Constraining the discovery space for artificial interstellar signals

Author:
Morrison, Ian

Publication Date:
2017

DOI:
https://doi.org/10.26190/unsworks/19848

Constraining the discovery space for artificial interstellar signals

Ian S. Morrison

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Physics

Faculty of Science

July 2017

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Morrison**

First name: **Ian** Other name/s: **Stuart**

Abbreviation for degree as given in the University calendar: **1890 Physics**

School: **Physics** Faculty: **Science**

Title: **Constraining the discovery space for artificial interstellar signals**

Abstract:

After more than 50 years of searching the skies across the electromagnetic spectrum, no evidence has yet been found for the existence of extraterrestrial life, let alone a signature that could be attributed to an intelligent and technological extraterrestrial civilisation. Proposed explanations range from the non-existence of such putative civilisations, to the likelihood that the search has barely "scratched the surface" in terms of covering the entire discovery space. The present work is premised on the latter position, and its central thesis is that most searches conducted thus far have been disadvantaged by sub-optimal design. In general, they have (1) not targeted the most appropriate signal types; (2) not targeted the region of the electromagnetic spectrum that should be preferred for interstellar communications; and (3) not consistently concentrated on the regions of the sky that hold the highest likelihood for the emergence of extraterrestrial intelligence. The very nature of past searches provides a credible explanation for the null result to date.

It is impossible to predict the specific technologies and communications methodologies likely to be adopted by extraterrestrial civilisations, particularly when one considers how much older and more technologically advanced than ourselves they may be. However, beginning with a single key assumption – that energy efficiency is a concern to those wishing to transmit signals across interstellar space – this work shows how the application of fundamental principles of astrophysics and information theory can lead to meaningful constraints on the discovery space for artificial interstellar signals. It is shown why the search for extraterrestrial intelligence should focus on intentionally transmitted wideband signals occupying the upper end of the microwave range of the radio spectrum - a region largely ignored by past searches. Furthermore, searches should concentrate on the inner Galaxy rather than our own solar neighbourhood. A new signal processing algorithm is introduced for the blind detection of wideband signals of a class that is attractive for interstellar communications, offering high detection sensitivity while making minimal assumptions about the precise signal format.

The findings of this thesis suggest a range of new priorities and approaches to the search for extraterrestrial intelligence, aimed at increasing its chances of success.

COPYRIGHT STATEMENT

Date 16/8/17

AUTHENTICITY STATEMENT

Signed .....

Date 16/8/17

Originality Statement

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed ……..

Date ……..

Table of Contents

List of Abbreviations

BEF	Bandwidth Expansion Factor
CMB	Cosmic Microwave Background
DM	Dispersion Measure
EIRP	Effective Isotropic Radiated Power
FFT	Fast Fourier Transform
GHZ	Galactic Habitable Zone
ICH	Interstellar Coherence Hole
IGM	Intergalactic Medium
ISI	Inter-Symbol Interference
ISM	Interstellar Medium
METI	Messaging to Extraterrestrial Intelligence
RFI	Radio Frequency Interference
SETI	Search for Extraterrestrial Intelligence
SKA	Square Kilometre Array
SN	supernova
SNe	supernovae
S/N	Signal-to-Noise Ratio

Acknowledgements

I will be forever grateful for the guidance and support provided by my supervisory team; Chris Tinney, Carol Oliver, Malcolm Walter and James Benford. A special mention must go to Malcolm who, as Director of the Australian Centre for Astrobiology at the time I commenced my candidature, went out on a limb to accept me as the Centre's first ever SETI student – hopefully not the last. I've also benefited enormously, both professionally and personally, from countless fruitful exchanges with friends and colleagues, most particularly David Messerschmitt, Michael Gowanlock, Jill Tarter, David Flannery, Andrew Siemion, Gerry Harp, and Willem van Straten. I would like to make special mention of David Messerschmitt and Jill Tarter for their constant support and encouragement, and for being endless sources of inspiration.

I had considered dedicating this work to my late father, Ned Morrison. I loved him dearly, but no doubt he would have dismissed this work (along with all astrobiology) as complete hogwash. Still, in his own way, he inspired me more than anyone else to pursue science.

But I can really only dedicate this to my wife Sarah, for her love and mostly unwavering support, which is surely more than I deserve. Most fortunate that our paths crossed back in Barcelona 1994, and I'm so grateful that we've been able to ride the same roller-coaster together ever since. I definitely got the better end of the deal.

Publications and presentations arising from this work

I.S. Morrison, “Detection of Antipodal Signalling and its Application to Wideband SETI”, Acta Astronautica, vol. 78, pp. 90-98, 2012.

D.G. Messerschmitt and I.S. Morrison, “Design of Interstellar Digital Communication Links: Some Insights from Communication Engineering'',” Acta Astronautica, vol. 78, pp. 80-89, 2012.

I.S. Morrison, “Interstellar Beacons Should Transmit at 50 GHz”, Astrobiology Science Conference (AbSciCon), Atlanta, Georgia, USA, April 2012.

I.S. Morrison, “Efficient Interstellar Communications and Implications for SETI (poster)”, Australian Astrobiology Meeting, Sydney, Australia, July 2013.

I.S. Morrison and M.G. Gowanlock, “Extending Galactic Habitable Zone Modelling to Include the Emergence of Intelligent Life”, Exoplanets, Biosignatures and Instruments (EBI2014), Tucson, USA, March 2014.

I.S. Morrison and M.G. Gowanlock, “Extending Galactic Habitable Zone Modeling to Include the Emergence of Intelligent Life,” Astrobiology, vol. 15, no. 8, pp. 683-696, 2015.

How often at night when the heavens are bright
With the light from the glittering stars
Have I stood here amazed and asked as I gazed
If their glory exceeds that of ours.

Taken from *Home on the Range* (John A. Lomax version), adapted from the poem *The Western Home* by Brewster Higley.

1 Introduction

1.1 Are there signals to find?

At this point in human history there have been no confirmed discoveries of any form of life – even the simplest microbial life – beyond Earth, let alone any evidence for the existence of intelligent extraterrestrial civilisations. However, it remains a very real possibility that such a discovery will be made in the near future; a view that has been buoyed by recent advances in the field of astrobiology, including discoveries of new and diverse extremophiles [1] [2] and the detection of large numbers of extrasolar planets [3] [4] (a small but significant fraction of which are classified as "Earth-like"). Despite the current uncertainty as to whether intelligent civilisations exist beyond Earth, the importance of this question suggests that an effort should be made to seek them out. In the words of Giuseppe Cocconi and Philip Morrison [5], "The probability of success is difficult to estimate; but if we never search the chance of success is zero".

Cocconi and Morrison's 1959 paper [5] is generally credited as marking the beginning of the scientific search for extraterrestrial intelligence (SETI). Their landmark paper maps out principles and methodologies that are still predominant in SETI today. Specifically they advocate the approach of searching in the radio part of the electromagnetic spectrum for 'narrowband' emissions – emissions in which the energy is concentrated within a narrow range

of frequencies. The first practical SETI experiment, "Project Ozma", conducted by Frank Drake in 1960 [6] followed these principles, as have the majority of other major search programmes conducted since that time. (Rather than list past and present SETI projects here, the reader is referred to the excellent surveys provided in [7] and [8].) The SETI Institute's [9] recent project that commenced in 2011 to search all of the "Kepler Worlds"[1] is understood to have employed the same basic methodology. None of these searches has been successful in making a confirmed discovery of a signal of extraterrestrial origin.

Should the lack of success in SETI to date be viewed as cause to question the basic postulate that life exists beyond the Earth? There is currently insufficient scientific evidence to make a confident statement either way regarding the existence or otherwise of extraterrestrial life. It is plausible that simple extraterrestrial life exists, but complex or intelligent extraterrestrial life does not. It is also plausible that extraterrestrial intelligences *do* exist and they are simply not transmitting signals in our direction, or there *are* transmissions but our search methods are inappropriate or lack the capability to discover these signals. In terms of the latter scenario, potential explanations for SETI's lack of success include: (1) we may not have been focussing our searches on the right parts of the sky; (2) we may not have been focussing our searches on the right parts of the electromagnetic spectrum; and (3) we may not have been focussing our searches on the types of signal waveforms likely to be employed for interstellar communications. This thesis addresses each of these points and concludes that, by their very design, conventional SETI searches are sub-optimal for the discovery of interstellar communications signals.

[1] The Earth-like exoplanets discovered by NASA's Kepler mission.

In conducting the assessments reported in this thesis, it has been the goal to refer as much as possible to fundamental science rather than the current technological status of our own civilisation. It can be argued that attempting to predict the specific technologies and communications methodologies likely to be adopted by extraterrestrial civilisations is a futile exercise, particularly when one considers how much older and more technologically advanced than ourselves they may be. However, it is not unreasonable to assume that extraterrestrial life will be constrained by the same laws of physics and mathematics as ourselves – and this presents an opportunity to apply a degree of objectivity in assessing current or newly-proposed approaches to SETI.

1.2 A new paradigm for SETI

There has been a high degree of anthropocentrism involved in reaching SETI's status quo. It is not uncommon to hear SETI researchers posit that the best place to look for extraterrestrial sources is our local galactic neighbourhood, on the basis that conditions for the emergence of intelligent life must be the most favourable in this region. Another typical SETI assumption is that any putative extraterrestrial signal would occupy a part of the spectrum that ensures it experiences low attenuation when passing through Earth's atmosphere [5]. Yet another example of anthropocentric reasoning is the commonly held view that SETI should look for the types of signals that on Earth have historically involved the strongest transmissions[2], which is suggestive of the narrowband signal components associated with radar or television/radio transmissions [10]. Furthermore, the preference for searching for narrowband tones has been, at least partly, driven by technological limitations. Akin to looking for lost keys under a street-light, the premise is to concentrate on searching for signal sources and types that our current technology is good at detecting. The technology to efficiently detect narrowband

[2] Those with the highest peak powers.

tones has existed since the earliest days of SETI, and the sensitivity of telescopes is less of a concern if signals are from nearby sources and are not attenuated by our atmosphere. This philosophy is perhaps best captured by "Dyson's Dictum", which suggests that the focus of SETI should be to "look for what's detectable, not for what's probable" [11].

While there is logic to Dyson's Dictum – it makes no assumptions about the nature of the extraterrestrial signal – it does not take advantage of pertinent information that can be expected to influence the design of any extraterrestrial signal source, or usefully constrain its spatial or spectral origins. It therefore misses the opportunity to provide useful guidance to those designing SETI experiments.

So, while traditional SETI is focussed on looking for signals that our current technology can efficiently detect, this thesis examines the very opposite strategy, which involves asking "what can we logically expect to be the most likely origin and form of interstellar communications signals?", and then orienting searches towards those types of signals, regardless of the challenges that may be faced in doing so. The problem should be approached with as few anthropocentric attitudes as possible, and our thinking should not be overly influenced by the constraints imposed by current Earth technology. These constraints will inevitably fall away as technology advances, and in the case of expanding the search beyond solely narrowband signals, have already largely evaporated with the inexorable march to date of 'Moore's Law' [12].

However, SETI already suffers from the challenge of multiple simultaneous search dimensions; spatial, temporal, spectral and polarisation. Allowing a further dimension – arbitrary signal formats – clearly adds to the scale of the search problem. Is there some way that the search-space can be narrowed? It is surely easier to find a needle in a smaller haystack.

One of the greatest challenges with interstellar communication is that it offers no scope for explicit coordination between the designers of the transmitting and receiving components of the system. However, David Messerschmitt (a communications engineer and SETI researcher) has suggested that this lack of coordination can be partially compensated by paying close attention to the underlying constraints, objectives and principles of communication link design. This can allow a form of *implicit coordination* to be achieved. Messerschmitt asserts that there is reason to be confident that the transmitter and receiver designers, addressing a common set of physical laws and propagation characteristics/impairments, will arrive at similar conclusions as to the basic elements of an end-to-end system design. This theme was explored in some detail by the author in collaboration with Messerschmitt, as reported in [13]. This paper has been included in full in Appendix A.

The work in this thesis relies heavily on the principle of implicit coordination to guide efforts to narrow the SETI search space in the dimensions of signal format and spectral location. In particular, the work has been primarily motivated by the implicit design goal of achieving *efficient communications*, which can be defined as communicating successfully at minimal cost in terms of resource consumption. For interstellar communications the dominant resource requirement is transmission energy, so efficiency translates to minimising the required energy per bit of information communicated.

On Earth, we are naturally concerned with efficiency because energy is a finite resource. However, it is possible to imagine there may be extraterrestrial civilisations with such advanced technology that they can access an effectively unlimited supply of energy. If this were the case, then energy efficiency would not be a concern to them. However, we have some justification following the past 50 years of SETI to speculate that such civilisations may

be rare. In the words of Gerry Harp of the SETI Institute[3]: "*Our past observations do provide some information... most importantly, we know that their transmitters are not arbitrarily powerful. If they were then we would have found them already... It says that ET has some limits on the power they can or will use to drive their transmitters... But if ET has some limitations, no matter how high they are, they will naturally think about optimizations that conserve resources.*"

A concern with managing energy consumption is only one of many possible explanations for SETI's lack of detections to date. It may be somewhat anthropocentric to suggest a desire for energy efficiency is universal. However, it is an assumption worth exploring that it could be a desire of the *majority* of extraterrestrial civilisations wishing to communicate with us. In these cases, we may expect the designers of their interstellar communications systems to pay close attention to achieving an efficient solution. This applies to both the transmitter and receiver designers since the overall system cost is shared by both.

Fortunately, there are fundamental laws of physics and information theory that can be applied to reliably gauge the comparative energy efficiency of different communications methods. Of course, we are constrained here by our current technological knowledge, which limits the types of communications methods that we can imagine. We are unable to assess the efficiency of methods not yet conceived. However, we *can* compare the relative efficiency of those different methods of which we are cognisant. We may also refer to the laws of information theory to assess how close a given design is to the fundamental optimum efficiency. There may be no need to invoke exotic unknown technologies if a simple method, already known to us, provides an efficiency that is already very close to the theoretical limit.

[3] Email correspondence, 25 February 2016.

1.3 Thesis scope and structure

The goal of this work is to review the past and current practices of SETI with a critical eye, to revisit many of the working assumptions, and challenge the "conventional wisdom" where appropriate. Overall this thesis aims to address the following fundamental SETI questions:

1. What to look for?

2. Where to look for it?

On the subject of *what* to look for, Chapter 2 compares two fundamentally different approaches to SETI: eavesdropping versus searching for intentional beacons. The origins and rationale of the so-called 'narrowband assumption' for beacons are reviewed, its weaknesses discussed, and a case for the alternative 'wideband assumption' is put forward. Consideration of the communications channel – the interstellar medium (ISM) – and other information-theoretic considerations lead to conclusions as to preferred waveform types for interstellar communications. An explanation is also presented for why SETI should take account of the possibility that signals could be transient (i.e. non-persistent) in their nature. Chapter 2 then explores techniques for detecting wideband signals, with emphasis on the low signal-to-noise ratio (S/N) regime that we anticipate for SETI. Existing methods are reviewed and a new method is proposed that aims to overcome the shortcomings of the existing methods. Significantly, this new method is a "blind detector" (capable of detection without knowledge of the signal waveform), which is essential for SETI. It also offers high sensitivity to a signal class that we postulate would be attractive for interstellar communications. Given sufficient observing time and processing power, the method can approach the theoretical maximum sensitivity, as would be achieved with an ideal 'matched filter' that has knowledge of the target signal format. Analysis and simulation results are presented that confirm the performance of the method.

Chapters 3 and 4 address “where to look?” in two key respects: (1) where in space, and (2) where in the electromagnetic spectrum. At the present time the SETI community does not have the resources to search all regions of space, at all times, for signals of every type, across every frequency band. An exhaustive search may be possible in the future, but until that time the logical strategy is to prioritise, i.e. to focus on those regions of the multi-dimensional search space that are more likely to contain extraterrestrial signal sources.

In terms of the spatial search dimension, research was conducted jointly with Michael Gowanlock to extend modelling of the Galactic Habitable Zone (GHZ) to include the emergence of intelligence and the type of technological civilisation detectable by SETI. The findings of this work were published in 2015 [14] and that paper has been reproduced in its entirety here as Chapter 3. This work confirmed quantitatively that the inner region of the Milky Way galaxy should provide the greatest chance for a successful SETI discovery, due to a higher density of habitable planets and greater opportunities for intelligence to emerge in that region. For targets in the inner Galaxy, the distance from Earth precludes the detection of unintentional (leakage) radio emissions, so in such a search, SETI should concern itself with the discovery of intentionally transmitted signals – backing up a conclusion already drawn in Chapter 2.

In terms of the spectral search dimension, Chapter 4 examines the question of frequency band selection for intentional beacon signals. The key starting assumption is that the energy supply available to the beacon-builder is finite and therefore energy has an associated cost to them; a cost they would naturally seek to manage. A model is developed to assess the end-to-end system cost as a function of operating frequency, for a range of scenarios. The model shows that there are efficiency advantages to operating in the high end of the microwave band, from approximately 30 to 90 GHz. The majority of past SETI searches have been conducted between 1 and 10 GHz, which is not the region of the spectrum that is most efficient for

interstellar communications. This may be a contributing factor to SETI's lack of success to date.

The thesis concludes with Chapter 5; a summary of the key findings of the work, and recommendations to improve the effectiveness of future SETI searches.

2 Artificial interstellar signals

In the absence of a confirmed discovery, it is not currently known how or where artificial interstellar signals may be discovered or, indeed, whether such signals even exist. It is the purpose of SETI to address these questions by searching for direct evidence, i.e. by seeking to discover artificial signals of extraterrestrial origin. However, at the present time the SETI community does not have the resources to search exhaustively all regions of space, for signals of every type, across every frequency band. A comprehensive search may be possible in the future, but until that time the logical strategy is to prioritise, i.e. to focus on those regions of the multi-dimensional discovery space that are more likely to contain extraterrestrial signal sources. This chapter examines the dimension of *signal type* and asserts that, to maximise the chances of a successful discovery, SETI should focus its searches on those artificial signal types judged most likely to be employed for interstellar communications. Traditionally, SETI has concentrated on searching for narrowband tones but, as will be explained, there are good reasons to eschew this approach in favour of wideband signal types.

2.1 Eavesdropping versus intentional beacons

There are essentially two distinct modes of conducting 'electromagnetic SETI'[4], distinguished by the type of electromagnetic signature for which detection is being attempted: (1) unintentional leakage radiation, or (2) intentionally transmitted signals. In the first case the search methodology is often termed *eavesdropping* for obvious reasons. The second case may involve either information-bearing (i.e. communication) signals or non-information-bearing signals. The term "beacon" is generally associated with the non-information bearing case, but

[4] Here we are considering only SETI by searching for artificial electromagnetic signatures, i.e. forms of electromagnetic radiation that are not recognised as the result of natural phenomena.

is sometimes used to refer to any type of deliberate interstellar signal transmission – which is how "beacon" will be used herein.

Proponents of eavesdropping generally cite the fact that Earth's current civilisation is known to produce a variety of electromagnetic radiations that 'leak out' into space. Therefore, it is reasonable to suppose that other extraterrestrial civilisations may generate similar signatures, and these may be detectable remotely by telescopes on Earth. But there are two fundamental issues with this hypothesis:

1. The radiation that unintentionally leaks from a planet inhabited by a technological civilisation is likely to consist of the sum of a multitude of sources at different frequencies, bandwidths, amplitudes, modulations and polarisations, and from different geographical locations with different radiation patterns and experiencing different reflections, diffractions, scatterings, etc. From a distance, this incoherent sum results in what is essentially broadband noise. There is likely to be no discernible structure to the radiation. To establish the existence of this cacophony of sources, one approach would be to measure the total electromagnetic energy emanating from the target planetary system, and if this is higher than expected for this type of system, then potentially there is something of interest happening. If one were to observe periodic variations in the level of emissions corresponding to a host planet's rotation period, or a reduction in emissions corresponding to times when a planet experiences a stellar occultation in the line of sight to the Earth, that would lend support to the possibility that a planet in the system is host to a technological civilisation. However, a system radiating an unusually high level of broadband noise does not automatically constitute evidence for a technological origin: it is difficult to distinguish between artificial and natural sources of un-structured noise.

2. The nature of leakage radiation from Earth has been changing over past decades with advances in technology. For example, there are fewer high power broadcasts using analogue modulation methods where the signal contains narrowband components, and an increasingly large number of relatively low power digitally-modulated wideband sources, such as mobile telephone base stations – and this trend is likely to continue. As noted in point 1 above, the incoherent sum of a multitude of different wideband sources resembles broadband noise. Eavesdropping is more likely to be successful if there exists a small number of individual sources of exceptionally high power, where that power is concentrated in either time (i.e. pulsed) or frequency (i.e. narrowband), such that the sources are more visible against the background noise of the observer's receiving system. An intelligent civilisation may typically spend very little of its technological history generating these types of emissions.

The most readily detected of Earth's emissions would be our highest power radar or television/radio transmissions. The detectability of such emissions has been studied by numerous researchers, including Loeb and Zaldarriaga [15], Forgan and Nichol [16], Billingham and Benford [17] and Siemion et al. [10]. Like any electromagnetic radiation, leakage radiation reduces in flux density as a function of distance squared. Taking all of Earth's current radiation sources together, the most optimistic of estimates suggest this signature could be detectable using an SKA-sized[5] receiver out to approximately 300 ly distance from Earth – and likely to be much less according to Billingham and Benford [17]. This suggests searches from Earth for extraterrestrial leakage from an Earth-like civilisation would be unlikely to succeed beyond a

[5] The Square Kilometre Array (SKA) is the largest next-generation radio telescope, due to be constructed in the early 2020s, which will ultimately have a collecting area in the order of one square kilometre (10^6 m^2).

few hundred light years. Unfortunately there are relatively few stars this close to Earth. For example, there are only ~450,000 stars within 300 ly of Earth, compared to an estimated ~300 billion total stars in the Milky Way. This limit to the discovery range significantly diminishes the chances of a successful SETI discovery from eavesdropping. As will be shown in Chapter 3, the vast majority of our Galaxy's stars capable of hosting a technological civilisation are likely to reside in the inner region of the Galaxy, many thousands of light years from Earth where eavesdropping will be ineffective. Notwithstanding, interest in eavesdropping on narrowband leakage signals remains strong within the SETI community, as evidenced by the attention given to this strategy in proposals for future SETI with the SKA [10].

The basic issue with eavesdropping is that the target emissions are typically leaking from their source in all directions of the sky, with energy that is incoherent and not generally concentrated in either time or frequency. This makes remote detection much more challenging. By contrast, an intentionally transmitted beacon signal can be coherent and concentrated in time and/or frequency. The signal can be directed in a tight beam by means of a high-gain transmitter antenna. The antenna directivity and transmit power level can be chosen to ensure the transmitted signal is detectable with practically-sized receiving antennas, across the entire breadth of the Galaxy[6], or indeed across intergalactic distances. Figure 2-1 illustrates the disparity between the discovery range for leaked emissions (within the red circle) versus intentional transmissions (within the orange circle, or even beyond it for extragalactic sources).

[6] Without delving into the details here, it can be shown that pan-Galactic communication at low information rate is possible using Arecibo-sized (305 m diameter) transmit and receive antennas and a transmission power of megawatt order, which would be straightforward to implement with current Earth technology.

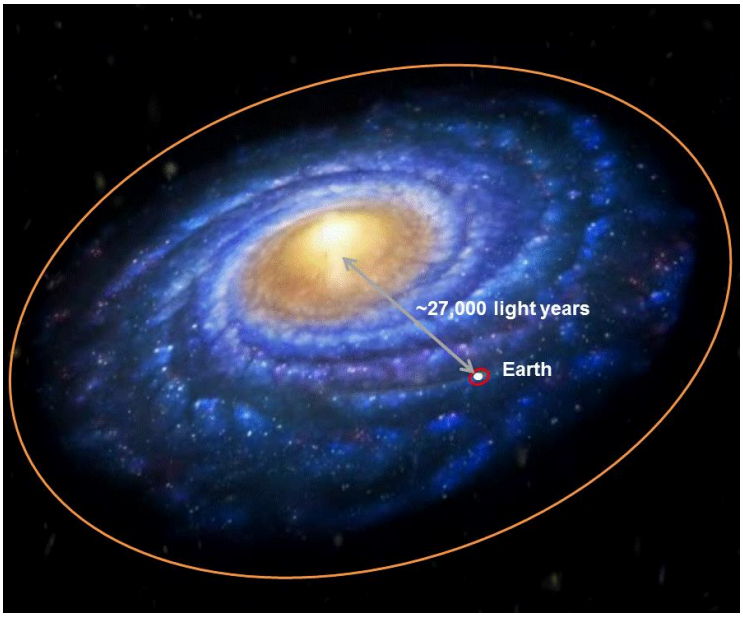

Figure 2-1: Comparison of typical discovery ranges for eavesdropping of leaked radiation (within the red circle) and detection of intentional beacons (within the orange circle).

[Milky Way illustration credit: Mark A. Garlick]

2.2 The narrowband assumption and its weaknesses

A narrowband signal of a given power is generally more detectable under Fourier spectral analysis against a background of broadband noise than a wideband signal of the same power (as illustrated in Figure 2-2). This has led to a perception that the inherent discoverability is higher for narrowband signals of a given S/N than for wideband signals at the same S/N [18]. This view has been further bolstered by the proponents of eavesdropping, where narrowband transmissions can reasonably be assumed to represent one of the most easily detected types of leakage radiation.

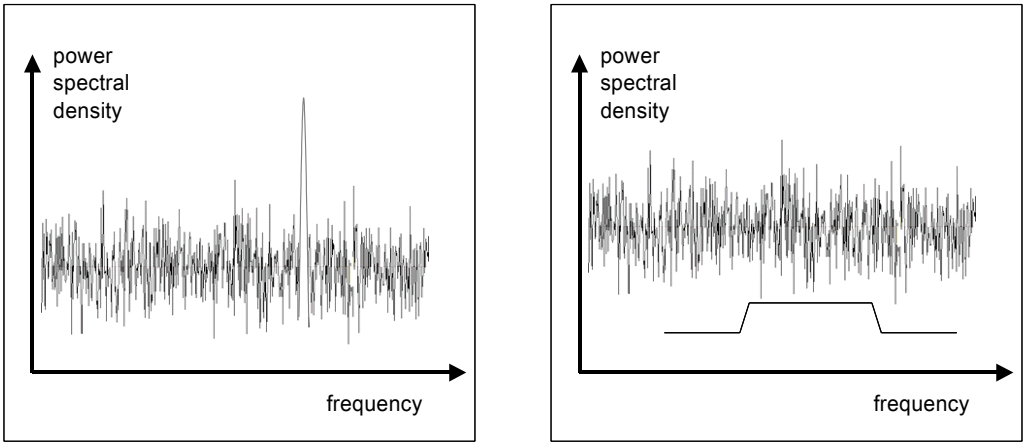

Figure 2-2: Illustrative power spectral densities for narrowband (left) and wideband (right) signals embedded in noise. The narrowband signal clearly protrudes above the receiver background noise level and can be detected using standard spectral analysis techniques. The wideband signal on the right is intended to have roughly the same total integrated energy as the narrowband signal on the left. Its noise-free spectrum is shown below the background noise for illustration purposes. However, its presence could not be inferred from standard spectral analysis because, being spread across a wider bandwidth, its peak power spectral density is much lower – in this case below the background noise density.

In the case of an intentionally transmitted beacon, an assumed advantage to the beacon builder in choosing a narrowband signal format is that it will make it easier for the intended recipient to distinguish the signal from natural emissions. Furthermore, if it were true that the discoverability for a given S/N is higher, this would imply a beacon source could function at a lower transmit power to achieve the desired discoverability objective. It is also generally assumed that narrowband signals are easier to generate at the high power levels needed for an interstellar beacon transmitter. Finally, the dispersive effects of propagation through the ISM (discussed further in Section 2.11.6) can be rendered insignificant if the signal bandwidth is made sufficiently small, meaning that dispersion will not be a complication to the discovery process.

Taken together, this reasoning has led to a preoccupation with searching for narrowband signals that has pervaded SETI from the very beginning and persists to the present day. We refer to this as the ***narrowband assumption***.

However, there are various reasons to challenge the narrowband assumption, including:

- **Narrowband signals are not *conclusively* of artificial origin**. In the SETI context, "narrowband" generally refers to signals of a few Hz of bandwidth or less. This is much narrower than the narrowest currently known natural signals; those from astrophysical masers, which have been found to have spectral components with a linewidth as low as 550 Hz [19]. However, the future discovery of natural phenomena with even narrower emissions cannot be ruled out.

- **Narrowband signals are not *fundamentally* more discoverable than wideband signals.** The high sensitivity of spectral analysis to narrowband signals arises because the detection algorithm (e.g. a Fast Fourier Transform (FFT) [20]) is based on separating the signal waveform into sinusoidal components, which results in the detector closely approximating a matched filter[7] for a narrow tone. Any waveform type, including an arbitrary wideband signal, can be detected with the same sensitivity by its corresponding matched filter. However, the design of the matched filter requires precise knowledge of the transmitted waveform, which is a key issue for interstellar communications where explicit coordination between the transmitter and

[7] The optimum detector for any signal (narrowband or wideband) is a 'matched filter'. The matched filtering process can be thought of as correlation with the expected (noise- and distortion-free) signal waveform, so by definition it requires knowledge of the transmitted waveform [32].

receiver is not possible. But, fundamentally, if the waveform is known, it makes no difference to discoverability what bandwidth the signal occupies [21].

- **Generation of very high-powered signals is not easier for narrowband signals than wideband signals.** Based on current Earth technology, the most efficient devices for generating microwave power emit incoherently over wide fractional bandwidths (>10% of the centre frequency) [22]. For generating narrowband or *coherent* wideband emissions, the preferred approach is to employ a module that uses a lower power device that is more linear in operation, followed by a linear amplifier. Combining the power from a large number of such modules (which would be a natural architecture for a multi-element array antenna) would equally support either narrowband or coherent wideband high power emission.

