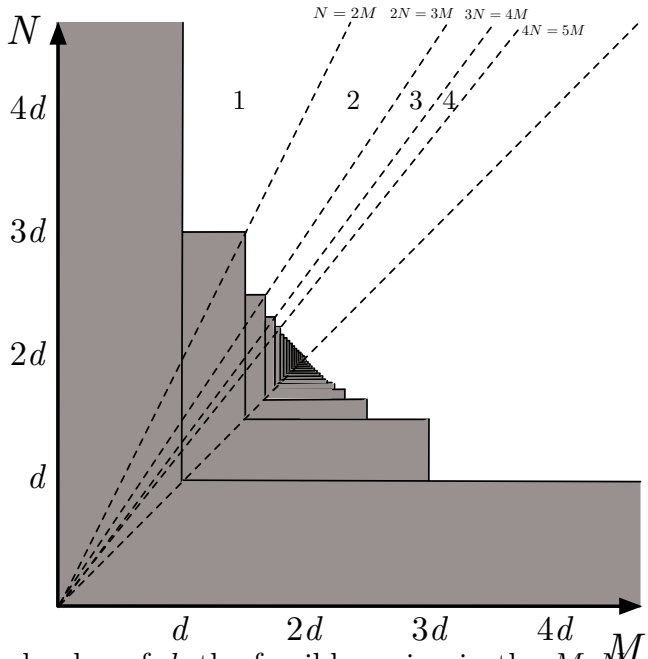


Figure 3.1: For a fixed value of d, the feasible region in the M, N plane is white while the infeasible region is shaded. The labels $1, 2, 3, 4, ...$ indicate the maximum alignment depth for M, N in the corresponding region.

receivers can always distinguish the desired signal. This is why the decoding constraint which appeared so frequently in Chapter 2, requiring that at each receiver the interference space is complementary to the desired signal space, does not appear in (3.1). It is only necessary to ensure that the interference space is not too large.

A simple example illustrates what restricts alignment in the MIMO situation. Suppose $2M < N$, e.g. $M = 3, N = 8$. Then the images $\mathrm{Im}(\mathbf{H}_{13}), \mathrm{Im}(\mathbf{H}_{23})$ of the channels from the pair of transmitters 1 and 2 to receiver 3, have trivial intersection. This means that no alignment whatsoever is possible: vectors cannot be selected at the transmitters in order to overlap at the interfered receiver. Since all three dimension d signal spaces are complementary at each receiver, this leads to the constraint $3d \leq N$. In terms of alignment depth, we see that a maximum depth of 1 is possible. This style of reasoning can be extended to produce all the constraints in Theorem 10, and is discussed at a heuristic level in Section 3.3 (with proofs in later sections).

Next, we generalize to K-users, continuing our investigation of feasibility. The work of Yetis et al. [16] proposed comparing the number of variables and equations in the system of bilinear equations (3.1) in order to determine when it has solutions. We make this precise by showing that the feasible solutions are an algebraic variety of the "expected" dimension, when the channel matrices are generic. Thus, we have the following necessary condition for interference alignment:

Theorem 11 (General necessary condition)**.** *Fix an integer K and integers d_i, M_i, and N_i for $1 \leq i \leq K$ and suppose the channel matrices $\mathbf{H}_{ij}$ are generic. If, for any subset $A \subset \{1, \ldots, K\}$, the quantity*

$$t_A = \sum_{i \in A} \Big(d_i(N_i - d_i) + d_i(M_i - d_i) \Big) - \sum_{i,j \in A, i \neq j} d_i d_j.$$

is negative, then the alignment equations (3.1) are infeasible. Moreover, if there are feasible solutions, then $t_{\{1,\ldots,K\}}$ is the dimension of the variety of solutions.

The constraint on t_A was obtained independently and simultaneously by Razaviyayn et al. [17, 18]. We note that the dimension of the variety of solutions is important, because when multiple strategies are feasible, we might wish to optimize over the feasible strategies according to some other criterion, such as the robustness of the system.

The necessary condition from Theorem 11 is not sufficient. For example, one additional requirement for there to even exist vector spaces $U_i \subseteq \mathbb{C}^{M_i}$ and $V_i \subseteq \mathbb{C}^{N_i}$ is that $d_i \leq M_i$ and $d_i \leq N_i$ for each i. However, in the fully symmetric (square) case that $M = N$, the necessary condition of Theorem 11 is also a sufficient condition.

Theorem 12 (Sufficiency for fully symmetric case)**.** *Suppose that $K \geq 3$ and furthermore that $d_i = d$ and $M_i = N_i = N$ for all users i. Then, for generic channel matrices $\mathbf{H}_{ij}$, the space of feasible schemes is non-empty and has dimension $Kd(2N - (K+1)d)$, if this quantity is non-negative, and is empty if it is negative. Thus, alignment is feasible if and only if*

$$N \geq \frac{d(K+1)}{2}.$$

We emphasize that all our results apply only to generic matrices. This means that there exists an open dense subset of the space of matrices (in fact the complement of an algebraic hypersurface) on which these statements hold. In particular, matrices chosen from a non-singular probability distribution will be sufficiently generic with probability one. On the other hand, specific matrices, such as $\mathbf{H}_{ij} = 0$ for $i \neq j$, may lead to a different answer.

Rearranging the inequality of Theorem 12, we have that the number of transmit dimensions satisfies $d \leq \frac{2N}{K+1}$, so the total normalized dof is $\frac{Kd}{N} = \frac{K}{N} \left\lfloor \frac{2N}{K+1} \right\rfloor \leq 2\frac{K}{K+1}$.

Corollary 13 (Fully symmetric achievable DoF)**.** *The maximum normalized dof is given by*

$$\mathsf{DoF} = \frac{K}{N} \left\lfloor \frac{2N}{K+1} \right\rfloor \leq 2\frac{K}{K+1}.$$

In sharp contrast to the $\frac{K}{2}$ total normalized degrees of freedom achievable for infinitely many parallel channels in [1], for the MIMO case we see that at most 2 degrees of freedom (normalized by the single-user performance of N transmit dimensions) are achievable for any number of users K and antennas N.

Theorem 12 suggests an engineering interpretation for the performance gain from increasing the number of antennas. Depending on whether $N < d(K+1)/2$ or not, there are two types of performance benefit from increasing N: (1) *alignment gain* or (2) *MIMO gain*. To illustrate these concepts, suppose that there are $K = 5$ users. If $N = 1$, i.e. there is only a single antenna at each node, then no alignment can be done and only one user can communicate on a single dimension, giving 1 total degree of freedom (dof). Increasing to $N = 2$ antennas allows three users to communicate with one dimension each, giving a total normalized dof $Kd/N = 3/2$. Similarly, increasing to $N = 3$ antennas allows each of the five users to use $d = 1$ dimension, giving a normalized dof $= 5/3$. Thus, each increase in N until $N = (K+1)/2 = 3$ leads to additional users able to transmit, and a gain of two dimensions per additional antenna; this is *alignment gain*. From here, however, increasing N has a different effect. If we double N to $N = 6$, there are still only 5 users, and each can now transmit along $d = 2$ dimensions instead of one, but the normalized dof remains at $5 \cdot 2/6 = 5/3$. The total dof increases at a slower rate: the increase is not due to more alignment being possible, but simply because more total dimensions are available. This is *MIMO gain*.

We note that unlike Theorem 10, Theorem 12 does not provide a way of computing the solutions. Instead of the linear algebra used to prove Theorem 10, Theorem 12 proves only the existence of solutions, using algebraic geometry. Nonetheless, solutions may be found numerically using homotopy continuation software. In addition to algebraic methods of root finding, others have proposed heuristic algorithms, mainly iterative in nature (see [19], [20], [21], [22], and [23]). Some have proofs of convergence, but no performance guarantees are known. Schmidt et al. [23], [24] study a refined version of the single-transmit dimension problem, where for the single-transmit dimension case ($d = 1$) they attempt to choose a good solution among the many possible solutions.

We briefly review the related work before giving an outline of the chapter.

Related work

The problem we consider, of maximizing degrees-of-freedom using linear strategies for the K-user MIMO IC, has received significant attention in the last several years. Jafar and Fakhereddin [25] determined the degrees of freedom of the two-user MIMO IC with an arbitrary number of antennas at each of the four terminals. Cadambe and Jafar [1] considered the problem for $K = 3$ users and $N = 2$ antennas, and showed that 3/2 dof was achievable. For more than 3 users or $N > 2$ they assumed infinite time or frequency diversity and applied their main $K/2$ result. As noted earlier, Gomadam et al. [15,19], posed the problem of determining feasibility of linear alignment in the constant channel setting and developed a heuristic iterative numerical algorithm, but left the problem unanswered.

Razaviyayn et al. [17], [18], have independently and simultaneously found results related to ours. They prove a necessary condition which corresponds to our necessary condition in Theorem 11, and they also have a matching sufficient condition for the special case where

$d_i = d$, and d divides M_i and N_i for each i. The symmetric square case $M_i = N_i = N, d_i = d$ we consider in the present chapter is not covered by their result.

In a different direction of inquiry, Razaviyayn et al. [21] show that checking the feasibility of alignment for general system parameters is NP-hard. Note that their result is not in contradiction to ours, since our simple closed-form expression applies only to the fully symmetric case.

For the symmetric three-user channel with M transmit and N receive antennas, Amir et al. [26] have independently proposed a similar achievable strategy for critical M, N satisfying both (3.2) and $M+N = 4d$. [26] is limited to critical values of M, N and contains no converse arguments beyond the equation counting bound of [27] and [17]. Also independently, Wang et al. [28] very recently posted a paper to the Arxiv containing many similar results. Their converse is information theoretic and, unlike ours, is not limited to linear strategies.

We emphasize that in this chapter we restrict attention to vector space schemes, where the effect of finite channel diversity can be observed. Interfering signals can also be aligned on the signal scale using lattice codes (first proposed in [29], see also [30], [31], [32], [33]), however the understanding of this type of alignment is currently at the stage corresponding to infinite parallel channels in the vector space setting. In other words, essentially "perfect" alignment is possible due to the infinite channel precision available at infinite signal-to-noise ratios. Recent progress on signal scale alignment at finite SNR includes [34], [35], [36], and [37].

Ghasemi et al. [33] apply alignment on the signal scale to the K-user symmetric $M \times N$ MIMO IC. The converse arguments in that paper are obtained by forming a two-user interference channel with two users transmitting and decoding jointly; they obtain the inequality $3d \leq \max(N, 2M)$ corresponding to $r = 1$ in (3.2) of the present chapter.

Outline

The rest of the chapter is organized as follows. Section 3.2 describes the channel model and vector space communication schemes. Section 3.3 is devoted to the three-user channel, and contains a heuristic description and proof of Theorem 10. We generalize to K-users in Section 3.4, proving Theorems 11 and 12.

3.2 Formulation

In this section we describe the MIMO interference channel model, vector space communication schemes, and degrees of freedom. The formulation of the alignment feasibility problem is at the end of the section.

Interference channel model

The K-user MIMO interference channel has K transmitters and K receivers, with transmitter i having M_i antennas and receiver i having N_i antennas. For $i = 1, \ldots, K$, receiver i wishes to obtain a message from the corresponding transmitter i. The remaining signals from transmitters $j \neq i$ are undesired interference. The channel is assumed to be constant over time, and at each time-step the input-output relationship is given by

$$\mathbf{y}_i = \mathbf{H}_{ii}\mathbf{x}_i + \sum_{\substack{1 \leq j \leq K \\ j \neq i}} \mathbf{H}_{ji}\mathbf{x}_j + \mathbf{z}_i\,, \quad 1 \leq i \leq K\,. \tag{3.3}$$

Here for each user i we have $\mathbf{x}_i \in \mathbb{C}^{M_i}$ and $\mathbf{y}_i, \mathbf{z}_i \in \mathbb{C}^{N_i}$, with $\mathbf{x}_i$ the transmitted signal, $\mathbf{y}_i$ the received signal, and $\mathbf{z}_i \sim \mathcal{CN}(0, I_{N_i})$ is additive isotropic white Gaussian noise. The channel matrices are given by $\mathbf{H}_{ji} \in \mathbb{C}^{N_i \times M_j}$ for $1 \leq i, j \leq K$, with each entry assumed to be independent and with a continuous distribution. We note that this last assumption on independence can be weakened significantly to a basic non-degeneracy condition but we will not pursue this here. For our purposes this means the channel matrices are generic. Each user has an average power constraint, $E(||\mathbf{x}_i||^2) \leq P$.

Vector space strategies and degrees-of-freedom

We restrict the class of coding strategies to *vector space* schemes. Suppose transmitter j wishes to transmit a vector $\hat{x}_j \in \mathbb{C}^{d_j}$ of d_j data symbols. These data symbols are modulated on the subspace $U_j \subseteq \mathbb{C}^{M_j}$ of dimension d_j, giving the input signal $\mathbf{x}_j = \mathbf{U}_j\hat{x}_j$, where $\mathbf{U}_j$ is a $M_j \times d_j$ matrix whose column span is U_j. The signal $\mathbf{x}_j$ is received by receiver i through the channel as $\mathbf{H}_{ji}\mathbf{U}_j\hat{x}_j$ The dimension of the transmit space, d_j, determines the number of data streams, or degrees-of-freedom, available to transmitter j. With this restriction to vector space strategies, the output is given by

$$\mathbf{y}_i = \mathbf{H}_{ii}\mathbf{U}_i\hat{x}_i + \sum_{\substack{1 \leq j \leq K \\ j \neq i}} \mathbf{H}_{ji}\mathbf{U}_j\hat{\mathbf{x}}_j + \mathbf{z}_i\,, \quad 1 \leq i \leq K\,. \tag{3.4}$$

The desired signal space at receiver i is thus $\mathbf{H}_{ii}U_i$, while the interference space is given by $\sum_{j \neq i} \mathbf{H}_{ji}U_j$, i.e. the span of the undesired subspaces as observed by receiver i.

In the regime of asymptotically high transmit powers, in order that decoding can be accomplished we impose the constraint at each receiver i that the desired signal space $\mathbf{H}_{ii}U_i$ is complementary to the interference space $\sum_{j \neq i} \mathbf{H}_{ji}U_j$. Equivalently, there must exist subspaces V_i with $\dim V_i = \dim U_i$ such that

$$\mathbf{H}_{ji}U_j \perp V_i\,, \quad 1 \leq i, j \leq K, \quad i \neq j\,, \tag{3.5}$$

and

$$\dim(\mathrm{Proj}_{V_i}\mathbf{H}_{ii}U_i) = \dim U_i\,. \tag{3.6}$$

Here $\mathbf{H}_{ji}U_j \perp V_i$ is interpreted to mean that V_i belongs to the dual space $(\mathbb{C}^{N_i})^*$ and V_i annihilates $\mathbf{H}_{ji}U_j$. Alternatively, $(\mathbf{V}_i)^\dagger \mathbf{H}_{ji}\mathbf{U}_j = 0$, where $\mathbf{V}^\dagger$ denotes the Hermitian transpose of $\mathbf{V}$ and $\mathbf{V}_i$ is a matrix with column span equal to V_i. Note that implicitly the transmit dimensions are assumed to satisfy the obvious inequality $d_i \leq \min(M_i, N_i)$. If each direct channel matrix $\mathbf{H}_{ii}$ has generic entries, then the second condition (3.6) is satisfied assuming $\dim V_i = d_i$ for each i (this can be easily justified—see [15] for some brief remarks). Hence we focus on condition (3.5).

The goal is to maximize degrees of freedom, i.e. choose subspaces $U_1, \ldots, U_K$, $V_1, \ldots, V_K$ with $d_i \leq \min(M_i, N_i)$ in order to

$$\begin{aligned} &\text{maximize} \quad d_1 + d_2 + \cdots + d_K \\ &\text{subject to} \quad \mathbf{H}_{ji}U_j \perp V_i\,, \quad 1 \leq i, j \leq K, \quad i \neq j\,, \end{aligned}$$

To this end, it is sufficient to answer the following feasibility question: given number of users K, number of antennas $M_1, \ldots, M_K$, $N_1, \ldots, N_K$, and desired transmit subspace dimensions $d_1, \ldots, d_K$, does there exist a choice of subspaces $U_1, \ldots, U_K$ and $V_1, \ldots, V_K$ with $\dim U_i = \dim V_i = d_i, 1 \leq i \leq K$, satisfying (3.5)?

3.3 Three-user channel

We begin with an informal overview of the arguments, followed by a proof of the converse and then achievability.

Heuristic description

We attempt to give an intuitive argument for the constraints in Theorem 10 in terms of the depth (or length) of alignment paths.

A given vector u_i in the signal space of transmitter i is said to initiate an alignment path of depth $r+1$ if there exists a sequence of vectors $u_{i+1}, u_{i+2}, \ldots, u_{i+r} \in \mathbb{C}^M$, such that

$$\mathbf{H}_{i,i-1}u_i = \mathbf{H}_{i+1,i-1}u_{i+1}, \ldots, \mathbf{H}_{i+r-1,i+r-2}u_{i+r-1} = \mathbf{H}_{i+r,i+r-2}u_{i+r}\,.$$

Here channel indices are interpreted modulo 3. For example, a vector u_2 at transmitter 2 initiating an alignment path of depth 3 means that there exist vectors u_3 and u_1 such that $\mathbf{H}_{21}u_2 = \mathbf{H}_{31}u_3$ and $\mathbf{H}_{32}u_3 = \mathbf{H}_{12}u_1$.

The feasible region of Figure 3.1 is divided up into sub-regions labeled with the maximum depth of an alignment path; this number depends on M and N through the incidence geometry of the images of the channel matrices $\mathrm{Im}(\mathbf{H}_{ij})$. We begin by examining sub-region 1, and then look at how things generalize to the other sub-regions.

The point of departure is the obvious constraint $d \leq M$ in order to have a d-dimensional subspace of an M dimensional vector space. Continuing, assuming $M \geq d$, suppose $2M \leq N$, so (M, N) lies in sub-region 1 of Figure 3.1. At receiver one, the images $\mathrm{Im}(\mathbf{H}_{21})$ and $\mathrm{Im}(\mathbf{H}_{31})$

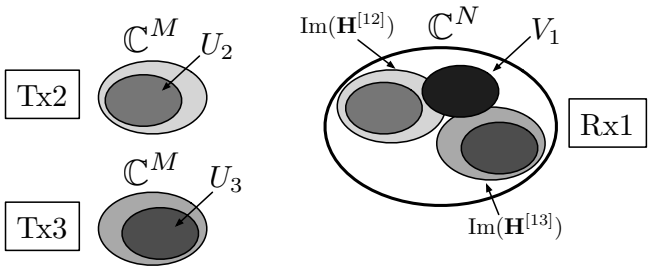


Figure 3.2: Sub-region 1: The figure indicates that no alignment is possible when $2M \leq N$, since $\text{Im}(\mathbf{H}_{12})$ and $\text{Im}(\mathbf{H}_{13})$ are complementary. Since the three subspaces $V_1, \mathbf{H}_{12}U_2, \mathbf{H}_{13}U_3$ are each of dimension d, complementary, and lie in $\mathbb{C}^N$ at receiver 1, we obtain the constraint $3d \leq N$.

of the channels from transmitters two and three are in general position and therefore their intersection has dimension $[2M - N]^+ = 0$; in other words, *alignment is impossible* in sub-region 1. Figure 3.2 shows pictorially that because alignment is not possible here, we have the constraint $3d \leq N$. Mathematically, we see that alignment is not possible because the map from $\mathbb{C}^{2M}$ to $\mathbb{C}^N$ given by the matrix $\begin{pmatrix}\mathbf{H}_{21} & \mathbf{H}_{31}\end{pmatrix}$ is injective.

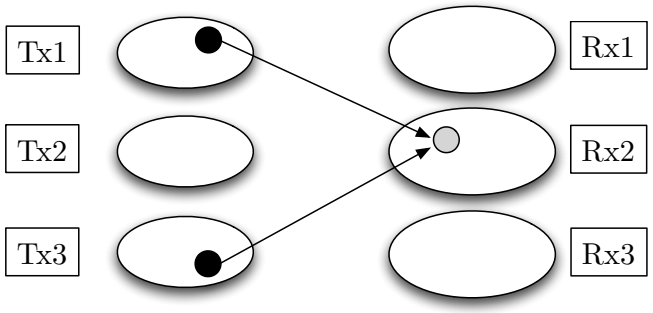


Figure 3.3: Sub-region 2: Alignment is possible here. The figure denotes an alignment path of depth 2.

Moving onward to sub-region 2, we have $2M > N$ and thus alignment *is* possible. This means that alignment paths of depth 2 are possible (Fig 3.3), with up to $2M - N$ interference dimensions overlapping at each receiver. Thus, the interference space $\mathbf{H}_{21}U_2 + \mathbf{H}_{31}U_3$ at receiver one occupies at least $2d - (2M - N)$ dimensions, and we have the constraint $3d \leq 2M$. However, because $3M \leq 2N$, no vector at (say) transmitter three can be *simultaneously* aligned at both receivers one and two, as indicated in Figure 3.4. One can also see that no simultaneous alignment is possible by changing change perspective to that of a combined

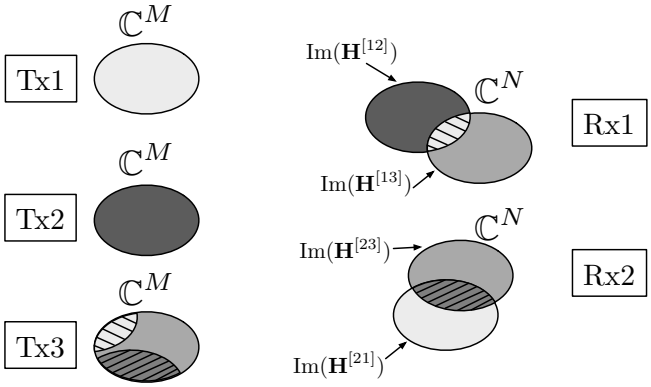


Figure 3.4: Sub-region 2: The striped regions at receivers one and two each denote the dimension $2M - N$ portion of the space in which alignment can occur. From transmitter three's perspective, one sees that *simultaneous* alignment is not possible for $2(2M-N) \leq M$, or equivalently, $3M \leq 2N$.

receiver one and two. One may check directly that (as a special case of Lemma 14), the map

$$\begin{pmatrix} \mathbf{H}_{21} & \mathbf{H}_{31} \\ & \mathbf{H}_{32} \quad \mathbf{H}_{12} \end{pmatrix} \tag{3.7}$$

from the three transmitters to $\mathbb{C}^{2N}$ is injective; analogously to the case in sub-region 1, this is interpreted to mean that no alignment is possible in the combined receive space $\mathbb{C}^{2N}$ (see Fig. 3.5). Thus, five complementary d-dimensional subspaces lie in $\mathbb{C}^{2N}$ and we obtain the constraint $5d \leq 2N$.

As far as achievability goes, the basic rule-of-thumb is to create alignment paths of maximum depth. Thus, in sub-region 2, where alignment is possible, the achievable strategy aligns as many vectors as possible and the remaining ones (if $d > 2(2M-N)$) are not aligned.

Both the necessary conditions and achievability arguments extend in a natural way. On the achievability end, alignment paths of maximum depth are used. For example, in sub-region 4, alignment paths of depth four are used (Fig 3.6). For the converse, a generalization of the matrix in (3.7) is shown to be full-rank in Lemma 14, giving the constraints in (3.2).

Proof of converse

We now proceed with the proof of the converse of Theorem 10. We begin with a key lemma, which we will also use for the achievability direction. We introduce two notational conveniences to be used throughout this section. As noted before, we interpret the indices modulo three, so that $\mathbf{H}_{12} = \mathbf{H}_{42}$ and so on. Since the indices can always be chosen to differ by exactly one, we will adopt the shorthand $\mathbf{H}_{i,+}$ and $\mathbf{H}_{i,-}$ for $\mathbf{H}_{i,i+1}$ and $\mathbf{H}_{i,i-1}$ respectively.

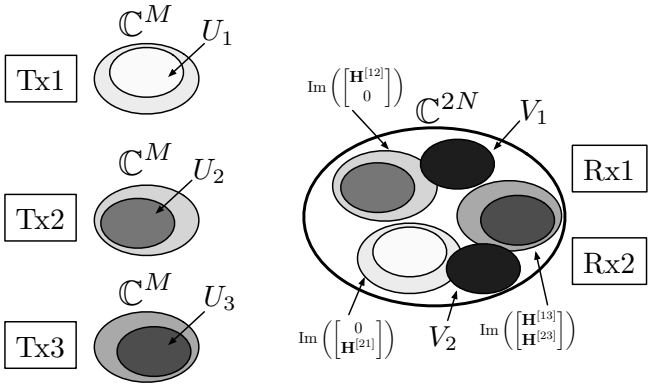


Figure 3.5: Sub-region 2: Considering the dimension $2N$ receive space formed by receivers one and two together, along with the map defined (3.7) from the three transmitters, shows that no alignment is possible in this combined space. Since there are five complementary subspaces of dimension d we obtain the constraint $5d \leq 2N$.

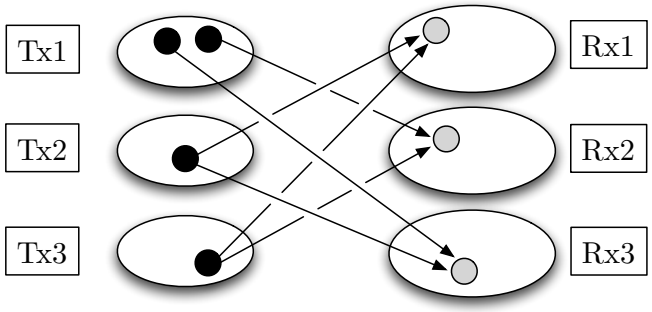


Figure 3.6: Sub-region 4: Alignment paths of depth four are denoted here, initiated by vectors at transmitter 1.

Lemma 14. *Suppose $N \geq M$. For any $r \geq 1$ define the $rN \times (r+1)M$ block matrix*

$$\mathbf{A}_r = \begin{pmatrix} \mathbf{H}_{21} & \mathbf{H}_{31} & & & & \\ & \mathbf{H}_{32} & \mathbf{H}_{12} & & & \\ & & \mathbf{H}_{13} & \mathbf{H}_{23} & & \\ & & & \ddots & & \\ & & & & \mathbf{H}_{r+1,-} & \mathbf{H}_{r+2,+} \end{pmatrix}, \tag{3.8}$$

where the indices are interpreted as described above. For generic channel matrices $\mathbf{H}_{ij}$, the matrix $\mathbf{A}_r$ has full rank, $\min(rN, (r+1)M)$.

Proof. In order to prove that $\mathbf{A}_r$ has full rank for generic channel matrices, it is sufficient to prove that it does for one particular set of matrices (see e.g. [38]). We specialize to the

matrices

$$\mathbf{B} := \mathbf{H}_{1,-} = \mathbf{H}_{2,-} = \mathbf{H}_{3,-} = \begin{pmatrix} \mathbf{I}_M \\ \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{C} := \mathbf{H}_{1,+} = \mathbf{H}_{2,+} = \mathbf{H}_{3,+} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_M \end{pmatrix},$$

where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix and the $\mathbf{0}$ denotes a block of 0s of size $(N-M) \times M$.

We will prove that, with these specializations, $\mathbf{A}_n$ has full rank by simultaneous induction on r, N, and M. If $r = 0$, then $\mathbf{A}_r$ is a $0 \times M$ matrix, which trivially has full rank. If $N \geq 2M$, then every row vector is a unit vector and all such unit vectors appear in some row, so the matrix has full rank.

Now we suppose that $N < 2M$. We permute the rows and columns of $\mathbf{A}_n$ as follows. We extract the first block of N rows, followed by the last $N-M$ rows of each of the subsequent $r-1$ blocks. We put these rows last, after the remaining rows, each in their induced order. Similarly, we take the first block of M columns, followed by the last $N-M$ columns of each of the other r column blocks, and place these to the right of all the other columns. This creates a square matrix of size $M + r(N-M)$ in the lower right, and we will prove that this submatrix is the identity.