- **A narrowband beacon signal cannot encode significant information content.** A truly monochromatic signal has zero bandwidth and cannot contain embedded information content. One may say that by its presence it provides information on the source's location on the sky, together with some other useful information uncovered through the detection process[8]. In message terms it effectively conveys just one bit of information: "you are not alone". If we assume there may be many interstellar beacons, any given beacon will typically not be the first detected by any given recipient – in which case there is little incremental value in being told "you are not alone", aside from helping to establish the prevalence of such sources. A signal of narrow but non-zero bandwidth is able to convey information at a low rate,

[8] The detection process may reveal amplitude scintillation or bandwidth broadening due to propagation of the signal through the ISM. Along with sky location, this information may help to associate the source with a specific star system.

commensurate with its bandwidth. However, as explained in Section 2.4, it cannot do this efficiently unless the data rate is a small fraction of the bandwidth. For example, a narrowband SETI detector that assumes a signal bandwidth of 1 Hz (typical of current narrowband SETI searches) may only be effective for discovering beacons with an information rate less than 0.1 bits per second (bit/s) – an extreme and unwanted limit on the search coverage of the full discovery space. For intentional transmissions, the defence of SETI's preoccupation with narrowband sources relies on the following assumptions: (1) that the beacon builder has no interest in sending a complex message containing more than a few hundred bits per hour to its target receivers, and/or (2) the signal is intended only as an 'attractor beacon' to aid discovery of an associated wideband communication signal.

- **Leaked radio emissions cannot be assumed to contain narrowband components**. In the eavesdropping scenario, it is true that the presence of narrowband spikes among a civilisation's leakage would make detection more straightforward. However, as pointed out by Shostak [23], it should not be assumed that extraterrestrial communications signals will contain narrowband components. On Earth our communications (and radar) signals have tended to become spectrally wider and flatter as technology has advanced, and there may only be a short period in a technological civilisation's history during which any significant level of narrowband radio emission is present.

- **Narrowband signals of high power are 'jammers'.** A narrowband signal (or pulsed wideband signal) concentrates its energy in a narrow frequency (or time) range, resulting in a high peak power spectral density. A powerful pan-Galactic or intergalactic narrowband beacon would saturate its frequency band in the direction of

transmission[9], thus rendering this part of the spectrum unusable for any other purpose. It would have a major impact on astronomical observations in that band, for potentially a significant fraction of the Galaxy. Such a signal has all the hallmarks of a 'jammer', such as is used by military forces to disrupt an enemy's communications capabilities. It has previously been suggested that beacons should operate at or close to important astrophysical spectral line frequencies, but this may be regarded as poor 'astronomy etiquette' by any civilisation that values astronomical science (as we do on Earth). The very opposite conclusion – that beacons should utilise a frequency that is well away from any spectral lines of interest – may well be more appropriate. It is arguably even more preferable for beacons to completely avoid transmitting 'spikey' signals in favour of those that are spread over wider bandwidths and possess lower peak power spectral density.

- **Narrowband signals are highly susceptible to electromagnetic interference**, known in the radio band as radio frequency interference (RFI). On Earth, even the most remote telescopes are exposed to terrestrial and satellite sources of RFI that can at times resemble the artificial signal types that SETI is seeking to discover. Furthermore, this RFI is typically at substantially higher flux densities than the weak signals being targeted. This severely impacts the efficacy and efficiency of Earth-based SETI, particularly narrowband SETI because of the higher peak power spectral densities associated with narrowband RFI sources. It is reasonable to assume that beacon builders will be conscious of this local issue for receivers located within technologically

[9] This would mean all directions for an isotropic transmitter. Even with a directional transmitter, saturation would occur in all directions in the vicinity of the transmitter due to emissions via the antenna side-lobes.

active civilisations and choose a signalling waveform that is less susceptible to RFI. As explained by Messerschmitt [24], spread-spectrum signalling [25], in which the signal appears like white noise, represents a compelling choice of waveform type. Diluting a signal's power across both frequency and time is the most effective way to maximise immunity to unknown sources of noise and interference at the receiver.

Of all the above considerations, the overriding concern must surely be the question of **information content**. Given the sizable investment needed to build and operate an interstellar beacon transmitter, to human sensibilities it seems incomprehensible to build a beacon that does not embed information in its signal. Discussions on the wider impacts of future successful SETI discoveries here on Earth invariably refer to the message content to be extracted from the signal (despite the obvious incongruity with the narrowband assumption so prevalent in observational SETI searches). It is perhaps anthropocentric to assume that extraterrestrial civilisations share humankind's motivation to communicate. But if they have gone as far as deciding to transmit an interstellar beacon signal, it seems probable that at least some such beacon builders will embed a message. Those wishing to send a message may still, for technical reasons, decide to assist discovery by transmitting a narrowband attractor beacon alongside their wideband communication signal. However, SETI should not rely on such assistance and instead be willing to search for beacons that consist of a single transmitted signal that communicates information. Furthermore, there seems little reason for a beacon builder to select a signalling format for this signal that is not *inherently discoverable*. The overall energy efficiency of the system should be higher if the communication signal itself can be discovered, without needing to expend energy transmitting a separate attractor signal. In this case, a SETI receiver has only the communication signal itself with which to achieve discovery. As pointed out by Clancy [26] and elaborated by Jones [27] and Messerschmitt [21], we should expect this communication signal to be wideband, since this follows directly from

efficiency considerations and the fundamental laws of information theory. This will be explained more fully in Section 2.4, where we also address the question "how wide is 'wide'?"

We will refer to the conclusion that *intentionally transmitted interstellar beacons* will be wideband as the ***wideband assumption***.

2.3 Isotropic versus directional beacons

Energy considerations also lead naturally to another conclusion about the nature of interstellar beacons: *they are likely to be transmitted using highly directional beams*.

An isotropic beacon is a compelling idea, because it requires no assumptions about the location of target receivers, aside from them falling within the design range of the beacon. However, as pointed out by Benford et al. [28] [29], the energy levels required to transmit a detectable signal *isotropically* over galactic-scale distances would be enormous. They suggest that pan-Galactic beacons would require an Effective Isotropic Radiated Power (EIRP) (the power transmitted in the direction of the target) of at least 10^{17} W. This implies a power source of at least 10^{17} W when using an isotropic antenna. A simple dipole antenna located near the Galactic centre would provide omnidirectional coverage in azimuth, with ~2 dB gain[10] in the direction of the Galactic plane (it would radiate less strongly perpendicularly to the Galactic plane, which is desirable). The required transmitter source power to achieve a 10^{17} W EIRP would be ~6×10^{16} W. By contrast, if a directional antenna of gain 100 dB[11] were used, this would achieve the same pan-Galactic range with a source power of just 10 MW – well within

[10] Antenna gain is discussed further in Chapter 4. It can roughly be equated with the inverse of the fraction of the total sky solid angle illuminated by the antenna beam.

[11] By way of example, 100 dB gain is achieved with an Arecibo-sized (305 m diameter) antenna of aperture efficiency 0.7 operating at 40 GHz.

the capabilities of current Earth technology. The drawback, of course, is that such a highly directional antenna produces an extremely narrow beam that will cover just 10^{-10} of the solid angle of the entire sky. Using such a beamwidth would require 10^{10} different successive pointings (or 10^{10} simultaneous beams) to achieve complete isotropic coverage. In the case of a beacon that cycled through successive pointings, each individual target receiver would be illuminated by the beacon signal for only 10^{-10} of the time – making discovery very difficult. The alternative strategy is for the beacon to cycle between a smaller number of targets that have been assessed as potential hosts for intelligent life. For example, for a target list of 10,000 stars hosting planetary systems, the total energy required for the directional beacon system of this example would be 1 million times less than the isotropic case. As the number of targets is increased, the trade-off against an isotropic transmitter becomes less favourable, being equivalent when there are 10^{10} targets (in this example). However, even with as many as a million targets, the directional beacon would require only 0.01% of the energy of an isotropic transmitter. It has now become a generally accepted assumption in SETI that beacons will utilise directional beams – the so-called “lighthouse analogy”.

Note that with a directional beacon system, if there are fewer simultaneous beams than the number of targets, this also implies that each target will not be illuminated continuously, but instead will see the beacon as a transient source with characteristic revisit and dwell times (that are unknown prior to discovery) [30].

Of course, the directional beacon strategy pre-supposes that the beacon-building civilisation has been able to conduct a sufficiently detailed survey of the Galaxy to enable them to produce an informed target list. The Kepler survey results suggest that most stars are host to at least one small rocky planet, meaning potentially billions of target stars. However, this list could be narrowed through astronomical observation to select only stars/planets exhibiting certain specified characteristics, such as stars with at least one planet within their circumstellar

habitable zone, or evidence of a planetary atmosphere containing a biosignature such as an over-abundance of oxygen (suggestive of the presence of microbial life). Completing such a Galactic-scale survey would require advanced astronomy capabilities[12] and incur a substantial cost in its own right that should be offset against the reduced energy costs of a directional beacon design. However, it seems reasonable to assume that any civilisation that desires to advertise itself by means of an interstellar beacon is likely to be a civilisation that takes a scientific interest in astronomy and cosmology – in which case they might naturally be expected to pursue this type of astronomical survey irrespective of what intentions they may have for operating an interstellar beacon, i.e. they are likely to compile the data needed for an informed beacon target list naturally over time, as and when their technology allows.

In this work we accept and adhere to the directional beacon assumption, i.e. that any civilisation choosing to invest in building an interstellar beacon will seek to benefit from the efficiencies of employing a directional transmitter, and they have the technical capability to short-list target systems more likely to harbour advanced life. This assumption is one of the foundations of the analysis presented in Chapter 4.

2.4 Wideband signalling for efficient communication of information

In Section 1.2 an argument was presented for why energy efficiency is likely to be a key factor in the design of interstellar beacon systems. Once operational, the dominant energy cost for an interstellar beacon will be that associated with generating the high-power transmitted signal. Maximising energy efficiency translates to minimising the required energy per transmitted information bit. For a given energy budget, this allows more data to be

[12] This is consistent with the common assumption in SETI that any other technological civilisation we may encounter is statistically likely to be more advanced technologically than ourselves, given we have only acquired the technology to perform observational astronomy in the past few hundred years.

transmitted, or the same amount of data to be sent over a greater distance or to a larger number of target recipients.

The limits on the minimum energy required to transmit information were established by Claude Shannon in the 1940s [31]. Shannon developed a mathematical expression for the maximum information rate that can be achieved over a channel of given bandwidth and S/N – referred to as 'channel capacity' *C*. Using the terminology from [32]:

$$\boldsymbol{C = W\,log_2\left(1 + \left(\frac{P}{WN_0}\right)\right)} \qquad (1)$$

where *W* is the total channel bandwidth, *P* is the signal power and N_0 is the noise power spectral density in the channel, assuming additive white Gaussian noise. The minimum energy per bit is achieved when the information rate matches *C*. Given the purely mathematical underpinnings of Shannon's work (i.e. involving no a priori assumptions), there is no reason to doubt that it represents a fundamental physical limit – one that will also be well known to any extraterrestrial beacon builder.

It is instructive to normalise Equation (1) by the S/N. Something similar was done by Tse and Viswanath in [33] where they normalised just the capacity, but it is useful to normalise both the capacity and bandwidth to generalise with respect to *P*. We define the signal-to-noise-density-ratio, SNDR = $\frac{P}{N_0}$ and define the capacity-per-unit-SNDR as:

$$\boldsymbol{\frac{C}{\mathrm{SNDR}} = \left(\frac{W}{\mathrm{SNDR}}\right) log_2\left(1 + \left(\frac{\mathrm{SNDR}}{W}\right)\right)} \qquad (2)$$

This is plotted in Figure 2-3 as a function of the normalised bandwidth $\left(\frac{W}{\mathrm{SNDR}}\right)$. It is seen that the normalised capacity asymptotes to $\log_2 e$ as the normalised bandwidth increases towards infinity. That is, there is a maximum information rate per unit of S/N that cannot be exceeded

no matter how much bandwidth is available. At low bandwidths the capacity falls rapidly and much larger S/N values are required for reliable communications.

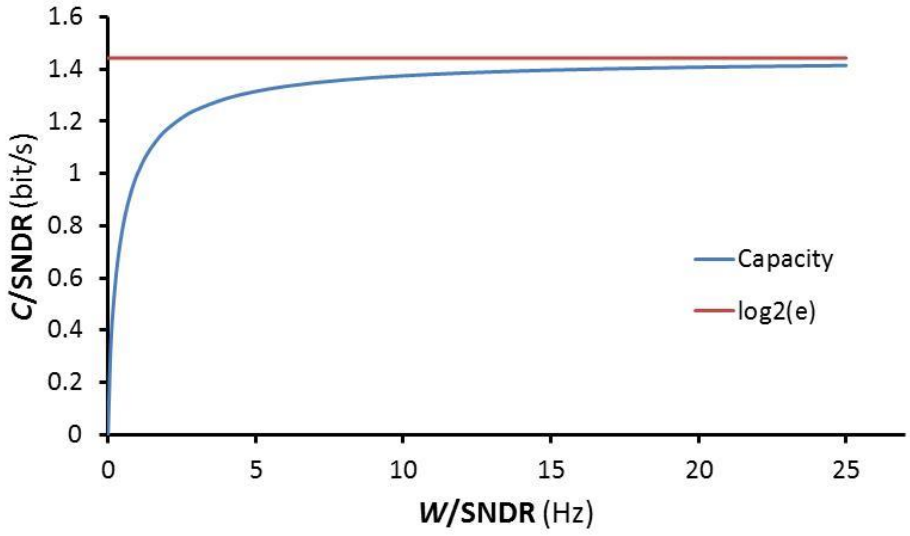

Figure 2-3: Capacity of the additive white Gaussian noise channel, normalised by the signal-to-noise-density-ratio, SNDR = $\frac{P}{N_0}$.

The dependence of required S/N on bandwidth is better illustrated by re-arranging Equation (1) to provide an expression for the minimum required E_b/N_0[13]. In practice, the communication rate, R, must always be less than C. We define the 'Bandwidth Expansion Factor' (BEF) as the ratio W/R. Then $E_b = \frac{P}{R} = BEF\left(\frac{P}{W}\right)$. From Equation (1) we have

$$\boldsymbol{R < W \log_2\left(1 + \left(\frac{P}{WN_0}\right)\right)}$$

$$\boldsymbol{2^{\left(\frac{R}{W}\right)} < 1 + \left(\frac{E_b}{BEF.N_0}\right)}$$

$$\boldsymbol{\left(\frac{E_b}{BEF.N_0}\right) > 2^{\left(\frac{1}{BEF}\right)} - 1}$$

[13] E_b is the energy per information bit. E_b/N_0 is equivalent to the SNDR per bit.

$$\frac{E_b}{N_0} > \boldsymbol{BEF}.\left(\mathbf{2}^{\left(\frac{\mathbf{1}}{\boldsymbol{BEF}}\right)} - \mathbf{1}\right) \qquad (3)$$

This expression has been plotted in Figure 2-4. In the limit as BEF → ∞, E_b/N_0 asymptotes to ln(2) = 0.6931 (-1.59 dB), known as the 'Ultimate Shannon Limit'. Regardless of the available bandwidth, reliable communications is only possible for E_b/N_0 values above this limit.

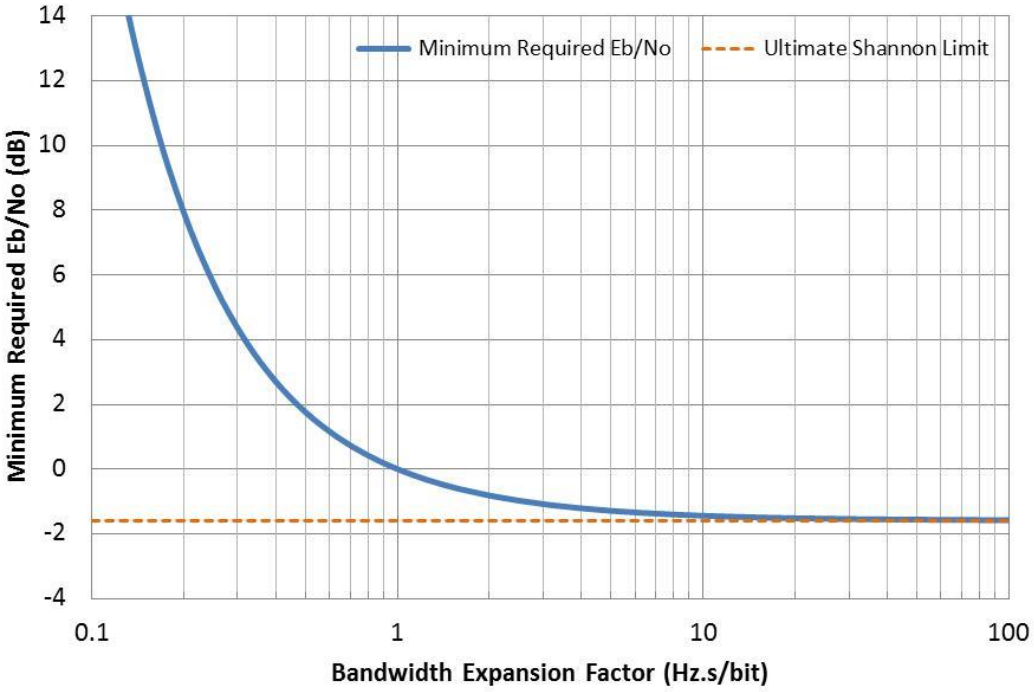

Figure 2-4: The minimum required E_b/N_0 as a function of the Bandwidth Expansion Factor, BEF = W/R (solid curve). Also shown is the asymptotic limit as BEF → ∞ (dashed curve), equal to -1.59 dB and known as the 'Ultimate Shannon Limit'.

The key observation from Figure 2-4 is that the lowest E_b/N_0 (highest energy efficiency) is only achieved with large BEFs. This means that **energy-efficient interstellar communications signals must be wideband.**

For interstellar communications, where transmitter power levels are potentially extremely large, there is a strong incentive to operate as close as possible to the Ultimate Shannon Limit.

Not only will this minimise the energy consumption during operation of the transmitter, but by reducing the power output requirement, the cost and complexity of building the transmitter system will also be reduced. Fortuitously, there is virtually unlimited bandwidth available for interstellar communications, particularly when operating at a centre frequency of tens of GHz (as recommended in Chapter 4). This enables a beacon builder to trade bandwidth for increased energy efficiency and, in principle, closely approach the Ultimate Shannon Limit. To get close to the limit, the BEF needs to be at least an order of magnitude higher than the information rate. Increasing the BEF further will give small incremental gains, but it is a case of diminishing returns. The additional cost to implementing more complex encoding/decoding processing required for larger BEFs may deter beacon builders from exceeding a BEF of ~10. Even so, this means the transmitted signal can be expected to have a bandwidth that is at least 10 times wider than the information rate (expressed in Hz).

There is a second reason to expect interstellar communications signals to be wideband, already mentioned in Section 2.1: spreading the signal power over frequency and time will make the signal detection at the target receiver less susceptible to RFI [24]. It may be desirable to increase the overall BEF (through a combination of encoding and spreading) to hundreds or thousands to maximise the benefits. Very large BEFs will increase implementation cost and can significantly complicate signal discovery, so there must be a trade-off when designing the waveform for a beacon transmitter.

2.5 Discovery versus communications

It is worthwhile noting at this point that the Ultimate Shannon Limit refers to the minimum S/N required to achieve reliable *communication*, i.e. to extract the information bits from a signal. We may refer to this as the 'decoding S/N'.

In SETI we are initially more concerned with *discovering* the existence of an artificial extraterrestrial signal. Recovering the information content of the signal can be attempted subsequently. It is typical within SETI to assume our current telescopes and receivers may not have sufficient sensitivity for extracting information. But for many signal types and detection algorithms it is possible to trade observation time for sensitivity. If we can stay on-target for long enough, it may be possible to integrate a detection metric over many communications symbols, such that our detector achieves what may be called the 'discovery S/N', at which point we can assert the *presence* of a signal.

Perhaps counter-intuitively, the 'discovery S/N' may need to be significantly higher than the 'detection S/N', otherwise there may be excessive false-positive detections. However, when there is a detection metric for a given signal class that accumulates with increasing observation time, it is within our control to achieve any desired 'discovery S/N', given sufficient observing time. Extracting a detection metric of this kind is more straightforward for some signal types than others. It is posited that the signalling method for an interstellar beacon should be chosen with this consideration in mind. In this way, *discoverability* can provide another form of implicit coordination.

2.6 Challenges of wideband SETI

Having established that SETI should concern itself with wideband signals, what are the implications? There are a number of aspects that make searching for wideband signals more challenging than narrowband signals. These include:

- **Many degrees of freedom**. As with a narrowband beacon, the operating frequency for a wideband interstellar beacon is unknown. Additionally for wideband beacons, the type of signal modulation and its associated parameters are also unknowns. Whilst a high-resolution spectrogram can detect narrowband signals with high sensitivity

(because it is essentially a matched filter for sinusoids), this will not provide high sensitivity for wideband signals. Achieving maximum detection sensitivity requires matched filtering against the precise waveform shape of the target signal, which is unknown. Searching for wideband signals requires either (i) matched filtering over a large number of trial-and-error guesses of the signal format, or (ii) the use of a detection algorithm that is tolerant of unknown signal parameters – so-called 'blind detection' – which will generally incur a loss of sensitivity.

- **Spread-spectrum.** As discussed in Section 2.4, for maximisation of power efficiency, the encoding and modulation of the signal is likely to result in a bandwidth that is substantially higher than the symbol rate. There are also compelling reasons to deliberately widen the signal bandwidth even further through some type of spread-spectrum technique [25], to improve tolerance to local sources of RFI at the receiver. This can be done without sacrificing power efficiency when the specific spreading process is known to the receiver, since the receiver is able to benefit from the processing gain associated with de-spreading. Once a spread-spectrum signal has been discovered, analysis of the signal can reveal the parameters of the spreading process. However, prior to discovery, the spreading process is an additional and very significant unknown. Spreading has the potential to render the signal virtually invisible because it can result in a very low power spectral density and a resemblance to white noise. Prior to de-spreading, the spectral density of the signal could be well below the noise floor of the receiver, or potentially even below the sky noise floor due to the cosmic microwave background (CMB), significantly complicating discovery.

- **Low S/N; long integration times.** When SETI is attempting discovery at low S/N, it becomes necessary to integrate a detection metric over a long observation time to reveal the presence of the signal. When integration times exceed minutes or hours, an

additional detection issue arises. The gain stability of practical telescopes is imperfect, even with rigorous processes for gain calibration. Also, depending on the frequency and line-of-sight dispersion measure (DM), the signal will experience some degree of amplitude scintillation, in both the time and frequency domains. When observations are longer than the characteristic timescales of the instrumental gain variations and scintillation, it becomes very difficult to disentangle the desired detection metric from these variations, particularly if the detection metric is gain-related (as is the case with an energy detector). A wideband signal, especially one of a spread-spectrum format, will have a much lower peak power spectral density than a narrowband signal of the same power, so this issue is greatly exacerbated with wideband SETI. (In Section 2.11 an alternative wideband detection metric is presented that avoids this issue.)

- **Interstellar channel impairments.** Wideband signals are affected in more complex ways than narrowband signals by propagation through the ISM. For example, Doppler shifts and accelerations (drifts) can arise due to the orbital motions of the host planets of the transmitter and receiver, and motion of the intervening ISM. A static shift causes a time-dilation effect, which for a sinusoid results in a simple frequency shift, whereas for wideband signals the entire signal shape is stretched in time. Doppler drift results in a dynamic time-dilation effect, which can severely impact the sensitivity possible with many wideband detection methods. As another example, ISM dispersion affects a sinusoid by introducing a simple path delay related to its frequency, whereas for a wideband signal, each component frequency experiences a different delay, resulting in a time smearing that distorts the shape of the waveform in both time and frequency. The impact of ISM propagation on wideband signals has been studied in detail by Messerschmitt [21]. One of the key conclusions of that work is that, for a given operating frequency, any line of sight through the ISM will provide a propagation

channel with a characteristic coherence time and coherence bandwidth. These two parameters define what Messerschmitt calls the Interstellar Coherence Hole (ICH). If the instantaneous signal bandwidth is less than the coherence bandwidth, and the detection process instantaneously operates on a time segment of the signal shorter than the coherence time, then the distortive effects of the ISM can essentially be ignored. As explained in [21], coherence bandwidths are typically tens of kHz to MHz, and coherence times are typically seconds to minutes.

Combined with the spatial search dimension, the many unknowns and challenges of wideband SETI have given rise to a common perception that wideband searches are impractical – which has been one of the main reasons for SETI's preoccupation with narrowband searches. Central to the work of this thesis is the demonstration that there is scope to constrain the size of the discovery space to a substantial degree through implicit coordination. Even so, there will necessarily be more degrees of freedom with a wideband search compared to a narrowband search.

2.7 Blind detection

One of the fundamental trade-offs when designing a signal detector is that between *specificity* and *sensitivity*. The "most blind" of detectors (e.g. an energy detector) will generally have the lowest sensitivity in the sense that they require a higher discovery S/N. At the other extreme is the matched filter, which requires the lowest discovery S/N for a given detection probability. For SETI, matched filtering can be employed by selecting a set of candidate waveform types and performing matched filtering for each waveform assumption. This can be computationally expensive, and there is also a significant likelihood that the actual beacon waveform does not fall within the list of candidates. More preferable for SETI is a blind detector that makes as few assumptions as possible about the waveform parameters while offering acceptable sensitivity. In [34] it has been shown how (largely) blind detection can be accomplished for a particular

class of waveforms (in this case spread-spectrum phase modulation) with a sensitivity that exceeds that of a simple energy detector. One of the core research challenges for wideband SETI is the development of new blind detection algorithms that maximise detection sensitivity while taking account of implementation practicalities. The latter concern relates to the computational complexity of the algorithms, which must be taken into account if real-time operation within radio telescope back-end processors is a goal. However, these concerns tend to diminish over time as processing capabilities inevitably increase, following Moore's law.

In Section 2.11, the "symbol-wise autocorrelation" (SWAC) algorithm that was introduced by this author in [34] is reviewed and a more advanced variant is described that trades computational complexity for increased sensitivity. It is shown that the advanced variant can approach the sensitivity of a matched filter for a certain class of target waveform, given sufficient observational data and computational resources. Whilst that level of sensitivity is not attained for all wideband signal types, it serves as an example direction for wideband SETI algorithm development: balancing specificity and sensitivity for signal classes that are attractive for interstellar communications. Crucially, it demonstrates that the commonplace perception of the intractability of wideband SETI is unfounded. There is no longer a practical reason for SETI to restrict itself to searching only for narrowband signals.

2.8 Examples of wideband signal types suited to interstellar communications

2.8.1 M-ary orthogonal modulation

Jones [27] and Messerschmitt [21] both recognised that interstellar communications signals are likely to occupy a wide bandwidth and operate close to the Ultimate Shannon Limit. They also both came to the conclusion that the best signalling method for interstellar communications is a scheme known as 'M-ary orthogonal modulation'. Messerschmitt favours a particular variant of M-ary orthogonal modulation that is a form of 'pulse-position

modulation', which can be applied in either the time dimension or the frequency dimension, or both. The frequency dimension case is illustrated in Figure 2-5. The scheme uses strong individual pulses that are transmitted in specific two-dimensional locations (centre frequency and time) chosen from a finite set of M possibilities. In this way, each pulse can convey $\log_2 M$ bits of information. The individual pulses consist of an energy burst that can either be relatively narrowband or wideband for improved RFI immunity. The bursts can occur over a wide total bandwidth, giving rise to a high BEF. This is a scheme that has been known and studied for decades [35], and is widely considered to be the simplest way (conceptually at least) to approach the Ultimate Shannon Limit when bandwidth is unconstrained.

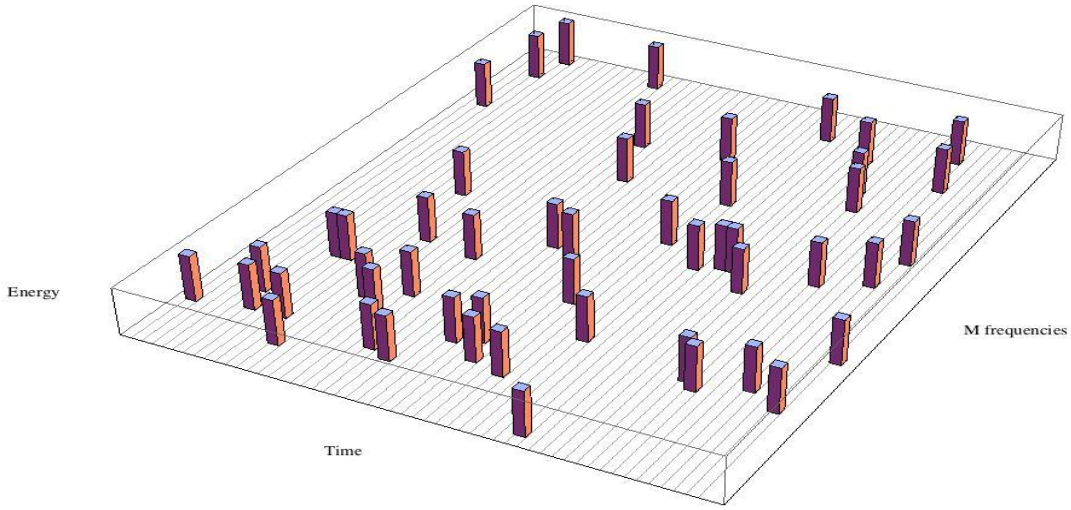

Figure 2-5: Illustration of M-ary pulse-position modulation, where information is encoded in the pulse location in the frequency dimension [credit: David Messerschmitt].

While this approach is simple and elegant (a useful trait from an implicit coordination point of view), it suffers from some drawbacks. One is that peak powers are high because the signal is sparse in frequency and/or time. This introduces implementation challenges for the transmitter, and also makes the signal more 'jammer-like', as discussed in Section 2.2. Another drawback is that this scheme approaches the fundamental performance limit relatively slowly with increasing bandwidth. This means the BEF may need to be many

thousands to get within a fraction of a dB of the Ultimate Shannon Limit, which may add cost to the transmitter implementation[14].

Another concern with a sparse signal of this type is its level of discoverability at low S/N. It relies on individual pulses exceeding the background noise floor of the receiver. As discussed in Section 2.5, when the S/N is too low for extracting the information content from a signal, it may still be possible to discover the *presence* of the signal by integrating a detection metric over a longer observation time. It is much more difficult to perform such long-duration integration with sparse signal types – which works against them as a choice for interstellar beacons.

2.8.2 Spread-spectrum phase modulation

The difficulties associated with detecting wideband signals at low S/N have led to concerns that information-bearing wideband beacon signals would need to be accompanied by a more easily discovered narrowband or pulsed 'attractor beacon'. However, these concerns have often overlooked the fact that wideband signals can be deliberately selected (or constructed) so as to aid discovery.

A wideband signalling approach that is inherently highly discoverable, and overcomes the concerns with sparse signalling methods, was suggested by this author in [34]. Referred to as 'antipodal spread-spectrum', it is a form of spread-spectrum binary phase modulation that is continuous and has a constant power envelope. With appropriate coding, it can be very power efficient and approach the Ultimate Shannon Limit in terms of required energy per information bit. As noted in [34], there also exists an effective blind detection mechanism suitable for low-

[14] Combining with error correction coding can allow the fundamental limit to be approached more rapidly with increasing bandwidth.

S/N discovery of such signals; the SWAC algorithm described in Section 2.11. Since the information-bearing signal itself is easily discovered, there is no need for a separate attractor beacon. The existence of this complementary signalling method and detection algorithm would appear to make antipodal spread-spectrum a compelling solution for the interstellar beacon application.

Antipodal modulation is a form of binary signalling in which there are only two members in the set of possible transmitted symbol values (i.e. the 'symbol alphabet'). With antipodal modulation the two members of the symbol alphabet are the inverse of one another. This alphabet can be represented as [A, -A]. Binary data can be mapped to this signal set by assigning one symbol type to represent 0 and the other to represent 1. A binary symbol alphabet represents the minimum possible size for a symbol set, which is attractive from an Occam's razor perspective. It can be argued that this makes the scheme more 'deducible' in the sense that it would be more likely to be considered as a candidate signalling scheme by a target civilisation developing its SETI strategy.

A simple example of an antipodal signal set is Binary Phase Shift Keying (BPSK), which can be described as signalling with the alphabet [1,-1]. The BPSK signal set is illustrated in Figure 2-6.

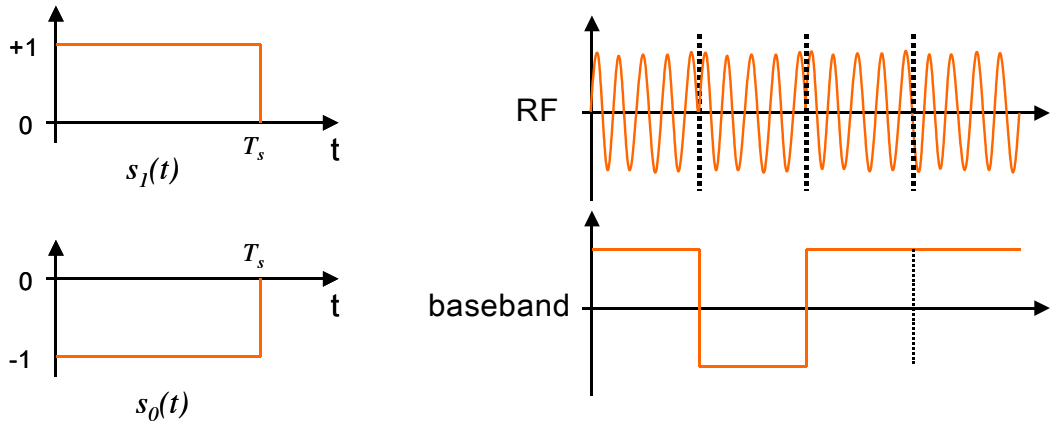

Figure 2-6: The BPSK signal set, and example waveforms at RF and baseband.

However, an antipodal signal set can potentially be much more complex than BPSK, with symbols having higher dimensionality. A higher dimensional symbol can be obtained, for example, by transmitting multiple 'chips' during the symbol interval T_s, where each chip is a single signalling unit in the two-dimensional complex plane, i.e. a single phase/amplitude over the smallest modulation interval; the chip interval T_c. The individual chips making up a symbol may vary in amplitude, phase or even width. Irrespective of the complexity of the symbol set, the antipodal constraint means that the waveform representing one alphabet member is precisely the negative of the other member at every point in the complex waveform representation (i.e. a point-wise 180° complex rotation of the first waveform). An example is shown in Figure 2-7; in this case a form of spread-spectrum BPSK where the waveforms have been modelled on segments of a pseudo-random binary sequence.

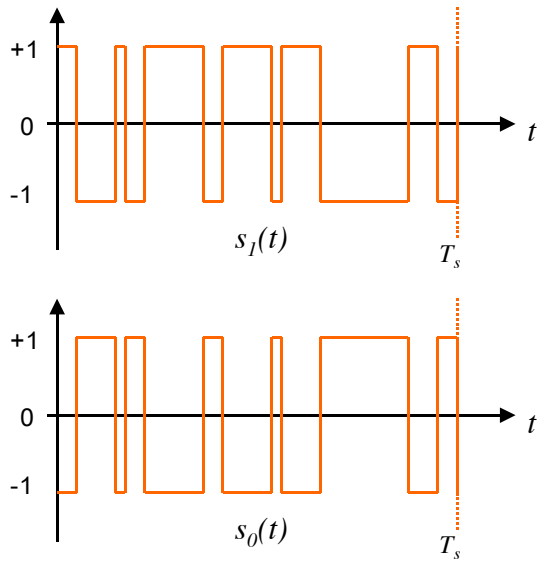

Figure 2-7: An illustrative spread-spectrum BPSK signal set – an example of a high-dimensionality binary antipodal alphabet.

A key feature of an antipodal signal set is that, regardless of the specific waveform shape for one alphabet member, the correlation of any one symbol with any other will always give either 1 (if they are the same symbol) or -1 if they are different symbols. This also means that when

successive symbols are transmitted on a channel, the correlation between adjacent symbols at the receiver will be 1 or -1 (following normalisation and ignoring noise and phase rotations for the present). This immediately suggests an autocorrelation process on the received signal could be used to reveal the presence of the modulation.

If we consider the autocorrelation of a sequence of randomly selected antipodal symbols, each of width T_s, then at delay T_s we see interesting behaviour. Over time the autocorrelation score averages to zero. This is because, on average, 50% of adjacent symbols are fully correlated and the other 50% are fully uncorrelated (i.e. produce a negative correlation score). On the face of it, this suggests autocorrelation would not be an effective method for detecting antipodal signals in noise. However, the autocorrelation behaviour of antipodal signal sets can be turned to an advantage when one recognises that every adjacent symbol pair produces a maximum *magnitude* correlation score, albeit a mix of positive and negative values. If the *absolute value* of each symbol-pair's correlation score is accumulated over the symbol sequence, then it can be seen that in fact antipodal signalling maximises the autocorrelation peak produced by the signal at delay T_s. This is the motivation behind the SWAC algorithm presented in Section 2.11.

Since the alphabet for antipodal modulation is binary, it conveys one bit per symbol. The overall BEF will be determined by a combination of the coding mechanism (specifically the ratio between redundant to non-redundant symbols) and the spreading factor (the number of chips per symbol) in the case of a spread-spectrum symbol waveform.

An attractive feature of the binary antipodal approach is that it can have a constant power envelope, that is, a peak-to-average power ratio of one. This would provide significant implementation advantages to an interstellar beacon transmitter, as it removes the need to

switch or modulate the very high power level being driven to the antenna system – in contrast to sparse modulation schemes such as that described in Section 2.8.1.

2.9 Data rates for interstellar communications

It has been typical of those contemplating the design of interstellar beacons to assume that such transmitters will operate at very low data rates. Recall the ICH concept introduced by Messerschmitt, defined by the ISM's coherence time and coherence bandwidth [21]. He makes the argument that there is nothing to gain from transmitting faster than one symbol per ICH, because it doesn't matter if a message takes years or centuries in the context of propagation times of hundreds or thousands of years, and it minimises the power. Fridman [36] takes this logic to extremes in suggesting data rates as low as 10^{-2} bit/s.

A counter-argument can be made. The transmitter power levels involved may not be impractically high when operating near the Ultimate Shannon Limit and when using a highly directional transmit antenna. What matters for overall system energy efficiency is the total energy required to communicate a message of a given length. This depends on the energy per bit and not the bit rate. So long as the energy per bit is kept low, there is no cost motivation to limit the data rate. If the length of the message is finite, it is equally appropriate to consider either (1) a short, fast transmission, or (2) a long, slow transmission. The latter approach will always require the lowest peak power, but the former approach may offer some advantages:

- **Communications efficiency.** Consider an interstellar beacon transmitter that employs antenna beams that cycle around multiple targets. If the data rate is high, even a short dwell time on each individual target can support communication of a large block of data. This may have benefits in the demodulation and decoding of message blocks. For optimum energy efficiency, code blocks need to be long (ideally many thousands of bits). Also, with many symbols in a contiguous sequence, it makes it possible to

accurately estimate the carrier phase of the signal, which allows the use of coherent demodulation methods that provide a performance gain over incoherent methods of typically ~3 dB.

- **Easier Discovery.** In any given observation interval at a target receiver, higher symbol rates will correspond to there being a larger number of contiguous symbols received during that observation interval. During discovery, it is likely that some form of detection metric must be accumulated during the observation, and each symbol contributes to the accumulation. Hence, higher symbol rates will translate directly to improved discovery S/N for any given length of observation.

- **Message Decoding Time.** Higher data rates mean faster delivery of messages. Long messages can be decoded within a 'project lifetime' rather than relying on multi-generational listening before anything interesting is received. It also allows more frequent updating of message content – potentially real-time 'streamed' content in the extreme.

Notwithstanding the arguments presented above, if we assume that multiple beacon transmitters exist, it is reasonable to suppose that across the totality of these systems there would be a range of different data rates in operation, covering the spectrum from low rates to high rates. Hence there is an argument for SETI to employ search strategies that accommodate a wide range of data rates.

2.10 Detection methods for wideband SETI

This section describes and compares a number of existing techniques that may be used to detect wideband signals of unknown form, and introduces a new method – symbol-wise autocorrelation (SWAC) – that aims to address the weaknesses of existing methods. SWAC is

then described in further detail in Section 2.11, including analysis of its sensitivity performance.

2.10.1 Matched filtering

Matched filtering is the optimal detection process for any signal type, and involves correlating the received signal with a clean reference version of the expected waveform. Figure 2-8 illustrates a matched filter detector for the case of a binary signalling alphabet [A,B]. The detector consists of two correlators: one correlating the input $r(t)$ against the waveform for symbol A, $s_A(t)$, producing output C_A; the other correlating $r(t)$ against the waveform for symbol B, $s_B(t)$, producing output C_B. The larger of C_A and C_B is used to decide which of $s_A(t)$ or $s_B(t)$ was assumed to have been transmitted in that symbol interval.

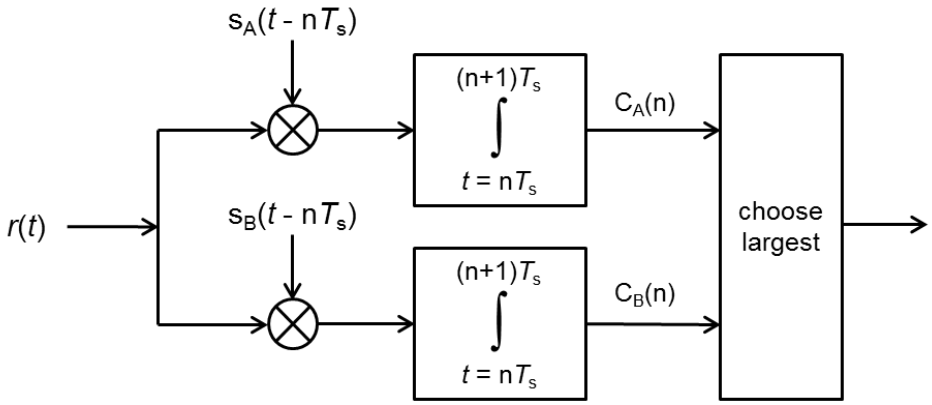

Figure 2-8: Matched filter detector structure for the binary symbol alphabet [A,B].

Fundamentally, matched filtering requires knowledge of the waveforms associated with the symbol alphabet. Narrowband SETI assumes one specific waveform (a sinusoid) and therefore matched filtering is straightforward for this case. It can be accomplished using a narrow band-pass filter tuned to the frequency of the sinusoid, or using an FFT with fine frequency resolution to simultaneously detect many frequencies. For wideband SETI we do not know the precise waveform(s) we are looking for, so matched filtering is only possible through a trial-

and-error approach on a series of assumed waveform types, as discussed in Sections 2.6 and 2.7. This approach clearly suffers from the limitation of a finite set of candidate waveforms. The parameter space for wideband signals is so vast that it is impossible to have a high confidence that any candidate list will contain the precise waveform actually transmitted by the beacon. Notwithstanding, if the candidate list did happen to contain the transmitted waveform(s), the maximum detection sensitivity would be achieved.

As an example, consider binary antipodal modulation with symbol alphabet [A, -A], as described in Section 2.8.2. The binary matched filter of Figure 2-8 can be simplified to a single correlator against A, since the second correlator will always produce the negative of the first correlator's output, i.e. $C_B = -C_A$. Assume that the symbol boundaries are known and the phase of the carrier signal has been resolved. Ignoring for the moment noise and other channel impairments, at each symbol the single correlator will return a normalised metric of C = +1 (if the input symbol was A) or C = -1 (if the input symbol was -A). This is a real-valued result, so we need only be concerned with the real component of the complex C value. Taking only the real component of C has the benefit of removing the imaginary component of the complex noise in the channel; a 3 dB reduction in noise power. The decision process involves measuring C and testing its sign to decide which of the two symbol values had been transmitted. When the input signal is noisy and the noise is complex additive white Gaussian noise of density N_0, the output S/N for the matched filter detector on individual symbols is given by the following well-known result [32]:

$$\boldsymbol{S/N_{MF} = 2.\frac{E_s}{N_0}} \qquad (4)$$

where E_s is the symbol energy (equal to the input signal power multiplied by the symbol period T_s). Note that here we are using the 'power definition' for S/N, i.e. signal power divided by the noise variance, which is common practice in the communications engineering discipline. In

astronomy, S/N is usually expressed in terms of the signal amplitude divided by the noise standard-deviation, which is simply the square-root of the power S/N.

A derivation for Equation (4) is provided in Appendix C. Note that, for a given noise density, S/N_{MF} depends only on the *energy* of each symbol. All waveform shapes perform equivalently with matched filtering.

The output S/N of a detector (S/N_{MF} in this case) can be used directly to establish the detector's performance in terms of its *miss probability* (the likelihood of a non-detection when a signal is present) and *false alarm probability* (the likelihood of reporting a detection when no signal is present). This is explained further in Appendix C, including a quantitative example showing how to determine the minimum required output S/N to achieve given specified miss and false alarm probabilities.

Now consider a sequence of symbols received at low S/N over which we wish to accumulate a **discovery** metric. If the values of each transmitted symbol are known by the receiver (or if the S/N is high, in which case the detector's decision on the transmitted symbol value will usually be correct), then it is possible to obtain a suitable decision metric for accumulation by appropriately sign-adjusting each C output according to the correct symbol value. (Without noise, this produces a positive metric of value |C| for every symbol, but with noise present, some values may be negative.) Accumulating this metric over *M* symbols will result in what may be termed the **"data aided"** (DA) matched filter output S/N, given by:

$$\boldsymbol{S/N_{MF,DA} = 2M.\frac{E_s}{N_0}} \qquad (5)$$

$S/N_{MF,DA}$ is seen to scale linearly[15] with M, meaning any target output S/N can be achieved given sufficient observing time.

However, it is important to note that this result **does not hold** for the SETI discovery scenario where one wishes to accumulate a metric over a large number of low-S/N symbols. In that scenario we do not know the transmitted symbol values, which we assume will resemble a random sequence. Furthermore, we assume the input S/N is so low that individual decisions by the matched filter are highly prone to error. At very low S/N, this error probability will tend to 0.5 and there is no way to reliably sign-adjust C values. If C values are not sign-adjusted, we can expect the equally positive and negative values to integrate towards zero over a long symbol sequence. To remove the effect of the random modulation of symbols and obtain a metric that accumulates, it is necessary to take the modulus (or square) of C for each symbol prior to accumulation. At first glance, accumulating |C| would appear no different to the sign-adjusting method that would be applied if the transmitted symbol values were known. However, it is not the same because here we perform the modulus operation on C (forcing all values to be positive), whereas, when the symbol values are known, the C values are multiplied by +1 or -1 depending on the corresponding correct symbol value. The impact of this additional modulus operation on the detector output S/N is significant. The sensitivity performance of this **"data blind"** method of matched filtering for SETI discovery purposes has not previously been reported. It has been quantified analytically in Appendix C, alongside the sensitivity analysis for energy detection and SWAC.

[15] If the output S/N was defined in terms of amplitude and standard deviation, it would scale with $\sqrt{M}$; the more familiar relationship to astronomers.

We define two variants of data-blind (DB) matched filtering, based on whether the modulation is removed by accumulating ABS(C) or SQR(C). The sensitivity of both variants is formulated in Appendix C, and the output S/N for each case are reproduced in Equations (6) and (7) below.

$$\boldsymbol{S/N_{\mathrm{MF,DB,ABS}}} = \frac{\boldsymbol{M}\left({}_1\boldsymbol{F}_1\left(-\frac{1}{2},\frac{1}{2},-\left(\frac{\boldsymbol{E_s}}{\boldsymbol{N_0}}\right)\right)-\boldsymbol{1}\right)^{\boldsymbol{2}}}{\left(\frac{\boldsymbol{\pi}}{\boldsymbol{2}}-\boldsymbol{1}\right)} \qquad (6)$$

where ${}_1F_1$ is the confluent hypergeometric function [37].

$$\boldsymbol{S/N_{\mathrm{MF,DB,SQR}}} = \boldsymbol{2M}.\left(\frac{\boldsymbol{E_s}}{\boldsymbol{N_0}}\right)^{\boldsymbol{2}} \qquad (7)$$

Note that these results apply specifically to binary antipodal modulation. This represents the best-case scenario for generating a cumulative discovery metric, and therefore these results represent an upper bound on the achievable sensitivity with any form of modulation.

Figure 2-9 compares the output S/N for the data-aided case ($S/N_{\mathrm{MF,DA}}$) and both data-blind cases ($S/N_{\mathrm{MF,DB,ABS}}$ and $S/N_{\mathrm{MF,DB,SQR}}$). It is seen that marginally better performance is obtained using the SQR data-blind variant across the entire input S/N range, so this is selected for use as the benchmark sensitivity curve in subsequent plots.

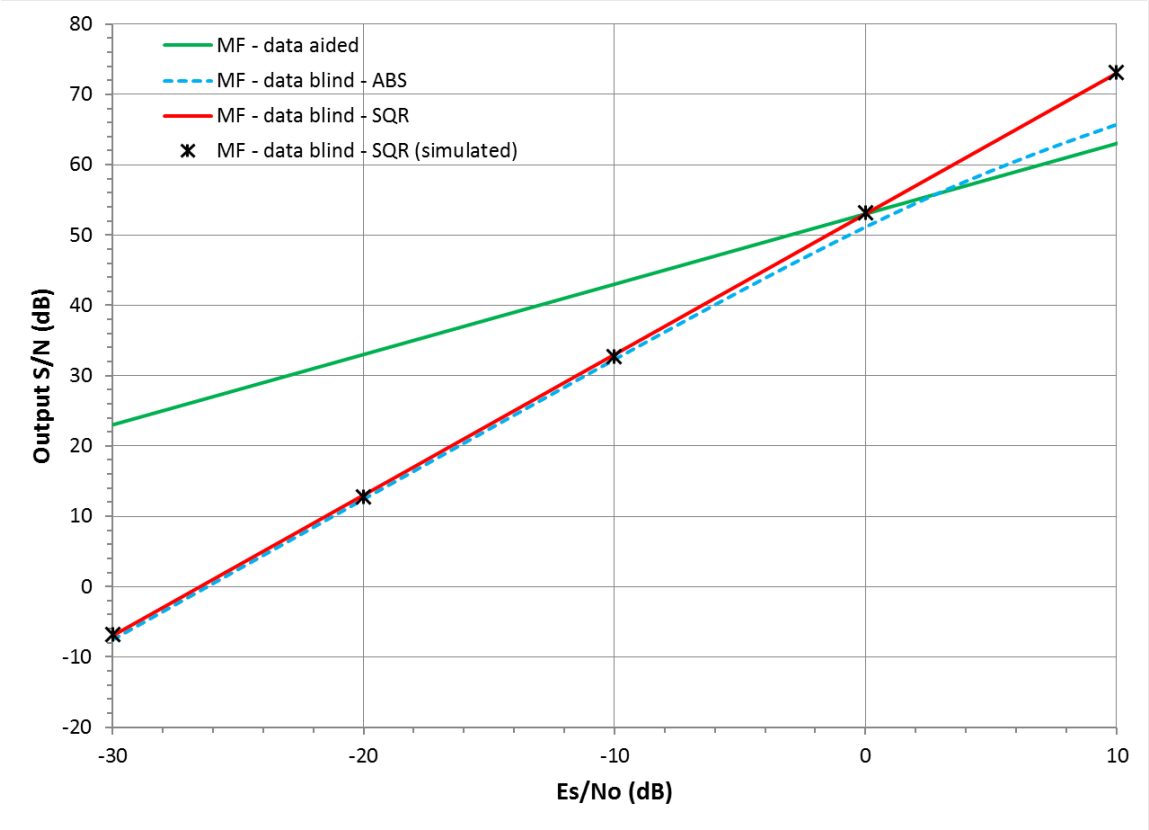

Figure 2-9: Discovery mode matched filter output S/N as a function of input S/N (expressed as E_s/N_0) for data-aided and data-blind matched filters, for an example scenario where the observation interval is M = 100,000 symbols. Simulation results for the SQR data-blind detector are also plotted. All the results here are for binary antipodal modulation, which represents the upper bound on the sensitivity performance achievable with any modulation method. *[Note that data-aided matched filtering is not available in the SETI context and the result is included here for comparison purposes only.]*

To validate the analytical results, the performance of the SQR detector variant was simulated in Matlab, processing a 100 s long test vector of binary antipodal spread-spectrum modulation immersed in additive white Gaussian noise to a specifiable E_s/N_0. The symbol rate was 1000 symbols/s, so the total number of symbols processed was M = 100,000. The spread-spectrum spreading factor was 64, and a random chip pattern was used, taken as the first 64 binary digits of π. Oversampling of 4 samples per chip resulted in a waveform with WT_s = 256 samples per symbol for this example. The "MF.DB.SQR" algorithm was run at each of five

different input E_s/N_0 values, with and without the signal component present at each noise level. The output S/N for each case was then determined according to Equation (21) in Appendix C. As seen in Figure 2-9, there is excellent agreement between the simulated and analytical results, confirming the correctness of the formulation of Equation (7).

Two features of the MF.DB.SQR result to note: (1) the output S/N (and hence sensitivity) at low input S/N is significantly less than what is normally assumed for matched filtering, and (2) the sensitivity at input E_s/N_0 values above 0 dB is actually superior to the data-aided sensitivity – although this is only of academic interest since for discovery we are concerned only with the low S/N regime.

To reiterate, the data-aided case is not available for SETI discovery mode, and has only been considered here for comparison purposes. Indeed, the data-blind forms of matched filtering are also unattractive for SETI because they still require precise knowledge of the modulation parameters, or numerous trial-and-error guesses at the actual symbol alphabet. However, the MF.DB.SQR data-blind result represents the best-case sensitivity performance possible for any detector in SETI discovery mode, and therefore it is the appropriate benchmark against which alternative detection methods should be judged.

2.10.2 Power spectral density

Perhaps the most obvious method for blind detection of wideband signals is to examine the power spectral density (PSD) of the band of interest, in the same way as is employed for the detection of narrowband signals. The PSD shows the distribution in the frequency domain of the measured power in the band, i.e. [signal+noise]. If [signal] is significantly more powerful than [noise], the spectrum of the modulated carrier will be clearly visible above the background noise level in the PSD. The more typical assumption for wideband SETI is that the target signal's power density is low and that its spectrum will not appear clearly above the

noise level, as depicted in Figure 2-2. However, even if [signal] is of significantly lower power density than [noise], then analysis of a sufficiently long observation interval will provide a high degree of noise averaging. In that case, if the processed bandwidth fully encompasses the signal bandwidth, the transition between regions of [noise] and [signal+noise] may be discernible. However, as discussed in Section 2.6, in the very low S/N scenario the veracity of this approach is impacted by instrumental and natural signal level variations during the observation interval, which may exceed the variation in power density one wishes to detect.

Note that if the measurement bandwidth is less than the signal bandwidth, there will be no frequency domain transitions in the PSD to detect, i.e. there will be no "null reference" provided by regions of background noise only. It may be possible to obtain such a null reference if there are time-domain variations in signal power – which we discuss below.

2.10.3 Energy detection

Similar to the PSD approach, energy detection involves integrating the [signal+noise] power over a given bandwidth and timespan to obtain a low-variance measure of the total energy in the entire observation band. This is simpler than computing the PSD, and provides the same quality of detection metric in cases where there is no structure observable in the PSD.

Energy detection is directly impacted by gain variations in the receiver signal chain, and by ISM scintillation, so suffers from the same issue for the low S/N scenario as examining the long-term average PSD. If the target signal is persistent in time, this approach suffers from a similar null reference issue to that described above for PSDs. In this case it is a time-domain issue as opposed to a frequency-domain issue. However, if the target signal happens to be transient in nature, it may be possible to extract a workable null reference by measuring the energy when the signal is not present. Unfortunately, for very low S/N, the transitions between [signal+noise] and [noise] may not be easily discernible. Nevertheless, if we assume a reliable

null reference can be established, measuring total power over long observation intervals may allow a sufficiently low measurement variance for a reliable detection.

For wideband emissions resembling incoherent noise (such as when many disparate radio sources combine incoherently), energy detection may be one of the few approaches that is feasible. However, for detecting a single coherent intentional beacon signal, energy detection will generally provide sub-optimal detection sensitivity compared to other methods. The sensitivity of energy detection is quantified analytically in Appendix C, alongside the analysis for matched filtering and SWAC. It is shown that for continuous cyclostationary signals, significantly higher sensitivity can be achieved with an appropriately designed autocorrelation detector like SWAC.

The output S/N for an energy detector used for SETI discovery is:

$$\boldsymbol{S/N_{ED}} = \frac{\boldsymbol{M}}{\boldsymbol{2WT_s}} \cdot \left(\frac{\boldsymbol{E_s}}{\boldsymbol{N_0}}\right)^{\boldsymbol{2}} \quad (8)$$

S/N_{ED} is maximised when the number of samples per symbol, WT_s = 1, i.e. when the measurement bandwidth is equal to the symbol rate. Reducing W further such that $WT_s < 1$ will actually decrease S/N_{ED} because it reduces the amount of signal energy detected, and S/N_{ED} is proportional the square of signal energy. If WT_s is larger than 1, such as will be the case with spread-spectrum modulation or when the signal is oversampled in relation to the symbol rate, the sensitivity is also degraded. In a practical SETI scenario, the signal's symbol rate will be unknown, so an energy detector will normally be operating sub-optimally. Trialling a range of measurement bandwidths is recommended when using energy detection for SETI discovery.

Figure 2-10 plots the output S/N for "data blind" matched filtering along with four energy detection scenarios: the non-spread case where *W* equals the symbol rate, and three spread-spectrum cases of *W* = 16, 256 or 4096 times the symbol rate.

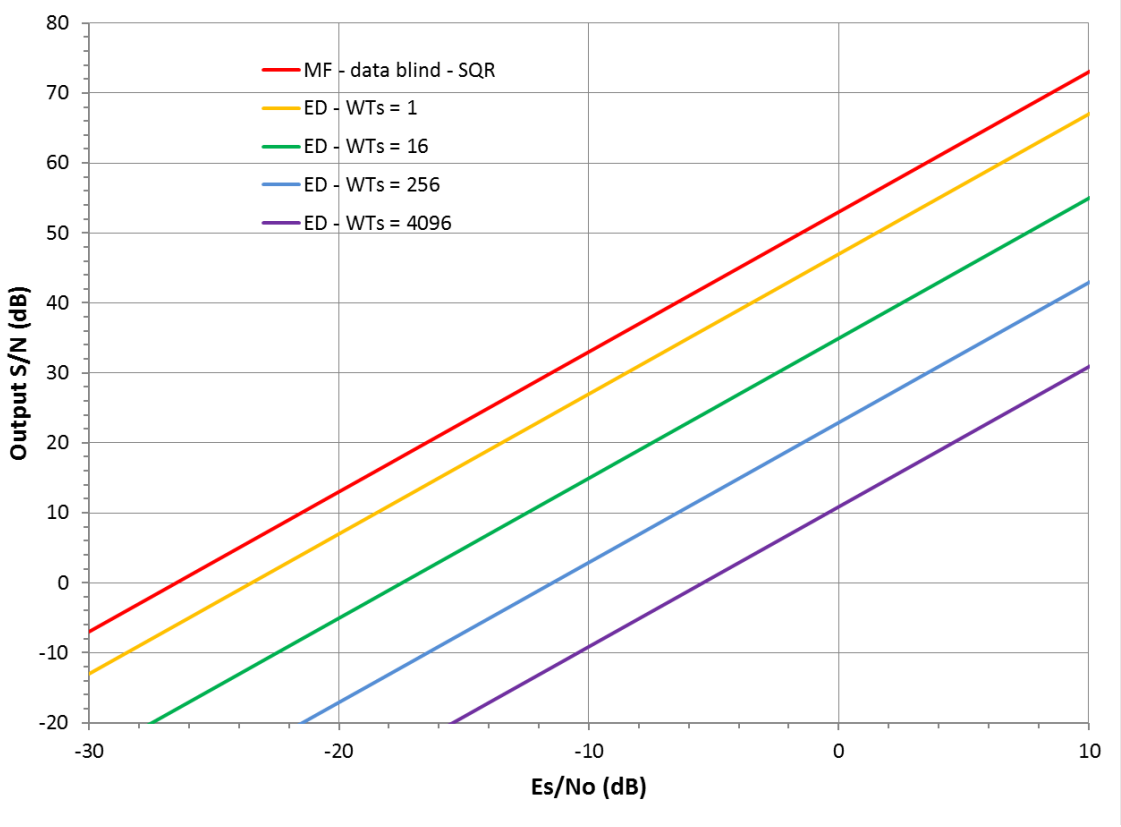

Figure 2-10: Energy detection output S/N as a function of input S/N (expressed as E_s/N_0) for various values of WT_s (the ratio of measurement bandwidth to symbol rate). Also shown is the result for data-blind matched filtering (SQR variant) from Figure 2-9, which represents the upper bound on achievable sensitivity in SETI discovery mode. In all cases the observation interval is M = 100,000 symbols.

Figure 2-10 clearly illustrates the dependency of energy detection sensitivity on measurement bandwidth. In the very best-case where $WT_s = 1$, S/N_{ED} is a constant 6 dB worse than matched filtering. For larger WT_s values the relative performance gets progressively worse. Energy detection is particularly poor for spread-spectrum signals because there is a high noise bandwidth but no de-spreading gain (as would be provided by a matched filter) to overcome it.