If we divide $\mathbf{B}$ and $\mathbf{C}$ into blocks by separating off the last $N-M$ rows and columns of each, then using our assumption that $N - M < M$, we get

$$\mathbf{B} = \begin{pmatrix} \tilde{\mathbf{B}} & \mathbf{B}' \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \mathbf{C} = \begin{pmatrix} \tilde{\mathbf{C}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N-M} \end{pmatrix},$$

where

$$\mathbf{B}' = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{N-M} \end{pmatrix} \quad \tilde{\mathbf{B}} = \begin{pmatrix} \mathbf{I}_{2M-N} \\ \mathbf{0} \end{pmatrix} \quad \tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{2M-N} \end{pmatrix}.$$

Therefore, the rearranged matrix has the form

$$\begin{pmatrix}
\tilde{\mathbf{B}} & \tilde{\mathbf{C}} & & & & \mathbf{B}' & & & & \\
 & \ddots & \ddots & & & & \ddots & & & \\
 & & \tilde{\mathbf{B}} & \tilde{\mathbf{C}} & & & & & \mathbf{B}' & \\
\tilde{\mathbf{C}} & & & & \mathbf{I}_M & & & & & \\
 & & & & \mathbf{0} & \mathbf{I}_{N-M} & & & & \\
 & & & & & \mathbf{0} & \mathbf{I}_{N-M} & & & \\
 & & & & & & \ddots & \ddots & & \\
 & & & & & & & \mathbf{0} & \mathbf{I}_{N-M}
\end{pmatrix}.$$

We can use the central $\mathbf{I}_M$, together with elementary column operations to clear the $\tilde{\mathbf{C}}$ on the left. Similarly, elementary row operations can use the diagonal $\mathbf{I}_{N-M}$ to clear the $\mathbf{B}'$s in the

upper right. The only remaining non-zero entries are in the $(M+r(N-M))\times(M+r(N-M))$ identity matrix in the lower right and the upper left block, with the copies of $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{C}}$. The latter matrix is just our specialized version of $\mathbf{A}_{r-1}$ with parameters M and N each decreased by $N-M$, and this matrix has full rank by the inductive hypothesis. □

The following proposition uses the preceding lemma to prove a new set of constraints.

Proposition 15 (Converse)**.** *Let $K=3$ and suppose $N \geq M$. Fix the dimensions per user $d_i = d$ and number of antennas $M_i = M, N_i = N$. Alignment is feasible only if*

$$(2r+1)d \leq \max(rN, (r+1)M), \quad \textit{for all } r \geq 0\,.$$

Remark 16. *Proposition 15 remains valid when allowing constant channel time extension, with M, N, and d appropriately normalized by the time extension value.*

Proof. We fix the value of $r \geq 0$, and omit dependence on r whenever convenient. Define the product of transmit spaces $\mathcal{U} = U_2 \times U_3 \times \cdots \times U_{r+2} \subset (\mathbb{C}^M)^{r+1}$, where as usual indices are interpreted modulo 3, and similarly let $\mathcal{V} = V_1 \times \ldots V_r \subset (\mathbb{C}^N)^r$. Note that each U_i and V_i has dimension d, so $\mathcal{U}$ and $\mathcal{V}$ have dimensions $(r+1)d$ and rd respectively.

First, suppose that $rN \geq (r+1)M$. Then Lemma 14 implies that the linear map $\mathbf{A}_r \colon (\mathbb{C}^M)^{r+1} \to (\mathbb{C}^N)^r$ is injective. By the orthogonality condition (3.5), we obtain $\mathcal{V} \perp \mathbf{A}_r\mathcal{U}$, and thus $rd + (r+1)d = \dim \mathcal{V} + \dim(\mathbf{A}_r\mathcal{U}) \leq \dim(\mathbb{C}^N)^r = rN$.

On the other hand, if $(r+1)M \geq rN$, the Hermitian transpose $\mathbf{A}_r{}^*$ is an injective linear map $\mathbf{A}_r{}^* \colon (\mathbb{C}^N)^r \to (\mathbb{C}^M)^{r+1}$. Again, the orthogonality conditions (3.5) imply that $\mathbf{A}_r{}^*\mathcal{V} \perp \mathcal{U}$ so $(2r+1)d \leq (r+1)M$. This proves the lemma. □

Note that when $r=0$ Proposition 15 reduces to the obvious constraint $d \leq M$ in order to have a d-dimensional subspace of an M-dimensional vector space. In fact, the proposition and its proof can be considered generalizations of this observation, with the inequality arising from the fact that the vector spaces $\mathcal{V} + \mathbf{A}_r\mathcal{U}$ or $\mathbf{A}_r{}^*\mathcal{V} + \mathcal{U}$ must be contained in $(\mathbb{C}^N)^r$ and $(\mathbb{C}^M)^r$ respectively.

Proof of Achievability

It remains to prove achievability in Theorem 10.

Theorem 17 (Achievability)**.** *Fix any M, N, and d satisfying* (3.2)*. Then alignment is feasible, i.e. there exists a choice of subspaces $U_1, U_2, U_3, V_1, V_2, V_3$ with $\dim U_i = \dim V_i = d$, for $1 \leq i \leq 3$, and $V_i \perp \mathbf{H}_{ij}U_j$ for $1 \leq i \neq j \leq 3$.*

Proof. The proof for the critical points satisfying $N + M = 4d$ is given as part of Proposition 19 below. The more general argument is similar, but tedious, and deferred to the appendix. □

Remark 18. *The achievable strategy specifies an explicit construction for the solutions in terms of the kernel of an appropriate matrix (or in terms of eigenvectors in the case $M = N$). This contrasts with the existence proofs for $K > 3$ from [27] (appearing in the latter part of this chapter) and [17], which do not provide a way to find solutions.*

Proposition 19. *Fix integers d and $N \geq M$ satisfying $N + M = 4d$. Then alignment is feasible if and only if either $N = M = 2d$ or the integer d is evenly divisible by $2d - M = N - 2d$.*

Proof. The necessity follows by some manipulations of Proposition 15. If $N \neq M$ and $d/(2d - M)$ is not an integer, then we set r to be the nearest integer to $M/(N - M)$, which is well-defined because of the equality:

$$\frac{M}{N-M} = \frac{d}{2d-M} - \frac{1}{2}.$$

Thus,

$$r = \frac{M}{N-M} + e$$

where e has absolute value strictly less than one half. Now, we get

$$\begin{aligned}(2r+1)d &= \frac{(N+M)^2}{4(N-M)} + \frac{e(N+M)}{2} \\ rN &= \frac{NM}{N-M} + eN \\ (r+1)M &= \frac{NM}{N-M} + eM.\end{aligned}$$

Which of the latter two is larger will depend on the sign of e. Assuming that e is positive, we can substitute and clear denominators to get that

$$(2r+1)d \leq \max\{rN, (r+1)M\}$$

is equivalent to

$$\begin{aligned}0 &\geq (N+M)^2 + 2e(N+M)(N-M) - 4NM - 4eN(N-M) \\ &= (N-M)^2 - 2e(N-M)^2,\end{aligned}$$

which will be false because e is less than one half. The case when e is negative works similarly.

We now turn to the sufficiency part of the proof. Suppose that $2d - M$ is positive and evenly divides d. We set $r = d/(2d-M)-1$, from which it follows that $M = d(2r+1)/(r+1)$ and $N = d(2r+3)/(r+1)$. For any integer i, we define shifted versions of the block matrix from (3.8):

$$\mathbf{A}_r^i = \begin{pmatrix} \mathbf{H}_{i,-} & \mathbf{H}_{i,+} & & & \\ & \mathbf{H}_{i+1,-} & \mathbf{H}_{i+1,+} & & \\ & & & \ddots & \\ & & & & \mathbf{H}_{i+r-1,-} \quad \mathbf{H}_{i+r-1,+} \end{pmatrix}$$

By Lemma 14, for generic channel matrices, $\mathbf{A}_r^i$ has full rank. Therefore, its kernel is a vector space of dimension $(r+1)M - rN = d/(r+1)$, and we denote this vector space by W_i. For $i+1 \le j \le i+r+1$, define $W_{i,j}$ to be the projection of W_i onto the $(j-i)$th block of coordinates. We claim that

$$\begin{aligned} U_j &= \sum_{i=j-1}^{j-r-1} W_{i,j}\,, \\ V_j &= \left(\mathbf{H}_{j,-} W_{j,j+1} + \sum_{i=j}^{j-r} \mathbf{H}_{j,+} W_{i,j+1} \right)^{\perp}\,, \end{aligned} \tag{3.9}$$

constitutes a feasible strategy for interference alignment. Before rigorously justifying this, we first do a naive dimension count to verify that

$$\dim U_j = (r+1)\dim W_{i,j} = d$$

and

$$\dim V_j = N - (r+2)\dim W_{i,j} = \frac{2r+3}{r+1}d - \frac{r+2}{r+1}d = d.$$

Any element of W_i consists of $r+1$ vectors $x_{i,j} \in \mathcal{C}^M$ for $i+1 \le j \le i+r+1$, and these vectors satisfy $\mathbf{H}_{j,+}x_{i,j+1} = -\mathbf{H}_{j,-}x_{i,j+2}$ for $i+1 \le j \le i+r$. First, since the channel matrices are injective, the only way for a subvector $x_{i,j}$ to be zero is for the whole vector to be zero, and thus each projection $W_{i,j}$ has the full dimension $d/(r+1)$. Second, these equations explain the apparent asymmetry in the definition of V_j, which can equivalently be defined as the complement of the sum over all applications of $\mathbf{H}_{j,-}$ and $\mathbf{H}_{j,+}$ to appropriate vector spaces W_*, but such vector spaces coincide. Indeed, this is the essence of the construction. From this observation, it follows that that $\mathbf{H}_{j,+}U_{j+1}$ and $\mathbf{H}_{j,-}U_{j-1}$ are orthogonal to V_j, which is what is required to be feasible.

The only thing remaining to be checked is that U_j and V_j actually have the expected dimensions. This is verified in Lemma 20 below.

Finally, we suppose that $M = N = 2d$. The channel matrices are square, and thus, generically, they are invertible, so we can define

$$\mathbf{S} = \mathbf{H}_{12}(\mathbf{H}_{32})^{-1}\mathbf{H}_{31}(\mathbf{H}_{21})^{-1}\mathbf{H}_{23}(\mathbf{H}_{13})^{-1}.$$

Again, generically, this matrix will have $2d$ distinct eigenvectors, and we choose V_1 to be the span of any d of them. Then we set

$$\begin{aligned} U_3^{\perp} &= (\mathbf{H}_{1,3})^{-1}V_1 \\ V_2 &= \mathbf{H}_{2,3}U_3^{\perp} \\ U_1^{\perp} &= (\mathbf{H}_{2,1})^{-1}V_2 \\ V_3 &= \mathbf{H}_{3,1}U_1^{\perp} \\ U_2^{\perp} &= (\mathbf{H}_{3,2})^{-1}V_3. \end{aligned}$$

These form a feasible strategy. □

Note that our constructions imply that the alignment solution is unique when $2d - M$ divides d, but there exist $\binom{2d}{d}$ solutions when $N = M = 2d$.

Lemma 20 below completes the proof of Proposition 19 by showing that subspaces U_j and V_j in the given construction have the expected dimension.

Lemma 20. *The subspaces U_j and V_j defined in* (3.9) *have dimension d.*

Proof. We first show that U_1 has dimension d; by symmetry of the construction, the dimensions of U_2 and U_3 will also be d.

The subspace $U_1 = \sum_{i=-r}^{0} W_{i,1}$ is the sum of $r+1$ subspaces $W_{i,j}$, which we claim are independent; suppose to the contrary, that there is some set of linearly dependent vectors $w_{i_1}, w_{i_2}, \ldots, w_{i_s}$, with $0 \le i_1 \le i_2 \le \cdots \le i_s \le r$, and $w_i \in W_{-i,1}$, satisfying $w_{i_s} - \sum_{\ell=1}^{s-1} \lambda_\ell w_{i_\ell} = 0$. Let s be the minimum such value, with all sets of subspaces $W_{i_1,j}, W_{i_2,j}, \ldots, W_{i_{s-1},j}$ for $j = 1, 2, 3$ being complementary.

Now, by the definition of the subspaces $W_{i,j}$, for each vector $w_{i_\ell} \in W_{-i_\ell,1}$ there is a sequence $u_{i_\ell}^2, \ldots, u_{i_\ell}^{q+1}$ of length $q := r + 1 - i_{s-1}$ satisfying $\mathbf{H}_{13} w_{i_\ell} = \mathbf{H}_{23} u_{i_\ell}^2, \ldots, \mathbf{H}_{qq+2,} u_{i_\ell}^q = \mathbf{H}_{q+1q+2,} u_{i_\ell}^{q+1}$. The linear combination $\sum_{\ell=1}^{s-1} \lambda_\ell w_{i_\ell}$ thus gives rise to a sequence $u^1, \ldots, u^{q+1}$ defined by $u^a = \sum_{\ell=1}^{s-1} \lambda_\ell u_{i_\ell}^a$ satisfying

$$\begin{aligned} \mathbf{H}_{13} w_{i_s} &= \mathbf{H}_{13} \left(\sum_{\ell=1}^{s-1} \lambda_\ell w_{i_\ell} \right) = \mathbf{H}_{23} u^2, \\ \mathbf{H}_{21} u^2 &= \mathbf{H}_{31} u^3 \\ &\vdots \\ \mathbf{H}_{q,q+2} u^q &= \mathbf{H}_{q+1,q+2} u^{q+1} . \end{aligned} \tag{3.10}$$

Note that by the minimality assumption of s, none of the u^j vectors are zero.

By the definition of $W_{-i_s,1}$, there is a length-$(i_s - 1)$ sequence of vectors preceding w_{i_s} satisfying alignment conditions similar to those in (3.10); together with w_{i_s} and the vectors in (3.10), this sequence can be extended to a sequence of vectors of total length $q + i_s = r + 1 + (i_s - i_{s-1}) > r + 1$, none of which are zero. Stacking the first $r+2$ of these vectors produces a nonzero element in the kernel of $\mathbf{A}_{r+1}^{i_s}$. However, $\mathbf{A}_{r+1}^{i_s}$ is full-rank by Lemma 14; the dimension of the kernel is $\Big[(r+2)M - (r+1)N\Big]^+ = M + d\Big(2r+1-2r-3\Big) = M - 2d < 0$, i.e. the kernel is trivial. This is the desired contradiction.

We now check that V_1 has dimension d, and again by symmetry, the dimensions of V_2 and V_3 will also be d. Note that if V_1 had dimension greater than d, we could choose a d-dimensional subspace and this would still satisfy the alignment equations (3.5). But V_1 is the orthogonal complement of the sum of $r+2$ subspaces $W_{i,j}$ of dimension $d/(r+1)$, so by subadditivity of dimension, we have the lower bound on dimension $\dim V_1 \ge N - (r+2) \dim W_{i,j} = d$. □

3.4 K-user fully symmetric channel

In this section we study the K-user fully symmetric MIMO IC. We first prove Theorem 11 giving a set of general necessary conditions. Next we prove Theorem 12, which shows that for the fully symmetric case in which $d_i = d$ and $M_i = N_i = N$ for all i, the necessary conditions are also sufficient.

The most natural view of the problem is to think of nature as fixing the channels with the engineer subsequently wishing to find the set of feasible communication strategies. However, reversing the picture turns out to be mathematically fruitful: we fix the communication strategy and study the set of channels for which the communication strategy is feasible. This approach, common in algebraic geometry, is the key to proving the results of the paper.

Some concepts from algebraic geometry will be necessary. For background see the texts by Hartshorne [38] or Shafarevich [39].

In algebraic geometry, the basic object of study is the solution set to a system of polynomial equations, called an *algebraic variety* or simply *variety*. The *Zariski topology* is defined by taking the closed sets to be the set of solutions to a system of polynomial equations. Any future reference to closed or open sets is with respect to the Zariski topology. A variety X is *reducible* if it can be written as a union of non-trivial subvarieties $X = X_1 \cup X_2$, where $X_1, X_2 \neq X$ and $X_1, X_2 \neq \varnothing$. A closed set X which is not reducible is *irreducible*. The constituent subsets $X_1, \ldots, X_n$ in an irreducible decomposition $X = X_1 \cup X_2 \cup \cdots \cup X_n$ are called the *components* of X. The *dimension* of an irreducible variety X is defined to be the maximum n such that there is a chain of irreducible varieties $Y_0, Y_1, \ldots, Y_{n-1}$ satisfying the strict inclusions $\varnothing \subsetneq Y_0 \subsetneq Y_1 \subsetneq \cdots Y_{n-1} \subsetneq X$. The *codimension* of a subvariety $Y \subseteq X$ is $\dim X - \dim Y$.

To represent the strategy space, we will be interested in the *Grassmannian* $G(d, N)$ of d-dimensional subspaces of N-dimensional affine space $\mathbb{C}^N$. The dimension of the Grassmannian $G(d, N)$ is $d(N-d)$. See [39] for more on Grassmannians. In particular, for each i, the transmit subspace U_i corresponds to a point in the Grassmannian, $U_i \in G(d_i, M_i)$, and similarly $V_i \in G(d_i, N_i)$. The strategy space is thus the product of the Grassmannians, $\mathcal{S} = \prod_{i=1}^K G(d_i, M_i) \times \prod_{i=1}^K G(d_i, N_i)$. Let $\mathcal{H} = \prod_{i \neq j} \mathbb{C}^{N_i \times M_i}$ denote the space of all cross channels H_{ij} for $i \neq j$. Concretely, $h \in \mathcal{H}$ is a length-$K(K-1)$ tuple of channel matrices $h = (\mathbf{H}_{12}, \mathbf{H}_{13}, \ldots, \mathbf{H}_{K,K-1})$.

In the product $\mathcal{S} \times \mathcal{H}$, define the incidence variety $\mathcal{I} \subseteq \mathcal{S} \times \mathcal{H}$ to be the set of ordered pairs (s, h) such that s is a feasible strategy for h. Each of $\mathcal{S}$, $\mathcal{H}$, and $\mathcal{I}$ is an algebraic variety. The dimensions of $\mathcal{S}$ and $\mathcal{H}$ are

$$\dim \mathcal{S} = \sum_{i=1}^{K} \Big(d_i(M_i - d_i) + d_i(N_i - d_i) \Big), \tag{3.11}$$

and

$$\dim \mathcal{H} = \sum_{\substack{1 \leq i,j \leq K \\ i \neq j}} M_i N_j \,, \tag{3.12}$$

and the dimension of $\mathcal{I}$ is computed in Lemma 22 below.

The following theorem can be thought of as the algebraic geometry analogue of the rank-nullity theorem from linear algebra (see e.g. Theorem 7 on page 76 of [39]). Given a map $f\colon X \to Y$, the *fiber* of a point y in Y is the inverse image of y under the map f, $f^{-1}(y) = \{x \in X : f(x) = y)\}$. A *polynomial map* is simply a map whose coordinates are given by polynomials.

Theorem 21 (Dimension of fibers)**.** *Let $f\colon X \to Y$ be a polynomial map between irreducible varieties. Suppose that f is dominant, i.e. the image of f is dense in Y. Let n and m denote the dimensions of X and Y respectively. Then $m \leq n$ and*

1. *$\dim Z \geq n - m$ for any $y \in f(X) \subset Y$ and for any component Z of the fiber $f^{-1}(y)$;*
2. *there exists a nonempty open subset $U \subset Y$ such that $\dim f^{-1}(y) = n - m$ for $y \in U$.*

We will apply this theorem to the projections of $\mathcal{I}$ to each of the factors $\mathcal{S}$ and $\mathcal{H}$. Projecting onto the first factor allows to find the dimension of $\mathcal{I}$.

Lemma 22. *$\mathcal{I}$ is an irreducible variety of dimension*

$$\sum_{i=1}^{K} \Big(d_i(M_i - d_i) + d_i(N_i - d_i)\Big) + \sum_{\substack{1 \leq i,j \leq K \\ i \neq j}} (M_i N_j - d_i d_j)$$

Proof. We consider the projection onto the first factor of our incidence variety, $p\colon \mathcal{I} \to \mathcal{S}$. For any point $s = (U_1, \ldots, U_K, V_1, \ldots, V_K) \in \mathcal{S}$, we claim that the fiber $p^{-1}(s)$ is a linear space of dimension

$$\dim p^{-1}(s) = \sum_{\substack{1 \leq i,j \leq K \\ i \neq j}} M_i N_j - d_i d_j.$$

To see this claim, we give local coordinates to each of the subspaces comprising the solution $s \in \mathcal{S}$. We write $u_a^{(i)}$ for the ath basis element of subspace U_i, where $u_a^{(i)}$ has zeros in the first d_i entries except for a 1 in the ath entry, and similarly for $v_b^{(j)}$ (this is without loss of generality). The orthogonality condition $V_j \perp \mathbf{H}_{ji} U_i$ can now be written as the condition $v_b^{(j)} \perp \mathbf{H}_{ji} u_a^{(i)}$ for each $1 \leq a \leq d_i$ and $1 \leq b \leq d_j$. Writing this out explicitly, we obtain

$$\begin{aligned}
0 = v_b^{(j)} \perp \mathbf{H}_{ji} u_a^{(i)} &= \sum_{\substack{1 \leq k \leq M_i \\ 1 \leq l \leq N_j}} v_b^{(j)}(k) \mathbf{H}_{ji}(k,l) u_a^{(i)}(l) \\
&= \sum_{\substack{1 \leq k \leq d_i \\ 1 \leq l \leq d_j}} v_b^{(j)}(k) \mathbf{H}_{ji}(k,l) u_a^{(i)}(l) + \sum_{k > d_i \text{ or } l > d_j} v_b^{(j)}(k) \mathbf{H}_{ji}(k,l) u_a^{(i)}(l) \\
&= \mathbf{H}_{ij}(a,b) + \sum_{k > d_i \text{ or } l > d_j} v_b^{(j)}(k) \mathbf{H}_{ji}(k,l) u_a^{(i)}(l)\,.
\end{aligned}$$

Note that this equation is linear in the entries of $\mathbf{H}_{ji}$. There are $d_i d_j$ such linear equations, and each one has a unique variable $\mathbf{H}_{ji}(a,b)$, so the equations are linearly independent and

each equation reduces the dimension by 1. The claim follows from the fact that in total there are $\sum_{i \neq j} d_i d_j$ equations and we began with $\dim \mathcal{H} = \sum_{\substack{1 \leq i,j \leq K \\ i \neq j}} M_i N_j$ dimensions (3.12).

We have shown that $\mathcal{I} \to \mathcal{S}$ is a vector bundle over the irreducible variety $\mathcal{S}$, and thus it is irreducible. Since $\dim p^{-1}(s)$ is the same for all $s \in \mathcal{S}$, Theorem 21 gives the relation

$$\dim \mathcal{I} = \dim \mathcal{S} + \dim p^{-1}(s)\,.$$

Since the dimension of $\mathcal{S}$ is exactly the first summation in the lemma statement, this proves the lemma. □

Proof of Theorem 11. We now consider the projection onto the second factor $q\colon \mathcal{I} \to \mathcal{H}$. If this map is dominant (i.e., generically the alignment problem is feasible), then by Theorem 21 the fiber $q^{-1}(h)$ for a generic $h \in \mathcal{H}$ has dimension

$$\dim q^{-1}(h) = \dim \mathcal{I} - \dim \mathcal{H}\,. \tag{3.13}$$

Since $\mathcal{H}$ has dimension equal to $\sum_{\substack{1 \leq i,j \leq K \\ i \neq j}} M_i N_j$, then Lemma 22 gives us the dimension in the statement of the theorem. Moreover, if the quantity in (3.13) is negative, then the fiber $q^{-1}(h)$ at a generic point must be empty. But the set of solutions to the tuple of channel matrices h is given by $p(q^{-1}(h))$, so for generic channel matrices this means there is no feasible strategy.

Now we turn to the other necessary conditions for the existence of a solution. The first necessary condition $d_i \leq \min(M_i, N_i)$ is obvious. Next, suppose that $d_i + d_j > N_i \geq M_j$ for some i and j. Since $\mathbf{H}_{ij}$ is a generic $N_i \times M_j$ matrix, its nullspace will be trivial. Thus, $\mathbf{H}_{ij} U_j$ will be a d_j-dimensional vector space. Since $d_i + d_j > N_i$, the vector spaces $\mathbf{H}_{ij} U_j$ and V_i cannot be orthogonal. If $d_i + d_j > M_j \geq N_i$, then the argument is similar, but with the roles of U_j and V_i reversed.

Finally, any feasible strategy for the full set of K transmitters and receivers, will, in particular be feasible for any subset. Therefore, a necessary condition for a general set of channel matrices to have a feasible strategy is that the same is true for any subset of the pairs. Since the number t_A is the dimension of the variety of solutions when restricted just to the transmitters and receivers indexed by $i \in A$, then t_A must be non-negative in order to have a feasible strategy. □

Now, we make the assumption that $N_i = M_i = N$ and $d_i = d$ for all $1 \leq i \leq K$, and also that $K \geq 3$, and we wish to prove a sufficient condition for the existence of a feasible strategy in Theorem 12. The following lemma reduces the problem of showing that almost all channel tuples $h \in \mathcal{H}$ have a solution to finding the dimension of the solution set for a single channel tuple $h \in \mathcal{H}$. Recall that q is the projection of the incidence variety $\mathcal{I}$ onto the second factor, and that q being dominant means that its image is dense in $\mathcal{H}$, i.e. generic channel matrices have a solution.

Lemma 23. *Suppose that there exists $h \in \mathcal{H}$ such that the dimension of $q^{-1}(h)$ is at most $Kd(2N - (K+1)d)$. Then q is dominant.*

Proof. Let $h \in \mathcal{H}$ be a point such that $q^{-1}(h)$ has at most the stated dimension. Let $\mathcal{Z}_0 = q(\mathcal{I})$ be the projection of $\mathcal{I}$ onto the second factor, and let $\mathcal{Z}$ denote the closure of $\mathcal{Z}$. By these definitions, the projection $q\colon \mathcal{I} \to \mathcal{Z}$ is dominant. Now, part 1 of Theorem 21 (dimension of fibers) gives

$$\dim q^{-1}(h) \geq \dim \mathcal{I} - \dim \mathcal{Z}\,,$$

from which it follows that

$$\dim \mathcal{Z} \geq \dim \mathcal{I} - \dim q^{-1}(h) = \dim \mathcal{H}\,.$$

But $\mathcal{Z} \subseteq \mathcal{H}$, so equality of dimensions and irreducibility of $\mathcal{H}$ implies $\mathcal{Z} = \mathcal{H}$ (see e.g. [39, Thm. 1, pg. 68]), or, in other words, that $q\colon \mathcal{I} \to \mathcal{H}$ is dominant. □

Using Lemma 23, proving Theorem 12 requires only that we find a tuple of channels $h \in \mathcal{H}$ so that the set of solutions has the correct dimension. This is provided by the following lemma.

Lemma 24. *Suppose that $K \geq 3$ and furthermore that $d_i = d$ and $M_i = N_i = N$ for all users i. If $Kd(2N - (K+1)d) \geq 0$, then there exists $h \in \mathcal{H}$ such that the dimension of $\dim q^{-1}(h)$ is at most $Kd(2N - (K+1)d)$.*

Proof. We consider the point $s_0 \in \mathcal{S}$ where each U_i and V_i is spanned by the first d standard basis vectors. Therefore, the set of channel matrices for which s_0 is a valid strategy are those $\mathbf{H}_{ij}$ such that

$$\begin{pmatrix} \mathbf{I}_d \\ 0 \end{pmatrix}^T \mathbf{H}_{ij} \begin{pmatrix} \mathbf{I}_d \\ 0 \end{pmatrix} = 0,$$

for i and j distinct integers between 1 and K. Here Id_n denotes the $n \times n$ identity matrix. It is clear that this implies that the upper left corner of H_{ij} must be zero, and thus we can write it in the form

$$\mathbf{H}_{ij} = \begin{pmatrix} 0 & \mathbf{F}^{[ij]} \\ \mathbf{G}^{[ij]} & \tilde{\mathbf{H}}^{[ij]} \end{pmatrix},$$

where $\mathbf{F}^{[ij]}$, $\mathbf{G}^{[ij]}$, and $\tilde{\mathbf{H}}^{[ij]}$ can be any matrices of size $d \times (N-d)$, $(N-d) \times d$, and $(N-d) \times (N-d)$ respectively. In a moment we will specify $\mathbf{F}^{[ij]}$ and $\mathbf{G}^{[ij]}$, but for now we assume they are fixed, but arbitrary.