This weakness was a prime motivation for the development of the SWAC algorithm described in Section 2.11.

Despite the poor sensitivity of energy detection for wideband signals, there may be some scenarios where it is the best approach available. It has very low specificity and can be utilised with any type of modulation, with the measurement bandwidth being the only parameter over which optimisation should be performed. Equation (8) shows that it is possible to overcome the poor sensitivity by means of accumulating over extended observation times (notwithstanding the null reference issue). S/N_{ED} scales in proportion to M, the number of symbols processed.

2.10.4 Statistical properties

More typically used in radio astronomy for detecting RFI, measurement of the statistical properties of a received signal can be used to distinguish signal from noise. Skewness and kurtosis of the received signal amplitude distribution are examples of statistics that are commonly used for this purpose [38]. Radio astronomical noise is very close to having a Gaussian amplitude distribution, with zero mean and power equal to the variance of the distribution. It is also very white across the measurement bandwidth. Any received signal that does not exhibit these characteristics may contain non-noise components, such as RFI or the desired signal.

However, a difficulty arises for wideband SETI at low S/N. Here the received signal is highly dominated by the background noise, so the measured statistics will closely match those of pure Gaussian noise. Unless the desired signal component is strong, the deviation from pure noise statistics may not be able to be disentangled from instrumental imperfections – in which case the statistical measure cannot be relied upon for reliable detection of a signal.

Furthermore, some signal types that might be employed for interstellar communications may themselves exhibit a white Gaussian amplitude distribution, and hence their presence would not alter the statistical measures observed with noise alone.

Nevertheless, the simplicity of gathering signal statistics on raw received telescope signals works in its favour. In some telescope back-ends, these statistics are already routinely computed for RFI mitigation purposes. Averaging of these statistics over long observation intervals may, in some circumstances, provide a useful metric for wideband SETI detection. The strength of the approach is that it makes no assumptions about the precise structure of the target signal, only that it has different statistics to astronomical noise. Its weakness is that it will be insensitive to signals that resemble noise. Also, it is not clear whether it can provide greater sensitivity than energy detection – this deserves future investigation. However, it has a major advantage over energy detection in that it comes with its own 'built-in' null reference, i.e. the default statistics expected for pure noise.

2.10.5 Cyclic spectral analysis

The detection of a wideband signal is made more difficult if its PSD is relatively flat across the signal bandwidth and contains no discrete spectral lines. This is generally the case when carriers are modulated using power-efficient modulation schemes, to avoid wasting energy on signal components that do not carry information. However, modulated carrier signals generally possess a statistical characteristic known as 'cyclostationarity'; a result of structural periodicity that gives rise to statistical properties that vary cyclically with time (see Section 2.10.6). In his seminal work on cyclostationarity, Gardner [39] points out that signals not having discrete spectral lines in their PSD may possess a second-order periodicity, which means that if a nonlinear process is applied to the signal, discrete spectral lines will be regenerated. In a generalisation of Fourier spectral analysis for periodic signals, Gardner has developed the concept of the 'cyclic spectrum' of a cyclostationary signal [40]. This method

produces a two-dimensional spectrum (in the axes of frequency and delay) that preserves the phase information of a signal. It can be used to obtain discrete spectral features that are not evident in a Fourier-generated spectrum. These features may, under some circumstances, be more easily distinguished from the noise than with a traditional PSD.

Cyclic spectral analysis does not require knowledge of the target waveform type; only that it possesses cyclostationarity. This low specificity makes it an attractive blind detector for wideband SETI. However, with low specificity generally comes a lower detection sensitivity. The SWAC algorithm described in Section 2.11 is also based on detection of cyclostationarity, and is conceptually simpler than cyclic spectral analysis. It is similar in the way that time-segments of the signal waveform are repeatedly 'folded' on themselves as part of the process[16]. However, SWAC generates discoverable features in a more direct way than cyclic spectral analysis and, as will be shown, can approach matched filter performance. This may lessen the motivation for utilising the more complex approach of cyclic spectral analysis when it cannot offer improved sensitivity. Nevertheless, while it is outside the scope of this thesis, a detailed analysis of the sensitivity of cyclic spectral analysis for SETI discovery deserves investigation.

2.10.6 Autocorrelation

Autocorrelation can be defined as the correlation of a waveform with a delayed version of itself. By computing the degree of correlation over a range of delay values, one can generate an 'autocorrelation spectrum' that essentially shows how 'self-similar' a waveform is over

[16] There is a similarity here also with the folding process applied to the analysis of pulsar emissions [43], although that method is concerned with analysing the power envelope and is typically unconcerned with instantaneous phase information in the signal. It that sense, the folding of pulsar signals may be thought of as 'incoherent folding' whereas SWAC and cyclic spectral analysis are examples of 'coherent folding' of the voltage signal.

time. If there exist any repeating patterns in the waveform, the autocorrelation spectrum will display peaks at the delays corresponding to the time separations between the repeated elements. As an example, if a waveform happened to consist of a contiguous sequence of duplicate waveform segments of length T_w, its autocorrelation would display a peak at delay T_w (and integer multiples of T_w).

While it is easy to see how a repetitive redundant signal can be detected with autocorrelation, it is less obvious that a recognisable autocorrelation signature can also be exhibited by signals where there is no redundant repetition. Autocorrelation can also reveal the presence of signals that contain some form of periodicity in their *structure*, with no requirement for repetition of the *content*[17] of the signal. As mentioned in Section 2.10.5, signals with structural periodicity exhibit cyclostationarity. This class encompasses virtually all digital modulation methods used in terrestrial communications systems. Any modulation approach that involves sending a sequence of symbols with a common symbol period T_s and chosen from a finite symbol set (alphabet) will display some degree of cyclostationarity. Even when specific symbol values in the sequence are selected randomly, over a sufficient length of time the finite alphabet ensures cyclostationarity. Specifically there exists periodicity in time T_s and so the autocorrelation of such a signal will display peaks at delay T_s and its multiples. However, the strength of the autocorrelation will depend on the size of the symbol alphabet and the distribution of symbol values in the waveform sample being analysed.

The potential to apply autocorrelation methods in SETI was recognised as early as 1965 by Drake [41]. More recently Harp et al. [42] have discussed a signalling method that can be

[17] Here the term 'content' is referring to the information that selects the specific sequence of symbols that makes up the transmitted waveform, i.e. the information that controls the modulation of the signal.

effectively detected by means of autocorrelation. Both of these examples consider a scenario where more than one signal is superimposed in either time or frequency, with autocorrelation used to detect the presence of *repetition*.

As explained in Section 2.8.2, conventional autocorrelation has a fundamental limitation in its applicability to wideband SETI, which arises because we assume the target signal is modulated with apparently random data. This may cause the autocorrelation metric to integrate to zero over a long observation interval with certain types of modulation, such as the constant-envelope binary antipodal modulation approach described in Section 2.8.2. In this case, conventional autocorrelation will only succeed if there is repetition of symbol *values* on a fixed time schedule, such as would occur if there was a regularly inserted 'frame synchronisation' pattern.

Another scenario where conventional autocorrelation can serve as an effective detector is when the signal exhibits envelope fluctuations. A pulsar emission is a good example of a signal that exhibits cyclostationarity through variations in its envelope related to its rotation cycle [43]. Autocorrelation is one method that can be used to discover pulsars – but this is accomplished using *incoherent* autocorrelation involving the 'Stokes I' power envelope rather than the complex voltage signal.

With communications signals, the envelope of the signal may vary, either because the alphabet symbols have different amplitudes, or due to slewing between different symbol values when the signal is bandlimited. Even with a randomly modulated signal, such envelope fluctuations will occur on a regular timescale, i.e. the symbol rate and its multiples. So here again, the autocorrelation spectrum of the power envelope can be expected to contain discrete components.

The SWAC algorithm, described in detail in Section 2.11, is also autocorrelation-based but it differs from conventional autocorrelation by taking account of assumed symbol boundaries in a modulated signal – hence the name *symbol-wise* autocorrelation. SWAC operates on the complex voltage signal and can extract a detection metric regardless of whether the power envelope is variable or constant. Also, detection is not conditional on the signal containing explicit repetition, which is attractive in a SETI context because it allows the possibility of signal discovery from any captured segment of an extraterrestrial signal without there needing to be a repeat of any content within the captured segment.

It is difficult to compare the detection sensitivity of autocorrelation with other methods such as matched filtering or energy detection because the behaviour of an autocorrelation detector is highly waveform-dependent. For many randomly modulated signals, conventional autocorrelation over long observation intervals is likely to yield worse sensitivity than SWAC or even energy detection over the same interval.

2.10.7 Karhunen-Loève Transform

The Karhunen-Loève Transform (KLT)[18] is an algorithm capable of detecting the presence of signals of arbitrary unknown structure that are embedded within noise. It performs an orthogonal linear transformation of [signal+noise], using an eigenvalue/eigenfunction computation to determine the optimal transformation axes for bringing the signal component 'into view'. The KLT is very computationally expensive but has moved from a theoretical curiosity to a potentially practical signal processing tool in recent years as real-time computing capabilities have increased. A good description of the KLT and the recent advances in its

[18] In other branches of science the KLT algorithm is variously known as 'Principal Component Analysis', the 'Hotelling Transform', 'Proper Orthogonal Decomposition' or 'Empirical Orthogonal Function Analysis'.

implementation can be found in Maccone's text [44]. However, despite these developments there remain concerns about the ability to process radio telescope data in real time with the KLT to make an initial discovery. Perhaps more realistic in the near-term is that the KLT will prove to be a powerful tool to analyse candidate 'events' once they have initially been discovered by other means. The potential of the KLT for discovering wideband extraterrestrial signals is left to others to explore.

2.11 Symbol-wise autocorrelation

In Section 2.10.1 a data-blind form of matched filtering was presented that could be used for SETI discovery if the modulation parameters happened to be known (not possible) or guessed (unlikely). Despite the impracticality of matched filtering for SETI discovery, its sensitivity performance represents an upper bound that is useful for assessing alternative detection methods. We now present an algorithm that has the benefit of being blind to both the modulation parameters and the data sequence, and which is shown can approach the same level of detection sensitivity as matched filtering, making it highly suited to wideband SETI discovery.

2.11.1 The "basic SWAC" algorithm

Consider a signal $s(t)$ consisting of a sequence of contiguous symbols s_k, as depicted in Figure 2-11. We assume the s_k are chosen from a finite alphabet.

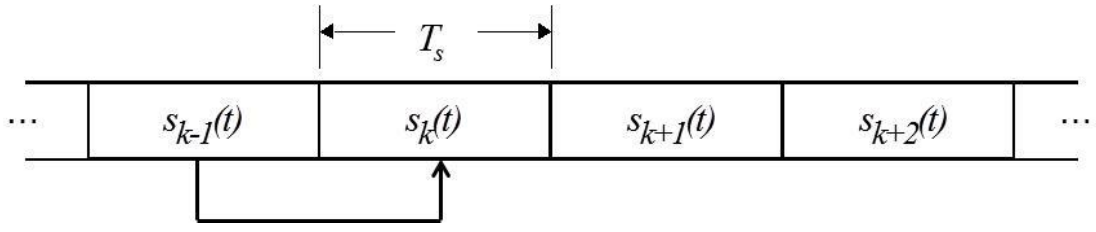

Figure 2-11: Inter-symbol correlation in a sequence of contiguous modulation symbols with symbol period T_s.

Consider now the autocorrelation of $s(t)$. At delay T_s the autocorrelation process is effectively measuring the degree of similarity between consecutive symbols. This is indicated in Figure 2-11 for the example of symbols $s_{k-1}(t)$ and $s_k(t)$. It is easy to see that the larger the symbol alphabet, the lower will be the average degree of similarity between consecutive symbols. Conversely, the average degree of similarity will be maximised with the smallest symbol alphabet: a binary alphabet. Even if the two possible symbol values are completely orthogonal (zero cross-correlation), for randomly selected symbol values there will be maximum correlation 50% of the time, when the adjacent symbols happen to be the same value. In general the cross-correlation between different members of a symbol set will *not* be zero, so this represents just one particular scenario. However, it illustrates how autocorrelation sensitivity is, in general, maximised with a binary alphabet.

The maximum magnitude of the autocorrelation of adjacent binary symbols will occur when the signal set is antipodal, i.e. the waveform representing a zero bit is the inverse of that representing a one bit (as described in Section 2.8.2). In this case the correlation of any one symbol with any other will always give either 1 (if they are the same symbol) or -1 if they are different symbols. This also means that when successive symbols are transmitted on a channel, the correlation between adjacent symbols at the receiver will be 1 or -1 (following normalisation and ignoring noise and phase rotations for the present). This immediately suggests an autocorrelation process on the received signal could be used to reveal the presence of the modulation.

As explained in Section 2.8.2, conventional autocorrelation of a sequence of randomly selected antipodal symbols will produce a score that averages to zero over time. However, every adjacent symbol pair produces a maximum *magnitude* correlation score, but a mixture of

positive and negative values. If one takes the *absolute value*[19] of each inter-symbol correlation score and accumulates this over the symbol sequence[20], then antipodal signalling maximises the autocorrelation peak produced by the signal at delay T_s.

As mentioned in Section 2.2.7, there are persuasive arguments in favour of utilising spread-spectrum modulation for interstellar signalling. We have already shown an example of such a signal set, for the antipodal case, in Figure 2-7. How does this type of signal set behave with autocorrelation? As explained above, any antipodal signal set will result in ±1 correlation scores between adjacent symbols, i.e. for an autocorrelation delay of T_s. What *does* change for the spread-spectrum case is the behaviour at other values of delay. Assuming the pseudo-noise sequence used for the spreading process has been selected appropriately, there will be a large reduction in the correlation score for sample delays that result in one or more chip intervals of time offset. This means that an autocorrelation spectrum for a spread-spectrum signal will display a sharper peak at T_s than the non-spread case. This helps to make the autocorrelation peak easier to distinguish amidst high levels of noise. In this way the use of spread-spectrum is highly beneficial for signal detection using autocorrelation methods.

The idea of using autocorrelation to detect unknown spread-spectrum signals is not new (e.g. [45]). However, previous approaches have not taken account of assumed symbol boundaries,

---

[19] Taking the absolute value is effectively 'stripping' the modulation of the signal. This can also be achieved by taking the square of the inter-symbol correlation scores.

[20] It is undesirable to take the absolute value of every *sample-wise* correlation because this will result in the noise energy adding incoherently. It is better to accumulate complex sample-wise correlation scores over the duration of a complete symbol – producing a *symbol-wise* correlation score – then take the absolute value before combining with the scores from other inter-symbol correlations. This will provide the maximum degree of noise averaging without compromising the signal component, thus maximising detection sensitivity.

which, as we have seen, can be exploited to better accommodate randomly modulated signals. We will also see in Section 2.11.5 that the symbol-wise approach allows very significant gains in detection sensitivity to be achieved.

The challenge in applying symbol-wise autocorrelations during SETI discovery is that, for a received waveform $y(t)$, we do not know the symbol boundaries (if indeed a modulated signal is present), nor the symbol period, T_s. Furthermore, we do not know the signal alphabet ([S, -S] in the antipodal case), nor the centre frequency of the modulated carrier.

However, if there was a signal component in $y(t)$ that happened to be modulated using an antipodal signal set, we can exploit the characteristics of antipodal signalling to reduce the search space dramatically. For a given segment of $y(t)$ we can perform a search over the symbol-period dimension without knowing the carrier frequency or signal alphabet. We denote as variable τ the trial symbol period values. We can make progressive calculations of the autocorrelation across a range of delays corresponding to the minimum symbol period τ_1 to maximum symbol period τ_2 under consideration in the search. For each trial τ we correlate an assumed sequence of noisy symbols $y(t)$ with a one-symbol-delayed version of $y(t)$ (i.e. $y(t+\tau)$), accumulating the absolute (or squared) value of each inter-symbol correlation score. If a signal is present, then at τ close to T_s, the autocorrelation score will peak at value D_{peak}. At other τ values the misalignment of symbol periods will produce a low average autocorrelation score[21]. A signal is deemed to be present if D_{peak} exceeds a specified threshold, which is set relative to the mean background (i.e. off-peak) autocorrelation score.

[21] This is important, as it provides a built-in null reference to aid detection. Those values of τ not related to the symbol period will generate an autocorrelation score, in the low S/N regime, that is almost

We call this algorithm **symbol-wise autocorrelation** (SWAC), which, in its discrete-time form, is expressed mathematically in equations (9), (10) and (11).

$$SWAC(\tau) = \sum_{n=1}^{M} \left| \sum_{k=k_0+(n-1)\tau}^{k_0+n\tau} \left(y_k \cdot \overline{y}_{k+\tau} \right) \right| \quad (9)$$

$$D = \max_{\tau \in [\tau_1, \tau_2], k_0 \in [0, \tau]} SWAC(\tau) \quad (10)$$

$$\hat{T}_s = \arg\max_{\tau \in [\tau_1, \tau_2], k_0 \in [0, \tau]} SWAC(\tau) \quad (11)$$

In Equation (9), y_k are the complex samples of waveform $y(t)$, M is the number of symbols processed, τ is the trial symbol period, and k_0 is the sample index corresponding to the first sample of each symbol. Equation (11) gives us the estimated symbol period of the signal embedded in $y(t)$, which will be useful for any subsequent processing to extract the information content of the signal.

It is worth emphasising that the search is over τ and k_0. One does not need to know the centre frequency, chip rate (bandwidth) or symbol alphabet (spreading codes).

A variation on Equation (9) that provides a worthwhile gain in detection sensitivity can be obtained by taking the absolute value of just the real component of each complex inter-symbol correlation score, as shown in Equation (12). This optimisation is only possible if the arbitrary

---

identical to that which is produced for noise only. This establishes a null reference, even when the signal is present. This is particularly valuable in cases where the signal is persistent in time.

phase shift between symbols in passband is successfully estimated and removed[22], which should be possible when a sufficient number of symbols are available to process (i.e. M > ~20).

$$SWAC(\tau) = \sum_{n=1}^{M} \left| \mathrm{Re}\left\{ \sum_{k=k_0+(n-1)\tau}^{k_0+n\tau} \left(y_k \cdot \overline{y}_{k+\tau} \right) \right\} \right| \qquad (12)$$

Note that SWAC can be used to detect any cyclostationary signal, but the algorithm achieves maximum sensitivity when the alphabet is of the binary antipodal form.

An expression for the detection sensitivity of the basic SWAC algorithm is derived in Appendix C for the binary antipodal case and assuming the optimised formulation of Equation (12). This represents the best-case scenario, which is useful to understand. It is also mathematically tractable, unlike cases where the cross-correlations between alphabet members are unknown. The degradation in sensitivity when detecting other modulation alphabets varies on a case-by-case basis and cannot easily be generalised. However, if we restrict our attention to binary spread-spectrum alphabets, then it can be shown that the sensitivity of SWAC will fall somewhere between the best-case figure and 6 dB below that, depending on the specifics of the alphabet[23].

---

[22] In the general case on a passband channel there will be an arbitrary number of cycles of the carrier during a symbol interval, hence the complex correlation between successive symbols will result in a complex score with the maximum magnitude and arbitrary unknown phase P (for like-valued symbols) or (P+180°) (for dissimilar symbols). Given a sufficient number of correlations in relation to the level of noise present, it should be possible to obtain a reasonably accurate estimate of P. If so, it can be removed by appropriately rotating each complex correlation score, placing all results on the Real axis (plus complex noise).

[23] The worst-case performance for a binary alphabet occurs when detecting binary orthogonal signalling, where the cross-correlation between the two alphabet members is zero, resulting in a loss of 6 dB in

2.11.2 Example

We illustrate the basic SWAC algorithm by way of an example. Assume an antipodal spread-spectrum BPSK waveform with a symbol rate of 2 symbol/s and a chip rate of 1000 chips per symbol. The PSD of a passband representation of this signal (centred at approximately 2.5 kHz) is shown in Figure 2-12.

Ignoring channel impairments, we now assume this signal is received embedded in noise at a low S/N such that the noise masks the presence of the signal in the PSD, as shown in Figure 2-13. Applying the SWAC algorithm to a 50 second burst of the received waveform generates the output of Figure 2-14, which shows a clear peak at the correct symbol period of 500 ms. The SWAC process has achieved this detection without knowledge of the centre frequency, symbol rate, chip rate, signal bandwidth, modulation method or symbol alphabet. It has performed a search over just one dimension: the symbol period.

---

sensitivity compared to the binary antipodal case. Alphabets having non-zero cross-correlation between members will experience less sensitivity loss.

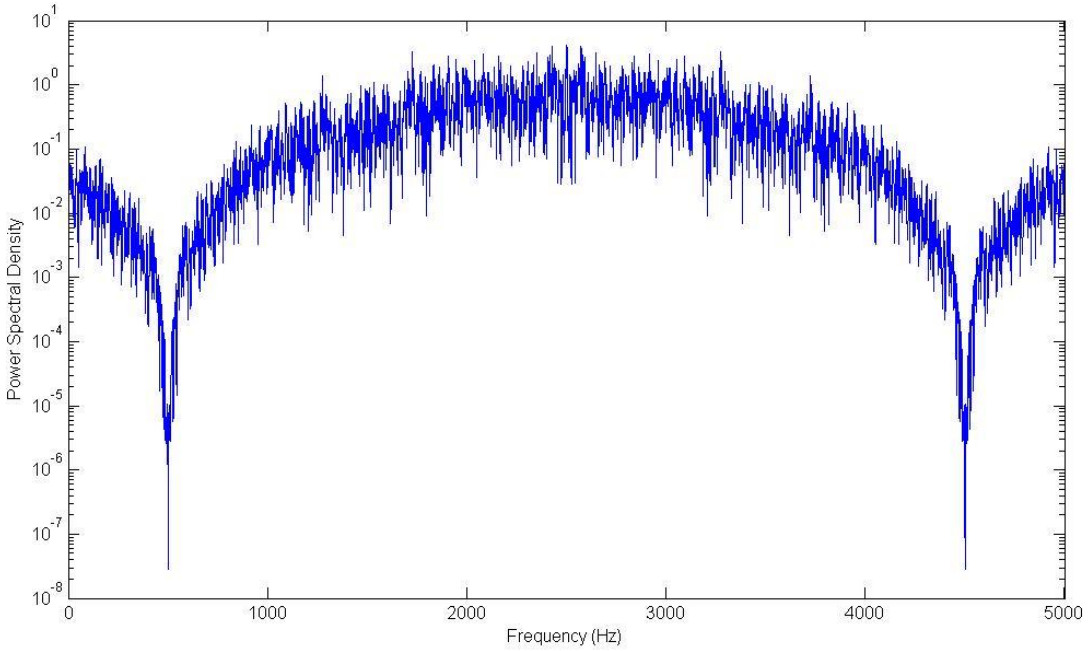

Figure 2-12: PSD for an illustrative antipodal spread-spectrum BPSK signal (2 symbol/s, 1000 chips/symbol, no noise).

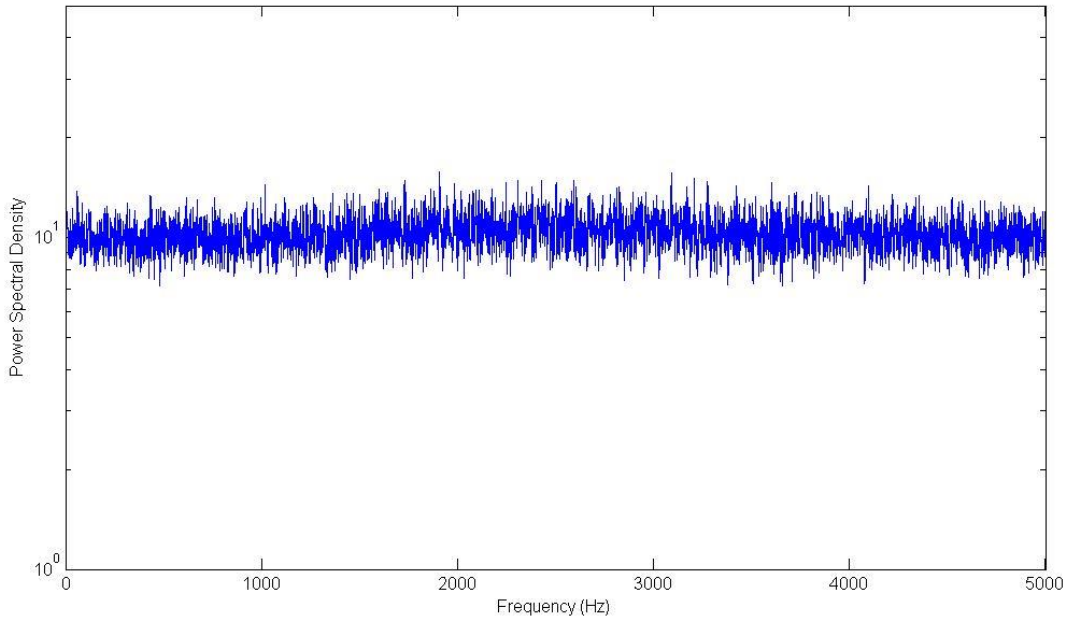

Figure 2-13: PSD for the illustrative signal of Figure 2-12 embedded in Gaussian noise (white across the measurement bandwidth).

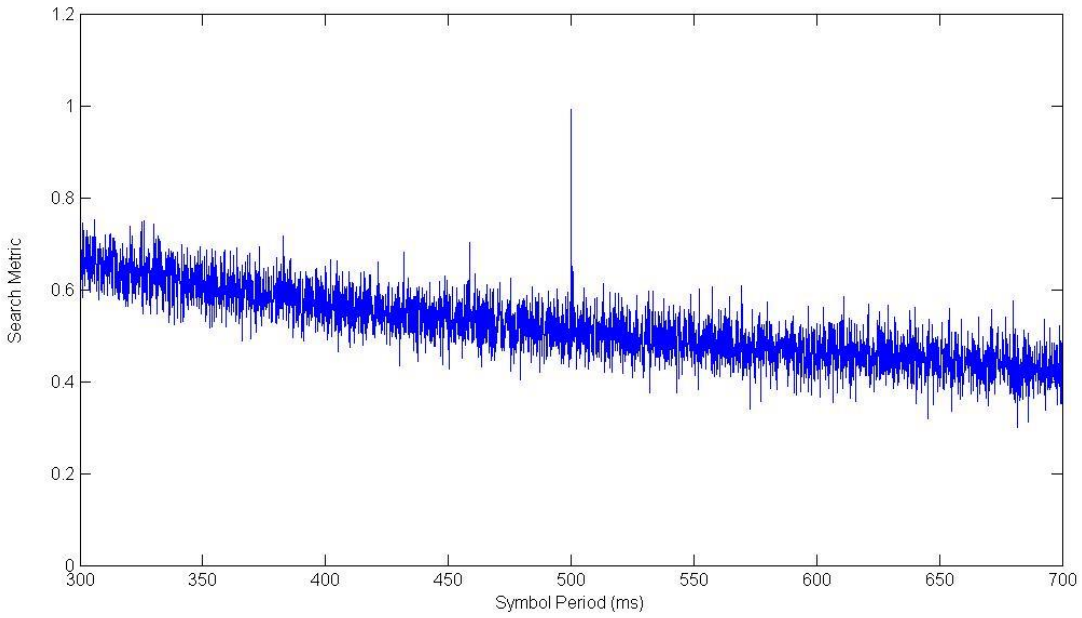

Figure 2-14: SWAC output as a function of assumed symbol period τ, with the waveform of Figure 2-13 as input.

2.11.3 Comments on detection sensitivity

In Appendix C a mathematical formulation for the detection sensitivity of SWAC is derived in terms of detector output S/N. It is worth noting here that the sensitivity increases proportionally with the time-span of signal processed. At a given symbol period this is the same as saying the sensitivity is proportional to M, the number of symbols processed[24].

The SWAC output plot of Figure 2-14 was obtained using the optimum value of k_0 in Equation (12). In addition to the search in the τ dimension, a search was also conducted over different k_0. It was found that the SWAC score is relatively insensitive to the k_0 assumption. This is seen in Figure 2-15, which is a pseudo-three-dimensional plot of the SWAC score as a function of

[24] Because we employ the power definition for S/N, the sensitivity increases proportionally with M rather than $\sqrt{M}$, as would be the case for an amplitude definition of S/N (as is typically used in astronomy).

both τ and k_0 (here shown as t_0; the starting time offset as a percentage of τ). Ten values of t_0 were tried at each trial τ, in steps of 10% of τ. There is an optimum value of t_0 (in this case 50%) but, regardless of the value of t_0 there is in all cases a distinct output peak at τ = 500 ms.

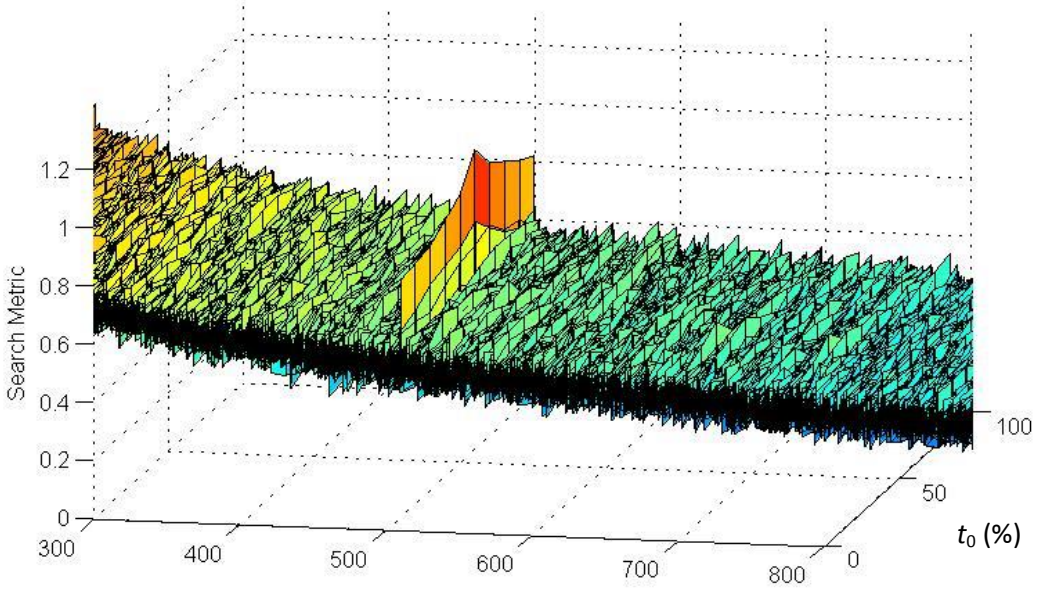

Figure 2-15: SWAC output as a function of assumed symbol period τ and symbol starting time offset t_0.