We now investigate the set of solution strategies for these fixed channel matrices H_{ij}. In local coordinates around the strategy s_0, the vector spaces U_i and V_i can be written as column spans as follows:

$$U_i = \mathrm{colspan} \begin{pmatrix} \mathbf{I}_d \\ \mathbf{U}_i \end{pmatrix}, \qquad V_i = \mathrm{colspan} \begin{pmatrix} \mathbf{I}_d \\ \mathbf{V}_i \end{pmatrix},$$

where $\mathbf{U}_i$ and $\mathbf{V}_i$ are $(N-d) \times d$ matrices of variables. In order to satisfy the orthogonality condition, we need that

$$\begin{pmatrix} \mathbf{I}_d \\ \mathbf{V}_i \end{pmatrix}^T \mathbf{H}_{ij} \begin{pmatrix} \mathbf{I}_d \\ \mathbf{U}_j \end{pmatrix} = \mathbf{V}_i^T \mathbf{G}^{[ij]} + \mathbf{F}^{[ij]} \mathbf{U}_j + \mathbf{V}_i^T \tilde{\mathbf{H}}^{[ij]} \mathbf{U}_j = 0.$$

We linearize this problem by dropping the final, quadratic term:

$$\mathbf{V}_i^T \mathbf{G}^{[ij]} + \mathbf{F}^{[ij]} \mathbf{U}_j = 0\,, \qquad 1 \le i, j \le K\,. \tag{3.14}$$

In algebraic geometry, the vector space defined by the linear equations (3.14) is known as the Zariski cotangent space, and its dimension gives an upper bound on the dimension of the variety at the given point [40, Thm. 9.6.8(ii)]. Hence we focus on the problem of computing the dimension of the set of solutions $(\mathbf{U}_j, \mathbf{V}_i)$ to (3.14).

We now give our construction of the matrices $\mathbf{F}^{[ij]}$ and $\mathbf{G}^{[ij]}$ and find the dimension of the Zariski cotangent space. We separate the construction into two cases: (1) K is odd, (2) K is even.

Case 1: K is odd. This case is relatively straightforward. Recall that $\mathbf{F}^{[ij]}$ is of size $d \times (N-d)$ and $\mathbf{G}^{[ij]}$ is of size $(N-d) \times d$. We write

$$\mathbf{F}^{[ij]} = \begin{pmatrix} \mathbf{A}^{[ij]}(\frac{K-1}{2}+1) & \mathbf{A}^{[ij]}(\frac{K-1}{2}+2) & \cdots & \mathbf{A}^{[ij]}(K-1) & 0 \end{pmatrix}, \quad \mathbf{G}^{[ij]} = \begin{pmatrix} \mathbf{A}^{[ij]}(1) \\ \mathbf{A}^{[ij]}(2) \\ \vdots \\ \mathbf{A}^{[ij]}(\frac{K-1}{2}) \\ 0 \end{pmatrix}, \tag{3.15}$$

where each $\mathbf{A}^{[ij]}(k)$ is a block of size $d \times d$ to be defined shortly, and the rightmost zero in $\mathbf{F}^{[ij]}$ is of size $d \times (N - d\frac{K+1}{2})$ while the bottom zero in $\mathbf{G}^{[ij]}$ is a block of zeros of size $(N - d\frac{K+1}{2}) \times d$. Note that the assumption $2Kd(N-d) \ge K(K-1)d^2$ is equivalent to $N \ge d\frac{K+1}{2}$, so the specification of $\mathbf{F}^{[ij]}$ and $\mathbf{G}^{[ij]}$ above makes sense.

Let

$$\mathbf{A}^{[ij]}(k) = \begin{cases} \mathbf{I}_d & \text{if } k = i - j \\ 0 & \text{otherwise} \end{cases}, \tag{3.16}$$

where addition of indices is modulo K. Now write

$$\mathbf{U}_i = \begin{pmatrix} \mathbf{X}_i(1) \\ \mathbf{X}_i(2) \\ \vdots \\ \mathbf{X}_i(\frac{K-1}{2}) \\ \mathbf{X}_i \end{pmatrix}, \quad \text{and} \quad \bar{V}_i = \begin{pmatrix} \mathbf{Y}_i(1) \\ \mathbf{Y}_i(2) \\ \vdots \\ \mathbf{Y}_i(\frac{K-1}{2}) \\ \mathbf{Y}_i \end{pmatrix}, \tag{3.17}$$

where $\mathbf{X}_i(t), \mathbf{Y}_i(t)$, $1 \le t \le \frac{K-1}{2}$ are $d \times d$ blocks of variables from $\mathbf{U}_i, \mathbf{V}_i$, respectively, and $\mathbf{X}_i$, $\mathbf{Y}_i$ are blocks of size $[N - d\frac{K+1}{2}] \times d$ containing the remaining variables.

With this notation and choice of matrices, the equations (3.14) defining the Zariski cotangent space read

$$\mathbf{X}_i(t) = 0, \quad 1 \le t \le \frac{K-1}{2}, \quad \text{and} \quad \mathbf{Y}_i(t) = 0, \quad 1 \le t \le \frac{K-1}{2}. \tag{3.18}$$

Considering these equations for all i, $1 \leq i \leq K$, we see that the codimension is $d^2 2K \left(\frac{K-1}{2}\right) = K(K-1)d^2$. Thus the set of solutions to the equations (3.14) have dimension $2Kd(N-d) - K(K-1)d^2$.

Case 2: K is even. The idea behind the argument is the same as in case 1, but the details are more involved. Put

$$\mathbf{F}^{[ij]} = \begin{pmatrix} \mathbf{A}^{[ij]}(\frac{K+2}{2}+1) & \mathbf{A}^{[ij]}(\frac{K+2}{2}+2) & \cdots & \mathbf{A}^{[ij]}(K-1) & \hat{\mathbf{F}}^{[ij]} & 0 \end{pmatrix}, \quad \mathbf{G}^{[ij]} = \begin{pmatrix} \mathbf{A}^{[ij]}(4) \\ \mathbf{A}^{[ij]}(5) \\ \vdots \\ \mathbf{A}^{[ij]}(\frac{K+2}{2}) \\ \hat{\mathbf{G}}^{[ij]} \\ 0 \end{pmatrix}, \tag{3.19}$$

where the matrices $\mathbf{A}^{[ij]}(k)$ are defined above (3.16), $\hat{\mathbf{F}}^{[ij]}$ is a block of size $d \times \left\lceil \frac{3d}{2} \right\rceil$ when j is odd and size $d \times \left\lfloor \frac{3d}{2} \right\rfloor$ when j is even, $\hat{\mathbf{G}}^{[ij]}$ is a block of size $\left\lceil \frac{3d}{2} \right\rceil \times d$ when i is odd and size $\left\lfloor \frac{3d}{2} \right\rfloor \times d$ when i is even, and any remaining entries are zero. Note that the assumption $2Kd(N-d) \geq K(K-1)d^2$ is equivalent to $N - d - d\frac{K-4}{2} \geq \left\lceil \frac{3d}{2} \right\rceil$, so the specification of $\mathbf{F}^{[ij]}$ and $\mathbf{G}^{[ij]}$ makes sense.

We write

$$\mathbf{U}_i = \begin{pmatrix} \mathbf{U}_i^0 \\ \mathbf{X}_i \\ \bar{\mathbf{X}}_i \end{pmatrix}, \quad \text{and} \quad \mathbf{V}_i = \begin{pmatrix} \mathbf{V}_i^0 \\ \mathbf{Y}_i \\ \bar{\mathbf{Y}}_i \end{pmatrix}, \tag{3.20}$$

where $\mathbf{U}_i^0, \mathbf{V}_i^0$ are of size $d\frac{K-4}{2} \times d$, $\mathbf{X}_i, \mathbf{Y}_i$ are of size $\left\lceil \frac{3d}{2} \right\rceil \times d$ for odd values of i and of size $\left\lfloor \frac{3d}{2} \right\rfloor \times d$ for even values of i , and $\bar{\mathbf{X}}_i, \bar{\mathbf{Y}}_i$ contain the remaining variables (if any). Now, exactly as in case 1 above, the choice of $\mathbf{F}^{[ij]}, \mathbf{G}^{[ij]}$ forces $\mathbf{U}_i^0 = \mathbf{V}_i^0 = 0$.

It remains to specify the matrices $\hat{\mathbf{F}}^{[ij]}, \hat{\mathbf{G}}^{[ij]}$. Let

$$\hat{\mathbf{F}}^{[ij]} = \begin{pmatrix} \mathbf{A}^{[ij]}(3) & \mathbf{B}^{[ij]} \end{pmatrix}, \quad \text{and} \quad \hat{\mathbf{G}}^{[ij]} = \begin{pmatrix} \mathbf{A}^{[ij]}(1) \\ (\mathbf{B}^{[ij]})^T \end{pmatrix}, \tag{3.21}$$

where again $\mathbf{A}^{[ij]}(k)$ is defined in (3.16). The matrices $\mathbf{B}^{[ij]}$ are either $d \times \left\lceil \frac{d}{2} \right\rceil$ or$d \times \left\lfloor \frac{d}{2} \right\rfloor$

depending on the indices; $\mathbf{B}^{[ij]}$ is given by

$$\mathbf{B}^{[ij]}(k) = \begin{cases} \begin{pmatrix} \mathbf{I}_{\lceil \frac{d}{2} \rceil} \\ 0 \end{pmatrix} & \text{if } i \text{ is even and } j = i+1 \text{ or } i \text{ is odd and } j = i+3 \\ \begin{pmatrix} \mathbf{I}_{\lfloor \frac{d}{2} \rfloor} \\ 0 \end{pmatrix} & \text{if } i \text{ is odd and } j = i+1 \\ \begin{pmatrix} 0 \\ \mathbf{I}_{\lceil \frac{d}{2} \rceil} \end{pmatrix} & \text{if } i \text{ is odd and } j = i+2 \\ \begin{pmatrix} 0 \\ \mathbf{I}_{\lfloor \frac{d}{2} \rfloor} \end{pmatrix} & \text{if } i \text{ is even and } j = i+3 \text{ or } i \text{ is even and } j = i+2 \\ 0 & \text{otherwise} \end{cases}, \tag{3.22}$$

where the 0 in (3.22) denotes a block of zeros of appropriate size to ensure that $\mathbf{B}^{[ij]}$ is rectangular with d rows. Here, again, addition of indices is modulo K.

Let

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_i^1 & \mathbf{X}_i^2 \\ \mathbf{X}_i^3 \end{pmatrix}, \quad \text{and} \quad \mathbf{Y}_i^T = \begin{pmatrix} \mathbf{Y}_i^1 & \mathbf{Y}_i^3 \\ \mathbf{Y}_i^2 \end{pmatrix}, \tag{3.23}$$

where for *even* values of i the blocks $\mathbf{X}_{i+1}^3, \mathbf{Y}_i^1, \mathbf{Y}_{i+1}^2$ are $\left\lceil \frac{d}{2} \right\rceil \times d$, $\mathbf{X}_i^3, \mathbf{Y}_i^2, \mathbf{Y}_{i+1}^1$ are $\left\lfloor \frac{d}{2} \right\rfloor \times d$, $\mathbf{X}_i^1, \mathbf{X}_{i+1}^2, \mathbf{Y}_{i+1}^3$ are $d \times \left\lceil \frac{d}{2} \right\rceil$, and $\mathbf{X}_i^2, \mathbf{X}_{i+1}^1, \mathbf{Y}_i^3$ is $d \times \left\lfloor \frac{d}{2} \right\rfloor$. Then our linear equation (3.14) implies

$$0 = \mathbf{Y}_i^T \hat{\mathbf{G}}^{[ij]} + \hat{\mathbf{F}}^{[ij]} \mathbf{X}_j = \begin{pmatrix} \mathbf{Y}_i^1 \\ \mathbf{Y}_i^2 \end{pmatrix} \mathbf{A}^{[ij]}(1) + \mathbf{Y}_i^3 (\mathbf{B}^{[ij]})^T + \mathbf{A}^{[ij]}(3) \begin{pmatrix} \mathbf{X}_j^1 & \mathbf{X}_j^2 \end{pmatrix} + \mathbf{B}^{[ij]} \mathbf{X}_j^3 .$$

For each even value of i we end up with the equations

$$0 = \begin{pmatrix} \mathbf{Y}_{i+3}^1 + \mathbf{X}_i^3 \\ \mathbf{Y}_{i+3}^2 \end{pmatrix}, \quad 0 = \begin{pmatrix} \mathbf{Y}_{i+4}^1 + \mathbf{X}_{i+1}^3 \\ \mathbf{Y}_{i+4}^2 \end{pmatrix}, \quad 0 = \begin{pmatrix} \mathbf{Y}_{i+3}^1 \\ \mathbf{Y}_{i+3}^2 + \mathbf{X}_{i+1}^3 \end{pmatrix},$$
$$0 = \begin{pmatrix} \mathbf{X}_i^1 + \mathbf{Y}_{i+1}^3 & \mathbf{X}_i^2 \end{pmatrix}, \quad 0 = \begin{pmatrix} \mathbf{X}_i^1 & \mathbf{X}_i^2 + \mathbf{Y}_{i+2}^3 \end{pmatrix}, \quad 0 = \begin{pmatrix} \mathbf{X}_{i+1}^1 & \mathbf{X}_{i+1}^2 + \mathbf{Y}_{i+2}^3 \end{pmatrix},$$

which implies that all variables appearing here are zero, i.e. $\mathbf{X}_i, \mathbf{Y}_i = 0$ for each i.

This proves that precisely $K(K-1)d^2$ entries of $\mathbf{U}_i, \mathbf{V}_i$ must be zero for this choice of matrices $\mathbf{F}^{[ij]}, \mathbf{G}^{[ij]}$, finishing case 2 and completing the proof of the lemma. □

Appendix

Proof of Theorem 17 (3-user achievability)

Here we prove Theorem 17 showing achievability for M, N, d satisfying (3.2). Let r be the (unique) integer such that

$$rN < (r+1)M \quad \text{and} \quad (r+1)N \geq (r+2)M . \tag{3.24}$$

Note that this implies, from equation 3.2, that

$$(2r+3)d \leq (r+1)N \tag{3.25}$$

and

$$(2r+1)d \leq (r+1)M\,. \tag{3.26}$$

We prove achievability by examining two cases: 1) $d \leq (r+1)[(r+1)M - rN]$ and 2) $d > (r+1)[(r+1)M - rN]$. Case 1 means that all of the signal space U_i can be obtained from alignment paths of length $r+1$ (up to integer rounding), whereas in case 2 we must use alignment paths of length r as well in order to attain the required d dimensions.

We first assume case 1 holds. Consider $\mathbf{A}_r^i$ as in the proof of Proposition 19, and let W_i be a dimension $\lfloor \frac{d}{r+1} \rfloor$ subspace in the kernel of $\mathbf{A}_r^i$. Let $d' := d - (r+1)\lfloor \frac{d}{r+1} \rfloor$, and if $d' > 0$ let w_i be a 1-dimensional subspace in $\ker \mathbf{A}_r^i \setminus W_i$. The projections $W_{i,j}$ are defined in Proposition 19 and the subspaces $w_{i,j}$ are defined analogously. The spaces w_1, w_2, w_3 are required in order to accommodate the remainder left when dividing d by $r+1$, and will together contribute d' dimensions to each signal space U_j. We put

$$U_j = \sum_{i=j-1}^{j-r-1} W_{i,j} + \sum_{i=j-1}^{j-d'} w_{i,j} \tag{3.27}$$

and

$$V_j = \left(\mathbf{H}_{j,+} W_{j,j+1} + \mathbf{H}_{j,+} w_{j,j+1} + \sum_{i=j}^{j-r} \mathbf{H}_{j-} W_{i,j+1} + \sum_{i=j}^{j-d'+1} w_{i,j} \right)^{\perp}. \tag{3.28}$$

If all of U_j's constituent subspaces are complementary, then U_j has dimension $(r+1)\left\lfloor \frac{d}{r+1} \right\rfloor + d' = d$; the justification for this statement is similar to the proof of Lemma 20 and omitted here. To see that V_j has dimension (at least) d, we observe that by subadditivity of dimension,

$$\dim V_j \geq N - (r+2)\left\lfloor \frac{d}{r+1} \right\rfloor - d' - e\,, \tag{3.29}$$

where $e = 0$ if $(r+1)|d$ and $e = 1$ otherwise. Plugging in the inequality (3.25) we obtain

$$\dim V_j \geq \frac{2r+3}{r+1} d - d - e - \left\lfloor \frac{d}{r+1} \right\rfloor = d + \frac{d}{r+1} - \left\lfloor \frac{d}{r+1} \right\rfloor - e \geq d\,.$$

Suppose now that case 2 holds, i.e. $d > (r+1)[(r+1)M - rN]$. This means that not all of the signal space U_i can be included in alignment paths of length $r+1$, so the remainder will be included in alignment paths of lenth r. Let $d' := d - (r+1)[(r+1)M - rN]$ and $d'' = d' - r\left\lfloor \frac{d'}{r} \right\rfloor$. As before, denote by W_i the kernel of the matrix $\mathbf{A}_r^i$, having dimension $(r+1)M - rN$. Denote by π the projection from $\mathbb{C}^{(r+1)M} \to \mathbb{C}^{rM}$ to the first rM coordinates. The space $\pi(\ker \mathbf{A}_r^i)$ is contained in $\mathbf{A}_{r-1}^i$. Let X_i for $i = 1, 2, 3$ each be a $\left\lfloor \frac{d'}{r} \right\rfloor$ dimensional

subspace in $\ker \mathbf{A}_{r-1}^i \backslash \pi(W_i)$, and let w_i be a 1-dimensional subspace in $\ker \mathbf{A}_{r-1}^i \backslash (\pi(W_i)+X_i)$. Put

$$U_j = \sum_{i=j-1}^{j-r-1} W_{i,j} + \sum_{i=j-1}^{j-r} X_{i,j} + \sum_{i=j-1}^{j-d''} w_{i,j} \tag{3.30}$$

and

$$V_j = \left(\mathbf{H}_{j-}(W_{j,j+1}+X_{j,j+1}+w_{j,j+1}) + \sum_{i=j}^{j-r} \mathbf{H}_j + W_{i,j+1} + \sum_{i=j}^{j-r+1} \mathbf{H}_{j+} X_{i,j+1} + \sum_{i=j}^{j-d'+1} w_{i,j} \right)^{\perp}. \tag{3.31}$$

As before, a naive count suggests that U_j should have dimension d, and this can be justified similarly to Lemma 20.

To see that V_j has dimension at least d we again use subadditivity of dimension to get

$$\begin{aligned} \dim V_j &\geq N - (r+2)[(r+1)M - rN] - (r+1)\left\lfloor \frac{d'}{r} \right\rfloor - d'' - e_1 \\ &= N - (r+2)[(r+1)M - rN] - \left\lfloor \frac{d'}{r} \right\rfloor - d' - e_1, \end{aligned}$$

where e_1 is zero if $r|d'$ and e_1 is one otherwise. Letting $e_2 := \frac{d'}{r} - \left\lfloor \frac{d'}{r} \right\rfloor$, we have

$$\begin{aligned} \dim V_j &\geq N - (r+2)[(r+1)M - rN] - \frac{d'}{r} - d' + e_2 - e_1 \\ &= N - \frac{(r+1)d}{r} + \frac{1}{r}[(r+1)M - rN] + e_2 - e_1 \\ &= d + \frac{(r+1)}{r} M - \frac{2r+1}{r} d + e_2 - e_1 \,. \end{aligned}$$

Substituting $\frac{r+1}{2r+1}M$ for d, the inequality (3.26) implies that

$$\dim V_j \geq d + e_2 - e_1 \,.$$

If $e_1 = 1$ then e_2 is strictly positive, so the fact that $\dim V_j$ is an integer implies $\dim V_j \geq d$.

Chapter 4

Towards optimal assembly for high-throughput shotgun sequencing

4.1 Introduction

DNA sequencing is the basic workhorse of modern day biology and medicine. Since the sequencing of the Human Reference Genome ten years ago, there has been an explosive advance in sequencing technology, resulting in several orders of magnitude increase in throughput and decrease in cost. Multiple "next-generation" sequencing platforms have emerged. All of them are based on the whole-genome shotgun sequencing method, which entails two steps. First, many short reads are extracted from random locations on the DNA sequence, with the length, number, and error rates of the reads depending on the particular sequencing platform. Second, the reads are assembled to reconstruct the original DNA sequence.

Assembly of the reads is a major algorithmic challenge, and over the years dozens of assembly algorithms have been proposed to solve this problem [5]. Nevertheless, the assembly problem is far from solved, and it is not clear how to compare algorithms nor where improvement might be possible. The difficulty of comparing algorithms is evidenced by the recent assembly evaluations Assemblathon 1 [41] and GAGE [42], where which assembler is "best" depends on the particular dataset as well as the performance metric used. In part this is a consequence of metrics for partial assemblies: there is an inherent tradeoff between larger continuous fragments (contigs) and fewer mistakes in merging contigs (misjoins). But more fundamentally, independent of the metric, performance depends critically on the dataset, i.e. length, number, and quality of the reads, as well as the complexity of the genome sequence. With an eye towards the near future, we seek to understand the interplay between these factors by using the intuitive and unambiguous metric of *perfect* reconstruction[1].

We note that the objective of reconstructing the original DNA sequence from the reads

[1]The notion of perfect reconstruction is only slightly more stringent than "finishing" a sequencing project as defined by the National Human Genome Research Institute [43], where finishing a chromosome requires at least 95% of the chromosome to be represented by contiguous sequence.

contrasts with the many *optimization-based* formulations of assembly, such as shortest common superstring (SCS) [44], maximum-likelihood [45], [46], and various graph-based formulations [47], [48]. When solving one of these alternative formulations, there is no guarantee that the optimal solution is indeed the original sequence.

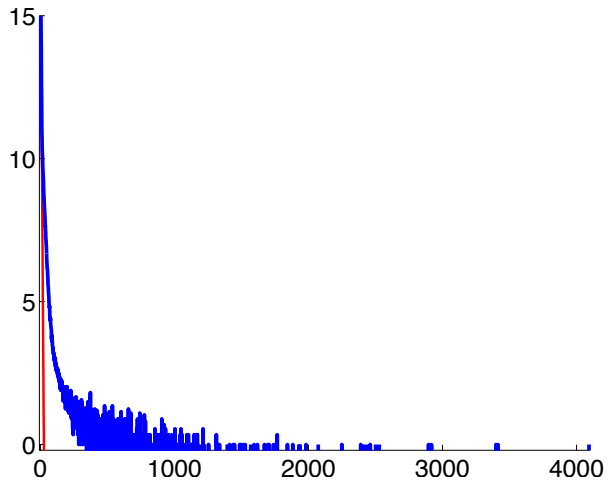
Because the goal of DNA sequencing is to reconstruct the original sequence, the most basic question is feasibility: given a set of reads, is it *possible* to reconstruct the original sequence? And second, if assembly is possible, which *algorithms* can successfully reconstruct? The feasibility question is a measure of the intrinsic *information* each read provides about the DNA sequence, and for given sequence statistics depends on characteristics of the sequencing technology such as read length and noise statistics. As such, it can provide an algorithm-independent basis for evaluating the efficiency of a sequencing technology. Equally important, algorithms can be evaluated based on their relative data requirements, and compared against the fundamental limit.

In studying these questions, we consider the most basic shotgun sequencing model where N noiseless reads of a fixed length L base pairs are uniformly and independently drawn from a DNA sequence of length G. By assumption the exact sequence to be assembled is unknown *a priori*, and in the Bayesian tradition we capture the uncertainty in the sequence through a probability distribution over possible sequences. We make the mild symmetry assumption that sequences related by certain simple transformations have similar prior probabilities. The optimal reconstruction is therefore given by the maximum a posteriori rule; maximum-likelihood is a special case resulting from a particular choice of prior distribution. Feasibility is thus rephrased as the question of whether, for given sequence statistics, the correct sequence can be reconstructed with probability $1 - \epsilon$ when N reads of length L are sampled from the genome. We note that answering the feasibility question of whether each N, L pair is sufficient to reconstruct is equivalent to finding the minimum required N (or the so-called *coverage depth* $c = NL/G$) as a function of L.

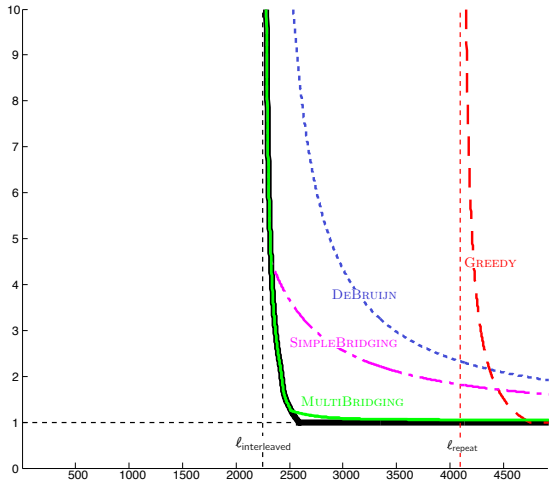
A lower bound on the minimum coverage depth needed was obtained by Lander and Waterman [49]. Their lower bound $c_{\text{LW}} = c_{\text{LW}}(L, \epsilon)$ is the minimum number of randomly located reads needed to cover the entire DNA sequence with a given target success probability $1 - \epsilon$. While this is clearly a necessary condition, it is in general not tight: only requiring the reads to cover the entire genome sequence does not guarantee that consecutive reads can actually be stitched back together to recover the original sequence. Characterizing when the reads can be reliably stitched together, i.e. determining feasibility, is an open problem.

As we will see, the ability to reconstruct depends crucially on the repeat statistics of the DNA sequence. We are interested in determining feasibility for statistics arising from a wide range of genomic sequences. Evaluating algorithms on statistics from existing genomes gives confidence in predicting whether the algorithms will be useful for an *unseen* genome with these statistics.

Our approach results in a pipeline, which takes as input a genome sequence and desired success probability $1 - \epsilon$, computes a few simple repeat statistics, and from these statistics produces a feasibility plot that indicates for which L, N reconstruction is possible. Fig. 4.1a displays the simplest of the statistics, the number of repeats as a function of the repeat



(a) For human chromosome 19, a log plot of number of repeats as a function of the repeat length ℓ.



(b) Feasibility plot. The thick black curve is a lower bound on the feasible N, L, and each colorful curve represents the lower boundary of feasible N, L for an algorithm.

Figure 4.1: Our pipeline takes as input a genome sequence (here human ch19) and desired success probability $1 - \epsilon$ (here 99%), computes a few simple repeat statistics including the one in Fig. (a), and from these statistics produces a feasibility plot as shown in Fig. (b).

length ℓ. Fig. 4.1b shows the resulting feasibility plot produced for human chromosome 19 (henceforth human ch19) with success probability 99%. The horizontal axis signifies read length L and the vertical axis signifies the coverage depth c normalized by c_{LW}, the coverage depth required as per Lander-Waterman [49] in order to cover the sequence. The normalized coverage depth $\bar{c} = c/c_{\text{LW}} = N/N_{\text{LW}}$ is also equal to the number of reads N normalized by the number of reads N_{LW} required to cover the sequence.

Since the coverage depth must satisfy $c \geq c_{\text{LW}}$, the normalized coverage depth satisfies $\bar{c} \geq 1$, and we plot the horizontal line $\bar{c} = 1$. Each colorful curve in the feasibility plot is the lower boundary of the set of feasible N, L pairs for an algorithm, and the thicker black curves depict lower bounds for *any* algorithm. As we discuss later, the green curve is achievable by a simple algorithm, and it nearly coincides with the lower bound. Thus Fig. 4.1b answers, up to a very small gap, the feasibility of assembly for human ch19, where successful reconstruction is desired with probability 99%. We produce similar plots for a dozen or so datasets, including those used in the recent GAGE assembly algorithm evaluation [42]. The curves in the feasibility plots are corroborated by simulations of the algorithms.