Since the SWAC score is relatively insensitive to incorrect k_0, rather than increase the computational complexity by a factor of 10 to search over k_0 it is actually more productive to set k_0 to zero and increase the length of data processed to compensate for the incorrect k_0 assumption, i.e. to overcome the reduced level of the SWAC peak at sub-optimal k_0. This is particularly useful when one realises that, in general, there will not be an exact integer multiple of waveform samples per symbol. Hence over a large M the assumed symbol boundaries will drift with respect to the actual boundaries, regardless of the initial choice for k_0. Processing a larger M will provide a higher detector output S/N and compensate for this effect. It has been found that a factor of two increase in M will overcome most of the loss due to incorrect k_0, with only a doubling of computational complexity.

Another aspect that affects detection sensitivity is the width of the autocorrelation peak. Spread-spectrum modulations produce a peak that is narrow in the τ axis whereas non-spread modulations produce broader peaks, making the discrimination from noise more difficult, and also the ability to quantify the precise symbol rate at which the peak occurs. This is seen clearly in the example plots shown in Figure 2-16 for non-spread and spread BPSK signals from GOES and GPS satellites respectively that were captured by the SETI Institute's Allen Telescope Array. This shows the benefit of using a spread-spectrum form of modulation as far as detection with SWAC is concerned.

It was explained previously how SWAC is better able to detect randomly modulated signals than conventional autocorrelation. This can be seen clearly in the example shown in Figure 2-17. Here the same noisy signal of length M = 20 symbols is analysed using conventional autocorrelation and SWAC. All output values in both plots were normalised by the same value, such that the peak SWAC output was value 1. Both methods show a peak at the correct symbol period, but the SWAC output is significantly stronger. This means that SWAC is able to achieve successful detection at lower input S/N values than conventional autocorrelation, i.e. it provides superior sensitivity. Over a longer observation interval, the difference between the two algorithms will become even more pronounced, as the signal-related peak from conventional autocorrelation will trend towards zero.

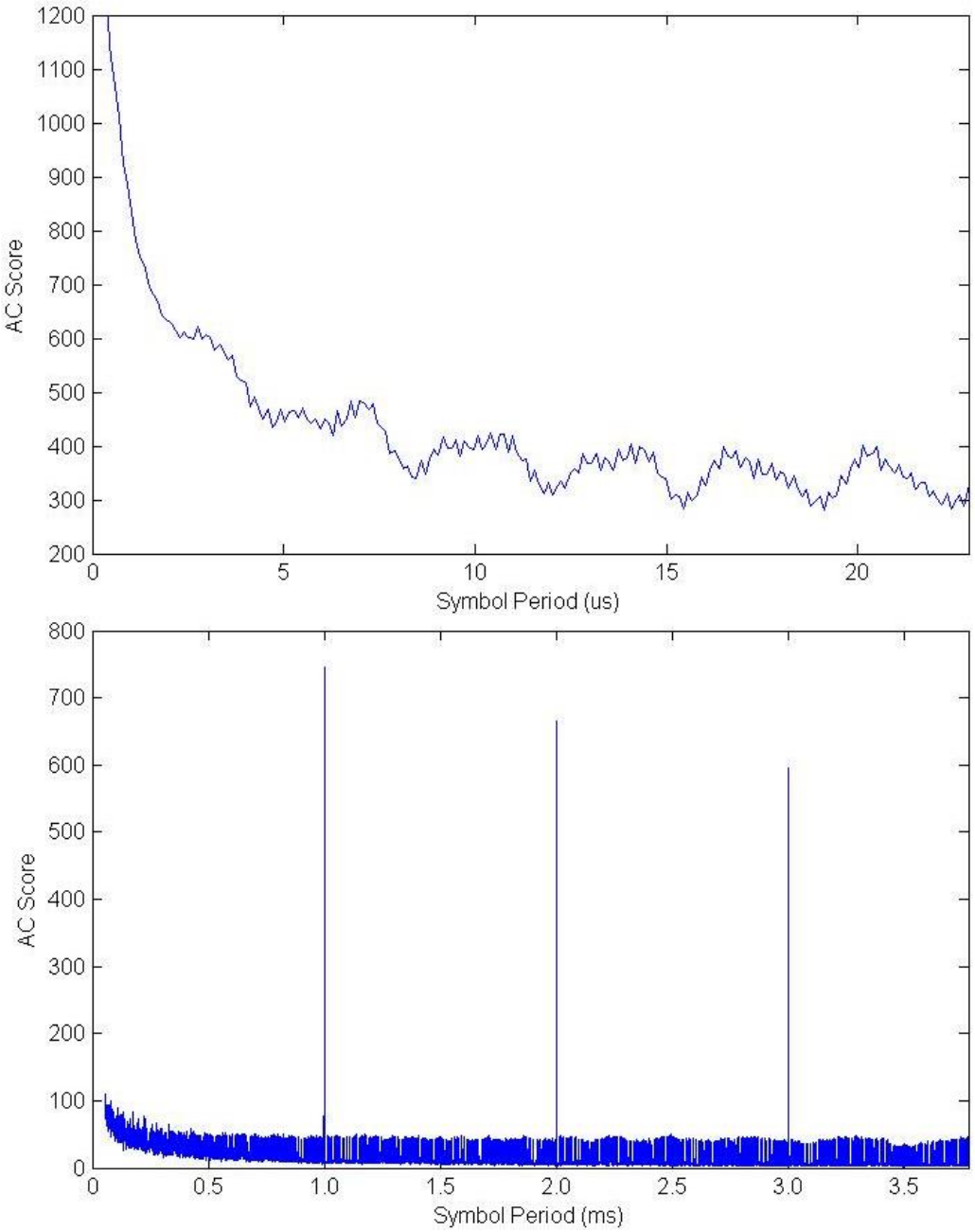

Figure 2-16: Comparison of SWAC outputs for a non-spread BPSK modulation from the GOES satellite (top) and a spread BPSK modulation from a GPS satellite (bottom).

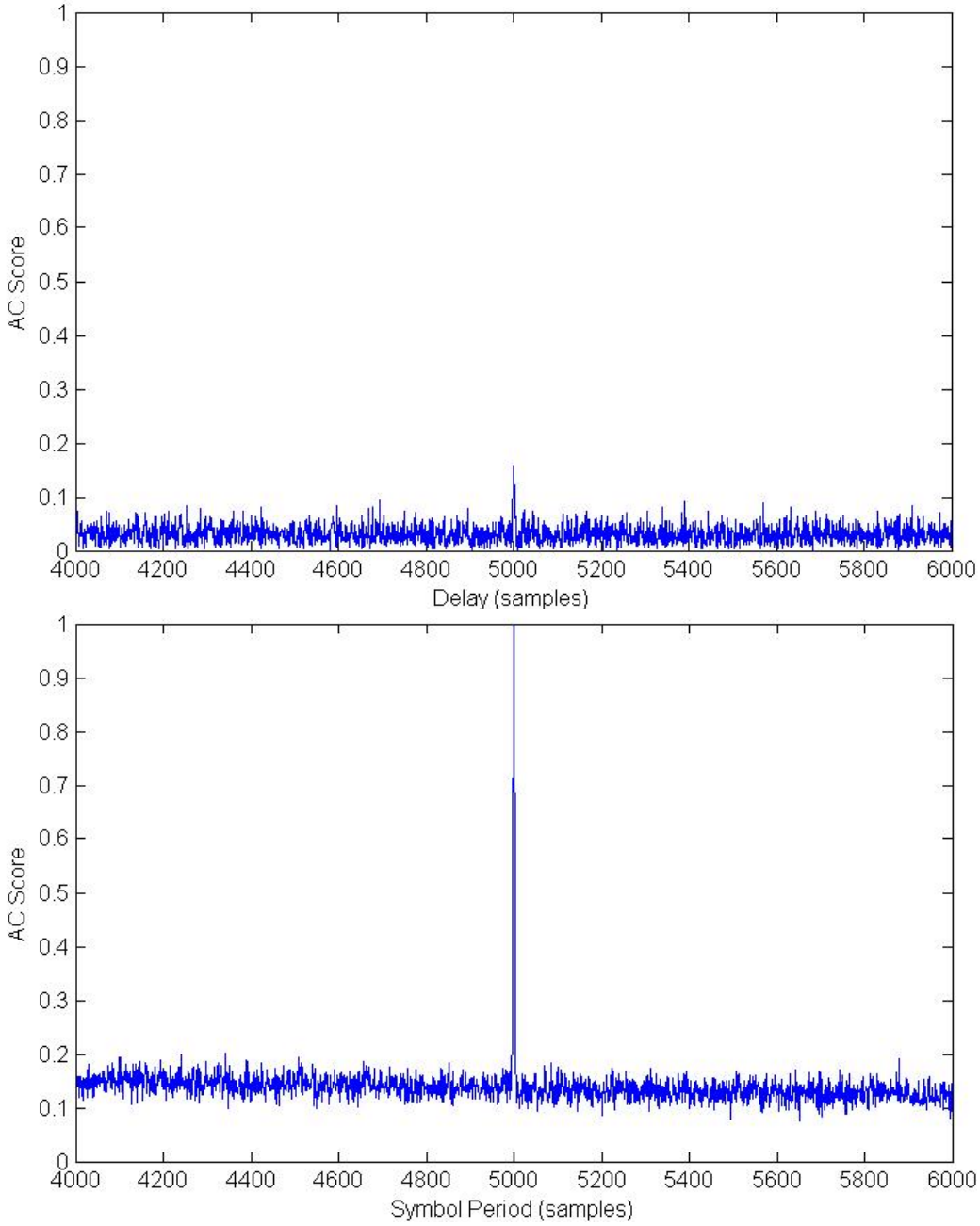

Figure 2-17: Comparison of conventional autocorrelation (top) with SWAC (bottom) for a spread-spectrum BPSK signal embedded in Gaussian noise with S/N = -10 dB.

2.11.4 SWAC sensitivity analysis

The sensitivity of SWAC is quantified analytically in Appendix C, alongside the analysis for matched filtering and energy detection. Both the ABS and SQR variants of SWAC are analysed, and the following expressions are derived for the output S/N in SETI discovery mode:

$$\boldsymbol{S/N_{SWAC,ABS}} = \frac{\boldsymbol{M}\left(\boldsymbol{A}\sqrt{\boldsymbol{2}\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{-1}}+\boldsymbol{WT_S}\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{-2}}}-\sqrt{\boldsymbol{WT_S}}\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{-1}}\right)^{\boldsymbol{2}}}{\boldsymbol{\pi}\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{-1}}+\boldsymbol{WT_S}\left(\frac{\boldsymbol{\pi}}{\boldsymbol{2}}-\boldsymbol{1}\right)\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{-2}}} \quad (13)$$

where

$$A = {}_1F_1\left(-\frac{1}{2},\frac{1}{2},\left(\frac{-1}{2\left(\frac{E_S}{N_0}\right)^{-1}+WT_S\left(\frac{E_S}{N_0}\right)^{-2}}\right)\right)$$

and where ${}_1F_1$ is the confluent hypergeometric function [37].

$$\boldsymbol{S/N_{SWAC,SQR}} = \frac{\boldsymbol{2M}}{\boldsymbol{(WT_S)^2}}\cdot\left(\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{4}}+\boldsymbol{2}\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{3}}+\left(\frac{\boldsymbol{E_S}}{\boldsymbol{N_0}}\right)^{\boldsymbol{2}}\right) \quad (14)$$

Note that these results apply specifically to binary antipodal modulation. This represents the best-case scenario for generating a cumulative discovery metric, and therefore these results represent an upper bound on the achievable sensitivity of SWAC with any form of modulation.

Figure 2-18 plots the output S/N for the ABS ($S/N_{SWAC,ABS}$) and SQR ($S/N_{SWAC,SQR}$) variants of SWAC. It is seen that marginally better performance is obtained using the SQR variant across the entire input S/N range, so this is selected for use as the benchmark sensitivity curve in subsequent plots.

Also plotted in Figure 2-18 for comparison are the curves for data-blind matched filtering and energy detection (for the WT_S = 256 case). SWAC is seen to out-perform energy detection for E_s/N_0 > 8 dB, and begins to approach matched filter performance at the high end of the E_s/N_0 range. However, at low E_s/N_0 SWAC is seen to be less sensitive than energy detection, by a factor of ~10 dB at an E_s/N_0 of 0 dB. This may appear discouraging, however this result is for the "basic SWAC" algorithm, and there are significant performance gains to be found with more sophisticated variants of SWAC, as demonstrated in Section 2.11.5. It is also crucial to

note that, at low S/N, energy detection has practicality issues due to the lack of a null reference. SWAC, on the other hand, provides an in-built null reference, making it usable even in the extremely low S/N regime where energy detection is unworkable.

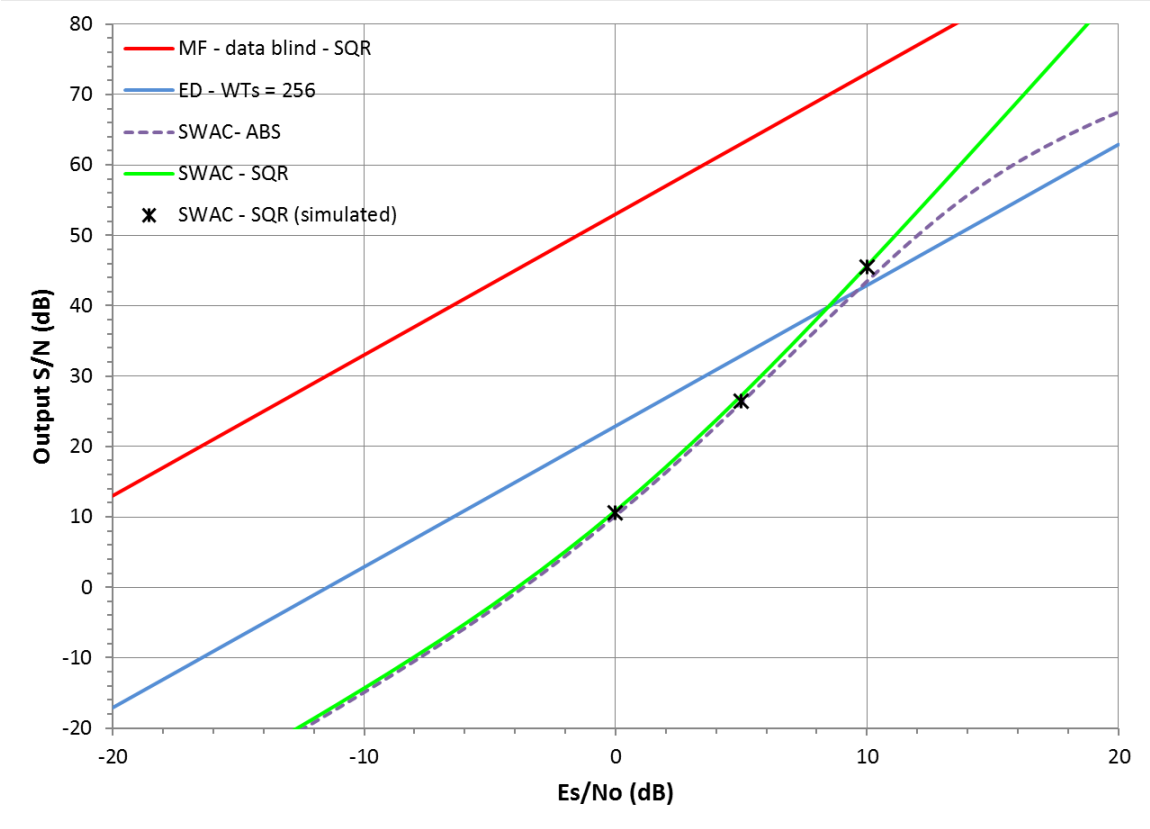

Figure 2-18: Basic SWAC output S/N as a function of input S/N (expressed as E_s/N_0) for the case of WT_s = 256, for both the ABS and SQR variants. Simulation results for the SQR SWAC variant are also plotted. Also shown for comparison are the results for data-blind matched filtering (SQR variant) and energy detection for WT_s = 256. In all cases the observation interval is M = 100,000 symbols.

To validate the analytical results, the performance of the SQR detector variant was simulated in Matlab, processing a 100 second long test vector of binary antipodal spread-spectrum modulation immersed in additive white Gaussian noise to a specifiable E_s/N_0. The symbol rate was 1000 symbols/s, so the total number of symbols processed was M = 100,000. The spread-spectrum spreading factor was 64, and a random chip pattern was used, taken as the first 64 binary digits of π. Oversampling of 4 samples per chip resulted in a waveform with WT_s = 256

samples per symbol for this example. The "SWAC.SQR" algorithm was run at each of three different input E_s/N_0 values, with and without the signal component present at each noise level. The output S/N for each case was then determined according to Equation (21) in Appendix C. As seen in Figure 2-18, there is excellent agreement between the simulated and analytical results, confirming the correctness of the formulation of Equation (14).

The simulation results are only reported over a limited E_s/N_0 range. At higher input E_s/N_0 values (above ~10 dB), the formulation for output S/N of Equation (21) in Appendix C will no longer accurately hold, since it implicitly assumes the signal energy is small in comparison to the noise energy. At the lower end of the E_s/N_0 range (below ~0 dB), the output S/N falls below 10 dB. When this low, the estimation of the detector output variance becomes unreliable for the length of data that was simulated, so the simulation output S/N results cannot be relied upon. However, the region of interest for the output S/N is likely to be from about 20 to 40 dB (for acceptable miss and false alarm probabilities – see Appendix C), so the detector behaviour below 10 dB output S/N is immaterial.

2.11.5 Enhanced sensitivity: near-neighbour SWAC

Note that the SWAC algorithm discussed to this point has been the "basic SWAC" approach that involves correlations between each test symbol and its immediately following adjacent symbol. Performing autocorrelation on a symbol-wise basis was seen to provide a means of countering the effect of random modulation (which weakens the detection sensitivity of conventional autocorrelation). But the real power of the symbol-wise approach comes from the ability to cross-correlate each test symbol with more than just one adjacent symbol. The concept can be generalised to include other pair-wise cross-correlations of symbols in close

proximity[25]. Figure 2-19 illustrates this with the example of correlating $s_{k-1}(t)$ with each of $s_k(t)$, $s_{k+1}(t)$ and $s_{k+2}(t)$.

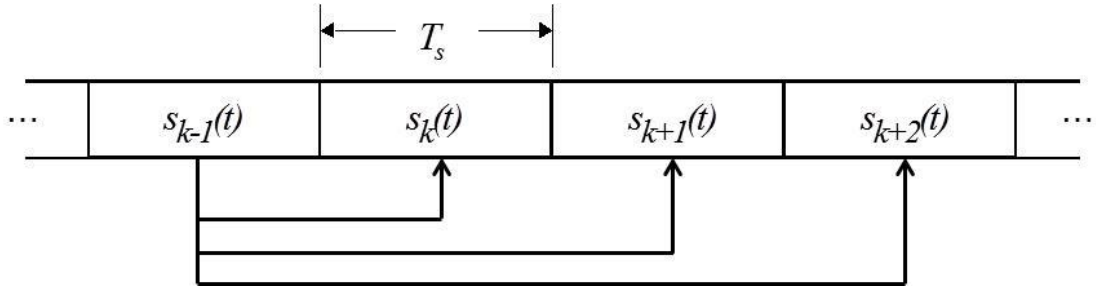

Figure 2-19: Multiple inter-symbol correlations in a sequence of contiguous modulation symbols with symbol period T_s.

The benefit of performing multiple inter-symbol cross-correlations on each test symbol can be understood in two ways:

1. Each additional cross-correlation effectively increases the value of *M* for a given time duration of observational data, or

2. Instead of correlating each test symbol with an equally noisy reference (its adjacent symbol), it is being correlated with a cleaner reference, derived from averaging multiple nearby symbols.

The second of these interpretations is particularly instructive. With an increasing number of cross-correlation pairs, *L*, the quality of the reference symbol improves, asymptotically approaching the shape of the noise-free reference waveform. This suggests that detection sensitivity for large *L* should approach that of a matched filter.

[25] The concept is analogous to improving the performance of a differential demodulator by exploiting multiple delays – so-called "Multiple Symbol Differential Detection" [60].

Note that L should not be increased indefinitely. Aside from the computational complexity aspect, it is important to limit the time separation of cross-correlation pairs to reduce the degradations due to Doppler drift and other time-varying channel characteristics (see Section 2.11.6). However, as long as the spacing of cross-correlation pairs remains within the coherence time of the propagation channel, each additional pair will increase the effective M and therefore improve the detection sensitivity. As an example, consider L = 3, where each test symbol is correlated with $1T_s$, $2T_s$ and $3T_s$ delayed versions of the signal. This will yield approximately a 3-fold effective increase in M, which will translate to an almost 5 dB improvement in detection sensitivity. Note that this is only *approximate* since the additional cross-correlation pairs are not completely independent, which leads to diminishing returns as L increases. The performance will approach that of a matched filter but can never exceed it.

We refer to the generalised SWAC algorithm as "near-neighbour SWAC" (NN-SWAC) to differentiate it from the basic algorithm, which might be referred to as "adjacent-symbol SWAC" (AS-SWAC).

Developing an exact expression for the detection sensitivity of NN-SWAC involves analysing the statistics of sums of non-independent variables. This is a work-in-progress, and the intention is for the result to be published by this author in a future paper.

However, to verify that NN-SWAC performance can approach that of a matched filter, a Matlab simulation program was developed. The program first generates a test vector with 1000 symbols of a randomly modulated spread-spectrum BPSK signal with WT_s = 256 samples per symbol (64 chips per symbol and 4 samples per chip) embedded in white Gaussian noise, with a parameterised input S/N. The NN-SWAC algorithm was executed on this test vector with different input S/N values, allowing a characterisation of the detector output S/N as a function input S/N. Multiple trials were conducted using different numbers of cross-correlation symbol

pairs: L = 1 (equivalent to basic SWAC), 10, and 100. The results are plotted in Figure 2-20. For comparison, the theoretical curves for (i) data-blind matched filtering, (ii) basic SWAC (WT_s = 256), and (iii) energy detection (WT_s = 256) are also plotted.

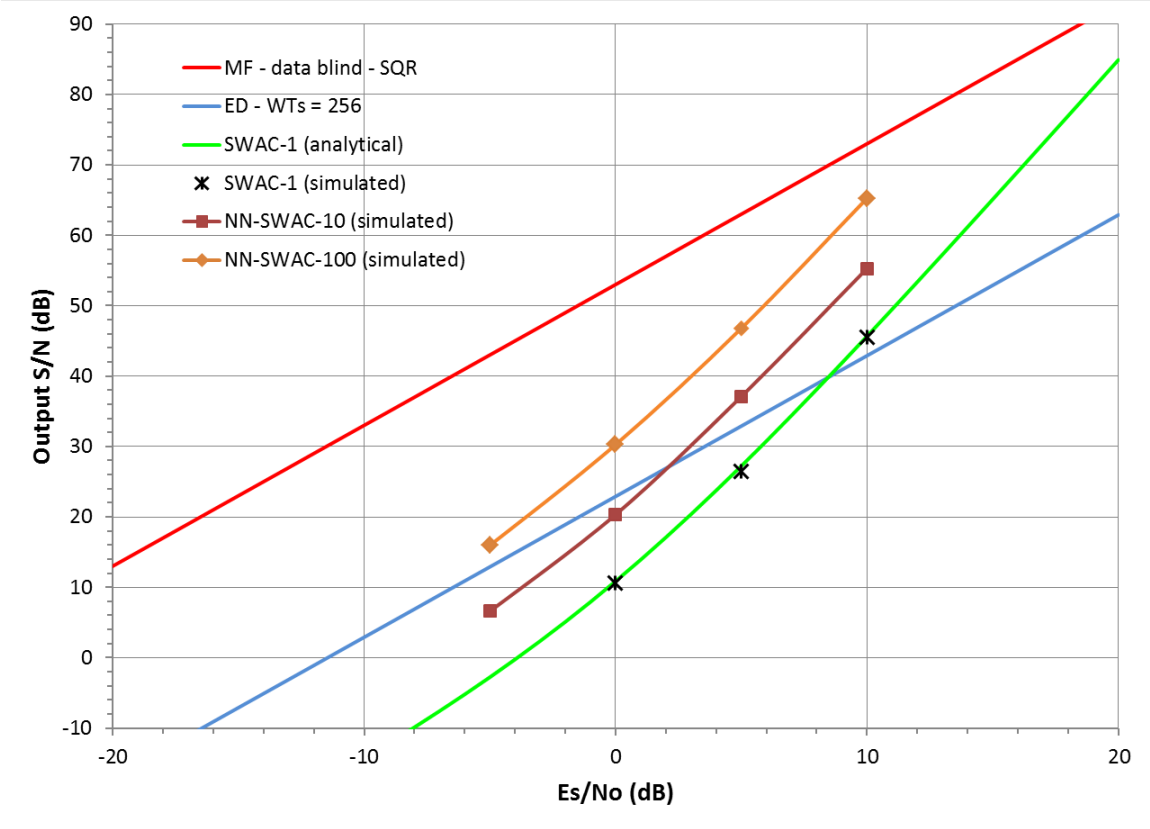

Figure 2-20: Simulated NN-SWAC output S/N as a function of input S/N (expressed as E_s/N_0) for the case of WT_s = 256, for L = 1, 10 and 100. Also shown for comparison are the results for data-blind matched filtering (SQR variant), energy detection for WT_s = 256, and the analytical result for L = 1 (basic SWAC). In all cases the observation interval is M = 100,000 symbols.

Figure 2-20 clearly illustrates the increasing sensitivity of NN-SWAC with increasing L, and convergence towards matched filter performance. Crucially, this ability to approach the theoretical optimal sensitivity is being achieved with a blind detector that has no a priori knowledge of the signal parameters.

It should be noted that this level of sensitivity is only achieved for the antipodal spread-spectrum modulation type; the sensitivity to other waveform types will be lower, depending

on their modulation alphabet. However, it does serve to illustrate that a wideband interstellar beacon can be designed such that, at a receiver with no knowledge of the precise signal parameters, it can be detected with equivalent sensitivity to a narrowband beacon.

2.11.6 Effect of channel impairments

A beacon signal emanating from outside of our solar system must propagate through the interstellar medium (ISM). If the transmitter is located outside of our galaxy then propagation through the intergalactic medium (IGM) is also involved. Propagation through the ISM/IGM will introduce various distortions and degradations to any communications signal. An excellent overview of these effects and their impact on wideband communications signals is presented by Messerschmitt in [21]. A detailed analysis of propagation effects on SWAC performance is outside the scope of this thesis, but is an intended future extension of the current research. For now we restrict ourselves to some qualitative remarks concerning the key effects of *Doppler*, *dispersion* and *scattering*.

Doppler

The relative motions of the transmitter, ISM/IGM and receiver give rise to Doppler effects observed by the receiver. Motions with constant velocity will result in a static time dilation, whereas for motions that involve any acceleration components, there will be a dynamic time dilation referred to as 'Doppler drift'.

SWAC is insensitive to static Doppler effects because all symbols are affected equally. The only consequence is that the value of τ at which the SWAC peak occurs will move very slightly, because of the lengthening (or shortening) of the symbol interval due to the time dilation. SWAC is, however, sensitive to Doppler *drift* because consecutive symbols experience slightly different degrees of time dilation. Over longer processing timespans there may be a 'smearing' of the SWAC peak across multiple delay bins. The frequency offset and phase shift

from one symbol to the next will also vary, which will reduce sensitivity and may preclude the use of the optimisation described in footnote 22. The effect is less significant for shorter symbol periods because near-neighbour symbols are closer together in time. For example, the effect of the Earth's rotation on a signal centred on 10 GHz will be insignificant for symbol rates greater than 100 symbol/s. However, such a constraint can be avoided completely if 'Doppler compensation' is employed. The transmitter and receiver are both aware of the component of their own acceleration along the line of sight. They can each therefore correct for this acceleration by appropriate frequency shifting processes synchronised to their known accelerations[26]. Doppler compensation is technically straightforward so there is a compelling case for it to be routinely employed for both Messaging to Extraterrestrial Intelligence (METI) and SETI. This would reduce the difficulties a receiver will face when attempting to detect signals of low symbol rate – and arguably it is the lower range of the symbol rate parameter space that is more important for SETI, since this corresponds to lower transmitter power requirements (which we assume would be desirable to beacon builders).

Dispersion

The ISM and IGM will also introduce a frequency-dependent delay known as *dispersion* [21] [43]. Dispersion is the result of interaction of the propagating signal with free electrons along the path of propagation, so the degree of dispersion (quantified by the Dispersion Measure, DM) depends on the sky direction and distance to the source. Dispersion can be

[26] Doppler compensation at the receiver can most simply be performed by chirping the local oscillator used for down-conversion. However, this method can only be used in a single-beam observing mode. In a multi-beam mode there will be different de-drift requirements in each beamformer. In this case it is necessary to perform the Doppler compensation separately for each beam, implemented digitally as part of the signal detection process – which increases the computational complexity.

algorithmically corrected if the DM is known, or through a series of trials with different DM estimates. If not corrected, dispersion results in delay-spread for wideband signals, which causes waveform distortion and inter-symbol interference (ISI) effects that are difficult to mitigate prior to discovery of the signal[27]. These effects can seriously compromise the performance of a matched-filter detector. However, autocorrelation-based detection such as SWAC is relatively immune to the effects of dispersion. Near-neighbour symbols all experience similar distortion to their waveforms, hence the high cross-correlation between them is retained. ISI can be more problematic but its significance is reduced when operating with longer symbol periods.

Importantly, dispersion effects decrease with increasing carrier frequency, and can essentially be discounted above ~30 GHz [21]. Chapter 4 makes the case for intentional beacon transmitters to operate in the range 30 to 90 GHz where dispersion effects are negligible and can be ignored. Indeed, this is a further advantage to operating a beacon in this region of the spectrum.

Scattering

The ISM and IGM will also introduce time-varying *scattering* effects that will degrade any signal propagating through them and make detection at a receiver more challenging [21] [43] [46]. Scattering arises from diffractive effects as radio emissions traverse the ISM. Scattering and dispersion together cause complicated, time-variable delay-spread behaviour for wideband signals, which not only results in waveform distortion and ISI, but also amplitude *scintillation*. In the regime of strong scattering (most sources beyond a few hundred pc, and frequency below a few tens of GHz), the scintillation observed on compact sources (i.e. those with

[27] After discovery it becomes possible to compensate for many of the distortions introduced by the ISM because the channel parameters can be estimated by the receiver and reversed [13].

angular size of order microarcseconds or smaller) has a variance that approaches unity (i.e. a modulation index of one). A similar form of scintillation also occurs in terrestrial wireless communications systems where it is referred to as "Rayleigh fading" [33]. A plot illustrating the typical variation in received signal strength of a source experiencing strong scattering is shown in Figure 2-21.

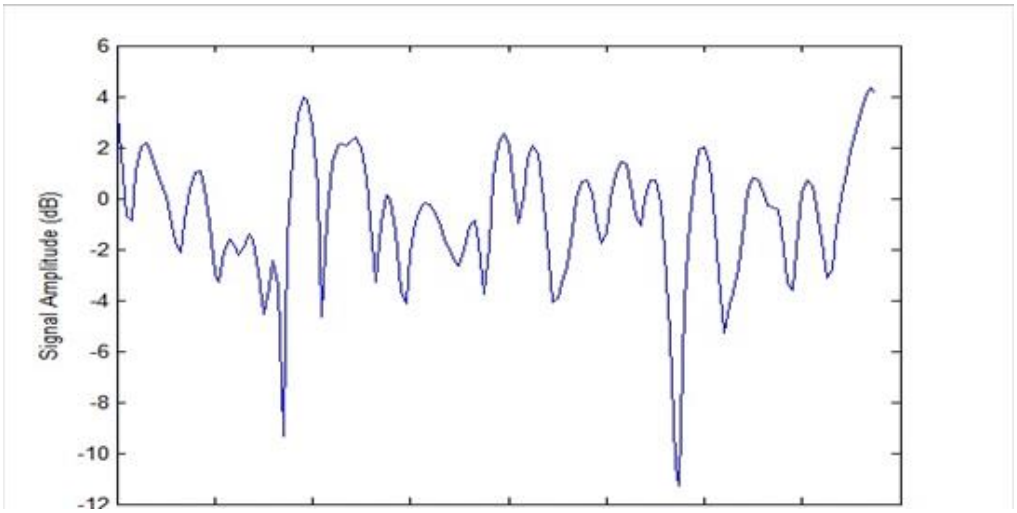

Figure 2-21: Illustration of ISM-induced scintillation. The plot shows the variation in received signal amplitude as a function of time on the horizontal axis. No specific timescale has been shown: depending on the sky direction and frequency of observation, the amplitude variations will occur on timescales of a few seconds to many minutes.

Rather than a hindrance, scintillation can actually be very helpful to SETI. There are times when the signal level is effectively amplified, which increases the S/N at the receiver during those times. However, the benefit will generally only be attainable if the signal has been designed such that its instantaneous bandwidth is no greater than the coherence bandwidth defined by the ICH, as discussed in Section 2.6. In addition, the detection process will not be adversely impacted if it instantaneously utilises only a span of signal no longer than the coherence time defined by the ICH – which is typically in the order of tens of seconds to

minutes for the frequencies associated with radio SETI. The reader is referred to [21] for further explanation of the ICH and how its characteristics vary spatially and spectrally.

SWAC is particularly suited to dealing with scattering/scintillation because it only operates on a finite window of near-neighbour symbols[28] while this window moves through the observed data containing *M* symbols, where *M* is typically much longer than *L*. Indeed, *L* should be chosen to be shorter than the coherence time of the channel, thus ensuring that the impact of scattering/scintillation is negligible. If the coherence time is many seconds or longer, then even at low symbol rates the signal will contain a large number of symbols within this timespan, thus presenting no real limitation on the choice of *L* (which anyway provides diminishing returns beyond *L* = ~100).

2.12 Chapter conclusions

This chapter has considered many aspects of the SETI problem and drawn conclusions concerning the types of artificial interstellar signals that SETI should anticipate discovering. The following is a summary of the key conclusions.

- Eavesdropping on a civilisation's radio emissions is not likely to be successful beyond distances of a few hundred light years, limiting the number of target sources. Intentional beacons may be discoverable on pan-galactic or even intergalactic distance scales, massively increasing the number of potential targets.

- Intentional interstellar beacons are likely to be information-bearing and wideband for maximum energy efficiency and RFI immunity. Their signalling methods are likely to

[28] For basic SWAC (aka AS-SWAC), the window length is 2 symbols. For the generalised NN-SWAC, the window length can be any finite length of *L* symbols ($L \leq M$).

operate close to the fundamental bound on energy efficiency, known as the Ultimate Shannon Limit.

- It is reasonable to assume that a beacon-builder will choose a signalling method that facilitates straightforward and efficient discovery by its intended recipients.

- A signalling scheme that is compelling for interstellar beacons has been identified: antipodal spread-spectrum binary phase modulation. With appropriate coding, it can achieve energy efficiency very close to the Ultimate Shannon Limit. Of equal importance, blind detection of this class of signals is possible, with sensitivity approaching that of matched filtering, given sufficient observing time and processing resources.

- An example blind detection algorithm for wideband SETI has been proposed – symbol-wise autocorrelation (SWAC) – based on detection of coherent cyclostationarity. Its sensitivity has been analysed and an advanced variant of the algorithm was shown via simulation to approach matched filter performance for the antipodal spread-spectrum modulation signal class.

- The veracity and practicality of wideband SETI has been demonstrated. There is no justification for limiting future SETI experiments to narrowband signal detection.

3 Extending galactic habitable zone modelling to include the emergence of intelligent life

Foreword

This chapter is a direct reproduction of a paper I co-authored with Michael G. Gowanlock, which was published in the journal *Astrobiology* in August 2015 [14]. The objective of this work was to take the simulation model of habitability of the Milky Way Galaxy developed previously by Gowanlock et al. [46] and extend it to include the emergence of intelligent life. The extended model then allows conclusions to be drawn as to the regions of the Galaxy with the highest propensity for intelligent life to emerge, and which therefore should logically offer the highest chances of success for SETI.

The research presented in the paper was conceived and planned jointly with Gowanlock. Data-sets produced during the previous simulation study involving Gowanlock [46] were utilised. To obtain the results presented here, new software programs were required to parse these data-sets and perform subsequent analysis. I was wholly responsible for the development and execution of this software. Both authors contributed to plotting the results of the analysis, and to preparing the text. As first author, I took responsibility for managing the submission and revision processes.

Abstract

Previous studies of the Galactic Habitable Zone have been concerned with identifying those regions of the Galaxy that may favour the emergence of *complex life*. A planet is deemed *habitable* if it meets a set of assumed criteria for supporting the emergence of such complex life. In this work, we extend the assessment of habitability to consider the potential for life to further evolve to the point of intelligence – termed the *propensity for the emergence of*

intelligent life, φ_I. We assume φ_I is strongly influenced by the time durations available for evolutionary processes to proceed undisturbed by the sterilising effects of nearby supernovae. The times between supernova events provide windows of opportunity for the evolution of intelligence. We developed a model that allows us to analyse these window times to generate a metric for φ_I, and we examine here the spatial and temporal variation of this metric. Even under the assumption that long time durations are required between sterilisations to allow for the emergence of intelligence, our model suggests that the inner Galaxy provides the greatest number of opportunities for intelligence to arise. This is due to the substantially higher number density of habitable planets in this region, which outweighs the effects of a higher supernova rate in the region. Our model also shows that φ_I is increasing with time. Intelligent life emerged at approximately the present time at Earth's galactocentric radius, but a similar level of evolutionary opportunity was available in the inner Galaxy more than 2 Gyr ago. Our findings suggest that the inner Galaxy should logically be a prime target region for searches for extraterrestrial intelligence, and that any civilisations that may have emerged there are potentially much older than our own.

3.1 Introduction

Recent studies of the habitability of the Milky Way Galaxy have given rise to the notion of a galactic habitable zone (GHZ), defined as the region (or regions) of the Galaxy that may favour the emergence of complex life (Gonzalez et al., 2001; Lineweaver et al., 2004; Gowanlock et al., 2011). This work has been motivated largely by the combined result of exoplanet detections (Perryman, 2012) and the discoveries of numerous extreme conditions under which life is found to thrive on Earth (Rothschild and Mancinelli, 2001; Cavicchioli, 2002). Together, they suggest that there may be many habitable planets in the Milky Way. The idea of the GHZ arises because, due to various underlying factors, life-supporting habitable planets are not distributed uniformly throughout space and time. The building blocks of terrestrial planets are

elements heavier than Helium, so sufficient time is required for these elements to form through stellar nucleosynthesis. Planets should also not occupy environments with a high frequency of transient radiation events, such as supernovae (SNe), that may endanger the long-term survival of complex life. Thus, the properties of the Milky Way and its stellar population will drive the regions where we may expect life to thrive on long timescales.

The motivation for the current work is to investigate how, and to what extent, consideration of the GHZ can assist in developing effective strategies for the search for extraterrestrial intelligence (SETI). In this work, we make an underlying assumption that intelligence (and potentially a technological civilisation) *can* emerge on a habitable planet, given time. Although we currently know of only the one example where this occurred on Earth, this is nonetheless an evidence-based assumption. The factors influencing the evolution of intelligence are not currently well understood. There is an ongoing scientific and philosophical debate as to whether its occurrence on Earth is an unlikely fluke (as advocated by contingency theorists such as Gould (1989) and Lineweaver (2005)), or may be inevitable (given sufficient time) in any habitable environment where life has originated (as defended by Ćirković (2012)). We take no side in this debate, contending that the rationality of conducting SETI does not depend on the number of potential target civilisations (a point also made by Ćirković). The probability of intelligence emerging in the Galaxy is clearly non-zero, and this is sufficient to conclude that other intelligences are *possible*, which in turn is sufficient justification to perform the SETI experiment. We concern ourselves only with the relative propensity for intelligence to emerge in different places/times in the Galaxy. This can tell us nothing about the absolute number of artificial sources of electromagnetic radiation we can expect to exist in the Galaxy, only which are the preferential directions to point our telescopes to maximise the chances of a detection. In formulating a metric for this relative propensity, we make no assumptions concerning the processes/causes/pressures involved in the evolution of intelligence, other than making the

weak (self-evident?) assumption that any event that takes time to happen will be more likely to happen if more time is available. Specifically, we examine only the general pre-conditions for intelligence that are known to have applied on Earth, and make the assumption that planets offering similar pre-conditions will make the evolution of intelligence possible on that planet. The more such planets, and the more time is available on those planets for evolutionary processes to proceed undisturbed, the greater will be the level of opportunity for intelligence to emerge. No matter how likely or unlikely the emergence of intelligence, it must surely be more likely where it is given more opportunity.

Early efforts to understand and quantify the potential for life and intelligence to arise throughout the Galaxy included the work of Drake in the 1960s, encapsulated by the "Drake equation" (Drake, 2003). However, the equation does not take account of the evolution of the physical properties of the Milky Way. The factors of the equation do not have a temporal dependence (Ćirković, 2004) or deal with the inherent parameter uncertainties through the application of probability distributions (Maccone, 2010; Glade et al., 2012). In terms of temporal considerations and the prioritisation of spatial search regions, the Drake equation can, therefore, provide little guidance to SETI.

The temporal and spatial aspects of galactic habitability were first quantified by Gonzalez et al. (2001), and later expanded to include dangers to the formation and habitability of terrestrial planets by Lineweaver et al. (2004), and then studied using a Monte Carlo simulation on the resolution of individual stars by Gowanlock et al. (2011). A comparison between the habitability of the Milky Way and M31 was made by Carigi et al. (2013). For an alternative perspective on these studies, see the work of Prantzos (2008).

The model described by Gowanlock et al. (2011) considers the stellar number density distribution and formation history of the Galaxy, planet formation mechanisms, and the

hazards to planetary biospheres as a result of supernova (SN) sterilisation events that take place in the vicinity of the planets. Based on timescales taken from the origin and evolution of life on Earth, the model suggests large numbers of potentially habitable planets may exist in our Galaxy (at least 1.2% of all stars in the Milky Way potentially host a habitable planet), with the greatest concentration likely being towards the inner Galaxy. This approach addresses the emergence of complex life (specifically land-based animal life), but it does not consider intelligence or the type of technological civilisation that can be detected by SETI.

Recent efforts to quantify the emergence of intelligent communicating civilisations within the Galaxy include those of Forgan (2009), Forgan and Rice (2010), and Hair (2011). The former two papers describe a Monte Carlo method to stochastically evaluate whether individual habitable planets reach a technological civilisation. They consider the impact of resetting events, albeit using a simplified model where resets occur at regular intervals. Their framework is very useful for understanding the constraints (both temporal and spatial) facing SETI. Hair (2011) modelled the absolute time of appearance of intelligence by means of a Gaussian distribution and proceeded to analyse the inter-arrival times of successive civilisations. Again, the findings provide useful insights into the co-temporality challenge of SETI. However, the model of Hair (2011) does not take into account the spatial and temporal variations of conditions conducive to the emergence of intelligence – a limitation also noted and discussed by Forgan (2011). Furthermore, in both models, the parameters assigned to their respective probability distributions are somewhat arbitrary, which is necessarily the case given that there is just one data point (the emergence of intelligence on Earth) with which to calibrate the models.

Given the challenges associated with modelling the emergence of civilisations, as described above, the goal in our work is not to estimate the absolute number of civilisations distributed historically throughout the Galaxy, but to analyse the *relative* propensity for intelligent life to

arise in different regions and epochs of the Galaxy. Relative numbers and distributions are sufficient to provide guidance to SETI. Until a first discovery is made, arguably the most effective SETI strategy (one that makes best use of limited resources) is to focus on those spatial regions likely to host the greatest number of potential extraterrestrial signal sources.

When considering potential target sources for SETI, their range must be taken into account, as well as the type of signal one is attempting to detect. There are essentially two distinct modes of conducting "electromagnetic SETI": (1) "eavesdropping" on unintentional leakage radiation, or (2) searching for intentionally transmitted beacon signals (which may or may not contain embedded information). Eavesdropping has the advantage that it does not rely on the cooperation of the radiating civilisation. However, the range over which such leakage radiation can be detected is limited; probably no more than a hundred pc, even assuming the presence of powerful pulsed or monochromatic sources (Forgan and Nichol, 2011). Therefore, eavesdropping may only be successful within the solar neighbourhood. In contrast, an intentional beacon signal can be highly directional and, with sufficient power, may be detectable over pan-galactic or even inter-galactic distances (Benford et al., 2010). Detecting such a beacon obviously relies on the existence of a beacon builder (and Earth being one of the beacon's targets), but it has the advantage that the higher permissible range dramatically increases the number of potential sources within the search space. These considerations have led to a series of works that address potential targets for SETI and habitable planets in light of current limitations and assumptions regarding other potential civilisations (Turnbull and Tarter, 2003; Beckwith, 2008; Kaltenegger et al., 2010). Assuming that SETI efforts advance with time, a body of work has been developed that considers the possibilities of technological civilisations in the astrobiological context beyond the technical limitations of SETI, which is the focus of the present study.

The approach adopted in the current work permits us to suggest guidelines for SETI that are grounded in evolutionary processes such as galactic chemical evolution, which in turn affect planet formation rates, thus avoiding approaches that assume uniform distributions of properties throughout the history of the Milky Way. Additionally, the self-consistent model ensures that the pressures on complex or intelligent life from biological extinction events (SNe in this work) are the result of the abovementioned evolutionary processes. Our objective is to account for the regulation of habitability and subsequent opportunities for intelligent life in the context of an evolving Galaxy.

Following this introduction, Section 3.3.2 describes the simulation model and analysis methodology. First, we provide an overview of the Monte Carlo simulation model of Gowanlock et al. (2011) on which the current work is based, including how trial planet populations are generated and how habitability is assessed. We then describe how this model is extended to assess the propensity for the emergence of intelligence (denoted φ_I) and the method of creating a metric for φ_I. Section 3.3.3 presents our results on the spatial and temporal variation of this metric and discusses their significance, with particular reference to SETI. Finally, Section 3.3.4 concludes the paper with a summary of our findings.

3.2 Methodology

3.2.1 Monte Carlo Habitability Model

The starting point for the present study was the model of the Milky Way developed by Gowanlock et al. (2011). In that model, various major observable properties of the disk of the Milky Way were used to populate stars and planets on an individual basis using Monte Carlo methods. To assess habitability, they modelled SNe as a function of the properties of the Milky Way, planet formation, and the time required for the emergence of complex life. Their modelling of the galactic disk incorporates a total stellar mass, an initial mass function (IMF), a

three-dimensional stellar number density distribution, a star formation history, and a galactic chemical evolution model. They only consider disk stars with galactocentric radii greater than 2.5 kpc, because of difficulties in accurately modelling the region inside 2.5 kpc, due to the complicated formation history of the bulge. Nevertheless, their model includes ~75% of the disk stars in the Galaxy, where the disk contains the majority of the stars in the Milky Way. Four variants of the model were proposed to assess sensitivity to variations in the parameters outlined above. In particular, two IMFs were utilised (Kroupa (2001) and Salpeter (1955)), and two stellar number density distributions (Jurić et al. (2008) and Carroll and Ostlie (2006)). A fixed total disk mass (Binney and Tremaine, 2008), star formation history, and associated galactic chemical evolution model (Naab and Ostriker, 2006) were employed.

All the models explored by Gowanlock et al. (2011) reproduced the same general behaviour and found that habitability was the greatest towards the inner Galaxy. In the present study, we concentrate only on the most pessimistic model (Model 4), based on a Kroupa IMF and found to have 1.2% of all stars hosting a habitable planet (of which 0.9% are tidally locked and 0.3% are non-locked to their host stars). For a detailed definition of the model and its associated parameters, see the work of Gowanlock et al. (2011).

Transient radiation events in the Milky Way create cosmic rays, X-rays, and gamma rays, which can deplete planetary atmospheres of ozone, expose planets to their host stars, and thus cause massive extinctions to land-based life (see Melott and Thomas (2011) for an overview of radiation hazards to our biosphere). The Gowanlock et al. (2011) model focuses on the ability of planets to survive SNe sterilisations. Given a total disk mass, an initial mass function, stellar number density distribution, and star formation history, type II supernovae (SNII) and type Ia supernovae (SNIa) were populated independently, which expresses differences in formation rates and sterilisation distances between these types of SN. It was assumed that planets nearby these SNe (the sterilisation distances of which reflect distributions of absolute

magnitudes of observations) will be uninhabitable for a finite time period after a sterilisation event occurs, and the planet can recover from the event.

SNe occur throughout the Milky Way, and there is even evidence of them occurring in recent geological history. Benítez et al. (2002) suggested that ~2 Myr ago a SN caused significant damage to Earth's ozone layer, which had an effect on the extinction of ocean life at the Pliocene-Pleistocene boundary. Furthermore, Bishop and Egli (2011) suggested that ~2.8 Myr ago, Earth was nearby a SN, as evidenced by ^{60}Fe in deep sea crust. In line with Gowanlock et al. (2011), we focus on SNe, which we assume to be the dominant danger to habitability.

Gowanlock et al. (2011) found that the highest density of habitable planets occurs in the regions with the highest stellar densities, and consequently highest frequency of SN events. As with other previous works on the habitability of galaxies (Lineweaver et al., 2004; Prantzos, 2008; Carigi et al., 2013), they did not account for stellar kinematics such as radial mixing, or oscillations above and below the midplane that may lead to varying levels of exposure to cosmic rays (Medvedev and Melott, 2007), on the basis that such motions were expected to have an insignificant overall negative impact on the fraction of habitable planets. Gowanlock et al. (2011) did not find a region in the Milky Way that was continuously sterilised, or sterilised at a sufficiently high frequency that planetary systems traveling through such a region would have a high probability of becoming sterilised. Should such a region have existed, then incorporating stellar motions above and below the midplane would have a greater impact on the results. Note that a star above or below the midplane that passes through it would be entering a region where there is a higher density of habitable planets (and hence cannot be significantly more hazardous to habitability). Therefore, oscillations above and below the midplane are unlikely to significantly decrease habitability, especially since oscillations of this type result in those stars still spending the majority of their time above or below the midplane. If such vertical stellar oscillations had been considered in Gowanlock et

al. (2011), the mixing would have produced a degree of averaging in the results for habitability above and below the midplane, slightly weakening the observed trends.

Gowanlock et al. (2011) populated the stars in their model by assigning each one a birth date, main sequence lifetime, and metallicity from the galactic chemical evolution model and star formation history, which assumes an inside-out formation history of the Milky Way. The metallicity-planet correlation of Fischer and Valenti (2005) was used, in combination with the population synthesis models of Ida and Lin (2005), to assign habitable planets to host stars in the model.

The timescales in Earth's history were adopted to calculate whether a planet is habitable. This is arguably the most speculative assumption in the model, as there is only Earth's pathway to complex (and intelligent) life to suggest such conditions on other planets. In light of focusing on dangers caused by SN events to planetary biospheres and, in particular, atmospheric ozone depletion, the focus is on the timescales for the build-up of ozone on Earth. The notion of planetary oxygenation time is adopted from the work of Catling et al. (2005), which proposes that, on Earth, a continuous duration of oxygenation is required for the emergence of complex life. Gowanlock et al. (2011) assumed that 1) any sterilisations that occur on the planets populated in the model before the ozone layer forms (this was approximately 2.3 Gyr ago on Earth) have no effect on habitability (since any life at this stage is assumed not to be surface-dwelling), and 2) the emergence of complex life requires a sufficient time period isolated from sterilisation events to allow for sufficient oxygenation for the emergence of complex life. Therefore, if a SN occurs within a threshold distance of d pc between the time period that the ozone formed and the period afterwards that is required for the emergence of complex life, then this has a resetting effect, and the planet must remain unsterilised for a time period before it is considered habitable.

Gowanlock et al. (2011) used the work of Gehrels et al. (2003), who found that a SNII will deplete the ozone layer and have a sterilising effect on planets at a distance of < 8 pc. The 8 pc distance was assumed to be just sufficient to sterilise a planet. By using the absolute magnitudes of SNII and SNIa events, a distribution of sterilisation distances was developed to reflect the notion that different magnitude events can occur and lead to varying sterilisation distances.

For SNII, *d* was selected from a probability distribution within the range of ~2-27 pc, and within the range ~14-27 pc for SNIa. Figure 3.1 shows an illustration of the timescales that demonstrate the interrelationship between sterilisations and the major events in Earth's history used to calculate the habitability of a planet in the model of Gowanlock et al. (2011). Note that region (C) in Figure 3.1 – the time after a planet becomes habitable – is the focus of the present work in extending the modelling to include the evolution from complex to intelligent life.

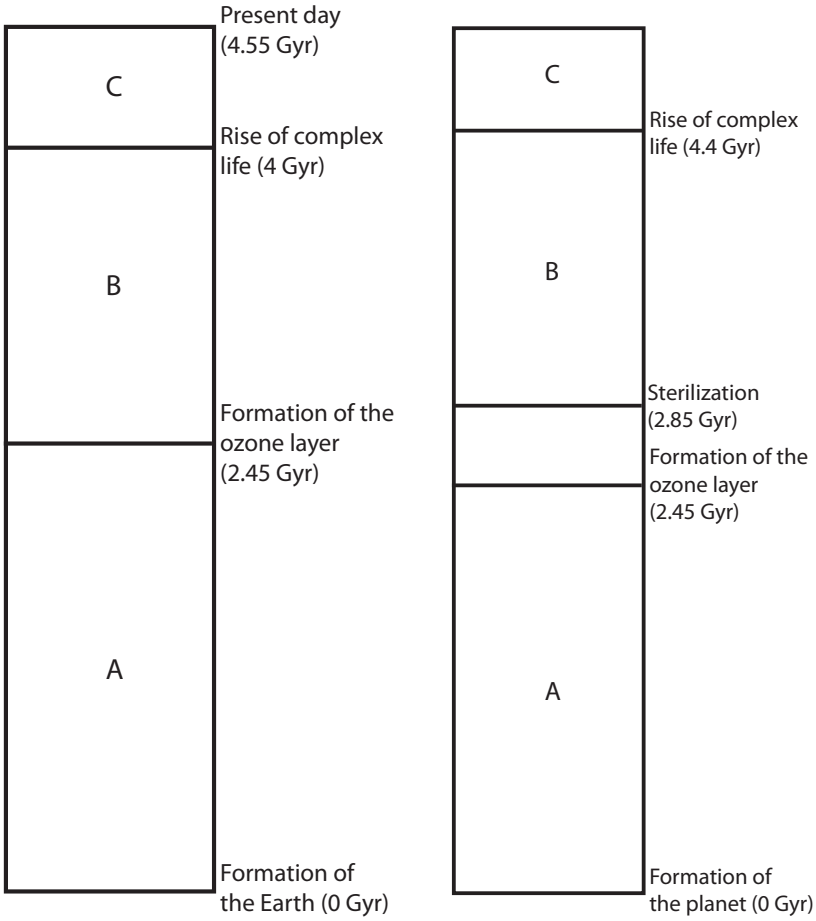

Figure 3.1. Left: Illustration of the major timescales on Earth used as criteria to calculate the habitability of a planet. For a given planet in our model, we assume that a sterilisation occurring before the formation of the ozone layer has no effect (A), whereas a sterilisation occurring during the period of continuous oxygenation does have an effect (B), and (C) shows the time since the rise of complex life to the present day. **Right:** Employing the timescales from Earth (on the left), we illustrate the effect a sterilisation has on a planet populated in our model. A sterilisation in (A) has no effect, since the ozone layer has not yet formed. However, the sterilisation in (B), shown at 2.85 Gyr, delays the possibility of complex life until 4.4 Gyr, as it disrupts the requirement of continuous oxygenation time. The time period (C) shows a hypothetical duration in which complex life has not been affected by any sterilisation events after the planet is considered habitable at 4.4 Gyr.

3.2.2 Gap Time Analysis

The methodology described above for assessing habitability is based on identifying planets that provide conditions conducive to the evolution of *complex land-based animal life*. We

assume that this represents the starting point for further stages of evolution that could lead to the emergence of intelligent life and, beyond that, to technological civilisations. In assessing the propensity for complex life to further evolve to intelligent life, φ_I, our fundamental assumption is that *time* is the primary barrier to this process. We reason that environmental conditions, at least at the beginning of the process, are favourable, given that they were deemed suitable for complex life to develop. We then consider the time period beyond that needed for the appearance of complex life to see whether sufficient time is available for further evolution to intelligence. We assume that this process would be disrupted by any nearby SNe, that is, if a SN occurs before intelligence is reached the process is reset. The time durations between such SNe are referred to as *gap times*, and we assume φ_I is strongly dependent on the number and length of these gap times. Without proposing a specific relationship between gap time length and its effect on φ_I, we suggest it is a reasonable assumption that longer gap times will provide greater opportunity for the emergence of intelligence.

Our analysis of gap times follows essentially the same methodology as employed by Gowanlock et al. (2011), but with an extended parameter range for the time duration between SNe. The goal was to assess whether these additional time requirements for the evolution of intelligence would alter the basic findings of Gowanlock et al. (2011), that is, to investigate the extent to which the regions of greatest propensity for intelligence corresponded to regions of greatest habitability (as defined for complex life). On Earth, the evolution from complex life to intelligence took just under 0.6 Gyr. Rather than apply this single figure, we acknowledge the lack of understanding of how the process works (and hence how long it typically takes) by considering a range of durations. As will be explained in Section 3.2.3, our metrics are based on cumulative gap times conditioned on a variable threshold value ranging from 0 to 2 Gyr. We prefer this approach over assigning a specific probability distribution to the time required

for intelligence to emerge for two reasons: (1) we have insufficient data to meaningfully ascribe a shape or mean value to this distribution, and (2) there are potential sensitivities that may be revealed by our model that could be masked by the averaging effect of applying a distribution.

At this point, it is important to note that, if a planet is assessed as "habitable," it does not mean that it will definitely become inhabited by complex life – only that conditions are favourable for this to happen. Likewise, if there is a long gap time between SNe on a habitable planet, it is not assured that intelligent life will emerge – only that this becomes a possibility. In this work, we do not attempt to quantify the percentage of planets that give rise to intelligence. We seek to produce a metric for φ_I that allows analysis of the *relative* likelihood of intelligence emerging in different regions and epochs of the Galaxy. A conservative position would be to assume that complex or intelligent life may only be able to arise on a small fraction of habitable planets. The opposite position would be that it is likely to arise on the majority of habitable planets. Either assumption, or any in between, can be made without affecting the veracity of any conclusions drawn from analysing relative propensities.

Figure 3.2 illustrates how gap times are related to the major events in a planet's timeline. For the illustrative example given, there are three gap times, labelled GAP_1, GAP_2, and GAP_3. GAP_1 is the time from the formation of a complete ozone layer to the first of two SNe, SN_1. GAP_2 is the time between SN_1 and SN_2. GAP_3 is the time from SN_2 to the present (or equally, death) time of the example planet. For the emergence of complex land-based animal life, we assume the same timeframe as observed on Earth, that is, 1.55 Gyr. Where a gap time exceeds 1.55 Gyr, this provides an opportunity for further evolution to intelligent life.

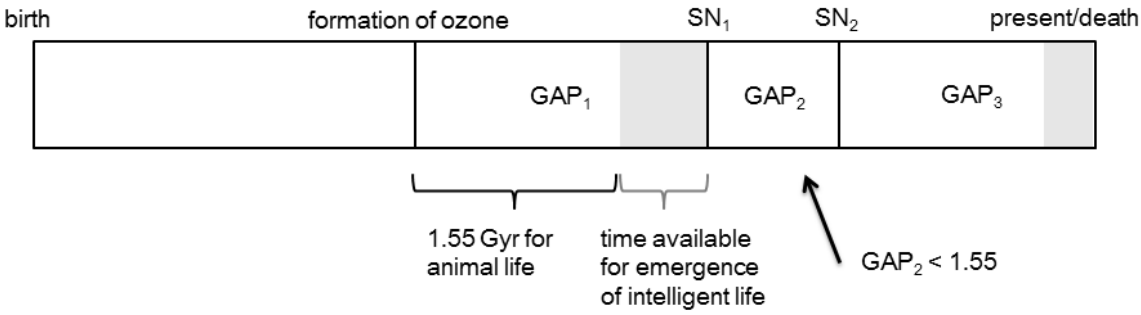

Figure 3.2. Illustrative planet timeline showing the major events from the birth (at left) to the present (or death) time (at right), and showing how "gap times" are calculated. In this example, there are two SNe, labelled SN1 and SN2. A gap time begins after the first formation of the ozone layer, or after an SN event. A gap time is ended by a SN, the death of the planet, or the present day, as we do not extrapolate beyond the age of the Universe. Any gap times exceeding 1.55 Gyr (the time assumed to be needed for the emergence of animal life) give rise to an opportunity for intelligent life to emerge. The shaded regions represent these "opportunity times," T_O, which are equal to the gap time less 1.55 Gyr.