An important consideration is that of computational complexity. Although most of the optimization-based formulations of assembly have been shown to be NP-hard, including SCS [50], [44], De Bruijn Superwalk [47], [51], and Minimum s-Walk on the string graph [48], [51], as pointed out by Nagarajan and Pop [52], typical instances of the problem may well be easier than the worst-case. Indeed, the performance of the algorithms described in

this paper have runtimes that depends explicitly on the repeat statistics, and are efficient for typical statistics, as discussed in Section 4.6, despite achieving performance close to the information theoretic limits.

Much of this paper focuses on algorithmic development, but our aim is not to propose new *practical* assembly algorithms that can operate on real-world read data. In particular, three important aspects of real data are missing in the basic model: 1) there is noise in the reads; 2) reads can come in mate-pairs; 3) the read arrival process is not uniform (nonuniform coverage depth). Designers of practical assembly algorithms spend much of their energy on dealing with these aspects. Rather, the goal of this work is to advocate a new systematic approach to the design of assembly algorithms with optimality or near-optimality guarantee. Current ongoing work builds on the foundation established in this paper to incorporate noise and mate pairs in the read data.

The read lengths required for perfect reconstruction are typically on the order of $500 - 3000$ base pairs (bp). This is substantially longer than the reads produced by Illumina, the current dominant sequencing technology, which produces reads of lengths 100-200bp; however, other technologies produce longer reads. PacBio reads can be as long as several thousand base pairs, and as demonstrated by [53], the noise can be cleaned by Illumina reads to enable near-perfect reconstruction. Thus our framework is already relevant to the current cutting edge technology.

Paper organization. The rest of the paper is outlined as follows. In the next section we highlight a few of the results. In Section 4.3 we discuss lower bounds, followed by Section 4.4 which analyzes algorithms in a progression towards optimality. Section 4.5 contains simulations. Section 4.6 discusses computational complexity, and derives analytical formulas for the critical window width and gap from optimality. Appendix 4.7 proves Lemma 37. Appendix 4.8 includes feasibility plots for a dozen or so datasets, including those used in the recent GAGE evaluation.

4.2 Results

In this section we present the results of our framework as applied to the simple read model considered in this paper. For this model we are able to provide an approximate answer to the feasibility problem as well as describe a near-optimal algorithm. Work in progress seeks to carry out a similar program for more elaborate read models.

Lower bounds. We begin by mentioning the lower bounds. Aside from the requirement $\bar{c} \geq 1$ for covering the sequence, there is another condition in terms of repeats. We define $L_{\text{crit}} = 1 + \max\{\ell_{\text{interleaved}}, \ell_{\text{triple}}\}$, where $\ell_{\text{interleaved}}$ is the length of the longest interleaved repeats in the DNA sequence (see Section 4.3 for a precise definition) and ℓ_{triple} is the length of the longest triple repeat. Reconstruction is impossible whenever the read length is below this threshold. This follows from a result of Ukkonen [54] in the context of Sequencing

by Hybridization, and we generalize Ukkonen's result to the shotgun sequencing setting by taking into account the randomness in the read process. This gives rise to the thick black nearly vertical line in Fig. 4.1b.

Towards optimal assembly. Informed by the lower bounds, we seek assembly algorithms that are close to optimal. It turns out that the required coverage depth for each algorithm depends only on simple repeat statistics extracted from DNA data, which may be thought of as a *sufficient statistic*. We briefly overview the algorithm progression.

Several of the first assemblers implemented a simple greedy algorithm, including TIGR [55], CAP3 [56], and more recently SSAKE [57]. The greedy algorithm, denoted here by Greedy, repeatedly merges reads with the highest overlap. Greedy cannot resolve simple repeats longer than the read length, due to the greedy nature of the algorithm. Thus, as seen in Fig. 4.1b, Greedy is limited by the condition $L > \ell_{\text{repeat}}$.

K-mer (de Bruijn) graph based algorithms (e.g. [58] and [47]) take a more global view via the construction of a K-mer graph, and can resolve simple repeats longer than the read length as long as they are not interleaved. However, the coverage depth needed for these algorithms is in general larger than the Lander-Waterman depth, because reads not only need to cover the DNA sequence but successive reads have to overlap by at least K to have a connected K-mer graph. A naive K-mer graph algorithm, which we denote by DeBruijn, simply constructs a K-mer graph and finds an Eulerian cycle. The performance of DeBruijn is plotted in Fig. 4.1b. DeBruijn requires K to be large, $K > L_{\text{crit}}$, and in turn requires a high coverage depth to ensure connectivity in the graph.

Existing algorithms use read information in a variety of distinct ways to resolve repeats. For instance, Pevzner et al. [47] observe that for graphs where each edge has multiplicity one, if one copy of a repeat is bridged, the repeat can be resolved through what they call a "detachment". The algorithm SimpleBridging described here is very similar, and resolves repeats with two copies if at least one copy is bridged. (A repeat is bridged if at least one copy is contained in a read extending to both sides beyond the repeat.)

Meanwhile, other algorithms are able to deal with higher edge multiplicities due to higher order repeats; IDBA (Iterative DeBruijn Assembler) [59] can successfully reconstruct if all copies of every repeat are bridged. But it is suboptimal to require that *all* copies of *every* repeat be bridged. We introduce MultiBridging, which combines these ideas to simultaneously allow for single-bridged double repeats, triple repeats in which all copies are bridged, and unbridged non-interleaved repeats.

The performance of MultiBridging, as shown in Fig. 4.1b, is nearly optimal for human ch19. For the observed DNA statistics there are two different situations, depending on the relative size of $\ell_{\text{interleaved}}$ and ℓ_{triple}. The statistics of human ch19 fall within the first case, $\ell_{\text{interleaved}} \gg \ell_{\text{triple}}$.

Dominant interleaved repeat: ($\ell_{\text{interleaved}} \gg \ell_{\text{triple}}$). MultiBridging matches the lower bound with respect to interleaved repeats: if there are unbridged interleaved repeats,

reconstruction is impossible. The algorithm is therefore optimal as long as the all-bridging triple repeat constraint is not active, i.e. $\ell_{\text{interleaved}} \gg \ell_{\text{triple}}$. The statistics of human ch19 fit within this case, and MultiBridging nearly matches the lower bound. The interleaved repeat constraint being dominant, and near-optimality of MultiBridging also holds for a majority of the other datasets we examine.

An interesting feature of the feasibility plots is that for typical repeat statistics exhibited by DNA data, the minimum coverage depth is characterized by a *critical phenomenon*: If the read length L is below $L_{\text{crit}} = \ell_{\text{interleaved}}$, reliable reconstruction of the DNA sequence is impossible no matter what the coverage depth is, but if the read length L is slightly above L_{crit}, then covering the sequence suffices, i.e. $\bar{c} = c/c_{\text{LW}} = 1$.

The sharpness of the critical phenomenon is described by the size of the *critical window*, which refers to the range of L over which the transition from one regime to the other occurs. Let L^* denote the minimum L for which coverage of the sequence suffices, i.e. the knee in the feasibility plot. The critical window size is thus $L^* - L_{\text{crit}}$. We can compute the size of the critical window under the assumption that the longest interleaved repeat dominates (c.f. Section 4.6). Letting

$$r := \frac{\log \frac{G}{L_{\text{crit}}}}{\log \epsilon^{-1}}\,, \tag{4.1}$$

it turns out that we have to a very good approximation

$$\frac{L^*}{L_{\text{crit}}} \approx \frac{2(r+1)}{2(r+1)-1}\,. \tag{4.2}$$

Let us evaluate (4.2) for human ch19 in Fig. 4.1b. The relevant parameters are $G = 55{,}808{,}983$, $L_{\text{crit}} = \ell_{\text{interleaved}} = 2248$, $\ell_{\text{triple}} = 1766$, and $\epsilon = 1\%$. Plugging into (4.2) gives $L^*/L_{\text{crit}} \approx 1.19$.

Dominant triple repeat: ($\ell_{\text{triple}} \gg \ell_{\text{interleaved}}$). If the largest triple repeat dominates, i.e. $\ell_{\text{triple}} \gg \ell_{\text{interleaved}}$, then MultiBridging has a gap to the lower bound (see Fig. 4.2). The gap means that the feasibility problem is not completely answered, but we are able to nevertheless *bound* the size of the gap. Subject to the largest triple repeat dominating, we derive an estimate on the worst-case gap in normalized coverage depth required by MultiBridging as compared to the lower bound. We find, moreover, that the critical phenomenon continues to hold and we derive an estimate of the size of the critical window.

The gap, given by a ratio between upper and lower bound may be estimated as:

$$\frac{N^{\text{MultiBridging}}}{N^{\text{lower}}} = 3 \cdot \frac{\log 3\epsilon^{-1}}{\log \epsilon^{-1}} \approx 3.72 \quad \text{for } \epsilon = 10^{-2}\,. \tag{4.3}$$

This means that if the longest triple repeat is dominant, then for L slightly larger than ℓ_{triple}, MultiBridging needs a coverage depth approximately 3.72 times higher than required by

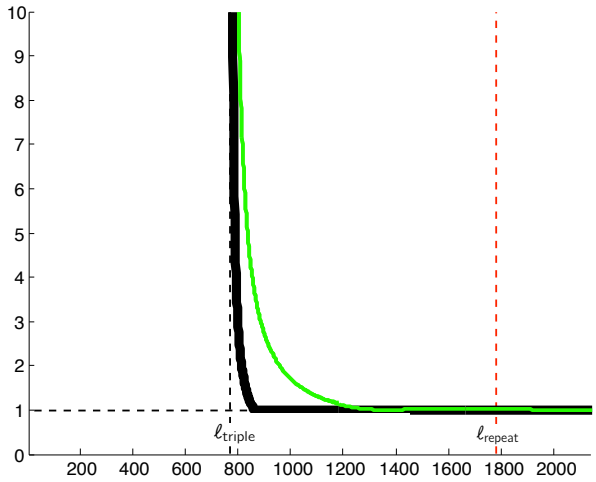


Figure 4.2: Performance of MULTIBRIDGING on *P Marinus*. *The effect of the longest triple repeat is dominant here, so for L slightly larger than* ℓ_{triple}, MULTIBRIDGING *needs a coverage depth higher than the coverage depth required by our lower bound, as predicted by equation* (4.3).

our lower bound. This is along the lines of a worst-case approximation ratio for MULTIBRIDGING. It is possible to show that this is a worst-case ratio over *all* possible repeat statistics, but this is not pursued here.

The size of the critical window is different for the lower and upper bounds. For the lower bound we obtain

$$\frac{L^*}{L_{\text{crit}}} = \frac{3(r+1)}{3(r+1)-1} \approx 1.06$$

for the example with $G \sim 10^8$, $L_{\text{crit}} \sim 1000$, and $\epsilon = 5\%$. Changing ϵ to 10^{-5} makes $\frac{L^*}{L_{\text{crit}}} \approx 1.17$, and as ϵ (and hence also r) tends to zero, $\frac{L^*}{L_{\text{crit}}} \to \frac{3}{2}$.

The analogous computation for L^*/L_{crit} for the upper bound due to MULTIBRIDGING yields

$$\frac{L^*}{L_{\text{crit}}} = \frac{r+1}{r + \frac{\log 3}{\log \epsilon^{-1}}} \approx 1.12\,, \tag{4.4}$$

for the example with $G \sim 10^8$, $L_{\text{crit}} \sim 1000$, and $\epsilon = 5\%$. The critical window size of the upper bound is about twice as large as that of the lower bound for typical values of G and L_{crit}, with ϵ moderate. But as $\epsilon \to 0$, we see from (4.4) that $L^*/L_{\text{crit}} \to \infty$, markedly different to the $L^*/L_{\text{crit}} \to \frac{3}{2}$ observed for the lower bound.

Remark 25. *An earlier work [60] has established the critical phenomenon for DNA with i.i.d. statistics, and showed that in this case* GREEDY *is optimal. However, real genomes, particularly those of eukaryotes, have many long repeats which are not well described by the i.i.d. model. Our results show that even for more complex DNA sequences, the critical behavior persists. However, in general the greedy algorithm is far from optimal and a more*

sophisticated K-mer-based assembly algorithm, MultiBridging*, is needed to approach optimality.*

4.3 Lower bounds

In this section we discuss lower bounds, due to coverage analysis and certain repeat patterns, on the required coverage depth and read length. The style of analysis here is continued in Section 4.4, in which we embark on our search for assembly algorithms that perform close to the lower bounds.

Coverage analysis

We start by presenting a lower bound on the coverage depth needed by *any* algorithm. Lander and Waterman's coverage analysis [49] gives the well known condition for the number of reads N_{LW} required to cover the entire DNA sequence with probability at least $1-\epsilon$. We may assume that the starting locations of the N reads are given according to a Poisson process with rate $\lambda = N/G$, and thus each spacing has an exponential(λ) distribution. A gap between two successive reads is equivalent to a spacing larger than L, an event with probability $e^{-\lambda L}$. The probability of coverage is equal to the probability of no gap between all $N-1$ pairs of successive reads, i.e.

$$(1-e^{-\lambda L})^{N-1}\,.$$

Solving for the smallest N such that this quantity is greater than $1-\epsilon$ yields N_{LW}, which is to a very good approximation given by the solution to the equation

$$N_{\mathrm{LW}} = \frac{G}{L}\log\frac{N_{\mathrm{LW}}}{\epsilon}\,. \tag{4.5}$$

The corresponding coverage depth is $c_{\mathrm{cov}} = N_{\mathrm{LW}}L/G$. This is our baseline coverage depth against which to compare the coverage depth of various algorithms. For each algorithm, we will plot

$$\bar{c} := \frac{c}{c_{\mathrm{cov}}} = \frac{N}{N_{\mathrm{LW}}}\,,$$

the coverage depth required by that algorithm normalized by c_{cov}. Note that $\bar{c}$ is also the ratio of the number of reads N required by an algorithm to N_{LW}. The requirement $\bar{c} \geq 1$ corresponds to the lower bound on the number of reads obtained by Lander-Waterman coverage condition.

Ukkonen's condition

Not only must there be enough reads, but the reads must have sufficient length. A lower bound on the read length L follows from Ukkonen's condition [54]: if there are *interleaved*

repeats or *triple repeats* in the sequence of length at least $L-1$, then more than one sequence agrees with the reads and hence correct reconstruction is not possible (Fig. 4.3). (In order to rule out guessing between two options, we make the assumption that the desired probability of successful reconstruction $1-\epsilon$ is greater than half.)

We take a moment to carefully define the various types of repeats. Let $\mathbf{s}_t^\ell$ denote the length-ℓ subsequence starting at position t. A *repeat* is a subsequence appearing twice, at some positions t_1, t_2 (so $\mathbf{s}_{t_1}^{t_1+\ell} = \mathbf{s}_{t_2}^{t_2+\ell}$) that is maximal (i.e. $s(t_1-1) \neq s(t_2-1)$ and $s(t_1+\ell) \neq s(t_2+\ell)$). Similarly, a *triple repeat* is a subsequence appearing three times, at positions t_1, t_2, t_3, such that $\mathbf{s}_{t_1}^{t_1+\ell} = \mathbf{s}_{t_2}^{t_2+\ell} = \mathbf{s}_{t_3}^{t_3+\ell}$, and such that neither of $s(t_1-1) = s(t_2-1) = s(t_3-1)$ nor $s(t_1+\ell) = s(t_2+\ell) = s(t_3+\ell)$ holds. A *copy* is a single one of the instances of the subsequence's appearances. A *pair* of repeats refers to two repeats, each having two copies. A pair of repeats, one at positions t_1, t_3 with $t_1 < t_3$ and the second at positions t_2, t_4 with $t_2 < t_4$, is *interleaved* if $t_1 < t_2 < t_3 < t_4$ or $t_2 < t_1 < t_4 < t_3$. Ukkonen's condition implies a lower bound on the read length,

$$L > 1 + \max\{\ell_{\text{interleaved}}, \ell_{\text{triple}}\}\,.$$

Here $\ell_{\text{interleaved}}$ is the maximum length of the shorter of a pair of interleaved repeats and ℓ_{triple} is the length of the longest triple repeat.

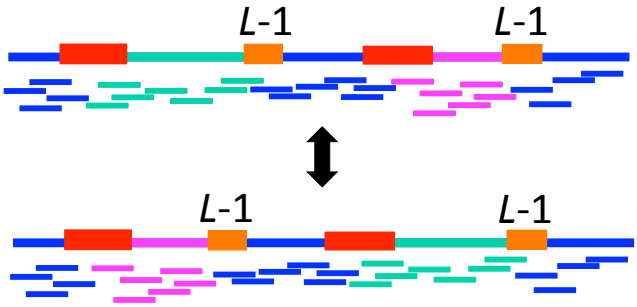


Figure 4.3: If there are interleaved repeats of length at least $L-1$, then two possible sequences (the green and magenta segments swapped) are consistent with the same set of reads and thus reconstruction is impossible.

Ukkonen's condition provides a lower bound on the read length, but it can be generalized to provide a lower bound on the coverage depth as follows. We say that a subsequence $\mathbf{s}_t^\ell$ of length ℓ starting at position t is *bridged* if there is a read strictly containing the subsequence, i.e. extending by at least one base pair to either side, and *unbridged* otherwise (see Figure 4.4). One observes that in Ukkonen's interleaved or triple repeats, the actual length of the repeated subsequences is irrelevant; rather, to cause confusion it is enough that all the pertinent repeats are unbridged. The *generalized Ukkonen's condition*, then, is the absence of any unbridged interleaved repeats or unbridged triple repeats. This condition is necessary for reconstruction.

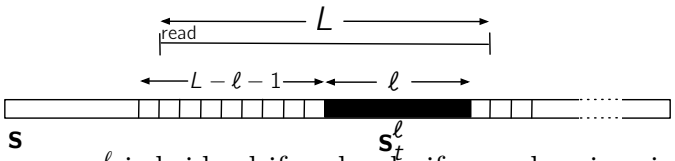


Figure 4.4: A subsequence $\mathbf{s}_t^\ell$ is bridged if and only if a read arrives in the preceding length $L-\ell-1$ interval. The probability the subsequence is unbridged is thus approximately $e^{-(N/G)(L-\ell-1)}$.

Theorem 26. *Given a DNA sequence* $\mathbf{s}$ *and a set of reads, if there is an unbridged pair of interleaved repeats or an unbridged triple repeat, then there is another sequence* $\mathbf{s}'$ *of the same length which is consistent with the set of reads.*

Given the repeat statistics of the DNA sequence to be reconstructed, Theorem 26 allows us to derive a necessary condition on N and L to meet a target probability of successful reconstruction $1-\epsilon$. We focus here on the condition that there are no unbridged interleaved repeats, with the analogous derivation for triple repeats relegated to subsection 4.7 in the appendix. Recall that a pair of repeats, one at positions t_1, t_3 with $t_1 < t_3$ and the second at positions t_2, t_4 with $t_2 < t_4$, is *interleaved* if $t_1 < t_2 < t_3 < t_4$ or $t_2 < t_1 < t_4 < t_3$. From the DNA we may extract a (symmetric) matrix of interleaved repeat statistics,

$$b_{mn} = \#\text{ pairs of interleaved repeats of lengths } m \text{ and } n\,.$$

We proceed by fixing both N and L and checking whether or not unbridged interleaved repeats occur with probability higher than ϵ. We will break up repeats into 2 categories: repeats of length at least $L-1$ (these are always unbridged), and repeats of length less than $L-1$ (these are sometimes unbridged). We assume that $L > \ell_{\text{interleaved}} + 1$, or equivalently $b_{ij} = 0$ for all $i, j \geq L-1$, since otherwise there are (with certainty) unbridged interleaved repeats and Ukkonen's condition is violated.

First, we estimate probability of error due to interleaved repeats of lengths $i < L-1$ and $j \geq L-1$. As noted before, the repeat of length j is too long to be bridged, so an error occurs if repeat i is unbridged. The probability that a subsequence of length ℓ is unbridged is approximately $e^{-\lambda(L-\ell-1)}$, equal to the probability of no Poisson arrivals in the interval of size $L-\ell-1$ before the subsequence (c.f. Figure 4.4). For a repeat, as long as the two copies' locations are not too nearby[2], each copy is bridged independently and hence the probability that both copes are unbridged is $e^{-2\lambda(L-\ell-1)}$. (Recall that a repeat is unbridged if *both* copies are unbridged.)

[2]More precisely, for the two copies of a a repeat of length ℓ to be bridged independently requires that no single read can bridge them both. This means their locations t and $t+d$ must have separation $d \geq L-\ell-2$.

A union bound estimate[3] gives a probability of error

$$P_{\text{error}} \approx \sum_{\substack{m<L-1 \\ n\geq L-1}} b_{mn} e^{-2\lambda(L-m-1)} \,. \tag{4.6}$$

Requiring the error probability to be less than ϵ and solving for L gives the necessary condition

$$L \geq \frac{1}{2\lambda} \log \frac{\gamma_1}{\epsilon} = \frac{G}{2N} \log \frac{\gamma_1}{\epsilon} \,, \tag{4.7}$$

where $\gamma_1 := \sum_{\substack{m<L-1 \\ n\geq L-1}} b_{mn} e^{2(N/G)(m+1)}$ is a simple function of the interleaved repeat statistic b_{mn}.

We now estimate the probability of error due to interleaved repeat pairs in which both repeats are shorter than $L-1$. In this case only one repeat of each interleaved repeat pair must be bridged. Again a union bound estimate gives

$$P_{\text{error}} \approx \sum_{m,n<L-1} b_{mn} e^{-2\lambda(L-m-1)} e^{-2\lambda(L-n-1)} \,.$$

Requiring the error probability to be less than ϵ gives the necessary condition

$$L \geq \frac{1}{4\lambda} \log \frac{\gamma_2}{\epsilon} = \frac{G}{4N} \log \frac{\gamma_2}{\epsilon} \,, \tag{4.8}$$

where $\gamma_2 := \sum_{m,n<L-1} b_{mn} e^{2(N/G)(m+n+2)}$ and similarly to γ_1 is computed from b_{mn}.

As discussed in this section, if the DNA sequence is not covered by the reads or there are unbridged interleaved or triple repeats, then reconstruction is not possible. But there is another situation which must be ruled out. Without knowing its length a priori, it is impossible to know how many copies of the DNA sequence are actually present: if the sequence $\mathbf{s}$ to be assembled consists of multiple concatenated copies of a shorter sequence, rather than just one copy, the probability of observing any set of reads will be the same. Since it is unlikely that a true DNA sequence will consist of the same sequence repeated multiple times, we assume this is not the case throughout the paper without further mention. Equivalently, if $\mathbf{s}$ does consist of multiple concatenated copies of a shorter sequence, we are content to reconstruct a single copy. If available, knowledge of the approximate length of $\mathbf{s}$ would then allow to reconstruct.

The necessary conditions (4.5), (4.7), and (4.8) can be applied to lower bound the coverage depth NL/G for any DNA sequence. We next turn to evaluating the performance of assembly algorithms, starting with the greedy algorithm.

[3]The union bound on probabilities gives an *upper bound*, so its use here is only an approximation. To get a rigorous lower bound we can use the inclusion-exclusion principle, but the difference in the two computations is negligible for the data we observed. For ease of exposition we opt to present the simpler union bound estimate.

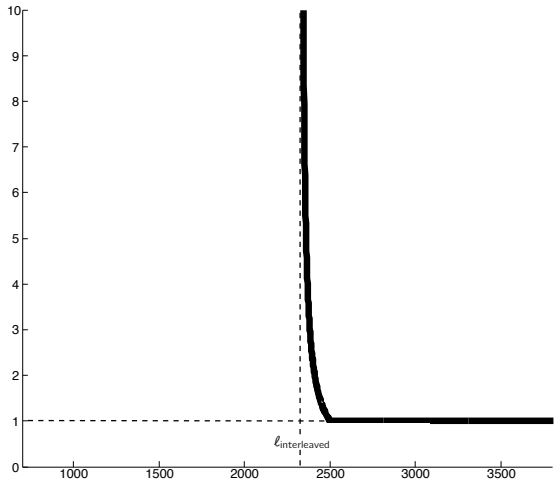


Figure 4.5: Lower bounds for human ch22: requiring coverage of the sequence implies $\bar{c} \geq 1$, and the generalized Ukkonen's condition imposes L be to the right of thick vertical line.

4.4 Towards an optimal assembly algorithm

We now begin our search for algorithms performing close to the lower bounds derived in the previous section. Algorithm assessment begins with obtaining deterministic sufficient conditions for success in terms of the positions of reads relative to repeats in the sequence. We then find the necessary N and L in order to satisfy these sufficient conditions with a target probability. The required coverage depth for each algorithm depends only on certain repeat statistics extracted from DNA data, which may be thought of as a *sufficient statistic*.

Greedy algorithm

The greedy algorithm was used by several of the most widely used genome assemblers for Sanger data, such as phrap, TIGR Assembler [55], and CAP3 [56]. SSAKE [57] is a more recent assembler that uses the greedy algorithm on high-throughput shotgun sequencing with short read data. The greedy algorithm's underlying data structure is the overlap graph, where each node represents a read and each (directed) edge $(\mathbf{y}, \mathbf{x})$ is labeled with the overlap $\text{ov}(\mathbf{y}, \mathbf{x})$ between the incident nodes' reads. The overlap of two reads $\text{ov}(\mathbf{y}, \mathbf{x})$ is defined to be the length of the longest prefix of $\mathbf{x}$ equal to a suffix of $\mathbf{y}$. For a node $\mathbf{v}$, the in-degree $d_{\text{in}}(\mathbf{v}) = |\{\mathbf{u} : (\mathbf{u}, \mathbf{v}) \text{ is an edge}\}|$ is the number of edges in the graph directed towards $\mathbf{v}$ and the out-degree $d_{\text{in}}(\mathbf{v}) = |\{\mathbf{u} : (\mathbf{v}, \mathbf{u}) \text{ is an edge}\}|$ is the number of edges directed away from $\mathbf{v}$. The algorithm is described as follows.

Algorithm 1 GREEDY. Input: reads $\mathcal{R}$. Output: sequence $\hat{\mathbf{s}}$.

1. For each read with sequence $\mathbf{x}$, form a node with label $\mathbf{x}$.

Greedy steps 2-3:

2. Consider all pairs of nodes $\mathbf{x}_1, \mathbf{x}_2$ in $\mathbb{G}$ satisfying $d_{\text{out}}(\mathbf{x}_1) = d_{\text{in}}(\mathbf{x}_2) = 0$, and add an edge $(\mathbf{x}_1, \mathbf{x}_2)$ with largest value $\text{ov}(\mathbf{x}_1, \mathbf{x}_2)$.
3. Repeat Step 2 until no candidate pair of nodes remains.

Finishing step:

4. Output the sequence corresponding to the unique cycle in $\mathbb{G}$.

Theorem 27. *Given a sequence* $\mathbf{s}$ *and a set of reads,* GREEDY *returns* $\mathbf{s}$ *if every repeat is bridged.*

Proof. We prove the contrapositive. Suppose GREEDY makes its first error in merging reads $\mathbf{r}_i$ and $\mathbf{r}_j$ with overlap $\text{ov}(\mathbf{r}_i, \mathbf{r}_j) = \ell$. Now, if $\mathbf{r}_j$ is the successor to $\mathbf{r}_i$, then the error is due to incorrectly aligning the reads; the other case is that $\mathbf{r}_j$ is not the successor of $\mathbf{r}_i$. In the first case, the subsequence $\mathbf{s}_{t_j}^{\ell}$ is repeated at location $\mathbf{s}_{t_i+L-\ell}^{\ell}$, and no read bridges either repeat copy.