In the example of Figure 3.2, there are two such *opportunity times*, T_O, which we may calculate as $T_{On} = (GAP_n - 1.55)$ when $GAP_n \geq 1.55$, and zero otherwise. Since GAP_2 is less than 1.55 Gyr, there is assumed to be no opportunity for the emergence of complex or intelligent life during that interval and hence $T_{O2} = 0$.

The Monte Carlo simulation described in Section 3.2.1 provides a hypothetical population of planets, along with pertinent data for each planet, including its location coordinates, birth/death dates, and a list of dates the planet was sterilised by SNe. Locations are specified by *x*, *y*, and *z* coordinates relative to the galactic centre, as shown in Figure 3.3. The *x* and *y* coordinates define the position projected onto the galactic midplane, and *z* is the height above (when positive) or below (when negative) the midplane. The galactocentric radius, *r*, is given by $(x^2 + y^2)^{1/2}$. We elected to work with a subset of the entire galactic dataset to take advantage of azimuthal symmetry, specifically a 15° sector of the full 360° dataset, as illustrated in Figure 3.3. Even with this fraction, our model included in excess of 70 million

planets, which is sufficient to allow statistical sampling errors to be ignored, and to safely assume that the chosen 15° sector would produce the same results as any other 15° sector. The metrics generated from this dataset (as described in Section 3.2.3) were scaled by a factor of (360/15) to obtain results that represent the entire Galaxy.

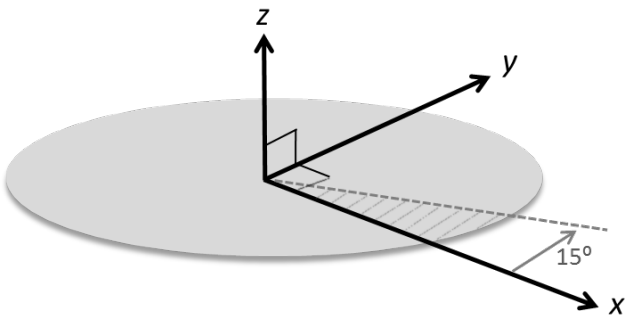

Figure 3.3. The coordinate system employed in the simulation model of Gowanlock et al. (2011) for defining planet locations relative to the galactic centre at (*x*,*y*,*z*) = (0,0,0). The model generates data for the entire 360° of "azimuth" in the (*x*,*y*) plane. To simplify processing for the present study, only a 15° subset was analysed, exploiting the model's azimuthal symmetry.

From the planet dataset, one can assess which planets are habitable (according to the criteria set by Gowanlock et al. (2011)) and additionally calculate the gap times experienced on each habitable planet. For some planets, no gaps exceeding 1.55 Gyr occur; for others, one or more such gaps occur. We treat multiple gaps on a single planet in an equivalent way to single gaps on multiple planets, that is, as independent opportunities for life to evolve. The total number and length of all gap times for all habitable planets produced by the simulation are accumulated, binned according to spatial location and temporal epoch, from which further analysis can be conducted.

3.2.3 Propensity Metric

We are primarily interested in examining how the propensity for the emergence of intelligent life, φ_I, varies as a function of spatial location and epoch within the Galaxy. To investigate this,

we create a metric for φ_{I} and observe, for our simulated planet population, the variability of this metric over time and as a function of r and z.

A straightforward metric, which we term φ_{Iu}, is the accumulated sum of all opportunity times, T_{On}, for the planets existing within a specified spatial bin. (For a single planet, this would correspond to summing the lengths of time represented by the grey shaded regions in Figure 3.2.) That is,

$$\boldsymbol{\varphi_{\mathrm{Iu}}(r_{\mathrm{j}}) = \sum_{n} T_{On}}$$

where r_{j} is the centre value of the j$^{\mathrm{th}}$ spatial bin. For example, if the data are binned according to galactocentric radius using bins of width w, then the radius range corresponding to r_{j} is $[(r_{\mathrm{j}}-w/2) \leq r < (r_{\mathrm{j}}+w/2)]$.

This method of computing φ_{Iu} is equivalent to assuming a uniform probability distribution for the required time for intelligent life to emerge. That is, the required time is assumed to be a uniformly distributed random variable, and hence the total probability will be proportional to the total cumulative time.

A variation for computing φ_{Iu} involves setting a threshold time, T_{thresh}, for the T_{O} values, and only those $T_{\mathrm{O}} \geq T_{\mathrm{thresh}}$ are included in the summation. For example, the rise of animal life on Earth occurred when the planet was ~4 Gyr old, and it was a further ~0.6 Gyr for the rise of intelligent life. If we assume these timescales, that is, that 0.6 Gyr is the minimum time for intelligence to emerge after a planet can support complex life, then only those $T_{\mathrm{O}} \geq 0.6$ Gyr are included in the summation.

We do not know the precise relationship between the value of T_O and the probability that intelligence will emerge during a time window of that length. It seems likely that the process of evolving intelligence requires a number of essential sub-processes, each occurring in sequence and each having its own specific time-distribution. This assumption was made by Carter (2008) and Forgan (2009). If this were the case, the overall time to achieve intelligence would be a random variable with a distribution approaching Gaussian (following the Central Limit Theorem)[28]. The uniform and Gaussian models for the probability distribution of the time for evolving intelligence are illustrated in Figure 3.4.

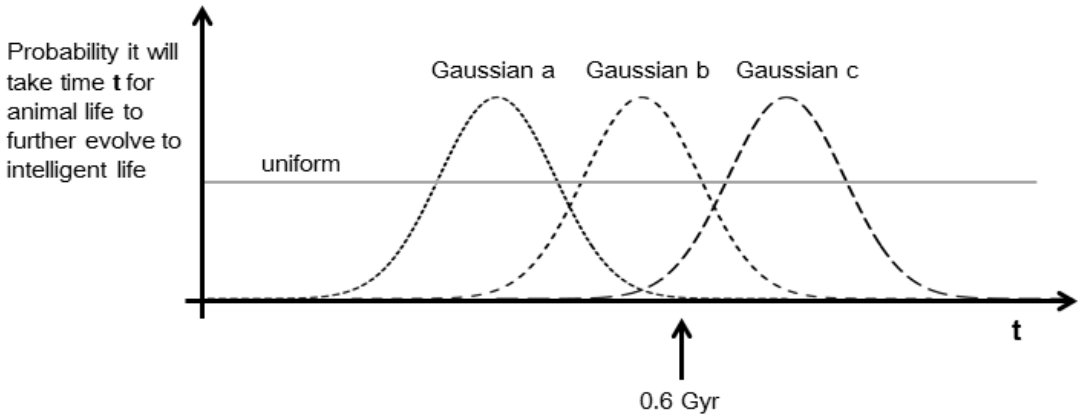

Figure 3.4. Alternative models for the probability distribution of the time taken for intelligence to evolve from animal life. The relationship between the propensity for the emergence of intelligent life, φ_I, and opportunity time T_O is dependent on this distribution. Shown are the uniform case and three different Gaussian cases of differing means relative to 0.6 Gyr (the time it took on Earth for animal life to evolve intelligence).

Although the Gaussian model may be more appropriate than the uniform model, it has the difficulty that we do not know the scale factor on the time axis. We do not know where the

[28] Given that evolutionary sub-process times are all positive random variables, it may be more accurate to model the distribution of the sum of sub-process times using the log-normal distribution. Regardless, for a large number of sub-processes, the log-normal distribution will converge towards Gaussian.

mean of the distribution lies relative to the 0.6 Gyr that was required on Earth. The three example Gaussian distributions in Figure 3.4 ("Gaussian a", "Gaussian b," and "Gaussian c") illustrate alternative timescales. If "Gaussian a" was an accurate representation, this would suggest that intelligence arose late on Earth. Conversely, if "Gaussian c" was an accurate representation, this would suggest that intelligence arose early on Earth. Without a calibrated timescale, we cannot assess the sensitivity of φ_I to changes in T_O. For example, if typical T_O values are to the left of the Gaussian bell-curve, then a small incremental increase in T_O will result in a large increase in φ_I. However, if the T_O are to the right of the bell-curve, then an incremental increase in T_O will have little effect on φ_I. Because of these uncertainties, there are difficulties in developing a φ_I metric that derives from a Gaussian (or any non-uniform) distribution. Furthermore, if we accept the premise that time is the primary determinant for φ_I, then a φ_I metric based on the summation of available time is not unreasonable. Therefore, we elect to employ the uniform propensity metric, φ_{Iu}, when generating the results reported in Sections 3.3.1, 3.3.2, and 3.3.3 below. Additionally, in Section 3.3.4, we propose an alternative method of analysis that obviates the difficulties of having to make any assumptions regarding the probability distributions for the time required for the emergence of intelligence. That methodology and its results are described in Section 3.3.4.

3.3 Model Results

3.3.1 Propensity Metric – Uniform Model

For the uniform model described in Section 3.2.3, the propensity for intelligent life is modelled as being proportional to total opportunity time, that is, the sum of all $T_O \geq T_{thresh}$. Figure 3.5 presents the results for φ_{Iu} for five alternative values of T_{thresh} over the radius range 2.5 to 15 kpc. The vertical axis represents the radial *density* of φ_{Iu}, that is, φ_{Iu} per parsec of radius.

The results exhibit two main features:

1. For all values of T_{thresh}, φ_{lu} is greatest towards the inner disk of the Galaxy; and

2. For all radii, φ_{lu} tends to decrease as T_{thresh} is increased.

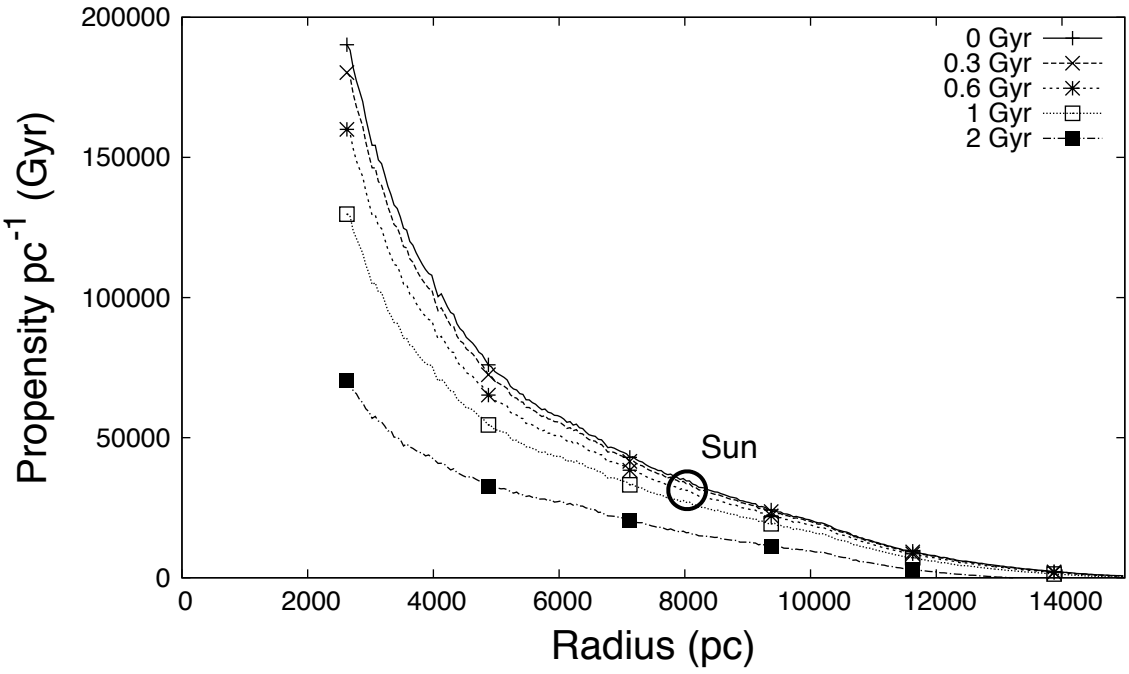

Figure 3.5. Propensity metric φ_{lu} as a function of r, for five time threshold values, T_{thresh}, ranging from 0 to 2 Gyr. The vertical axis is the radial density of φ_{lu}, i.e., φ_{lu} per parsec of radius. For the uniform model, φ_{lu} is modelled as being proportional to total opportunity time, i.e., the sum of all $T_O \geq T_{thresh}$.

The first observed feature is consistent with the trends in habitable planet density reported by Gowanlock et al. (2011). To assess whether this result simply tracks the habitable planet density, we also examine the average of φ_{lu} per habitable planet, which has been plotted in Figure 3.6. It is seen that the average propensity does indeed vary with radius, displaying a region of maximum average propensity between about 6 to 10 kpc. At smaller radii, the average propensity is marginally lower, which can be attributed to the higher rate of SN events. At larger radii there is a rapid decline in average propensity, which can be attributed

to the reducing average age of habitable planets with increasing *r*. This in turn is due to the later epochs at which the critical metallicity for habitable planet formation occurs with increasing *r*. (The variation of φ_{lu} with epoch time is discussed further in Section 3.3.5). Despite the variations in average propensity per planet, the overall favourability of the inner Galaxy, as seen in Figure 3.5, is due to the sheer number of habitable planets predicted by the model in this region.

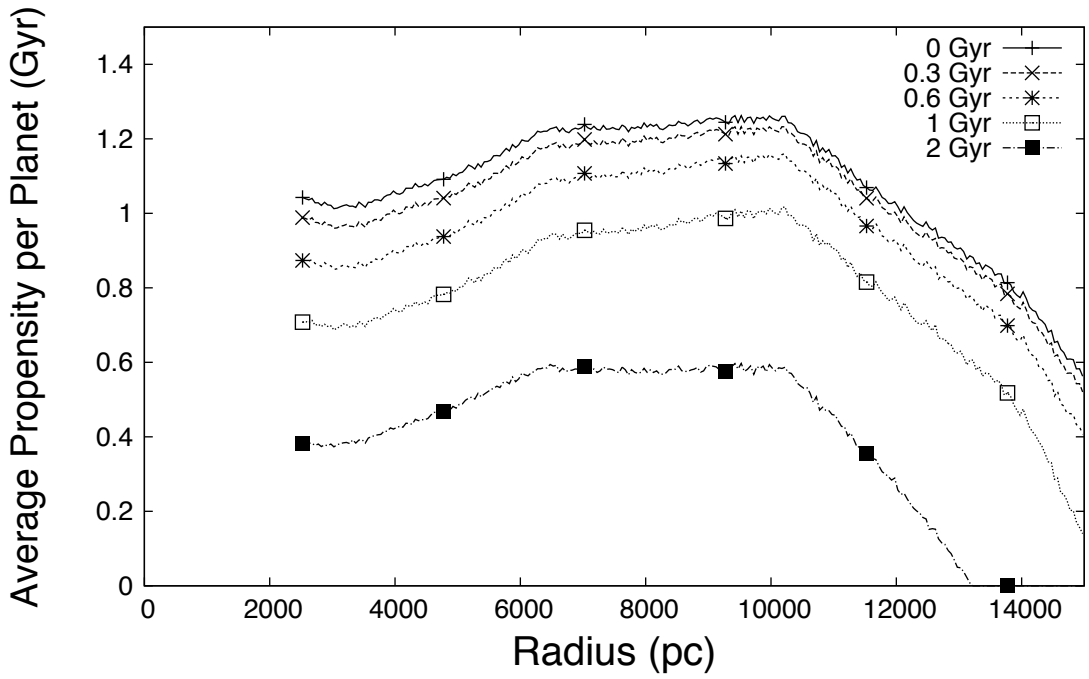

Figure 3.6. Average of φ_{lu} per habitable planet as a function of *r*, for five T_{thresh} values ranging from 0 to 2 Gyr. The vertical axis is the radial density of the per-planet average of φ_{lu}, i.e., the average of φ_{lu} per parsec of radius.

The results of Figure 3.5 suggest φ_{lu} will peak at some radius less than 2.5 kpc (but assumed to be greater than zero due to the proximity of the central black hole). The precise location of the peak cannot be determined, as it lies beyond the lower radius range of our model. However, the basic conclusion of the favourability of the inner Galaxy is not altered by the precise location of the peak; only the definition of "inner." Were the model of Gowanlock et

al. (2011) to be extended in future to lower radii, the analysis of this paper could be repeated to provide a closer bound on the radius of peak propensity.

The second observed feature in Figure 3.5 – the consistent reduction in φ_{lu} with increasing T_{thresh} – is predictable, given that larger values of T_{thresh} permit fewer T_O to be included in the metric summation.

For reference, we have marked on Figure 3.5 the circumstances that hold for the Sun and Earth (i.e., r = 8 kpc and T_{thresh} = 0.6 Gyr). The value of φ_{lu} for these parameters is ~35,000 Gyr per radial parsec. This is the aggregated propensity for all habitable planets occupying an annular ring of 1 pc width, at a galactocentric radius of 8 kpc (noting from Figure 3.6 that the average φ_{lu} per planet is ~1.1 Gyr in this region). It is seen that for lower radii, the density of φ_{lu} per radial parsec is up to 4 to 5 times higher. This is due to the higher number density of habitable planets in this region, rather than the average φ_{lu} per planet (which we see from Figure 3.6 is ~0.9 Gyr in this region). This may be interpreted as follows: we know that intelligent life can arise (it has arisen at least once) at 8 kpc, and there should be an even greater chance that it has arisen in regions closer to the galactic centre. This is seen even with larger T_{thresh} assumptions, up to 2 Gyr.

A different approach to examining the propensity for intelligence was taken by Forgan and Rice (2010), who used the Rare Earth hypothesis framework. They also found that the inner Galaxy should have the greatest number of intelligent civilisations. Despite major differences in model assumptions and goals between this work and theirs, the overall conclusions are in general agreement.

3.3.2 Propensity Above and Below the Midplane

We now consider the variation of φ_{Iu} in two spatial dimensions: r and z. We present the results in the form of contour maps, which show r on the horizontal axis, z on the vertical axis, and a color-coding of φ_{Iu} in the plot.

Figure 3.7 shows the φ_{Iu} contour maps for five values of T_{thresh}, ranging from 0 to 2 Gyr. A logarithmic scale is used for the color-coding, allowing greater detail to be seen in regions where the φ_{Iu} values are low. Each contour map represents a cross-sectional view of the Galaxy, approximately to scale. The figure illustrates that, for T_{thresh} values of 0, 0.3, and 0.6 Gyr, the inner Galaxy has the highest φ_{Iu} at the midplane, as φ_{Iu} is dominated by the number of planets in the region. For these T_{thresh} values we see the influence of SN sterilisations between $r \approx 5$ to $r \approx 9$ kpc, where φ_{Iu} is slightly higher above and below the midplane at these radial positions. For $T_{thresh} = 2$ Gyr, which assumes that the timescale for the rise of intelligence is more than three times that experienced on Earth, from 2.5 kpc to ~12 kpc, φ_{Iu} is always greater above and below the midplane.

We know that intelligent life has arisen at least once at $r = 8$ kpc near the galactic midplane, and there should be an even greater chance that it has arisen in those regions that are shown as “hotter” on the contour map, such as closer to the galactic centre. Furthermore, if intelligence typically takes longer to arise than it has on Earth, the model suggests SETI should prioritise targets above and below the midplane at our radial position and towards the inner Galaxy. However, if intelligence takes roughly the same time as it has on Earth, or less, then the model suggests SETI should target the midplane of the inner Galaxy.

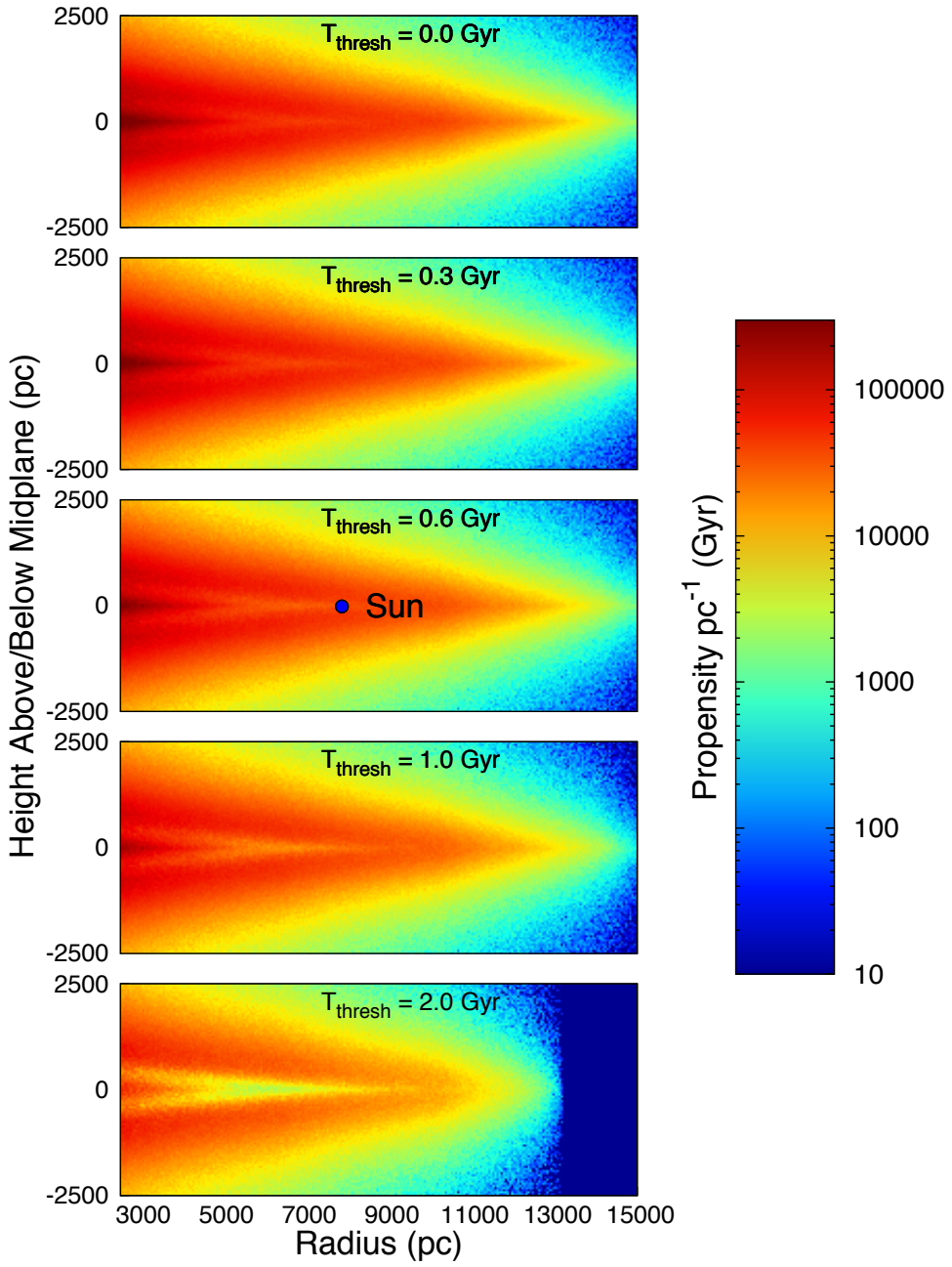

Figure 3.7. Contour map plots of φ_{lu} as a function of r and z. Five separate contour maps are provided, corresponding to T_{thresh} values of 0, 0.3, 0.6, 1, and 2 Gyr, respectively. A logarithmic scale is used for the color-coding, which is defined in the legend given at the right.

3.3.3 Propensity Expressed in Galactic Coordinates

For SETI, it is instructive to consider how φ_{lu} varies as a function of the pointing direction of Earth-based telescopes, specifically the variation of φ_{lu} over *galactic coordinates*. We show

this in Figure 3.8, where φ_{lu} (for the T_{thresh} = 0.6 Gyr case) has been plotted as a function of galactic longitude (l) and galactic latitude (b). A logarithmic color-coding has been used to show the total φ_{lu} per bin of area on the sky, where bins of approximately one square degree have been used across the whole sky[29]. The three panels correspond to range limits from the observer of 3, 4, and 5 kpc, each plotted with the same color-coding range for φ_{lu}. Each panel shows the entire sky in an equal-area sinusoidal projection, as seen from a vantage point of r = 8 kpc and z = 0: Earth's approximate location. As discussed earlier, the central bulge and inner disk (r < 2.5 kpc) are excluded in our model. For this reason, results beyond a range of 5.5 kpc from the observer are incomplete with our model, which is why only ranges below 5.5 kpc have been shown.

Figure 3.8 can be interpreted as showing the relative density of potential targets per antenna pointing as a function of location in the sky. For a range limit of 3 kpc, the density is relatively low, because this represents a small volume of sky that contains relatively few habitable planets. Within this range, there is a minor advantage to observing towards the galactic centre, and slightly above or below the midplane, consistent with our findings of Section 3.3.3. Increasing the range limit to 4 kpc increases the observed volume of sky, and also includes more of the inner Galaxy. Consequently, the number of potential targets is significantly higher. The advantage of observing above/below the midplane remains. At a range limit of 5 kpc, the observed volume of sky is larger again and includes a significant fraction of the inner Galaxy. The density of potential targets is clearly the highest towards the inner Galaxy, in a region bounded roughly by $|l| \leq$ ~30° and $|b| \leq$ ~15°. The advantage of observing slightly above/below the midplane remains, but is now less pronounced, which is consistent with the

[29] This corresponds to ~1° in both l and b at the midplane. At higher latitudes the l range of each bin is increased to maintain roughly the same solid angle on the sky.

fact that more planets are now included that are closer to the galactic centre, where the highest density was found to be on the midplane (see Section 3.3.3).

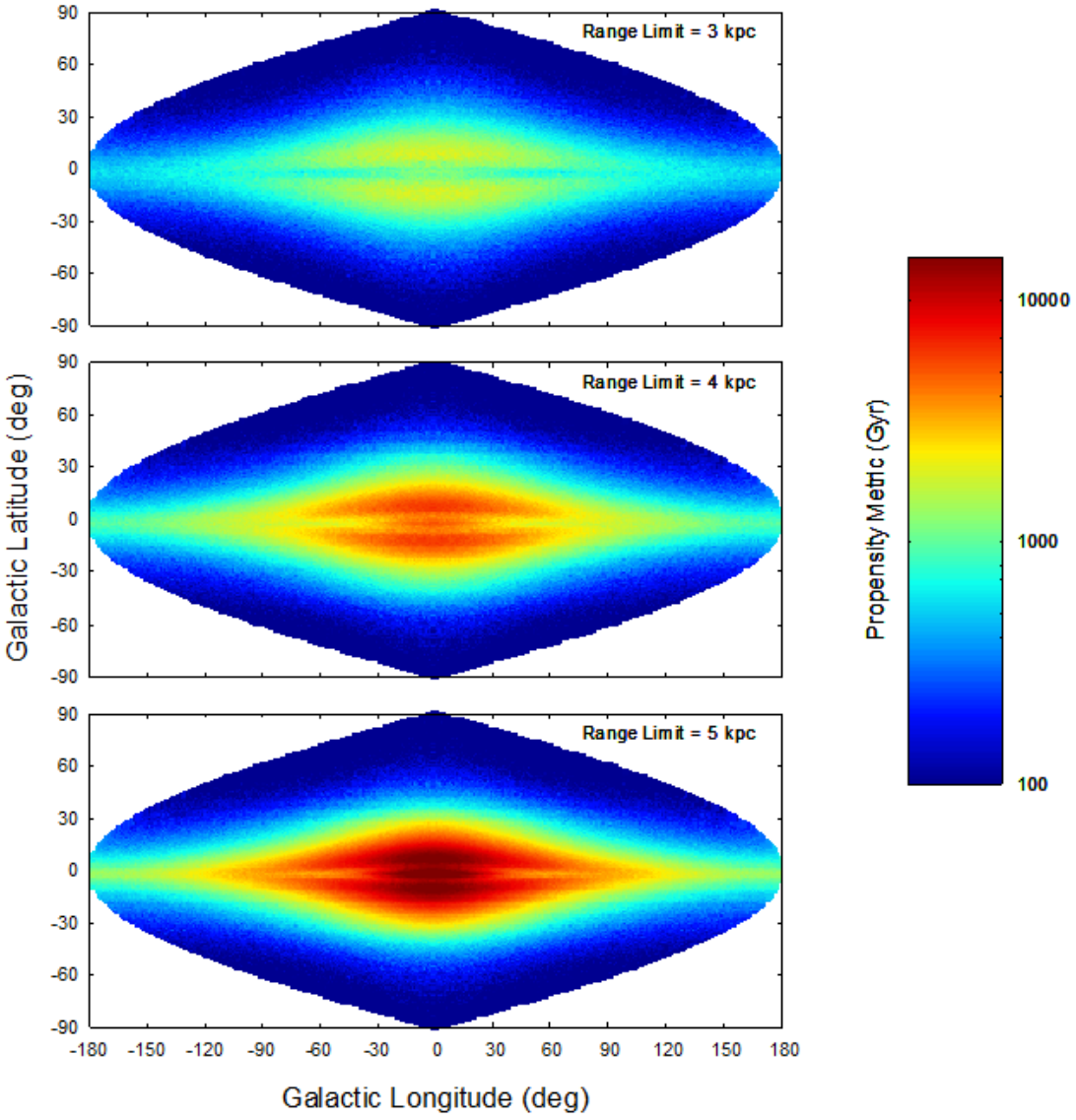

Figure 3.8. Contour map plots of φ_{lu} (for T_{thresh} = 0.6 Gyr) as a function of galactic longitude (l) and latitude (b), for three range limit cases: 3 kpc, 4 kpc, and 5 kpc. Note that the longitude scale shown applies only to b = 0. The plots employ an equal-area sinusoidal projection, where the scale of the horizontal axis varies with latitude, i.e., proportionally to cos(b).

The implication of Figure 3.8 for SETI is that a compelling strategy would appear to be a complete survey of a region of the sky centred on the galactic centre and spanning approximately 60° of longitude and 30° of latitude. Note that the majority of target planets in this region will be close to the galactic centre, so searches should focus on deliberate transmissions[30].

3.3.4 Opportunity Time Distributions

Although the statistical relationship between T_O and φ_I cannot be known precisely for the reasons discussed in Section 3.2.3, it is still possible to make meaningful statements concerning relative propensities in our model. Across the numerous habitable planets in the model, opportunities occur of varying durations, spanning a continuum of T_O values. We create histograms of the distribution of T_O durations in Figure 3.9 for seven different galactocentric radii. A radial bin size of 50 pc is used in each case, with the bin centres as listed in the figure legend.

The first feature to be noted in Figure 3.9 is that shorter T_O occur more frequently than longer T_O. The maximum count occurs at the shortest durations, and the count decreases monotonically with increasing duration. This is expected, as the SN resetting events make longer durations less probable.