In the second case, there is a repeat $\mathbf{s}_{t_j}^{\ell} = \mathbf{s}_{t_i+L-\ell}^{\ell}$. If $\mathbf{s}_{t_i+L-\ell}^{\ell}$ is bridged by some read $\mathbf{r}_k$, then $\mathbf{r}_i$ has overlap at least $\ell + 1$ with $\mathbf{r}_k$, implying that read $\mathbf{r}_i$ has already found its successor before step ℓ (either $\mathbf{r}_k$ or some other read with even higher overlap). A similar argument shows that $\mathbf{s}_{t_j}^{\ell}$ cannot be bridged, hence there is an unbridged repeat. □

Theorem 27 allows us to determine the coverage depth required by GREEDY: we must ensure that all repeats are bridged. A calculation similar to the one for bridging interleaved repeats (4.6) gives

$$P_{\text{error}} \approx \sum_m a_m e^{-2\lambda(L-m-1)}\,, \tag{4.9}$$

where a_m is the number of repeats of length m. Requiring $P_{\text{error}} \leq \epsilon$ and solving for L gives

$$L \geq \frac{1}{2\lambda} \log \frac{\gamma}{\epsilon} = \frac{G}{2N} \log \frac{\gamma}{\epsilon}\,, \tag{4.10}$$

where $\gamma := \sum_m a_m e^{2(N/G)(m+1)}$.

The performance obtained by the greedy algorithm is plotted in Fig. 4.6, and nearly matches the lower bound. Chromosome 22 is special, however, in that ℓ_{repeat} (which determines the performance of GREEDY) is not much larger than $\ell_{\text{interleaved}}$ which forms the lower bound.

Chromosome 19 has a large difference between $\ell_{\text{interleaved}} = 2248$ and $\ell_{\text{repeat}} = 4092$, and in Fig 4.7 we see a correspondingly large gap. GREEDY is evidently sub-optimal in handling interleaved repeats. Nevertheless, once the read length is slightly longer than ℓ_{repeat}, for Chromosome 19 GREEDY requires only $\bar{c} \geq 1$, i.e. as soon as the reads are longer than ℓ_{repeat}, coverage of the sequence is sufficient for GREEDY's successful reconstruction.

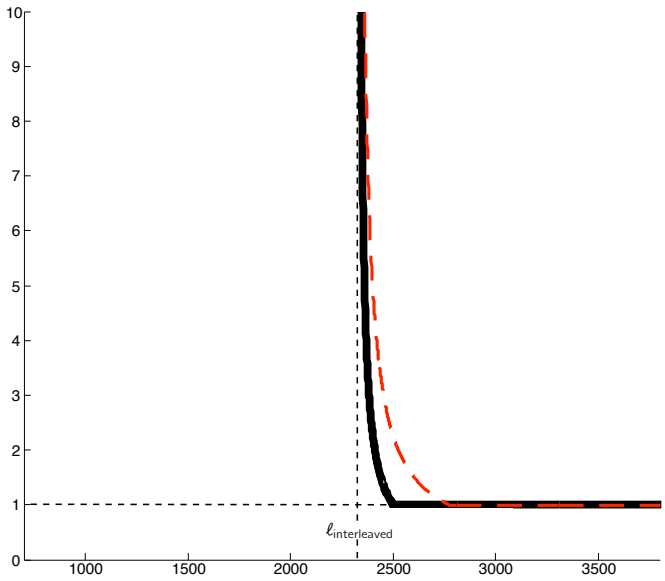


Figure 4.6: Greedy performance (in red) for Human chromosome 22.

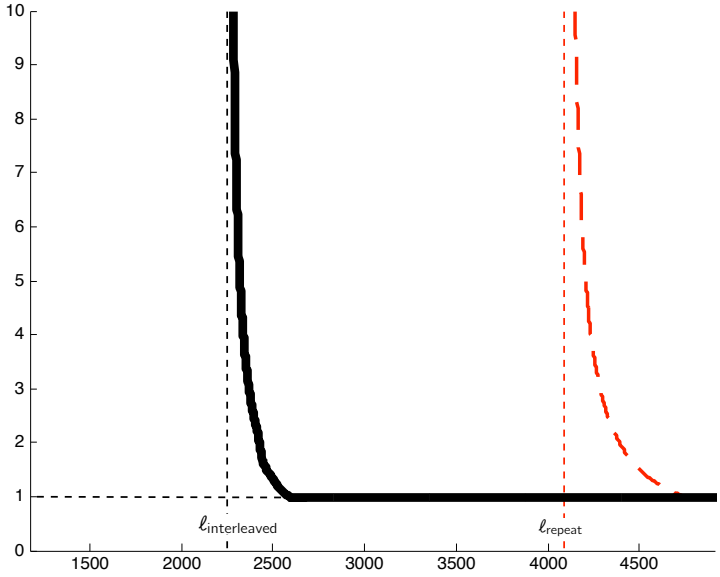


Figure 4.7: Greedy performance (in red) for Human chromosome 19.

K-mer algorithms

The greedy algorithm requires a read length much longer than called for by Ukkonen's condition: it fails when there are unbridged repeats, even if there are no unbridged *interleaved* repeats. As mentioned in the introduction, Ukkonen's condition was originally introduced in the context of SBH, where one observes all length L subsequences. (The set of all length

L subsequences is known as the L-spectrum and denoted by $\mathcal{S}_L$.) Since our lower bound generalizes Ukkonen's condition to shotgun sequencing, it makes sense to likewise turn to SBH for an assembly algorithm.

Sequencing by Hybridization

For the SBH model, where the set of reads is the L-spectrum $\mathcal{S}_L$, an optimal reconstruction algorithm was discovered by Pevzner [61] . Here optimality means it matches Ukkonen's necessary condition: if there are no interleaved or triple repeats of length at least $L-1$, then reconstruction is possible. Pevzner's algorithm is based on finding an appropriate cycle in a K-mer graph (also known as a de Bruijn graph) with $K = L-1$ (see e.g. [62] for an overview). A K-mer graph is formed from a sequence $\mathbf{s}$ by first adding a node to the graph for each unique K-mer (length K subsequence) found in the set of reads, and then adding an edge between any two nodes representing adjacent K-mers (two K-mers in a read are said to be *adjacent* if they overlap by $K-1$ nucleotides). Each edge is included only once, independent of how many reads designate its inclusion. Edges thus correspond to $(K+1)$-mers in $\mathbf{s}$ and paths correspond to longer subsequences obtained by merging the labels of the constituent nodes with offset one at each step. There exists a cycle corresponding to the original sequence $\mathbf{s}$, as shown in the following lemma, and reconstruction entails finding this cycle.

Lemma 28. *Fix an arbitrary K and form the K-mer graph from the $(K+1)$-spectrum $\mathcal{S}_{K+1}$. The sequence $\mathbf{s}$ corresponds to a unique cycle $\mathcal{C}(\mathbf{s})$ traversing each edge at least once.*

To prove the lemma, note that all $(K+1)$-mers in $\mathbf{s}$ correspond to edges and adjacent $(K+1)$-mers in $\mathbf{s}$ are represented by adjacent edges. An induction argument shows that $\mathbf{s}$ corresponds to a cycle. The cycle traverses all the edges, since each edge represents a unique $(K+1)$-mer in $\mathbf{s}$.

In both SBH and shotgun sequencing the number of times each edge e is traversed by $\mathcal{C}(\mathbf{s})$ (henceforth called the *multiplicity* of e) is unknown a priori, and finding this number is part of the reconstruction task. Repeated $(K+1)$-mers in $\mathbf{s}$ correspond to edges in the K-mer graph traversed more than once by $\mathcal{C}(\mathbf{s})$, i.e. having multiplicity greater than one. In order to estimate the multiplicity, previous works seek a solution to the so-called Chinese Postman Problem (CPP), in which the goal is to find a cycle of the shortest total length traversing every edge in the graph (see e.g. [63], [58], [47], [46]). It is not obvious under what conditions the CPP solution *correctly* assigns multiplicities in agreement with $\mathcal{C}(\mathbf{s})$. For our purposes, as we will see in Theorem 33, the multiplicity estimation problem can be sidestepped (thereby avoiding the CPP formulation) through a certain modification to the K-mer graph.

Ignoring the issue of edge multiplicities for a moment, Pevzner [61] showed for the SBH model that if the edge multiplicities are known with multiple copies of each edge included according to the multiplicities, and moreover Ukkonen's condition is satisfied, then there is a *unique Eulerian cycle* in the K-mer graph and the Eulerian cycle corresponds to the

original sequence. (An Eulerian cycle is a cycle traversing each edge exactly once.) Pevzner's algorithm is thus to find an Eulerian cycle and read off the corresponding sequence. Both steps can be done efficiently.

Lemma 29 (Pevzner [61])**.** *In the SBH setting, if the edge multiplicities are known, then there is a unique Eulerian cycle in the K-mer graph with $K = L - 1$ if and only if there are no unbridged interleaved repeats or unbridged triple repeats.*

Most practical algorithms (e.g. [58], [64], [65]) condense unambiguous paths (called unitigs by Myers [66] in a slightly different setting) for computational efficiency. The more significant benefit for us, as shown in Theorem 33, is that if Ukkonen's condition is satisfied then condensing the graph obviates the need to estimate multiplicities. Condensing a K-mer graph results in a graph of the following type.

Definition 30 (Sequence graph)**.** *A* sequence graph *is a graph in which each node is labeled with a subsequence, and edges $(\mathbf{u}, \mathbf{v})$ are labeled with an* overlap $a_{\mathbf{uv}}$ *such the subsequences $\mathbf{u}$ and $\mathbf{v}$ overlap by $a_{\mathbf{uv}}$ (the overlap is not necessarily maximal). In other words, an edge label $a_{\mathbf{uv}}$ between nodes $\mathbf{u}$ and $\mathbf{v}$ indicates that the $a_{\mathbf{uv}}$-length suffix of $\mathbf{u}$ is identical to the $a_{\mathbf{uv}}$-length prefix of $\mathbf{v}$.*

The sequence graph generalizes both the overlap graph used by Greedy in Section 4.4 (nodes correspond to reads, and edge overlaps are *maximal* overlaps) as well as the K-mer algorithms discussed in this section (nodes correspond to K-mers, and edge overlaps are $K - 1$).

We will perform two basic operations on the sequence graph. For an edge $e = (\mathbf{u}, \mathbf{v})$ with overlap $a_{\mathbf{uv}}$, *merging* $\mathbf{u}$ and $\mathbf{v}$ along e produces the concatenation $\mathbf{u}_1^{\text{end}}\mathbf{v}_{a_{\mathbf{uv}}+1}^{\text{end}}$. *Contracting* an edge $e = (\mathbf{u}, \mathbf{v})$ entails two steps (c.f. Fig. 4.8): first, merging $\mathbf{u}$ and $\mathbf{v}$ along e to form a new node $\mathbf{w} = \mathbf{u}_1^{\text{end}}\mathbf{v}_{a_{\mathbf{uv}}+1}^{\text{end}}$, and, second, edges to $\mathbf{u}$ are replaced with edges to $\mathbf{w}$, and edges from $\mathbf{v}$ are replaced by edges from $\mathbf{w}$. We will only contract edges $(\mathbf{u}, \mathbf{v})$ with $d_{\text{out}}(\mathbf{u}) = d_{\text{in}}(\mathbf{v}) = 1$.

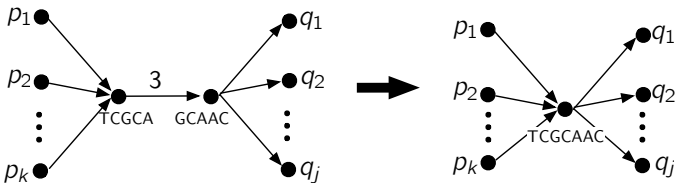


Figure 4.8: Contracting an edge by merging the incident nodes as described immediately before Defn. 31. Repeating this operation results in the condensed graph.

We now define what it means to condense a graph.

Definition 31 (Condensed sequence graph)**.** *The* condensed sequence graph *replaces unambiguous paths by single nodes. Concretely, any edge $e = (u, v)$ with $d_{out}(u) = d_{in}(v) = 1$ is contracted, and this is repeated until no candidate edges remain, yielding the* condensed sequence graph.

For a path $\mathcal{P} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q$ in the original graph, the corresponding path in the condensed graph is obtained by contracting an edge $(\mathbf{v}_i, \mathbf{v}_{i+1})$ whenever it is contracted in the graph, replacing the node $\mathbf{v}_1$ by $\mathbf{w}$ whenever an edge $(\mathbf{u}, \mathbf{v}_1)$ is contracted to form $\mathbf{w}$, and similarly for the final node $\mathbf{v}_q$. It is impossible for an intermediate node $\mathbf{v}_i$, $2 \leq i < q$, to be merged with a node outside of $\mathcal{P}$, as this would violate the condition $d_{\text{out}}(u) = d_{\text{in}}(v) = 1$ for edge contraction in Defn. 31.

In the condensed sequence graph $\mathbb{G}$ obtained from a sequence $\mathbf{s}$, nodes correspond to subsequences via their labels, and paths in $\mathbb{G}$ correspond to subsequences in $\mathbf{s}$ via merging the constituent nodes along the path. If the subsequence corresponding to a node $\mathbf{v}$ appears twice or more in $\mathbf{s}$, we say that $\mathbf{v}$ corresponds to a repeat. Conversely, subsequences of length $\ell \geq K$ in $\mathbf{s}$ correspond to paths $\mathcal{P}$ of length $\ell - K + 1$ in the K-mer graph, and thus by the previous paragraph also to paths in the condensed graph $\mathbb{G}$.

We record a few simple facts about the condensed sequence graph obtained from a K-mer graph.

Lemma 32. *Let $\mathbb{G}_0$ be the K-mer graph constructed from the $(K+1)$-spectrum of $\mathbf{s}$ and let $\mathcal{C}_0 = \mathcal{C}_0(\mathbf{s})$ be the cycle corresponding to $\mathbf{s}$. In the condensed graph $\mathbb{G}$, let $\mathcal{C}$ be the cycle obtained from $\mathcal{C}_0$ by contracting the same edges as those contracted in $\mathbb{G}_0$.*

1. *Edges in $\mathbb{G}_0$ can be contracted in any order, resulting in the same graph $\mathbb{G}$, so the condensed graph is well-defined. Similarly $\mathcal{C}$ is well-defined.*
2. *The cycle $\mathcal{C}$ in $\mathbb{G}$ corresponds to $\mathbf{s}$ and is the unique such cycle.*
3. *The cycle $\mathcal{C}$ in $\mathbb{G}$ traverses each edge at least once.*

The condensed graph has the useful property that if the original sequence was reconstructible, then there is an Eulerian cycle corresponding to the sequence:

Theorem 33. *Let $\mathcal{S}_{K+1}$ be the $(K+1)$-spectrum of $\mathbf{s}$ and $\mathbb{G}_0$ be the K-mer graph constructed from $\mathcal{S}_{K+1}$, and let $\mathbb{G}$ be the condensed sequence graph obtained from $\mathbb{G}_0$. If Ukkonen's condition is satisfied, i.e. there are no triple repeats or interleaved repeats of length at least K, then there is a unique Eulerian cycle $\mathcal{C}$ in $\mathbb{G}$ and $\mathcal{C}$ corresponds to $\mathbf{s}$.*

Proof. We will show that if Ukkonen's condition is satisfied, the cycle $\mathcal{C} = \mathcal{C}(\mathbf{s})$ in $\mathbb{G}$ corresponding to $\mathbf{s}$ (constructed in Lemma 32) traverses each edge exactly once in the condensed K-mer graph, i.e. $\mathcal{C}$ is Eulerian. Pevzner's [61] arguments show that if there are multiple Eulerian cycles then Ukkonen's condition is violated, so it is sufficient to prove that $\mathcal{C}$ is Eulerian. As noted in Lemma 32, $\mathcal{C}$ traverses each edge at least once, and thus it remains only to show that $\mathcal{C}$ traverses each edge *at most* once.

To begin, let $\mathcal{C}_0$ be the cycle corresponding to $\mathbf{s}$ in the original K-mer graph $\mathbb{G}_0$. We argue that every edge $(\mathbf{u}, \mathbf{v})$ traversed twice by $\mathcal{C}_0$ in the K-mer graph $\mathbb{G}_0$ has been contracted in the condensed graph $\mathbb{G}$ and hence in $\mathcal{C}$. Note that the cycle $\mathcal{C}_0$ does not traverse any node three times in $\mathbb{G}_0$, for this would imply the existence of a triple repeat of length K, violating

the hypothesis of the Lemma. It follows that the node $\mathbf{u}$ cannot have two outgoing edges in $\mathbb{G}_0$ as $\mathbf{u}$ would then be traversed three times; similarly, $\mathbf{v}$ cannot have two incoming edges. Thus $d_{\text{out}}(\mathbf{u}) = d_{\text{in}}(\mathbf{v}) = 1$ and, as prescribed in Defn. 31, the edge $(\mathbf{u}, \mathbf{v})$ has been contracted. □

Theorem 33 characterizes, deterministically, the values of K for which reconstruction from the $(K+1)$-spectrum is possible. We proceed with application of the K-mer graph approach to shotgun sequencing data.

Basic K-mer algorithm

Starting with Idury and Waterman [58], and then Pevzner et al.'s [47] EULER algorithm, most current assembly algorithms for shotgun sequencing are based on the K-mer graph. Idury and Waterman [58] made the key observation that SBH with subsequences of length $K+1$ can be *emulated* by shotgun sequencing if each read overlaps the subsequent read by K: the set of all $(K+1)$-mers within the reads is equal to the set of all $(K+1)$-mers in the sequence $\mathbf{s}$, i.e. the $(K+1)$-spectrum $\mathcal{S}_{K+1}$. Ignoring the reads gives rise to the algorithm DEBRUIJN described next, and algorithms using the reads to a greater extent are discussed in the next subsection.

Algorithm 2 DEBRUIJN. Input: reads $\mathcal{R}$, parameter K. Output: sequence $\hat{\mathbf{s}}$.

K-mer steps 1-3:

1. For each subsequence $\mathbf{x}$ of length K in a read, add a node to the graph $\mathbb{G}$ with label $\mathbf{x}$.
2. For each read, add edges between nodes representing adjacent K-mers in the read.
3. Condense the graph as described in Definition 31.
4. *Finishing step:* Find an Eulerian cycle in $\mathbb{G}$ and return the corresponding sequence.

The parameter K determines the *granularity* of the K-mer graph: repeats of length $K-1$ do not appear in the graph. Conditions (a) and (b) of Lemma 34 below ensure that K is large enough that neither interleaved nor triple repeats cause the graph to be tangled. At the same time, increasing K requires more reads as they must overlap by K: otherwise the K-mer graph is not connected. Condition (c) guarantees that the graph is connected.

As discussed below, the choice $K = 1 + \max\{\ell_{\text{triple}}, \ell_{\text{interleaved}}\}$ is optimal for DEBRUIJN, so the parameter K could be removed from the statement of Lemma 34. We keep K explicit for two reasons. First, algorithm DEBRUIJN does not actually require *a priori* knowledge of the repeat lengths and thus in a hypothetical practical use of the algorithm K would typically be larger than the optimum. Second, the sufficient conditions in Lemma 34 are organized in order to facilitate comparison with the improved K-mer algorithms in the next subsection.

Lemma 34. *Fix a sequence* $\mathbf{s}$. *Then* DEBRUIJN *with parameter choice* K *successfully returns the sequence* $\mathbf{s}$ *if the reads satisfy the following assumptions:*

(a) interleaved repeats: no interleaved repeats of length at least K, i.e. $K > \ell_{interleaved}$

(b) triple repeats: no triple repeats of length at least K, i.e. $K > \ell_{triple}$

(c) coverage depth: every read overlaps its successor by at least K.

Proof. By assumption (c) all $(K+1)$-mers in $\mathbf{s}$ are contained in reads, so the K-mer graph constructed by DeBruijn is the same as from the $(K+1)$-spectrum $\mathcal{S}_{K+1}$. Now conditions (a) and (b) state that $\mathbf{s}$ has no triple or interleaved repeats of length $\geq K$, so the hypotheses of Theorem 33 are met and there is a unique Eulerian cycle found by DeBruijn. □

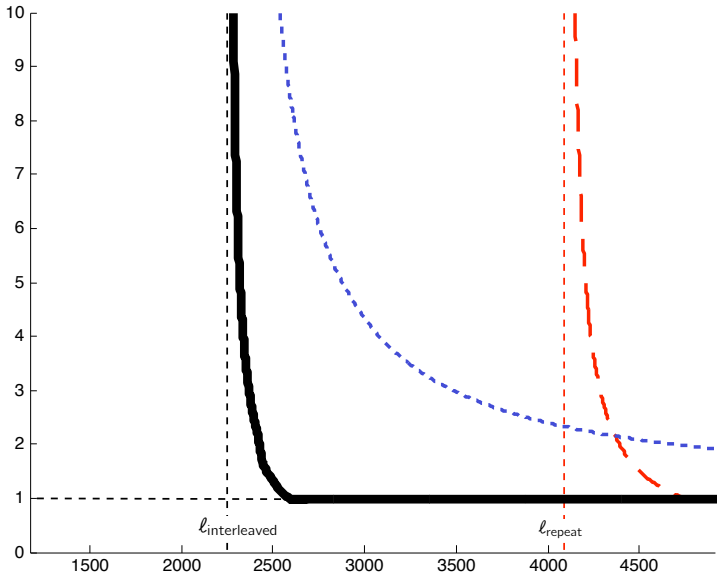


Figure 4.9: Performance of DeBruijn (in blue) on Chrom19.

Lander and Waterman's coverage analysis [49] applies also to Condition (c) of Lemma 34. The coverage depth required in order that each read overlap the subsequent read by K base pairs is approximately given by solving for $N_{K-\text{cov}}$ in the equation

$$N_{K-\text{cov}} = \frac{G}{L-K} \log \frac{N_{K-\text{cov}}}{\epsilon}.$$

Comparing to our baseline N_{LW}, we have

$$\frac{N_{K-\text{cov}}}{N_{\text{LW}}} = \frac{1}{1-\frac{K}{L}} \cdot \frac{\log N_{K-\text{cov}} - \log \epsilon}{\log N_{\text{LW}} - \log \epsilon} \approx \frac{1}{1-\frac{K}{L}}. \tag{4.11}$$

Since decreasing K reduces the coverage depth required, Lemma 34 identifies the optimal choice of K to be $K = 1 + \max\{\ell_{\text{triple}}, \ell_{\text{interleaved}}\}$. Plugging this value into (4.11), we have

$$\frac{N}{N_{\text{LW}}} \gtrsim \frac{1}{1-\frac{1+\max\{\ell_{\text{triple}},\ell_{\text{interleaved}}\}}{L}}. \tag{4.12}$$

The performance of DeBruijn is plotted in Fig. 4.9. DeBruijn significantly improves on Greedy by obtaining the correct first order performance: given sufficiently many reads, the read length L may be decreased to $1 + \max\{\ell_{\text{triple}}, \ell_{\text{interleaved}}\}$. Still, the number of reads required to approach this critical length is large and remains far from the lower bound. In order to reduce the number of reads required while staying within the K-mer graph framework, K must be reduced. We pursue this in the following subsection.

Improved K-mer algorithms

Algorithm DeBruijn ignores a lot of information contained in the reads, and indeed all of the K-mer based algorithms proposed by the sequencing community (including [58], [47], [67], [68], [64], [65]) use the read information to a greater extent than the naive DeBruijn algorithm. Better use of the read information, as described below in algorithms SimpleBridging and MultiBridging, will allow to relax the condition $K > \max\{\ell_{\text{interleaved}}, \ell_{\text{triple}}\}$ for success of DeBruijn, which in turn reduces the high coverage depth required by Condition (c).

Existing algorithms use read information in a variety of distinct ways to resolve repeats. For instance, Pevzner et al. [47] observe that for graphs where each edge has multiplicity one, if one copy of a repeat is bridged, the repeat can be resolved through what they call a "detachment". The algorithm SimpleBridging described here is very similar, and resolves repeats with two copies if at least one copy is bridged. Meanwhile, other algorithms are better suited to higher edge multiplicities due to higher order repeats; IDBA (Iterative DeBruijn Assembler) [59] creates a series of K-mer graphs, each with larger K, and at each step uses not just the reads to identify adjacent K-mers, but also all the unbridged paths in the K-mer graph with smaller K. Although it is not stated explicitly in their paper, we make the observation here that if all copies of every repeat are bridged, then this is sufficient to ensure reconstruction.

It is suboptimal to require that *all copies of* every repeat up to the maximal K be bridged. We introduce MultiBridging, which combines these ideas to simultaneously allow for single-bridged double repeats, triple repeats in which all copies are bridged, and unbridged non-interleaved repeats.

Resolving 2-repeats: SimpleBridging

SimpleBridging improves on DeBruijn by resolving bridged 2-repeats (i.e. a repeat with exactly two copies in which at least one copy is bridged by a read). Condition (a) $K > \ell_{\text{interleaved}}$ for success of DeBruijn (ensuring that no interleaved repeats appear in the initial K-mer graph) is updated to require only no *unbridged* interleaved repeats, which matches the lower bound. With this change, Condition (b) $K > \ell_{\text{triple}}$ forms the bottleneck for typical DNA sequences. Thus SimpleBridging is optimal with respect to interleaved repeats, but it is suboptimal with respect to triple repeats.

SimpleBridging deals with repeats by performing surgery on certain nodes in the sequence graph. In the sequence graph, a repeat corresponds to a node we call an *X-node*, a node with in-degree and out-degree each at least two (Fig. 4.10). A self-loop (e.g. Fig. 4.14b) contributes one each to the in-degree and out-degree. The cycle $\mathcal{C}(\mathbf{s})$ traverses each X-node at least twice, so X-nodes correspond to repeats in $\mathbf{s}$. The converse is false: not all repeats correspond to X-nodes. We call an X-node which is traversed exactly twice a 2-X-node; these nodes have in-degree and out-degree 2 and correspond to 2-repeats. An X-node is said to be bridged if the corresponding repeat in $\mathbf{s}$ is bridged.

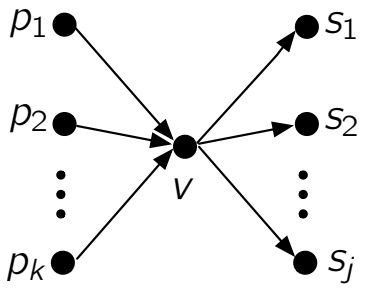


Figure 4.10: An X-node in a condensed sequence graph is a junction centered at a node v with $d_{\text{out}}(v) \geq 2$ and $d_{\text{in}}(v) \geq 2$. A self-loop (e.g. Fig. 4.14b) contributes one each to the in-degree and out-degree.

In the *bridging step* of SimpleBridging (illustrated in Fig. 4.11), bridged 2-X-nodes are duplicated in the graph and incoming and outgoing edges are inferred using the bridging read, reducing possible ambiguity. Since bridging reads extend one base to either end of a repeat, it will be convenient to use the following notation for extending sequences: Given an X-node $\mathbf{v}$ with an incoming edge $(\mathbf{p}, \mathbf{v})$ and an outgoing edge $(\mathbf{v}, \mathbf{q})$, let

$$\mathbf{v}^{\rightarrow \mathbf{q}} = \mathbf{v}_1^{\text{end}}\, \mathbf{q}_{a_{\mathbf{vq}}+1}^{a_{\mathbf{vq}}+1}, \quad \text{and} \quad {}^{\mathbf{p}\rightarrow}\mathbf{v} = \mathbf{p}_{\text{end}-a_{\mathbf{pv}}}^{\text{end}-a_{\mathbf{pv}}} \mathbf{v}_1^{\text{end}}\,. \tag{4.13}$$

Here $\mathbf{v}^{\rightarrow \mathbf{q}}$ denotes the subsequence $\mathbf{v}$ appended with the single next base in the merging of $\mathbf{v}$ and $\mathbf{q}$ and ${}^{\mathbf{p}\rightarrow}\mathbf{v}$ the subsequence $\mathbf{v}$ prepended with the single previous base in the merging of $\mathbf{p}$ and $\mathbf{v}$. For example, if $\mathbf{v} = \text{ATTC}$, $\mathbf{p} = \text{TCAT}$, $a_{\mathbf{pv}} = 2$, $\mathbf{q} = \text{TTCGCC}$, and $a_{\mathbf{vq}} = 3$, then $\mathbf{v}^{\rightarrow \mathbf{q}} = \text{ATTCG}$, ${}^{\mathbf{p}\rightarrow}\mathbf{v} = \text{CATTC}$, and ${}^{\mathbf{p}\rightarrow}\mathbf{v}^{\rightarrow \mathbf{q}} = \text{CATTCG}$. The idea is that a bridging read is consistent with only one pair ${}^{\mathbf{p}\rightarrow}\mathbf{v}$ and $\mathbf{v}^{\rightarrow \mathbf{q}}$ and thus allows to match up edge $(\mathbf{p}, \mathbf{v})$ with $(\mathbf{v}, \mathbf{q})$.

Lemma 35. *Suppose $\mathcal{C}$ corresponds to a sequence $\mathbf{s}$ in a condensed sequence graph $\mathbb{G}$. If a read $\mathbf{r}$ bridges an X-node $\mathbf{v}$, then there are unique edges $(\mathbf{p}, \mathbf{v})$ and $(\mathbf{v}, \mathbf{q})$ such that ${}^{\mathbf{p}\rightarrow}\mathbf{v}$ and $\mathbf{v}^{\rightarrow \mathbf{q}}$ are adjacent in $\mathbf{r}$.*

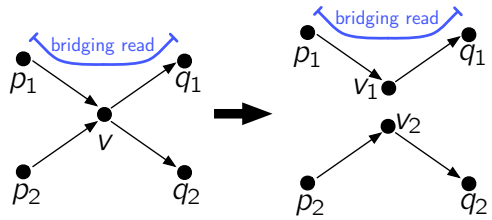


Figure 4.11: When a read bridges a repeat with two copies, the 2-X-node corresponding to the repeat is duplicated and potential ambiguity is reduced.

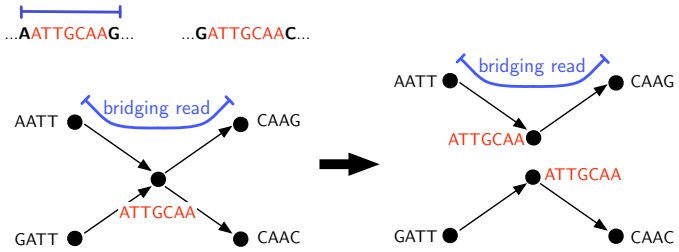


Figure 4.12: An example of the bridging step in SimpleBridging.

Algorithm 3 SimpleBridging. Input: reads $\mathcal{R}$, parameter K. Output: sequence $\hat{\mathbf{s}}$.

K-mer steps 1-3:

1. For each subsequence $\mathbf{x}$ of length K in a read, form a node with label $\mathbf{x}$.
2. For each read, add edges between nodes representing adjacent K-mers in the read.
3. Condense the graph as described in Definition 31.
4. *Bridging step:* See Fig. 4.11. While there exists an X-node $\mathbf{v}$ with $d_{\text{in}}(\mathbf{v}) = d_{\text{out}}(\mathbf{v}) = 2$ bridged by some read $\mathbf{r}$: (i) Remove $\mathbf{v}$ and edges incident to it. Add duplicate nodes $\mathbf{v}_1, \mathbf{v}_2$. (ii) Choose the unique $\mathbf{p}_i$ and $\mathbf{q}_j$ such that $\mathbf{p}_i{}^{\rightarrow}\mathbf{v}$ and $\mathbf{v}^{\rightarrow}\mathbf{q}_j$ are adjacent in $\mathbf{r}$, and add edges $(\mathbf{p}_i, \mathbf{v}_1)$ and $(\mathbf{v}_1, \mathbf{q}_j)$. Choose the unused $\mathbf{p}_i$ and $\mathbf{q}_j$, and add edges $(\mathbf{p}_i, \mathbf{v}_2)$ and $(\mathbf{v}_2, \mathbf{q}_j)$. (iii) Condense the graph.
5. *Finishing step:* Find an Eulerian cycle in the graph and return the corresponding sequence.

Lemma 36. *Fix a sequence* $\mathbf{s}$. *The algorithm* SimpleBridging *with parameter choice* K *correctly reconstructs* $\mathbf{s}$ *if the reads satisfy the following assumptions:*

(a) interleaved repeats: no ***unbridged*** *interleaved repeats*

(b) triple repeats: no triple repeats of length $\geq K$, *i.e.* $K > \ell_{triple}$

(c) coverage depth: every read overlaps its successor by at least K.

Proof. The proof is very similar to that of Lemma 37 below and is omitted. □

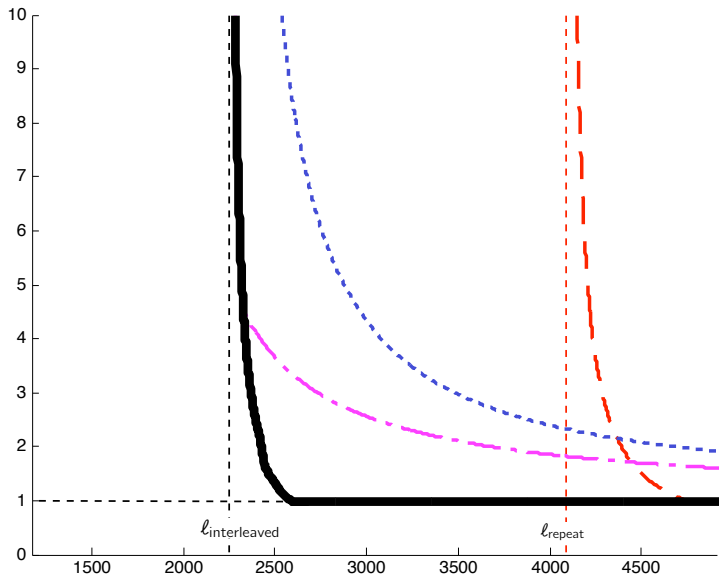


Figure 4.13: Performance of SimpleBridging (in magenta) on Chrom19.

Figure 4.13 plots the performance of SimpleBridging. We produce the plot by translating the conditions of Lemma 36 into statements relating G, N, L, and ϵ. Condition (a) that there be no unbridged interleaved repeats was already derived as part of the lower bound in Section 4.3, yielding (4.8). Condition (b) requires $K > \ell_{\text{triple}}$, and therefore the best choice for K is $K = 1 + \ell_{\text{triple}}$. Plugging into Condition (c) as given by (4.11), computed in the context of DeBruijn, we have

$$\frac{N}{N_{\text{LW}}} \gtrsim \frac{1}{1 - \frac{1+\ell_{\text{triple}}}{L}} . \tag{4.14}$$

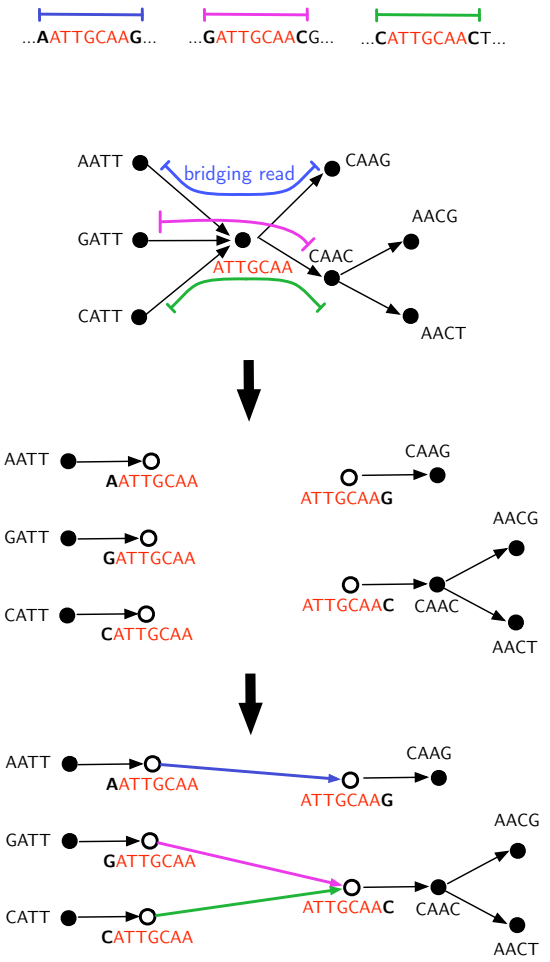
The plot is obtained by choosing the minimum N and L satisfying the two conditions (4.8) and (4.14).

Resolving triple repeats: MultiBridging

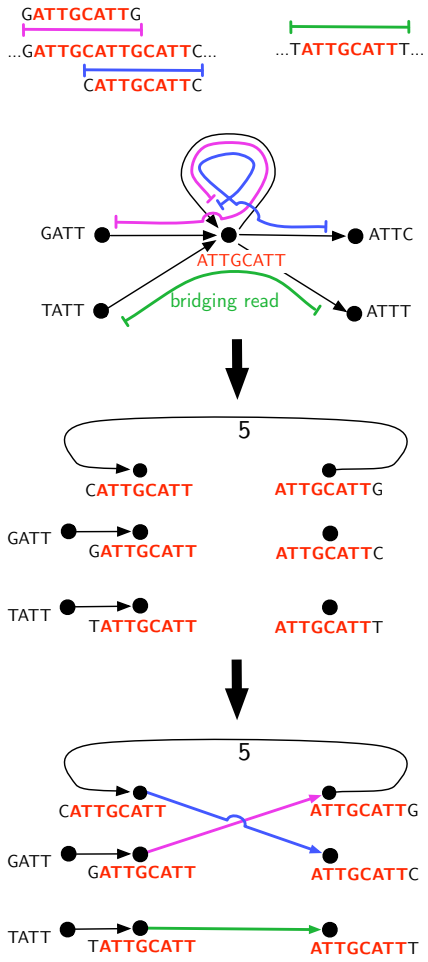
We now turn to triple repeats. As previously observed, it can be challenging to resolve repeats with more than one copy [47], because an edge into the repeat may be paired with more than one outgoing edge. As discussed above at the top of the section, our approach here shares elements with IDBA [59].

The algorithm MultiBridging has a more sophisticated *bridging step II*, which resolves higher order repeats. As noted in the previous subsection, repeats correspond to nodes in the sequence graph we call *X-nodes* (c.f. Fig. 4.10), nodes with in-degree and out-degree each at least two. All X-nodes correspond to repeats, but not all repeats correspond to X-nodes. A repeat is said to be *all-bridged* if *all* repeat copies are bridged, and an X-node is called all-bridged if the corresponding repeat is all-bridged.

The requirement that triple repeats be all-bridged allows them to be resolved *locally*. The X-node resolution procedure given in Step 4 of MultiBridging can be interpreted in the K-mer graph framework as increasing K locally so that repeats do not appear in the graph. An outline of the X-node resolution step is as follows. First, new nodes are added with labels one base longer than the repeat length, one for each incoming or outgoing edge of the X-node. (An extra edge is added for self-loops.) Edges are then added for adjacent subsequences using the bridging reads, as per Lemma 35. Repeating this process increases the node lengths so that repeats are eventually contained safely in the interior of nodes, where they cause no ambiguity.



(a) The X-node with label ATTGCAA corresponds to a nested repeat.



(b) X-node with a self-loop

Figure 4.14: X-node resolution for two different examples.

Algorithm 4 MULTIBRIDGING. Input: reads $\mathcal{R}$, parameter K. Output: sequence $\hat{\mathbf{s}}$.

K-mer steps 1-3:

1. For each subsequence $\mathbf{x}$ of length K in a read, form a node with label $\mathbf{x}$.
2. For each read, add edges between nodes representing adjacent K-mers in the read.
3. Condense the graph as described in Definition 31.
4. *Bridging step II:* While there exists a bridged X-node $\mathbf{v}$ (Fig. 4.10): (i) For each edge $(\mathbf{p}_i, \mathbf{v})$ create a new node $\mathbf{u}_i = {}^{\mathbf{p}_i\rightarrow}\mathbf{v}$ and edge $(\mathbf{p}_i, \mathbf{u}_i)$ with weight $a_{\mathbf{p}_i,\mathbf{v}} + 1$ and for each edge $(\mathbf{v}, \mathbf{q}_j)$ create a new node $\mathbf{w}_j = \mathbf{v}^{\rightarrow\mathbf{q}_j}$ and edge $(\mathbf{v}, \mathbf{q}_j)$ with weight $a_{\mathbf{v},\mathbf{q}_i} + 1$ (notation defined in (4.13)). (ii) If $\mathbf{v}$ has a self-loop $(\mathbf{v}, \mathbf{v})$ with weight $a_{\mathbf{v},\mathbf{v}}$, add an edge $(\mathbf{v}^{\rightarrow\mathbf{v}}, {}^{\mathbf{v}\rightarrow}\mathbf{v})$ with weight $a_{\mathbf{v},\mathbf{v}} + 2$ (c.f. Fig. 4.14b). (iii) Remove node $\mathbf{v}$ and all incident edges. (iv) For each pair $\mathbf{u}_i, \mathbf{w}_j$ adjacent in a read, add edge $(\mathbf{u}_i, \mathbf{w}_j)$. If exactly one each of the $\mathbf{u}_i$ and $\mathbf{w}_j$ nodes have no added edge, add the edge. (v) Condense the graph.
5. *Finishing step:* Find an Eulerian cycle in the graph and return the corresponding sequence.

Lemma 37. *Fix a sequence* $\mathbf{s}$*. The algorithm* MULTIBRIDGING *successfully reconstructs the sequence* $\mathbf{s}$ *if the reads satisfy the following assumptions:*

(a) interleaved repeats: no ***unbridged*** *interleaved repeats*

(b) triple repeats: all triple repeats are ***all-bridged***

(c) the sequence is covered by the reads.

Proof. The proof is contained in the appendix. □

Unlike the previous K-mer algorithms, DEBRUIJN and SIMPLEBRIDGING, it is unnecessary to specify a parameter K for MULTIBRIDGING. Implicitly MULTIBRIDGING uses $K = 1$, which makes the condition that reads overlap by K equivalent to coverage of the genome. Depending on the repetitiveness of the genome, building the initial K-mer graph with K around 20 or 40 seems to provide a good trade-off between computational efficiency and required number of reads.

Figure 4.15 plots the performance of MULTIBRIDGING, obtained by solving for the relationship between G, N, L, and ϵ in order to satisfy the conditions of Lemma 37. Condition (a) is already dealt with in (4.8), and Condition (c) is simply the requirement that $\frac{N}{N_{\text{LW}}} \geq 1$.

We turn to Condition (b) that all triple repeats are all-bridged. Let c_m denote the number of triple repeats of length m. A union bound estimate over triple repeats for the event that one such triple repeat fails to be all-bridged gives

$$P_{\text{error}} \approx \sum_m 3 \cdot c_m e^{-\lambda(L-m-1)}, \tag{4.15}$$

and requiring $P_{\text{error}} \leq \epsilon$ and solving for L yields

$$L \geq \frac{1}{\lambda} \log \frac{\gamma_3}{\epsilon} = \frac{G}{N} \log \frac{\gamma_3}{\epsilon}, \tag{4.16}$$

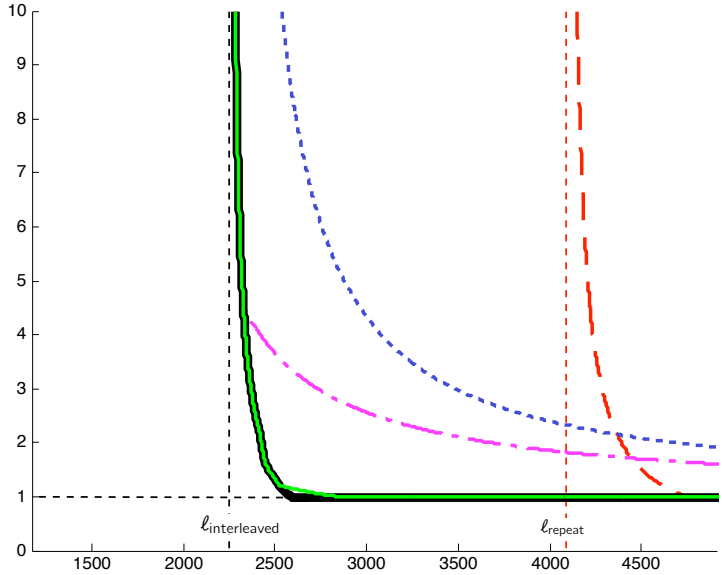


Figure 4.15: Performance of MultiBridging (in green) on Chrom19.

where $\gamma_3 := \sum_m 3c_m e^{(N/G)\cdot(m+1)}$ is computed from the triple repeat statistics c_m.

In order to understand the cost of all-bridging triple repeats, compared to simply bridging one copy as required by our lower bound, it is instructive to study the effect of the single longest triple repeat. Setting $c_{\ell_{\text{triple}}} = 1$ and $c_m = 0$ for $m \neq \ell_{\text{triple}}$ makes $\gamma_3 = 3e^{(N/G)\cdot(\ell_{\text{triple}}+1)}$ in (4.16) and

$$L \geq L_3^{\text{all}} := \ell_{\text{triple}} + 1 + \frac{1}{\lambda}\log 3\epsilon^{-1} = \ell_{\text{triple}} + 1 + \frac{G}{N}\log 3\epsilon^{-1}\,. \tag{4.17}$$

Bridging the longest triple repeat, as shown in Section 4.7, requires

$$L \geq L_3 := \ell_{\text{triple}} + 1 + \frac{1}{3\lambda}\log \epsilon^{-1} = \ell_{\text{triple}} + 1 + \frac{G}{3N}\log \epsilon^{-1}\,. \tag{4.18}$$

Solving for N in equations (4.18) and (4.17) gives

$$N_3 \geq \frac{G}{3}\cdot\frac{\log \epsilon^{-1}}{L - \ell_{\text{triple}} - 1} \quad \text{and} \quad N_3^{\text{all}} \geq G\cdot\frac{\log \epsilon^{-1} + \log 3}{L - \ell_{\text{triple}} - 1}\,. \tag{4.19}$$

The ratio is

$$\frac{N_3^{\text{all}}}{N_3} = 3\cdot\frac{\log 3\epsilon^{-1}}{\log \epsilon^{-1}} \approx 3.72 \quad \text{for } \epsilon = 10^{-2}\,. \tag{4.20}$$

This means that if the longest triple repeat is dominant, then for L slightly larger than ℓ_{triple}, MultiBridging needs a coverage depth approximately 3.72 times higher than required by our lower bound.

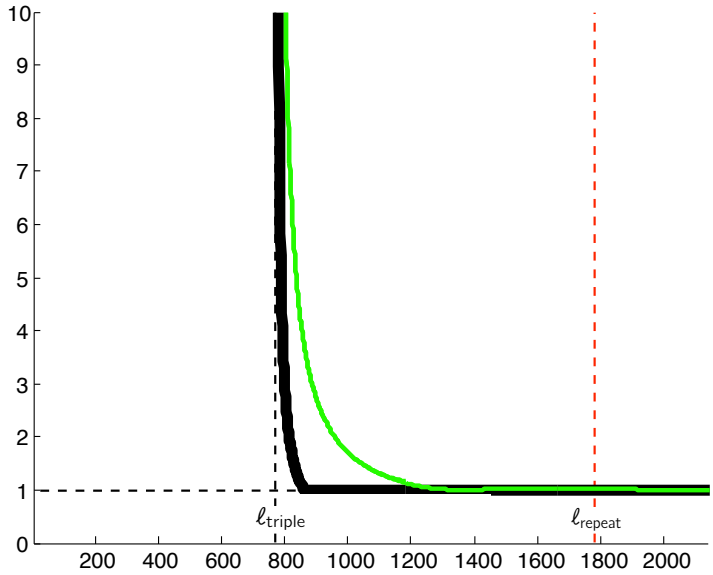


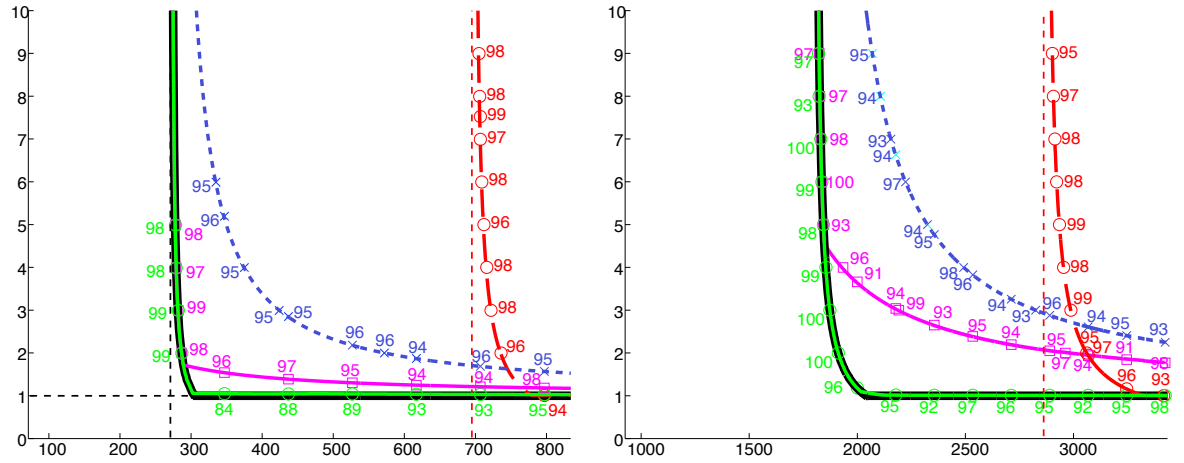
Figure 4.16: Performance of MultiBridging on *P. Marinus*. *The effect of the longest triple repeat is dominant here, so for L slightly larger than ℓ_{triple},* MultiBridging *needs a coverage depth higher than the coverage depth required by our lower bound, as predicted by equation* (4.20).

4.5 Algorithm simulations

In order to verify the performance predictions for the algorithms, we ran simulations of each algorithm along the curve at which $< 5\%$ error was predicted (Fig. 4.17). Values of N and L were sampled at regular intervals along the vertical $\bar{c} = N/N_{\text{LW}}$ axis and along the horizontal L axis, and projected onto the curve. For each point (L, N), we simulated 100 datasets and ran the various algorithms, recording how many times successful reconstruction was achieved.

This was done for two of the three data sets in the GAGE assembly evaluation for which complete reference sequence are available, namely Staphylococcus aureus and Rhodobacter sphaeroides. Runtimes on Human Chromosome 14 were too long to be able to obtain informative simulation results. Simulation results support the predictions based on repeat statistics: for each algorithm the numbers, e.g. 93, 98, 95, indicate the number of successful reconstructions out of 100 trials, and agree with the expected 5% error rate (allowing for random fluctuations).

We note that we ran MultiBridging with $K = 40$ in order to have reasonable runtimes; the dependence on runtime is further discussed in Section 4.6. This results in less than a 95% success rate for R. sphaeroides in the coverage-limited regime (i.e. camping from the line $\bar{c} = 1$).



(a) Simulation results for R. sphaeroides

(b) Simulation results for S. Aureus with $\epsilon = 5\%$.

Figure 4.17: Simulation results for R. sphaeroides and S. Aureus on each of the four algorithms, both with $\epsilon = 5\%$. For each algorithm the numbers, e.g. 93, 98, 95, indicate the number of successful reconstructions out of 100 trials with randomly sampled reads for each trial.

4.6 Discussion

Computational complexity

Computational complexity is an important consideration in the design of assembly algorithms. We note that most of the optimization-based formulations of assembly have been shown to be NP-hard, including SCS [3], [7], De Bruijn Superwalk [22], [12], and Minimum s-Walk on the string graph [15], [12]. The computational hardness results have led to heuristic-based algorithm development emphasizing computational efficiency. However, as pointed out by Nagarajan and Pop [18], typical instances of a problem may well be easier than the worst-case. Indeed, on typical repeat statistics, the algorithm MultiBridging achieves performance close to the information theoretic limits while being efficient.

We now compute the run-time of MultiBridging. The algorithm MultiBridging has two phases: the K-mer graph formation step, and the repeat resolution step. The K-mer graph formation runtime can be easily bounded by

$$O((L-K)NK),$$

assuming $O(K)$ look-up time for each of the $(L-K)N$ K-mers observed in reads. This step is common to all K-mer graph based algorithms, so previous works to decrease the practical runtime or memory requirements are applicable.

The repeat resolution step depends on the repeat statistics and choice of K. It can be loosely bounded as

$$O\Big(\sum_{\ell=K}^{L} L \sum_{\substack{\text{max repeats } x \\ \text{of length } \ell}} d_x\Big).$$

The second sum is over distinct maximal repeats x of length ℓ and d_x is the number of (not necessarily maximal) copies of repeat x. The bound comes from the fact that each maximal repeat of length $K < \ell < L$ is resolved via exactly one bridged X-node, and each such resolution requires examining at most the Ld_x distinct reads that contain the repeat. We note that

$$\sum_{\ell=K}^{L} L \sum_{\substack{\text{max repeats } x \\ \text{of length } \ell}} d_x < \sum_{\ell=K}^{L} La_\ell\,,$$

and the latter quantity is easily computable from our sufficient statistics. Creating the list of reads containing each repeat can be done as part of the K-mer graph formation step, and maintaining it does not add complexity to the resolutions. For our data sets, with appropriate choice of K, the bridging step is much simpler than the K-mer graph formation step: for RSPHAEROIDES we use $K_0 = 40$ to get $\sum_{\ell=K}^{L} La_\ell = 412$; in contrast, $N > 22421$ on our L-range is much larger. Similarly, for HumanChr14, using $K = 100$, $\sum_{\ell=K}^{L} La_\ell = 81284$ while $N > 733550$ for the relevant range of L; for SAureus, $\sum_{\ell=K}^{L} La_\ell = 558$ while $N > 8031$.

Size of critical window

In Section 4.2 we discussed the critical behavior in read length L:

1. If the read length L is below L_{crit}, reliable reconstruction of the DNA sequence is impossible no matter what the coverage depth is.

2. If the read length L is slightly above L_{crit}, then covering the sequence suffices, i.e. $c^* = c_{LW}$.

The first part follows from Ukkonen's condition, which requires that $L > L_{\text{crit}} = 1 + \max\{\ell_{\text{interleaved}}, \ell_{\text{triple}}\}$. One can observe the second part from the plots in Section 4.8, but in this subsection we seek to understand just how much larger than L_{crit} must L be in order that covering the sequence suffices for reconstruction, and furthermore how this depends on the parameters ϵ and G.

Critical window size if $\ell_{\text{interleaved}} \gg \ell_{\text{triple}}$

Let us focus here on the bound due to interleaved repeats (rather than triple repeats), and furthermore assume that the effect of the single largest interleaved repeat is dominant. In this case $\ell_{\text{interleaved}} = L_{\text{crit}} - 1$ is the length of the shorter of the pair of interleaved repeats,

and let ℓ_1 be the length of the longer of the two. For $L_{\text{crit}} < L \leq \ell_1 + 1$, we are in the setting of (4.7) but with a redefined $\gamma_1 = e^{2(N/G)(L_{\text{crit}}-1)}$. Thus,

$$L \geq L_{\text{crit}} + \frac{G}{2N} \log \epsilon^{-1} \,, \tag{4.21}$$

and solving for N gives

$$N_{\text{repeat}} = \frac{G}{2} \frac{\log \epsilon^{-1}}{L - \ell_2 - 1} \tag{4.22}$$

Let L^* be the value of L at which the curve described by constraint (4.22) intersects the Lander-Waterman coverage value, i.e. $N_{\text{repeat}}(L^*) = N_{\text{LW}}(L^*) := N^*$. This is the minimum read length for which coverage of the sequence suffices for reconstruction.

We now solve for $\frac{L^*}{L_{\text{crit}}}$. First, the Lander-Waterman equation (4.5) at $N = N^*$ is

$$N^* = \frac{G}{L^*} \log \frac{N^*}{\epsilon} \,, \tag{4.23}$$

and setting equal the right-hand sides of (4.23) and (4.22) at $L = L^*$ gives

$$\frac{G}{L^*} \log \frac{N^*}{\epsilon} = \frac{G}{2} \frac{\log \epsilon^{-1}}{L^* - \ell_2 - 1} \,.$$

A bit of algebra yields

$$\frac{L^*}{L_{\text{crit}}} = \frac{2}{2 - x} \,, \tag{4.24}$$

where

$$x := \cdot \frac{\log \epsilon^{-1}}{\log N^* + \log \epsilon^{-1}} \,. \tag{4.25}$$

Since $x \leq \frac{1}{2}$, equation (4.24) implies $L^* \leq 2L_{\text{crit}}$, and combined with the obvious inequality $L^* \geq L_{\text{crit}}$, we have $L_{\text{crit}} \leq L^* \leq 2L_{\text{crit}}$. Thus

$$N_{\text{LW}}(2L_{\text{crit}}) \leq N^* \leq N_{\text{LW}}(L_{\text{crit}}) \,, \tag{4.26}$$

and applying the Lander-Waterman fixed-point equation (4.5) yet again gives

$$\frac{G}{2L_{\text{crit}}} \log \frac{N_{\text{LW}}(2L_{\text{crit}})}{\epsilon} \leq N^* \leq \frac{G}{L_{\text{crit}}} \log \frac{N_{\text{LW}}(L_{\text{crit}})}{\epsilon} \,. \tag{4.27}$$

$$\frac{\log \epsilon^{-1}}{\log \frac{G}{L_{\text{crit}}} + \log \log \frac{N_{\text{LW}}(L_{\text{crit}})}{\epsilon} + \log \epsilon^{-1}} \leq x \leq \frac{\log \epsilon^{-1}}{\log \frac{G}{L_{\text{crit}}} - 1 + \log \log \frac{N_{\text{LW}}(2L_{\text{crit}})}{\epsilon} + \log \epsilon^{-1}} \,. \tag{4.28}$$

and this can be relaxed to

$$\frac{\log \epsilon^{-1}}{\log \frac{G}{L_{\text{crit}}} + \log \epsilon^{-1} + \log \log \frac{G}{\epsilon L_{\text{crit}}}} \leq x \leq \frac{\log \epsilon^{-1}}{\log \frac{G}{L_{\text{crit}}} - 1 + \log \epsilon^{-1}} \,. \tag{4.29}$$

Letting

$$r := \frac{\log \frac{G}{L_{\text{crit}}}}{\log \epsilon^{-1}}\,, \tag{4.30}$$

we have to a very good approximation

$$\frac{L^*}{L_{\text{crit}}} \approx \frac{2(r+1)}{2(r+1)-1}\,. \tag{4.31}$$

For $G \sim 10^8$, $L_{\text{crit}} \sim 1000$, and $\epsilon = 5\%$, we get $\log \frac{G}{L_{\text{crit}}} \approx 13.8$ and $\log \epsilon^{-1} \approx 3.0$, so $r \approx 4.6$ and

$$\frac{L^*}{L_{\text{crit}}} = \frac{2(r+1)}{2(r+1)-1} \approx 1.1\,.$$

From (4.30) we see that the relative size of $\log \epsilon^{-1}$ and $\log \frac{G}{L_{\text{crit}}}$ determines the size of the critical window. If in the previous example $\epsilon = 10^{-5}$, say, then $\frac{L^*}{L_{\text{crit}}}$ increases to 1.3. As ϵ tends to zero, r approaches zero as well and $\frac{L^*}{L_{\text{crit}}} \to 2$.

Critical window size if $\ell_{\text{triple}} \gg \ell_{\text{interleaved}}$

We now suppose the single longest triple repeat dominates the lower bound and estimate the width of the critical window. In this case $\ell_{\text{triple}} = L_{\text{crit}} - 1$ is the length of the longest triple repeat. Since we don't have matching lower and upper bounds for triple repeats, we separately compute the critical window size for each.

We start with the lower bound. For $L > L_{\text{crit}}$, the minimum value of N required in order to bridge the longest triple repeat is given by (4.19) and repeated here:

$$N_{\text{triples}} = \frac{G}{3} \cdot \frac{\log \epsilon^{-1}}{L - L_{\text{crit}}}\,. \tag{4.32}$$

As for the interleaved repeats case considered earlier, we let L^* be the value of L at which the curve described by constraint (4.32) intersects the Lander-Waterman coverage value, i.e. $N_{\text{triple}}(L^*) = N_{\text{LW}}(L^*) := N^*$. This is the minimum read length for which coverage of the sequence suffices for reconstruction.

A similar procedure as leading to (4.24) gives $L^*/L_{\text{crit}} = 3/(3-x)$ with x defined in (4.25). One can check that the estimates on x in (4.29) continue to hold, and we therefore get

$$\frac{L^*}{L_{\text{crit}}} \approx \frac{3(r+1)}{3(r+1)-1}\,. \tag{4.33}$$

For the same example as before, $G \sim 10^8$, $L_{\text{crit}} \sim 1000$, and $\epsilon = 5\%$, we get $r \approx 4.6$ and

$$\frac{L^*}{L_{\text{crit}}} = \frac{3(r+1)}{3(r+1)-1} \approx 1.06\,.$$

Changing ϵ to 10^{-5} makes $\frac{L^*}{L_{\text{crit}}} \approx 1.17$, and as ϵ (and hence also r) tends to zero, $\frac{L^*}{L_{\text{crit}}} \to \frac{3}{2}$.

The analogous computation for L^*/L_{crit} for the upper bound, as given by N_3^{all} in (4.19), yields

$$\frac{L^*}{L_{\text{crit}}} = \frac{r+1}{r+\frac{\log 3}{\log \epsilon^{-1}}} \approx 1.12\,, \tag{4.34}$$

for the example with $G \sim 10^8$, $L_{\text{crit}} \sim 1000$, and $\epsilon = 5\%$. The critical window size of the upper bound is about twice as large as that of the lower bound for typical values of G and L_{crit}, with ϵ moderate. But as $\epsilon \to 0$, we see from (4.34) that $L^*/L_{\text{crit}} \to \infty$, markedly different to the $L^*/L_{\text{crit}} \to \frac{3}{2}$ observed for the lower bound.

Improved lower bound for triple repeats

We show one case where a triple repeat only being single-bridged leads to confusion. Denote by $\mathbf{s}_1$ and $\mathbf{s}_2$ the two sequences depicted in Fig. 4.18. If the true sequence is $\mathbf{s}_1$, and only the copy of $\mathbf{x}$ between $\mathbf{u}$ and $\mathbf{y}$ is bridged, then $\mathbb{P}(\mathcal{R}|\mathbf{s}_2) > \mathbb{P}(\mathcal{R}|\mathbf{s}_1)$, and maximum likelihood assembly would make a mistake in calling $\mathbf{s}_2$. This demonstrates that the gap between the lower bound and upper bound based on the performance of MultiBridging is not exclusively a consequence of the algorithm being suboptimal– the lower bound is also loose.

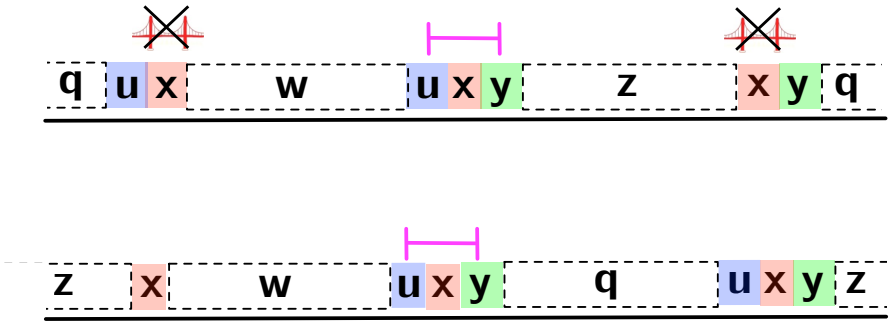


Figure 4.18: If only a single copy of the triple repeat is bridged as in the first instance, then the second sequence has higher likelihood and the maximum likelihood assembly is erroneous.

Dominant effect of longest repeats

The goal of this section is to find out which constraints are active for which parameter choices N, L, and thereby derive a simple expression bounding the performance of Greedy.

The coverage constraint can be expressed as

$$L \geq L_{\text{cov}} := \lambda^{-1} \log \frac{G}{\epsilon} = \lambda^{-1} \log G + \lambda^{-1} \log \epsilon^{-1}\,.$$

The performance of Greedy can be derived in terms of the requirement that all repeats be bridged. The probability that a particular repeat of length ℓ is unbridged is approximately

$e^{-2\lambda(L-\ell)}$. The requirement that the longest repeat be bridged gives

$$L \geq L_{\text{repeat}} := \ell_{\text{repeat}} + 1 + \frac{1}{2\lambda} \log \epsilon^{-1} .$$

A similar expression is obtained from applying the union bound over all repeats, resulting in the condition for GREEDY

$$L \geq L_{\text{GREEDY}} := \frac{1}{2\lambda} \log(\epsilon^{-1} \sum a_\ell e^{2\lambda\ell}) .$$

Now, suppose that $a_\ell \leq G^2 e^{-\alpha\ell}$, and $\alpha \geq 2\lambda+\epsilon$. Integrating gives $\sum a_\ell e^{2\lambda\ell} \leq G^2 \sum e^{\ell(2\lambda-\alpha)} \leq G^2 \frac{1}{\alpha-2\lambda}$. We obtain

$$L_{\text{GREEDY}} \leq \frac{1}{2\lambda} \log \frac{G^2}{\epsilon^2} \leq L_{\text{cov}} ,$$

which means that coverage of the sequence implies success for GREEDY. Alternatively, suppose that $\alpha \geq \frac{\log G^2}{\ell_{\text{repeat}}}$. Then

$$L_{\text{GREEDY}} = \frac{1}{2\lambda} \log \epsilon^{-1} + \frac{1}{2\lambda} \log \sum a_\ell e^{2\lambda\ell} \leq \frac{1}{2\lambda} \log \epsilon^{-1} + \frac{1}{2\lambda} \log \frac{G^2}{2\lambda - \alpha} e^{(2\lambda-\alpha)\ell_{\text{repeat}}} \leq L_{\text{repeat}} ,$$

and bridging the longest repeat implies success of GREEDY.

The difficulty occurs if a_ℓ is such that $\alpha < \min\{2\lambda + \epsilon, \frac{\log G^2}{\ell_{\text{repeat}}}\}$. The solution is to retain not only the longest repeat, but all repeats longer than some parameter x. Define $A := \sum_{\ell \leq x} a_\ell e^{2\lambda\ell}$ and $B = \sum_{\ell > x} a_\ell e^{2\lambda\ell}$. We have $L_{\text{GREEDY}} = \frac{1}{2\lambda} \log(A + B) + \frac{1}{2\lambda} \log \epsilon^{-1}$. Let us define

$$x := \max\{y : \max_{1 \leq \ell \leq y} a_\ell \leq G^2 e^{-(2\lambda+\epsilon)\ell}\} .$$

The implication is that if $A \geq B$, then $L_{\text{GREEDY}} \leq L_{\text{cov}} + \frac{1}{2\lambda}$, and if $B \geq A$, then $L_{\text{GREEDY}} \leq L_B + \frac{1}{2\lambda}$, where $L_B := \frac{1}{2\lambda} \log B + \frac{1}{2\lambda} \log \epsilon^{-1}$. Thus we replace the constraint due to bridging the longest repeat by L_B, bridging all the repeats longer than x.

Taking the sum in B only over repeats longer than x simplifies the numerical computations, allowing to factor out $e^{2\lambda x}$. This gives

$$\frac{1}{2\lambda} \log B = \log \left(e^{2\lambda x} \sum_{\ell > x} a_\ell e^{2\lambda(\ell - x)} \right) = 2\lambda x \log \left(\sum_{\ell > x} a_\ell e^{2\lambda(\ell - x)} \right) .$$

For the genomes we examined, x is fairly close to ℓ_{repeat}.

4.7 Proof of correctness for MULTIBRIDGING

Proofs for K-mer algorithms

We will use $m_{\mathcal{C}}(\mathbf{v})$ to denote the multiplicity (traversal count) a cycle $\mathcal{C}$ assigns a node $\mathbf{v}$. The multiplicity $m_{\mathcal{C}}(\mathbf{v})$ is also equal to the number of times the subsequence $\mathbf{v}$ appears in

the sequence corresponding to $\mathcal{C}$. For an edge e, we can similarly let $m_{\mathcal{C}}(e)$ be the number of times $\mathcal{C}$ traverses the edge. The following key lemma relates node multiplicities with the existence of X-nodes.

Lemma 38. *Let $\mathcal{C}$ be a cycle in a condensed sequence graph $\mathbb{G}$, where $\mathbb{G}$ itself is not a cycle, traversing every edge at least once. If $\mathbf{v}$ is a node with maximum multiplicity at least 2, i.e. $m_{\mathcal{C}}(\mathbf{v}) = \max_{u \in \mathbb{G}} m_{\mathcal{C}}(\mathbf{u}) \geq 2$, then $\mathbf{v}$ is an X-node. As a consequence, if $m_{\mathcal{C}}(\mathbf{v}) \geq 3$ for some $\mathbf{v}$, i.e. $\mathcal{C}$ traverses some node at least three times, then $m_{\mathcal{C}}(\mathbf{u}) \geq 3$ for some X-node $\mathbf{u}$.*

Proof. Let $\mathbf{v}$ be a node with maximum multiplicity $m_{\mathcal{C}}(\mathbf{v}) = \max_{\mathbf{u} \in \mathbb{G}} m_{\mathcal{C}}(\mathbf{u})$. We will show that $\mathbf{v}$ is an X-node, i.e. $d_{\text{out}}(\mathbf{v}) \geq 2$ and $d_{\text{in}}(\mathbf{v}) \geq 2$.

We prove that $d_{\text{out}}(\mathbf{v}) \geq 2$ by supposing that $d_{\text{out}}(\mathbf{v}) = 1$ and deriving a contradiction. Denote the outgoing edge from $\mathbf{v}$ by $e = (\mathbf{v}, \mathbf{u})$, where $\mathbf{u}$ is distinct from $\mathbf{v}$ since otherwise $\mathbb{G}$ is a cycle. If $d_{\text{in}}(\mathbf{u}) \geq 2$, then $\mathbf{u}$ must be traversed more times than $\mathbf{v}$, contradicting the maximality of $m_{\mathcal{C}}(\mathbf{v})$, and if $d_{\text{in}}(\mathbf{u}) = 1$, then the existence of the edge e contradicts the fact that $\mathbb{G}$ is condensed. The argument showing that $d_{\text{in}}(\mathbf{v}) \geq 2$ is symmetric to the case $d_{\text{in}}(\mathbf{v}) \geq 2$. □

Proof of Lemma 37. We assume that all triple repeats are all-bridged, that there are no unbridged interleaved repeats, and that all reads overlap their successors by at least 1 base pair. We wish to show that MultiBridging returns the original sequence.

Consider the condensed sequence graph $\mathbb{G}_0$ constructed in steps 1-3 of MultiBridging. Suppose all X-nodes that are either all-bridged or correspond to bridged 2-repeats have been resolved according to repeated application of the procedure in step 4 of MultiBridging, resulting in a condensed sequence graph $\mathbb{G}$. We claim that 1) $\mathbf{s}$ corresponds to a cycle $\mathcal{C}$ in $\mathbb{G}$ traversing every edge at least once, 2) $\mathcal{C}$ is Eulerian, and 3) $\mathcal{C}$ is the unique Eulerian cycle in $\mathbb{G}$.

Proof of Claim 1. Let $\mathbb{G}_n$ be the graph after n resolution steps, and suppose that $\mathcal{C}_n$ is a cycle in $\mathbb{G}_n$ corresponding to the sequence $\mathbf{s}$ and traversing all edges. We will show that there exists a cycle $\mathcal{C}_{n+1}$ in $\mathbb{G}_{n+1}$ corresponding to $\mathbf{s}$ and traversing all edges, and that $\mathbb{G}_t = \mathbb{G}$ for a finite t, so by induction, there exists a cycle $\mathcal{C}$ in $\mathbb{G}$ corresponding to $\mathbf{s}$ and traversing all edges. The base case $n = 0$ was shown in Lemma 28. Moving on to arbitrary $n > 0$, let $\mathbf{v}$ be an X-node in $\mathbb{G}_n$ labeled as in Fig. 4.10. The X-node resolution step is constructed precisely to preserve the existence of a cycle corresponding to $\mathbf{s}$. Each traversal of $\mathbf{v}$ by the cycle $\mathcal{C}_n$ assigns an incoming edge $(\mathbf{p}_i\mathbf{v})$ to an outgoing edge $(\mathbf{v}, \mathbf{q}_j)$, and the resolution step correctly determines this pairing by the assumption on bridging reads.

Note that all X-nodes in the graph $\mathbb{G}_{n+1}$ continue to correspond to repeats in $\mathbf{s}$. The process terminates: let $\mathcal{L}(\mathbb{G}_i) = \sum_{\mathbf{v} \in \mathbb{G}_i} m_{\mathcal{C}_i}(\mathbf{v}) \mathbf{1}_{m_{\mathcal{C}_i}(\mathbf{v}) > 1}$ and observe that $\mathcal{L}(\mathbb{G}_i)$ is strictly decreasing in i. Thus $\mathbf{s}$ corresponds to a cycle $\mathcal{C}$ in $\mathbb{G}$ traversing each edge at least once.

Proof of Claim 2. We next show that $\mathcal{C}$ is an Eulerian cycle. If $\mathbb{G}$ is itself a cycle, and $\mathbf{s}$ is not formed by concatenating multiple copies of a shorter subsequence (assumed not to be the case, see discussion at end of Section 4.3), then $\mathcal{C}$ traverses $\mathbb{G}$ exactly once and is an Eulerian cycle. Otherwise, if $\mathbb{G}$ is not a cycle, then we may apply Lemma 38 to see that any node with $m_{\mathcal{C}}(\mathbf{v}) \geq 3$ implies the existence of an X-node $\mathbf{u}$ with $m_{\mathcal{C}}(\mathbf{u}) \geq 3$. Node $\mathbf{u}$ must be all-bridged, by hypothesis, which means that an additional X-node resolution step can be applied to $\mathbb{G}$, a contradiction. Thus each node $\mathbf{v}$ in $\mathbb{G}$ has multiplicity $m_{\mathcal{C}}(\mathbf{v}) \leq 2$.

We can now argue that no edge $e = (\mathbf{u}, \mathbf{v})$ is traversed twice by $\mathcal{C}$ in the condensed sequence graph $\mathbb{G}$, as it would have been contracted. Suppose $m_{\mathcal{C}}(e) \geq 2$. The node $\mathbf{u}$ cannot have two outgoing edges as this implies $m_{\mathcal{C}}(\mathbf{u}) \geq 3$; similarly, $\mathbf{v}$ cannot have two incoming edges. Thus $d_{\text{out}}(\mathbf{u}) = d_{\text{in}}(\mathbf{v}) = 1$, but by Defn. 31 the edge $e = (\mathbf{u}, \mathbf{v})$ would have been contracted.

Proof of Claim 3. It remains to show that there is a *unique* Eulerian cycle in $\mathbb{G}$. All X-nodes in $\mathbb{G}$ must be unbridged 2-X-nodes (correspond to 2-repeats in $\mathbf{s}$), as all other X-nodes were assumed to be bridged and have thus been resolved in $\mathbb{G}$.

We will map the sequence $\mathbf{s}$ to another sequence $\mathbf{s}'$, allowing us to use the characterization of Lemma 29 for SBH with known multiplicities. Denote by $\mathbb{G}'$ the graph obtained by relabeling each node in $\mathbb{G}$ by a single unique symbol (no matter the original node label length), and setting all edge overlaps to 0. Through the relabeling, $\mathcal{C}$ corresponds to a cycle $\mathcal{C}'$ in $\mathbb{G}'$, and let $\mathbf{s}'$ be the sequence corresponding to $\mathcal{C}'$. Writing $\mathcal{S}_2'$ for the 2-spectrum of $\mathbf{s}'$, the graph $\mathbb{G}'$ is by construction precisely the 1-mer graph created from $\mathcal{S}_2'$, and there is a one-to-one correspondence between X-nodes in $\mathbb{G}'$ and unbridged repeats in $\mathbf{s}'$. Through the described mapping, every unbridged repeat in $\mathbf{s}'$ maps to an unbridged repeat in $\mathbf{s}$, with the *order of repeats preserved*.

There are multiple Eulerian cycles in $\mathbb{G}$ only if there are multiple Eulerian cycles in $\mathbb{G}'$ since the graphs have the same topology, and by Lemma 29 the latter occurs only if there are unbridged interleaved repeats in $\mathbf{s}'$, which by the correspondence in the previous paragraph implies the existence of unbridged interleaved repeats in $\mathbf{s}$. □

Lower bound due to triple repeats

We translate the generalized Ukkonen's condition prohibiting unbridged triple repeats into a condition on L and N. Let c_m denote the number of triple repeats of length m. Then a union bound estimate gives

$$\mathbb{P}(\mathcal{E}) \approx \sum_m c_m e^{-3\lambda(L-m-1)} . \tag{4.35}$$

Requiring $\mathbb{P}(\mathcal{E}) \leq \epsilon$ and solving for L gives

$$L \geq \frac{1}{3\lambda} \log \frac{\gamma_3}{\epsilon} = \frac{G}{3N} \log \frac{\gamma_3}{\epsilon} , \tag{4.36}$$

where $\gamma_3 := \sum_m c_m e^{3(N/G)(m+1)}$.

Truncation estimate for bridging repeats (GREEDY and MULTIBRIDGING)

The repeat statistics a_m and c_m used in the algorithm performance curves are potentially overestimates. This is because a large repeat family—one with a large number of copies f—will result in a contribution $\binom{f}{2} \approx f^2/2$ to a_m and $\binom{f}{3} \approx f^3/6$ to c_m.

We focus here on deriving an estimate for the required N, L for bridging all repeats with probability $1-\epsilon$. This upper bound reduces the sensitivity to large families of short repeats. The analogous derivation for all-bridging all triple-repeats is very similar and is omitted.

Suppose there are a_m repeats of length m. The probability that some repeat is unbridged is approximately, by the union bound estimate,

$$\mathbb{P}(\mathcal{E}) \approx \sum_m a_m e^{-2\lambda(L-m)}\,. \tag{4.37}$$

Requiring $\mathbb{P}(\mathcal{E}) \le \epsilon$ and solving for L gives

$$L \ge \frac{1}{2\lambda}\log\frac{\gamma}{\epsilon} = \frac{G}{2N}\log\frac{\gamma}{\epsilon}\,, \tag{4.38}$$

where $\gamma := \sum_m a_m e^{2(N/G)m}$. Now, if a_m *overcounts* the number of repeats for small values of m, the bound in (4.38) might be loose. In order for each read to overlap the subsequent read by x nucleotides, with probability of failure $\epsilon/2$, it suffices to take

$$L \ge L_{\text{K-cov}}\left(x, \frac{\epsilon}{2}\right) := x + \frac{1}{\lambda}\log\frac{2N}{\epsilon}\,. \tag{4.39}$$

Thus, for any $x < L$, we may replace (4.38) by

$$L \ge \min_x \max\{\frac{1}{2\lambda}\log\frac{2\gamma(x)}{\epsilon}, L_{\text{K-cov}}(x, \frac{\epsilon}{2})\}\,, \tag{4.40}$$

where $\gamma(x) = \sum_{m>x} a_m e^{2(N/G)m}$, and obtain a *looser* bound.

4.8 Feasibility Plots

In this section we display the output of our pipeline for 12 datasets. For each dataset we plot

$$\log(1 + a_\ell),$$

the log of one plus the number of repeats of each length ℓ. From the repeat statistics a_m, $b_{m,n}$, and c_m, we produce a feasibility plot. The thick black line denotes the lower bound on feasible N, L, and the green line is the performance achieved by MULTIBRIDGING.

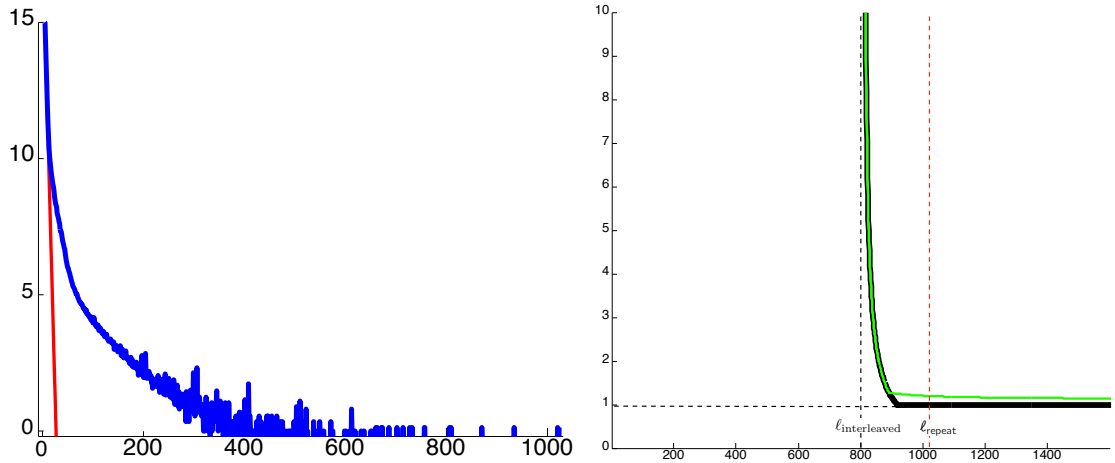


Figure 4.19: Human Chrom 14. $G = 88,289,540$, $\ell_{\text{triple}} = 611$, $\ell_{\text{interleaved}} = 805$, $\ell_{\text{repeat}} = 1022$.

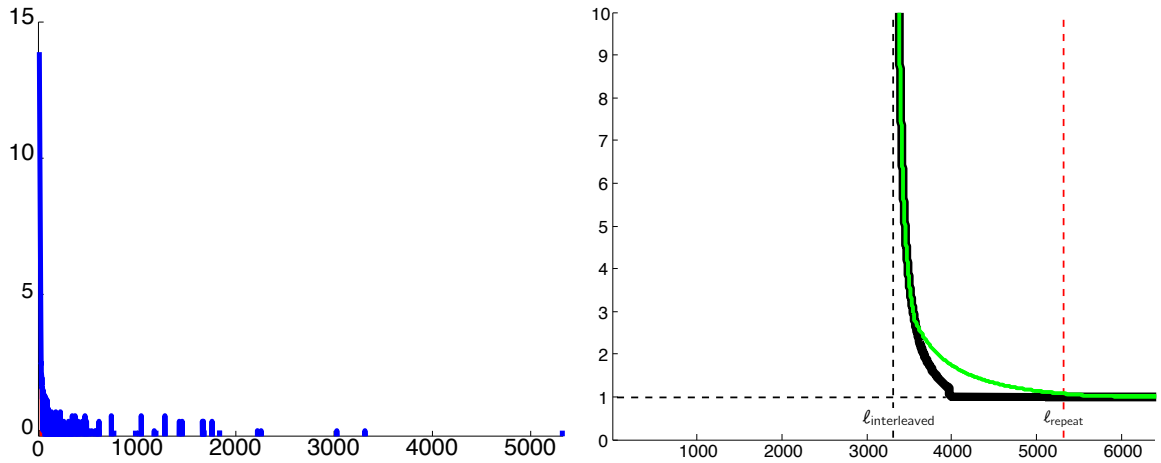


Figure 4.20: Lactofidus. $G = 2,078,001$, $\ell_{\text{triple}} = 3027$, $\ell_{\text{interleaved}} = 3313$, $\ell_{\text{repeat}} = 5321$.

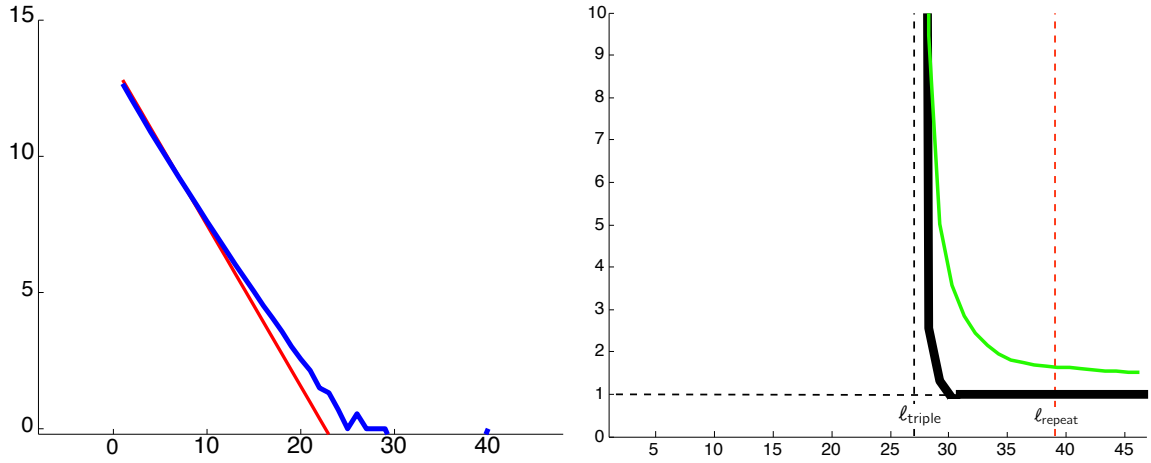


Figure 4.21: Buchnera. $G = 642,122$, $\ell_{\text{triple}} = 27$, $\ell_{\text{interleaved}} = 23$, $\ell_{\text{repeat}} = 39$.

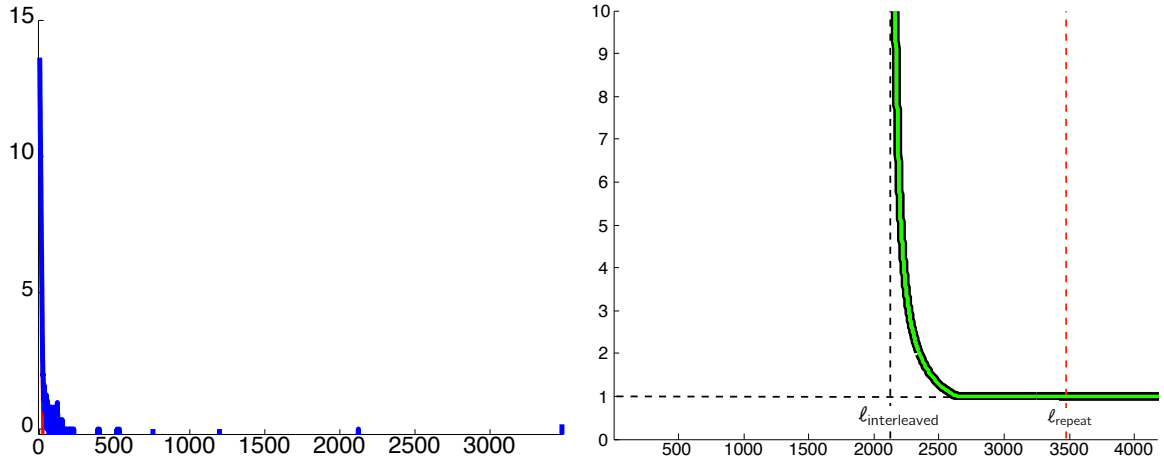


Figure 4.22: Heli51. $G = 1,589,954$, $\ell_{\text{triple}} = 219$, $\ell_{\text{interleaved}} = 2122$, $\ell_{\text{repeat}} = 3478$.

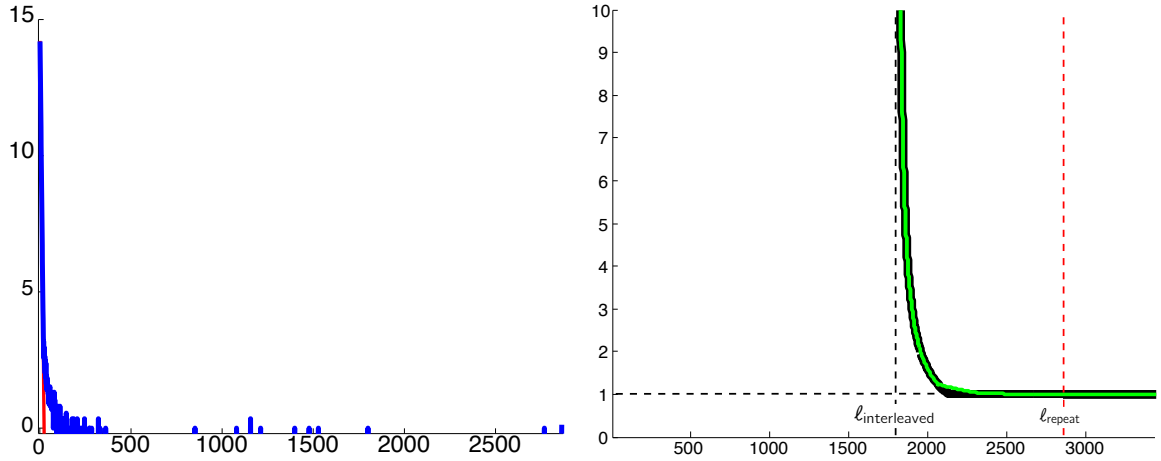


Figure 4.23: Staphylococcus Aureus. $G = 2,872,915$, $\ell_{\text{triple}} = 1397$, $\ell_{\text{interleaved}} = 1799$, $\ell_{\text{repeat}} = 2862$.

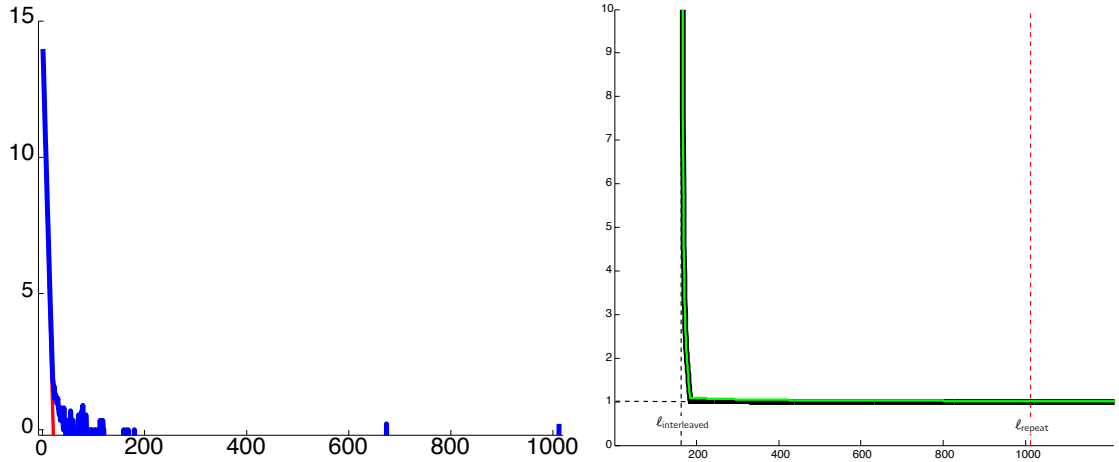


Figure 4.24: Salmonella. $G = 2,215,568$, $\ell_{\text{triple}} = 112$, $\ell_{\text{interleaved}} = 163$, $\ell_{\text{repeat}} = 1011$.

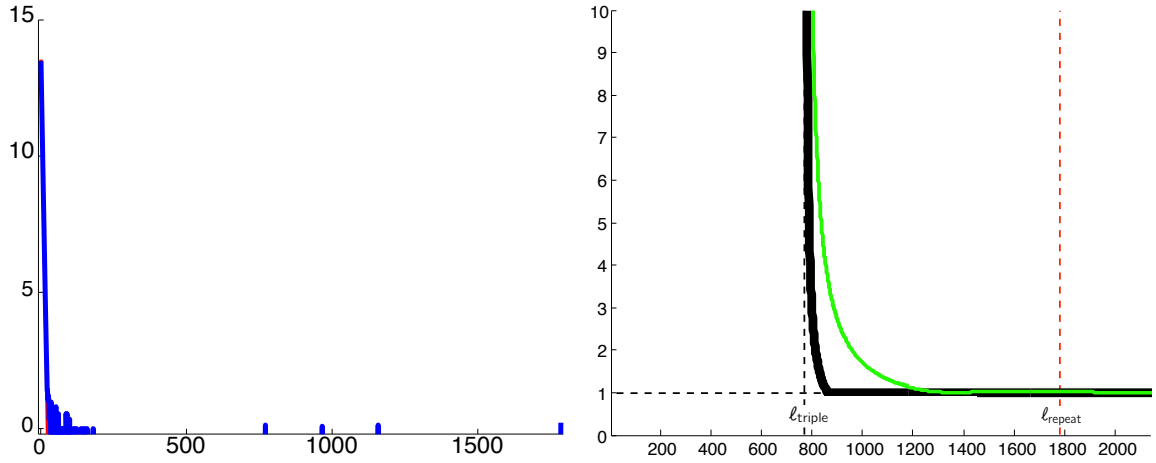


Figure 4.25: Perkinsus marinus. $G = 1,440,372$, $\ell_{\text{triple}} = 770$, $\ell_{\text{interleaved}} = 92$, $\ell_{\text{repeat}} = 1784$.

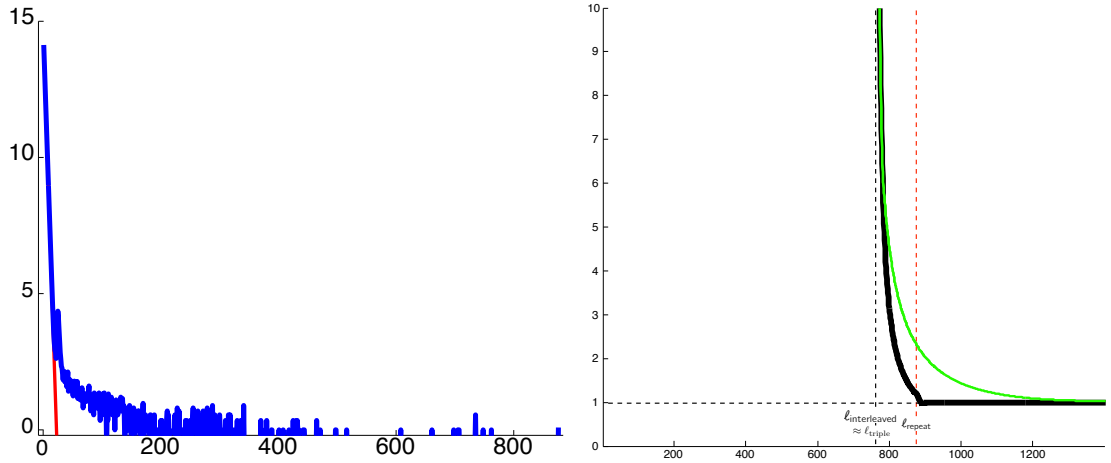


Figure 4.26: Sulfolobus islandicus. $G = 2,655,198$, $\ell_{\text{triple}} = 734$, $\ell_{\text{interleaved}} = 761$, $\ell_{\text{repeat}} = 875$.

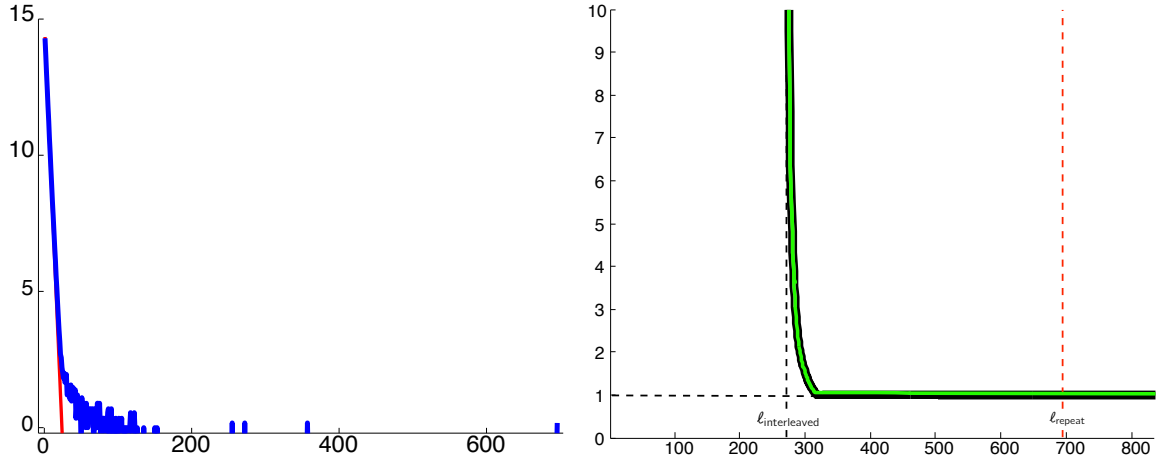


Figure 4.27: Rhodobacter sphaeroides. $G = 3,188,599$, $\ell_{\text{triple}} = 114$, $\ell_{\text{interleaved}} = 271$, $\ell_{\text{repeat}} = 695$.

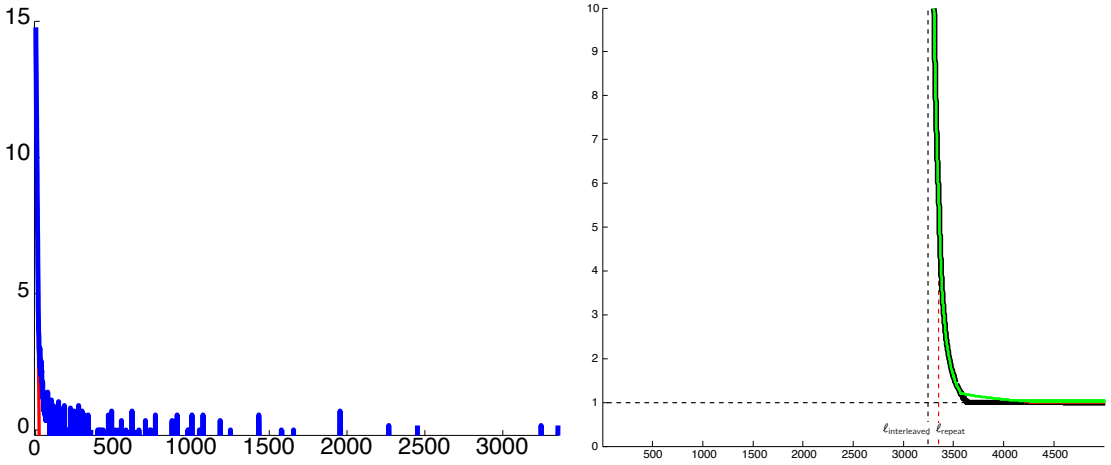


Figure 4.28: Ecoli536. $G = 4,938,920$, $\ell_{\text{triple}} = 2267$, $\ell_{\text{interleaved}} = 3245$, $\ell_{\text{repeat}} = 3353$.

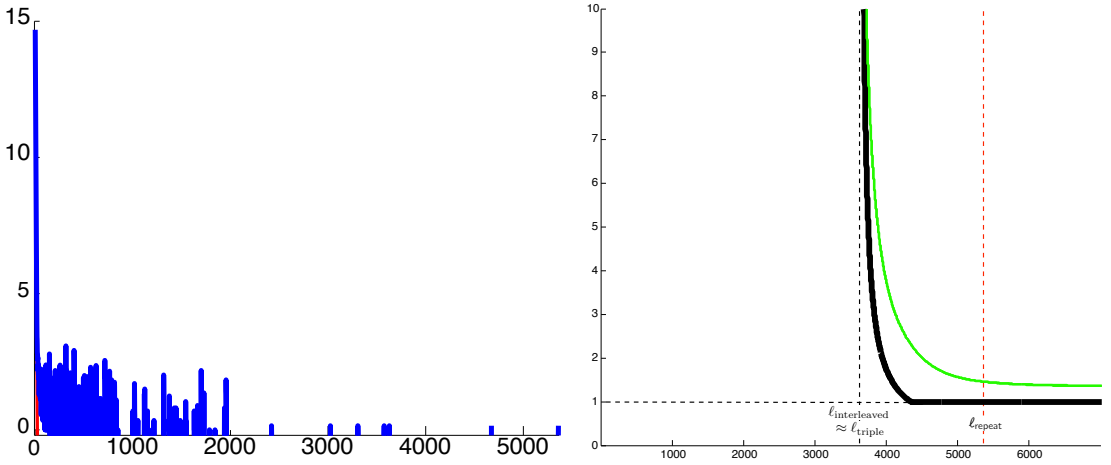


Figure 4.29: Yesnina. $G = 4,504,254$, $\ell_{\text{triple}} = 3573$, $\ell_{\text{interleaved}} = 3627$, $\ell_{\text{repeat}} = 5358$.

Bibliography

[1] V. Cadambe and S. Jafar, "Interference alignment and degrees of freedom of the k-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425 –3441, August 2008.

[2] R. Etkin, D. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *Information Theory, IEEE Transactions on*, vol. 54, no. 12, pp. 5534–5562, 2008.

[3] M. Maddah-Ali, A. Motahari, and A. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3457 –3470, August 2008.

[4] S. Jafar and S. Shamai, "Degrees of freedom region of the mimo x channel," *Information Theory, IEEE Transactions on*, vol. 54, no. 1, pp. 151–170, 2008.

[5] Wikipedia, "Sequence assembly — Wikipedia, the free encyclopedia," 2012, [Online; accessed Nov-20-2012] http://en.wikipedia.org/wiki/Sequence_assembly.

[6] C. Alkan, S. Sajjadian, and E. Eichler, "Limitations of next-generation genome sequence assembly," *Nature methods*, vol. 8, no. 1, pp. 61–65, 2010.

[7] M. Baker, "De novo genome assembly: what every biologist should know," *Nature methods*, vol. 9, no. 4, pp. 333–337, 2012.

[8] P. Gupta and P. Kumar, "The capacity of wireless networks," *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 388–404, 2000.

[9] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *Information Theory, IEEE Transactions on*, vol. 53, no. 10, pp. 3549–3572, 2007.

[10] U. Niesen, "Interference alignment in dense wireless networks," *Information Theory, IEEE Transactions on*, vol. 57, no. 5, pp. 2889–2901, 2011.

[11] A. Ozgur and D. Tse, "Achieving linear scaling with interference alignment," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*. IEEE, 2009, pp. 1754–1758.

[12] B. Nazer, S. Jafar, M. Gastpar, and S. Vishwanath, "Ergodic interference alignment," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on.* IEEE, 2009, pp. 1769–1773.

[13] L. Grokop, D. Tse, and R. Yates, "Interference alignment for line-of-sight channels," *Information Theory, IEEE Transactions on*, vol. 57, no. 9, pp. 5820–5839, 2011.

[14] V. Cadambe, S. Jafar, and C. Wang, "Interference alignment with asymmetric complex signaling—settling the høst-madsen–nosratinia conjecture," *Information Theory, IEEE Transactions on*, vol. 56, no. 9, pp. 4552–4565, 2010.

[15] K. Gomadam, V. Cadambe, and S. Jafar, "Approaching the capacity of wireless networks through distributed interference alignment," in *IEEE GLOBECOM*, December 2008, pp. 1 –6.

[16] C. Yetis, T. Gou, S. Jafar, and A. Kayran, "On feasibility of interference alignment in MIMO interference networks," *IEEE Trans. Signal Processing*, vol. 58, no. 9, pp. 4771 –4782, September 2010.

[17] M. Razaviyayn, L. Gennady, and Z. Luo, "On the degrees of freedom achievable through interference alignment ina MIMO interference channel," private communication, submitted to SPAWC 2011.

[18] ——, "On the degrees of freedom achievable through interference alignment in a MIMO interference channel," private communication, Draft manuscript.

[19] K. Gomadam, V. Cadambe, and S. Jafar, "A distributed numerical approach to interference alignment and applications to wireless interference networks," *Information Theory, IEEE Transactions on*, vol. 57, no. 6, pp. 3309 –3322, June 2011.

[20] S. Peters and R. Heath, "Interference alignment via alternating minimization," in *Acoustics, Speech and Signal Processing, 2009. IEEE International Conference on*, April 2009, pp. 2445 –2448.

[21] M. Razaviyayn, M. Sanjabi Boroujeni, and Z.-Q. Luo, "Linear transceiver design for interference alignment: Complexity and computation," in *Signal Processing Advances in Wireless Communications (SPAWC), 2010 IEEE Eleventh International Workshop on*, June 2010, pp. 1 –5.

[22] D. S. Papailiopoulos and A. G. Dimakis, "Interference alignment as a rank constrained rank minimization," in *IEEE GLOBECOM*, 2010.

[23] I. Santamaria, O. Gonzalez, R. Heath, and S. Peters, "Maximum sum-rate interference alignment algorithms for MIMO channels," in *IEEE GLOBECOM*, December 2010, pp. 1 –6.

[24] D. A. Schmidt, W. Utschick, and M. L. Honig, "Beamforming techniques for single-beam MIMO interference networks," in *Allerton Conference on Communication, Control, and Computing*, September 2010.

[25] S. Jafar and M. Fakhereddin, "Degrees of freedom for the MIMO interference channel," *IEEE Trans. Inf. Theory*, vol. 53, no. 7, pp. 2637 –2642, July 2007.

[26] M. Amir, A. E. Keyi, and M. Nafie, "A new achievable DoF region for the 3-user $M \times N$ symmetric interference channel," May 2011, preprint, arXiv:1105.4026v1.

[27] G. Bresler, D. Cartwright, and D. Tse, "Geometry of the 3-user MIMO interference channel," in *Allerton Conference on Communication, Control, and Computing*, September 2011, arXiv:1110.5092v1.

[28] C. Wang, T. Gou, and S. Jafar, "Subspace alignment chains and the degrees of freedom of the three-user MIMO interference channel," September 2011, arXiv:1109.4350v1.

[29] G. Bresler, A. Parekh, and D. Tse, "The approximate capacity of the many-to-one and one-to-many Gaussian interference channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, September 2010.

[30] V. Cadambe, S. Jafar, and S. Shamai, "Interference alignment on the deterministic channel and application to fully connected gaussian interference networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 269 –274, January 2009.

[31] R. Etkin and E. Ordentlich, "The degrees-of-freedom of the K-user gaussian interference channel is discontinuous at rational channel coefficients," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4932 –4946, November 2009.

[32] A. S. Motahari, S. Gharan, M. Maddah-Ali, and A. K. Khandani, "Real Interference Alignment: Exploiting the Potential of Single Antenna Systems," 2009.

[33] A. Ghasemi, A. Motahari, and A. Khandani, "Interference alignment for the K user MIMO interference channel," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, June 2010, pp. 360 –364.

[34] O. Ordentlich and U. Erez, "Interference alignment at finite SNR for time-invariant channels," in *Information Theory Workshop (ITW)*, October 2011.

[35] Y. Wu, S. Shamai, and S. Verdu, "Degrees of freedom of the interference channel: A general formula," in *IEEE International Symposium on Information Theory (ISIT)*, August 2011, pp. 1362 –1366.

[36] U. Niesen and M. Maddah-Ali, "Interference alignment: From degrees-of-freedom to constant-gap capacity approximations," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2077–2081.

[37] O. Ordentlich, U. Erez, and B. Nazer, "The approximate sum capacity of the symmetric gaussian k-user interference channel," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2072–2076.

[38] R. Hartshorne, *Algebraic Geometry*, 1st ed., ser. Graduate Texts in Mathematics. Springer, 1977.

[39] I. Shafarevich, *Basic Algebraic Geometry*, 2nd ed. Springer, 1995.

[40] D. A. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 3rd ed., ser. Undergraduate Texts in Mathematics. NJ, USA: Springer-Verlag, 2007.

[41] D. Earl, K. Bradnam, J. John, A. Darling, D. Lin, J. Fass, H. Yu, V. Buffalo, D. Zerbino, M. Diekhans *et al.*, "Assemblathon 1: A competitive assessment of de novo short read assembly methods," *Genome research*, vol. 21, no. 12, pp. 2224–2241, 2011.

[42] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke, "GAGE: A critical evaluation of genome assemblies and assembly algorithms," *Genome research*, vol. 22, no. 3, pp. 557–567, 2012.

[43] N. National Human Genome Research Institute, "Human genome sequence quality standards," Dec 2012, [Online; accessed Dec-12-2012] http://www.genome.gov/10000923.

[44] J. D. Kececioglu and E. W. Myers, "Combinatorial algorithms for DNA sequence assembly," *Algorithmica*, vol. 13, pp. 7–51, 1993.

[45] E. Myers, "Toward simplifying and accurately formulating fragment assembly," *Journal of Computational Biology*, vol. 2, no. 2, pp. 275–290, 1995.

[46] P. Medvedev and M. Brudno, "Maximum likelihood genome assembly," *Journal of computational Biology*, vol. 16, no. 8, pp. 1101–1116, 2009.

[47] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proc Natl Acad Sci USA*, vol. 98, pp. 9748–53, 2001.

[48] E. Myers, "The fragment assembly string graph," *Bioinformatics*, vol. 21, pp. ii79–ii85, 2005.

[49] E. Lander and M. Waterman, "Genomic mapping by fingerprinting random clones: A mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[50] J. Gallant, D. Maier, and J. Astorer, "On finding minimal length superstrings," *Journal of Computer and System Sciences*, vol. 20, no. 1, pp. 50–58, 1980.

[51] P. Medvedev, K. Georgiou, G. Myers, and M. Brudno, "Computability of models for sequence assembly," *Algorithms in Bioinformatics*, pp. 289–301, 2007.

[52] N. Nagarajan and M. Pop, "Parametric complexity of sequence assembly: theory and applications to next generation sequencing," *Journal of computational biology*, vol. 16, no. 7, pp. 897–908, 2009.

[53] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, and A. M. Phillippy, "Hybrid error correction and de novo assembly of single-molecule sequencing reads," *Nat Biotech*, vol. 30, pp. 693–700, 2012.

[54] E. Ukkonen, "Approximate string matching with q-grams and maximal matches," *Theoretical Computer Science*, vol. 92, no. 1, pp. 191–211, 1992.

[55] G. G. Sutton, O. White, M. D. Adams, and A. Kerlavage, "TIGR Assembler: A new tool for assembling large shotgun sequencing projects," *Genome Science & Technology*, vol. 1, pp. 9–19, 1995.

[56] X. Huang and A. Madan, "CAP3: A DNA sequence assembly program," *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[57] R. Warren, G. Sutton, S. Jones, and R. Holt, "Assembling millions of short DNA sequences using SSAKE," *Bioinformatics*, vol. 23, pp. 500–501, 2007.

[58] R. Idury and M. Waterman, "A new algorithm for DNA sequence assembly," *J. Comp. Bio*, vol. 2, pp. 291–306, 1995.

[59] Y. Peng, H. Leung, S. Yiu, and F. Chin, "IDBA–a practical iterative de Bruijn graph de novo assembler," in *Research in Computational Molecular Biology*. Springer, 2010, pp. 426–440.

[60] S. Motahari, G. Bresler, and D. Tse, "Information theory of DNA sequencing," 2012, http://arxiv.org/abs/1203.6233.

[61] P. A. Pevzner, "DNA physical mapping and alternating Eulerian cycles in colored graphs," *Algorithmica*, vol. 13, no. 1/2, pp. 77–105, 1995.

[62] P. Compeau, P. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly," *Nat Biotech*, vol. 29, no. 11, pp. 987–991, 11 2011. [Online]. Available: http://dx.doi.org/10.1038/nbt.2023

[63] P. Pevzner, "ℓ-tuple DNA sequencing: computer analysis," *J Biomol Struct Dyn.*, vol. 7, no. 1, pp. 63–73, 1989.

[64] I. Maccallum, D. Przybylski, S. Gnerre, J. Burton, I. Shlyakhter, A. Gnirke, J. Malek, K. McKernan, S. Ranade, T. P. Shea, L. Williams, S. Young, C. Nusbaum, and D. B. Jaffe, "Allpaths 2: small genomes assembled accurately and with high continuity from short paired reads," *Genome Biol*, vol. 10, no. 10, p. R103, 2009.

[65] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Res*, vol. 18, no. 5, pp. 821–9, May 2008.

[66] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, "A whole-genome assembly of drosophila," *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000. [Online]. Available: http://www.sciencemag.org/content/287/5461/2196.abstract

[67] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and İ. Birol, "ABySS: A parallel assembler for short read sequence data," *Genome Research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[68] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proceedings of the National Academy of Sciences*, vol. 108, no. 4, pp. 1513–1518, 2011. [Online]. Available: http://www.pnas.org/content/108/4/1513.abstract