The second feature of Figure 3.9 is that the T_O count for a given duration value tends to decrease with increasing r. This is explained by the decreasing habitable planet density with increasing r. Crucially, it is seen that the curves for each radius case do not cross, meaning that

[30] As discussed in Section 3.1, only intentional beacons are likely to be detectable over ranges exceeding a few hundred light years. Hence, for target sources in the vicinity of the inner Galaxy, this requires that SETI search for beacons.

this relationship holds for **all** T_O durations. That is, if one considers a particular T_O duration, then regardless of the duration value, there will always be a greater number of opportunities of that duration toward the inner Galaxy. Regardless of the statistical relationship that exists between T_O and φ_I, there are always more opportunities at each duration value in the inner Galaxy, so the overall φ_I must be higher in this region.

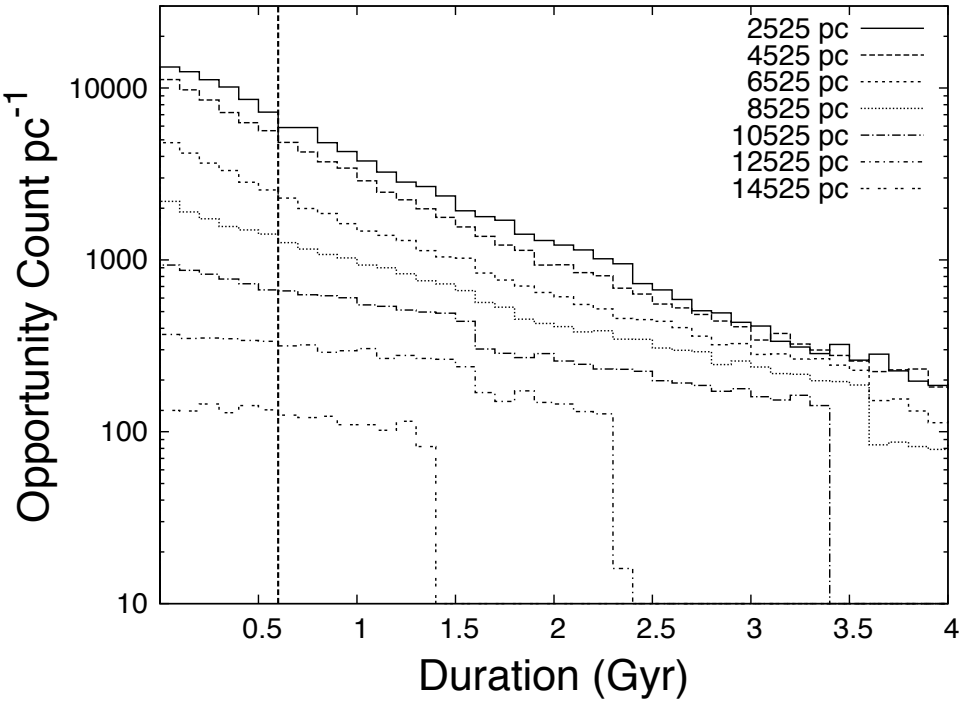

Figure 3.9. Histograms of T_O of differing durations, for seven specific values of *r*. The vertical axis shows the number of opportunities per parsec of radius (on a logarithmic scale) that have the duration given on the horizontal axis. The dashed vertical line crossing the horizontal axis at 0.6 Gyr corresponds to the T_O experienced when intelligence arose on Earth. The cut-offs are a result of the age distribution of stars across the disk, where no planets beyond a given *r* can have an associated duration because they are too young.

This has neatly allowed us to circumvent the calibration issue raised in Section 3.2.3. We may not be able to comment meaningfully on absolute values of φ_I, but we can make the robust assertion that φ_I values are relatively higher toward the inner Galaxy. For example, with reference to Figure 3.9, the opportunity count corresponding to Earth's scenario (T_O = 0.6 Gyr

and r = 8 kpc) is ~1,500 per parsec. At lower radii, toward the inner Galaxy, the opportunity count is seen to be greater than 6,000 per parsec. That is, the inner Galaxy presents more than four times the number of opportunities (of the duration needed on Earth for intelligence to emerge) than the region in which the Earth is located. This provides further support for the conclusion drawn in Sections 3.3.1 and 3.3.2, that is, that there is a greater likelihood that intelligence will arise in the inner Galaxy than at Earth's radius.

A further observation from Figure 3.9 is that, at high radii, there is a hard cut-off in the T_O distributions. Above the cut-off duration there are no opportunities for intelligence to emerge. For example, for r = 14,525 pc, there are no opportunities longer than approximately 1.4 Gyr. This is due to the lower age of planets at higher radii. In the case of 14,525 pc radius, there are no planets in the model that are old enough to provide a gap between SN events greater than (1.55 + 1.4) = 2.95 Gyr.

3.3.5 Opportunities By Epoch

We have seen that the abundance and duration of opportunities, as encapsulated by our metric φ_{lu}, varies with r and z. It is also instructive to investigate how opportunities vary by epoch, that is, as a function of time since the formation of the Galaxy. In Figure 3.10, we plot total opportunity counts versus epoch time for six values of r, for the case of T_{thresh} = 0.6 Gyr. The horizontal axis is the time since the formation of the Galaxy. At any point on the time axis, the corresponding count on the vertical axis represents the total number of habitable planets on which there is currently an "active" opportunity for intelligent life to evolve. Each opportunity contributes to the count value at all time values during the extent of the opportunity.

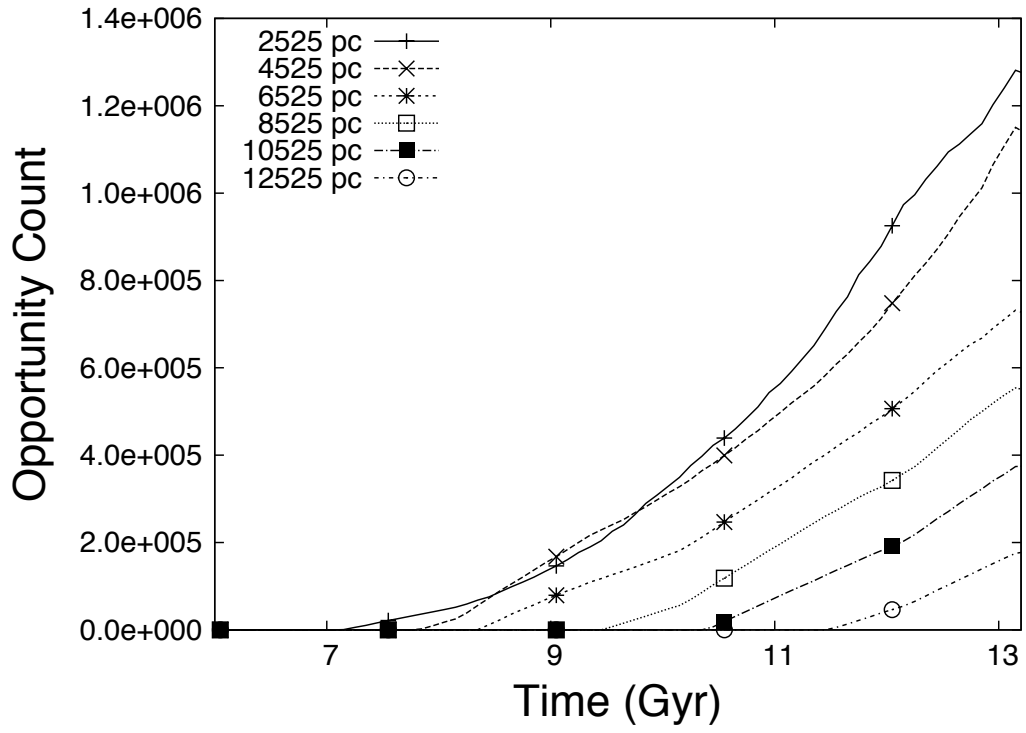

Figure 3.10. Opportunity counts versus epoch time (time since the formation of the Galaxy) for six cases of r and for T_{thresh} = 0.6 Gyr. At any point on the time axis, the count on the vertical axis represents the total number of habitable planets on which there is currently an "active" opportunity.

At all radii, the number of opportunities is seen to increase monotonically with time. For smaller radii, the counts are higher due to the higher number density of habitable planets. Each curve stays at zero for a certain duration before beginning to rise. The time at which the rise begins is earlier at smaller radii, which is due to the higher average age of planets toward the inner Galaxy. At higher radii, the first opportunities for the emergence of intelligence do not occur until later times, once the planet ages reach the necessary threshold.

For all values of radius, the number of active opportunities thus far in galactic history is at its maximum at the *present time*. That is, the likelihood of intelligence emerging is right now the highest it has ever been. Furthermore, the trend of increasing propensity will continue into the future, likely for the next few Gyr, as the metallicity increases throughout the disk, thus

supporting higher planet formation rates. Additionally, contributing to the increasing propensity is the star formation rate, which has not been in significant decline in the past few Gyr, and the fact that more time is available for the development of intelligent life on planets that currently exist in the Galaxy.

As previously mentioned, we focus on modelling planets in the galactic disk and have not considered the galactic bulge due to the complicated formation history and dynamical effects that may be important to consider in this region. Jiménez-Torres et al. (2013) modelled the dynamical effects of stellar flybys on planetary systems in discrete regions and found that different galactic environments may reduce the habitability of planets due to either a) strong gravitational interactions that may perturb planetary systems, or b) weaker interactions that may perturb primordial material left over from planet formation, such as Oort cloud-like objects, which may cause a flux of material to impact the inner planets and potentially a mass extinction event. Since we have ignored the galactic bulge, we have not attempted to model these effects in this work. Stellar flybys may reduce the habitability of planets in different galactic environments, and thus the propensity for intelligent life within the Milky Way.

We note that, while we have only considered the effects of SN sterilisations, there are other events that could decrease habitability in the near future, such as gamma ray bursts (GRBs). Piran and Jimenez (2014) examined the impact of GRBs on galactic habitability and concluded that, as GRBs are more likely to occur in the inner Galaxy and sterilise kpc-scale regions, the outer Galaxy is a better place to find life. We note that there are some assumptions made in their work that will bear further analysis, specifically that Long GRBs are found preferentially in low-metallicity dwarf galaxies, and that the assumption that the low-metallicity members of the disk population of the Milky Way can be equated with the low-metallicity dwarf hosts of GRBs in external galaxies may well not be true. Moreover, their analysis ignores the significant directional beaming of GRBs, which may allow the habitability of large regions of a galaxy in

the vicinity of a GRB to be unaffected. The rate at which GRBs occur is also important to include, since sterilisation by a GRB is not necessarily fatal to life in that region for the rest of Galactic history (as our simulations model for SNe extinctions). Finally, we note that the assumption that GRB rate scales with the stellar density is not dissimilar to the SNe rate scaling with stellar density. Our simulations of the impact of SNe show that, despite the higher SNe rate, the best place to search for intelligence is the inner Galaxy. The results of Piran and Jimenez (2014) suggest that a detailed simulation of the impact of GRBs could be a worthwhile future extension to the habitability models on which the present work is based. However, in the absence of detailed modelling, we can be confident in making the qualitative assertion that GRBs are unlikely to decrease habitability in the Milky Way to levels significantly lower than those currently experienced for two reasons: i) the frequency of such events and their destructive power are not sufficient to significantly inhibit the propensity for intelligence over a large spatial extent; and ii) the general increase over time in the propensity for intelligence (due to increasing planet age) would tend to offset the negative impact of GRBs.

A further observation from Figure 3.10 is that, at the time intelligence arose on Earth (approximately the present time), our model suggests a similar number density of active opportunities was present in the inner Galaxy more than 2 Gyr ago. This does not imply that other civilisations have actually emerged in the inner Galaxy, but it does offer some insight into the potential age of any such civilisations, should they exist.

3.4 Conclusions

A model has been developed to analyse the potential for the development of intelligent life in the Milky Way, one that considers the context of an evolving Galaxy, the formation of planets in this environment, and the occurrence of SN sterilising events that put pressure on the ability

of planets to host intelligent life. We created a metric, φ_{Iu}, to assess the propensity for the emergence of intelligence, and we examined the spatial and temporal variation of φ_{Iu}.

We conclude that the inner Galaxy[31] across all epochs appears to have the highest φ_{Iu}, as a result of the domination in this region of the number density of planets that meet our propensity metric criteria. Even if we vary the expected time for the emergence of intelligence to a value more than three times greater than that which was required on Earth, the inner disk of the Galaxy provides the greatest number of opportunities for intelligence to emerge, despite having a higher SN rate than all other locations in the disk. Further investigation of this relationship suggests that planet locations slightly above and below the midplane may be more favourable than locations precisely at the midplane between $r \approx 5$ to $r \approx 9$ kpc, due to increased exposure to SN events. This effect is more pronounced as the expected time for the emergence of intelligence increases. Interestingly, we find that the average φ_{Iu} per planet at Earth's radial position of $r = 8$ kpc is greater than the inner Galaxy. However, since there are fewer habitable planets at Earth's radial position, the overall value of φ_{Iu} is still lower.

We also find that, at all galactic radii, φ_{Iu} is increasing steadily with time. It is presently the highest it has been in galactic history, and it will continue to rise for several Gyr into the future. Our model provides an estimate of the number of active opportunities for the emergence of intelligence at the present time at Earth's radius. It also shows that a similar number of opportunities were available in the inner Galaxy more than 2 Gyr ago. If any civilisations have emerged in the inner Galaxy, they may be considerably older than our own.

[31] More specifically, the inner Galaxy at $r > 2.5$ kpc, since our model has only been constructed with disk stars at radii between 2.5 and 15 kpc.

While the inner Galaxy has a higher overall propensity for intelligent life, as we have defined in this study, we note that this does not imply any degree of actual inhabitancy. The emergence of life and intelligence may be truly rare events, and their occurrence on Earth may be a statistical outlier. It is possible that no other form of intelligence (or life of any kind) has arisen elsewhere in our Galaxy. However, the alternative – that life and intelligence does exist elsewhere in our Galaxy – is also possible, and the results of this study suggest this may be the more probable scenario. In this regard, our findings can be interpreted as optimistic for the prospects of SETI. They also suggest a high priority should be given to searching in the direction of the galactic centre.

Our work does not provide a means by which to estimate the absolute number of civilisations that may have arisen in the Galaxy or the rate at which new civilisations may emerge in the future. However, we can be confident in asserting that the potential for intelligence to emerge is becoming greater with time. There are likely to be more new civilisations emerging in the future than have emerged in our past.

Acknowledgments

The authors wish to acknowledge the support of the Australian Centre for Astrobiology at the University of New South Wales, and the NASA Astrobiology Institute at the University of Hawaii. This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA08DA77A issued through the Office of Space Science. The authors are grateful for valuable feedback on the manuscript provided by Chris Tinney, David Flannery, Malcolm Walter, Carol Oliver, James Benford, Antonia Rowlinson and the anonymous reviewers. Particular thanks to Chris Tinney for suggesting the inclusion of the galactic coordinate contour maps.

Author Disclosure Statement

No competing financial interests exist.

References

Beckwith, S.V.W. (2008) Detecting life-bearing extrasolar planets with space telescopes. *Astrophys. J.* 684:1404-1415.

Benford, G., Benford, J. and Benford, D. (2010) Searching for cost-optimized interstellar beacons. *Astrobiology* 10:491-498.

Benítez, N., Maíz-Apellániz, J., and Canelles, M. (2002) Evidence for nearby supernova explosions. *Phys Rev Lett* 88:081101.

Binney, J. and Tremaine, S. (2008) *Galactic Dynamics*, 2nd ed., Princeton University Press, Princeton, NJ.

Bishop, S. and Egli, R. (2011) Discovery prospects for a supernova signature of biogenic origin. *Icarus* 212:960-962.

Carigi, L., García-Rojas, J. and Meneses-Goytia, S. (2013) Chemical evolution and the galactic habitable zone of M31. *Revista Mexicana de Astronomía y Astrofísica* 49:253-273.

Carroll, B.W. and Ostlie, D.A. (2006) *An Introduction to Modern Galactic Astrophysics and Cosmology*, Addison-Wesley, San Francisco.

Carter, B. (2008) Five- or six-step scenario for evolution? *International Journal of Astrobiology* 7:177-182.

Catling, D.C., Glein, C.R., Zahnle, K.J., and McKay, C.P. (2005) Why O_2 is required by complex life on habitable planets and the concept of planetary "oxygenation time." *Astrobiology* 5:415-438.

Cavicchioli, R. (2002) Extremophiles and the search for extraterrestrial life. *Astrobiology* 2:281-292.

Ćirković, M. (2004) The temporal aspect of the Drake equation and SETI. *Astrobiology* 4:225-231.

Ćirković, M. (2012) *The Astrobiological Landscape*, Cambridge University Press, Cambridge, UK, Chapter 7: "SETI and its discontents."

Drake, F.D. (2003 September 29) The Drake equation revisited: part I. *Astrobiology Magazine*.

Fischer, D.A. and Valenti, J. (2005) The planet-metallicity correlation. *Astrophys. J.* 622:1102-1117.

Forgan, D.H. (2009) A numerical testbed for hypotheses of extraterrestrial life and intelligence. *International Journal of Astrobiology* 8:121-131.

Forgan, D.H. (2011) Spatio-temporal constraints on the zoo hypothesis, and the breakdown of total hegemony. *International Journal of Astrobiology* 10:341-347.

Forgan, D.H. and Nichol, R.C. (2011) A failure of serendipity: the Square Kilometre Array will struggle to eavesdrop on human-like extraterrestrial intelligence. *International Journal of Astrobiology* 10:77-81.

Forgan, D.H. and Rice, K. (2010) Numerical testing of the Rare Earth hypothesis using Monte Carlo realization techniques. *International Journal of Astrobiology* 9:73-80.

Gehrels, N., Laird, C.M., Jackman, C.H., Cannizzo, J.K., Mattson, B.J., and Chen, W. (2003) Ozone depletion from nearby supernovae. *Astrophys J* 585:1169-1176.

Glade, N. Ballet, P. Bastien, O. (2012) A stochastic process approach of the Drake equation parameters. *International Journal of Astrobiology* 11:103-108.

Gonzalez, G., Brownlee, D. and Ward, P. (2001) The galactic habitable zone: galactic chemical evolution. *Icarus* 152:185-200.

Gould, S.J. (1989) *Wonderful Life: The Burgess Shale and the Nature of History*, W.W. Norton, New York, Chapter 1: "The iconography of an expectation."

Gowanlock, M.G., Patton, D.R. and McConnell, S. (2011) A model of habitability within the Milky Way Galaxy. *Astrobiology* 11:855-873.

Hair, T.W. (2011) Temporal dispersion of the emergence of intelligence: an inter-arrival time analysis. *International Journal of Astrobiology* 10:131-135.

Ida, S. and Lin, D.N.C.. (2005) Toward a deterministic model of planetary formation. III. Mass distribution of short-period planets around stars of various masses. *Astrophys. J.* 626:1045-1060.

Jiménez-Torres, J. J., Pichardo, B., Lake, G., and Segura, A. (2013) Habitability in different Milky Way stellar environments: a stellar interaction dynamical approach. *Astrobiology* 13:491-509.

Jurić, M., Ivezić, Ž., Brooks, A., Lupton, R.H., Schlegel, D., Finkbeiner, D., Padmanabhan, N., Bond, N., Sesar, B., Rockosi, C.M., Knapp, G.R., Gunn, J.E., Sumi, T., Schneider, D.P., Barentine, J.C., Brewington, H.J., Brinkmann, J., Fukugita, M., Harvanek, M., Kleinman, S.J., Krzesinski, J., Long, D., Neilsen, Jr., E.H., Nitta, A., Snedden, S.A. and York, D.G. (2008) The Milky Way tomography with SDSS. I. stellar number density distribution. *Astrophys. J.* 673:864-914.

Kaltenegger, L., Eiroa, C., Ribas, I., Paresce, F., Leitzinger, M., Odert, P., Hanslmeier, A., Fridlund, M., Lammer, H., Beichman, C., Danchi, W., Henning, T., Herbst, T., Léger, A., Liseau, R., Lunine, J., Penny, A., Quirrenbach, A. Röttgering, H., Selsis, F., Schneider, J., Stam, D., Tinetti, G., and White, G.J. (2010) Stellar aspects of habitability – characterizing target stars for terrestrial planet-finding missions. *Astrobiology* 10:103-112.

Kroupa, P. (2001) On the variation of the initial mass function. *Mon Not R Astron Soc* 322:231-246.

Lineweaver, C.H., Fenner, Y. and Gibson, B.K. (2004) The Galactic habitable zone and the age distribution of complex life in the Milky Way. Science, 303:59-62.

Lineweaver, C.H. (2005) Book review: *Intelligent Life in the Universe: From Common Origins to the Future of Humanity*, by Peter Ulmschneider, Springer, 2003. *Astrobiology* 5:658-661.

Maccone, C. (2010) The statistical Drake equation. *Acta Astronaut* 67:1366-1383.

Medvedev, M.V. and Melott, A.L. (2007) Do extragalactic cosmic rays induce cycles in fossil diversity? *Astrophys. J.* 664:879-889.

Melott, A.L. and Thomas, B.C. (2011) Astrophysical ionizing radiation and the Earth: a brief review and census of intermittent intense sources. *Astrobiology* 11:343-361.

Naab, T. and Ostriker, J.P. (2006) A simple model for the evolution of disc galaxies: the Milky Way. *Mon Not R Astron Soc* 366:899-917.

Perryman, M. (2012) The history of exoplanet detection. *Astrobiology* 12:928-939.

Piran, T. and Jimenez, R. (2014) Possible role of gamma ray bursts on life extinction in the Universe. *Phys Rev Lett* 113, doi:10.1103/PhysRevLett.113.231102.

Prantzos, N. (2008) On the "galactic habitable zone." *Space Sci Rev* 135:313-322.

Rothschild, L.J. and Mancinelli, R.L. (2001) Life in extreme environments. *Nature* 409:1092-1101.

Salpeter, E.E. (1955) The luminosity function and stellar evolution. *Astrophys. J.* 121:161-167.

Turnbull, M.C., and Tarter, J.C. (2003) Target selection for SETI. I. A catalog of nearby habitable stellar systems. *Astrophys J Suppl Ser* 145:181-198.

4 Preferred frequency band for interstellar beacons

This chapter reviews past thinking and draws conclusions on preferred frequency bands for SETI, noting that end-to-end system efficiency has not generally been taken into account in past analyses, despite the high energy demands of communicating across interstellar distances. If one accepts the maxim that there will always be competing demands for finite resources, we should expect any extraterrestrial beacon system to be designed for optimum energy efficiency. We saw in Chapter 2 how efficiency considerations lead to favouring wideband signals over narrowband signals for interstellar communications. In this chapter we will see that a desire to maximise efficiency also provides guidance on preferred frequency bands. We put ourselves "in the shoes" of an interstellar beacon builder and develop an end-to-end system model, which shows that efficiency (or cost) is highly dependent on our choice of transmission frequency. A strong argument emerges to favour the high end of the microwave band from ~30 to ~90 GHz – a region of the spectrum that has largely been unexplored by SETI to date.

4.1 Historical thinking: the "cosmic water hole" idea

Cocconi and Morrison's 1959 paper [5] first introduced SETI as a scientific endeavour, and was quick to identify electromagnetic wave propagation as the most suitable technique to communicate information over the vast distances of interstellar space. Further, they constrained the preferred region of the electromagnetic spectrum to the radio band between approximately 1 MHz to 10 GHz, using energy and atmospheric absorption arguments to reject frequencies outside of this range. Aware of the technology limitations of their day, they made a further proposal that SETI should focus on one small segment of the band: the near vicinity of 1.42 GHz, which is the radio emission line of neutral hydrogen. They argued that every astronomer in the Universe would be intensely observing at this frequency, so locating a

beacon source in the vicinity of this frequency would improve the chances of the signal being detected by its intended recipient.

The rationale of Cocconi and Morrison was extended and elaborated by Bernard Oliver, first in the seminal SETI document, the 'Cyclops report' [18], and later in a contribution to a 1977 technical report to NASA [48] (which was later reproduced in [49]). Oliver provided a more detailed rationale for choosing electromagnetic waves over other methods for interstellar communication[33], and supported the conclusions of Cocconi and Morrison that the radio band is preferred. He presented a detailed analysis of the various sources of sky and receiver noise, summarised in plots that are still routinely referred to today, and which we include here as Figure 4-1 and Figure 4-2. These figures plot equivalent thermal noise temperature (a measure of noise power spectral density, i.e. power per unit frequency) that the different noise sources will generate in a telescope's front-end receiver as a function of frequency. They depict what is known as the "microwave window"; the region of the radio spectrum that is least affected by noise and attenuation (i.e. the most transparent to radio waves). Figure 4-1 assumes a receiver located above Earth's atmosphere while Figure 4-2 assumes a receiver on the Earth's surface, which is therefore affected by the molecular absorption characteristics of the atmosphere.

[33] In signalling via electromagnetic waves, the communication message is encoded as electromagnetic fluctuations where the waveform of the signal encapsulates in some way the message information – a process known as 'modulation' [22]. A modulated signal will span a range of frequencies in the electromagnetic spectrum and is generally classified according to its centre frequency and bandwidth (the span of lowest to highest frequencies over which energy content is present in the signal). This signal is transmitted from the source with an antenna, whereupon it propagates through interstellar space (the communications 'channel') before being received at the destination antenna. In the process of propagation, the signal will experience attenuation and distortions. By the time it reaches the receiver, the signal is likely to be weak and its ability to be detected accurately will be affected by various sources of noise and distortion present in any practical receiver implementation [22].

Diagram has been removed due to copyright restrictions

Figure 4-1: Free-space microwave window (credit: B. Oliver et al. [18])

Diagram has been removed due to copyright restrictions

Figure 4-2: Terrestrial microwave window (credit: B. Oliver et al. [18])

On the left of Figure 4-1 and Figure 4-2 the noise is dominated by 'galactic noise', which is the combination of all natural radio emitters in the Galaxy (primarily synchrotron emission), the absolute level of which varies with Galactic latitude, b.

In the left and centre of Figure 4-1 and Figure 4-2 is the isotropic noise of the cosmic microwave background (CMB), which can be modelled as a 'black body radiator' of temperature of 2.725 K[34]. Although the CMB's temperature is constant and is not frequency-dependent, its radiance varies with frequency according to the standard black-body characteristic, increasing approximately with f^2 in what is referred to as the 'Rayleigh-Jeans region' prior to reaching its peak, after which it declines exponentially. The frequency of peak spectral radiance is determined by the black body temperature, and is ~100 GHz in the case of the CMB at 2.725 K. As discussed further in Section 4.4.1, the CMB is isotropic so the equivalent thermal noise it generates in any receiver (with any antenna radiation pattern) will be the same as for an isotropic receiving antenna, which has a power gain that declines as f^2. Hence, in the Rayleigh-Jeans region the equivalent thermal noise density due to the CMB will be flat with frequency. Thermal noise density can be characterised by a temperature T_t according to $N_0 = kT_t$, where k is Boltzmann's constant. In the Rayleigh-Jeans region of the CMB (up to ~10 GHz), the equivalent thermal noise temperature is constant and equal to the CMB temperature. As the CMB radiance begins to flatten as it approaches its peak, the CMB's equivalent thermal noise temperature will begin to decline to less than its black body temperature, reaching ~ 1 K at 100 GHz. As the CMB radiance falls away beyond the peak, the equivalent thermal noise temperature rapidly declines into insignificance. A quantitative expression for the CMB equivalent thermal noise density as a function of frequency is derived in Section 4.5.3.

[34] The CMB temperature has been revised down slightly since Oliver's day as a result of more accurate observations.

To the right of Figure 4-1 and Figure 4-2 is a bound on minimum receiver noise: the 'quantum limit' of 'shot noise' (a source of noise intrinsic to any detector of electromagnetic radiation). Shot noise is discussed further in Section 4.4.

Figure 4-2 includes additional noise due to atmospheric absorption and re-emission above ~10 GHz. Combining these different noise sources, we see that there is a region of minimum total noise temperature, which is approximately between 1 GHz and 10 GHz – the "microwave window" where transmission of interstellar communication signals will be least affected by noise.

In considering the most favourable frequencies for interstellar communications, Oliver [18] further assumed that received signals will experience Doppler drift due to Earth's rotation, requiring a wider bandwidth receiver structure and thus degrading the detection sensitivity. From this he concluded it is favourable to operate at the lower end of the microwave window where absolute Doppler shifts are smaller, thus requiring less widening of receiver bandwidth. This happens to coincide with the location of the spectral lines for H (neutral hydrogen, 1.42 GHz) and OH (hydroxyl radical, 1.662 MHz). Noting that these are the dissociation products of water, Oliver made the suggestion that SETI should focus on this region on the basis that water is a known precursor for life and any intelligent species is likely to appreciate the romanticism of "meeting" at what he called the "cosmic water hole".

Although the elegance of Oliver's thinking is appealing, as we will see in Section 4.5, this rather anthropocentric argument is trumped by more practical energy efficiency considerations. It can in fact be shown that the higher end of the microwave window offers greater end-to-end energy efficiency. In addition, more advanced receiver designs can easily avoid suffering sensitivity loss due to Doppler drift, thus negating Oliver's argument for preferring the lower end of the microwave window.

The Cyclops report [18] cites another reason for preferring the lower end of the microwave window; the lower individual photon energy at lower frequencies. However, when the fundamental limit of sky noise – the CMB – is taken into account, sky noise is found to dominate over receiver shot noise at these frequencies. There is actually no advantage to lower photon energy until shot noise starts to dominate over the CMB, which is from approximately 40 GHz upwards. This is discussed further in Section 4.5.

In addition to the "cosmic water hole", various other "natural frequencies" for SETI have been proposed over the years, including multiples of various molecular hyperfine transition frequencies, where the multiplicative factor is an integer or fundamental constant such as π or *e* [50] [51]. However, as technological capabilities have improved, so has the ability to search over wider ranges of frequencies, and hence speculating about individual frequencies has become less important. Notwithstanding, another natural frequency is presented in Section 4.6 that has a compelling rationale and has hitherto gone almost completely untested by practical searches. It is particularly interesting because it falls within the optimum frequency band suggested by the efficiency analysis presented in the following sections.

4.2 Implicit coordination and the efficiency argument

In an end-to-end interstellar communications system, the transmitter and the receiver share the overall system cost. It is therefore in the interests of both the transmitter and receiver designers to consider end-to-end designs that are efficient in the sense that they achieve the communications objective with minimal use of resources (or cost). For a given resource budget, an efficient design will maximise the amount of information that can be communicated, or will allow a given amount of information to be sent over a greater distance or to a larger number of recipients. There would appear to be no benefit in deliberately designing a system that is inefficient in its use of resources. As pointed out in Section 1.2, this

shared efficiency goal can provide a form of *implicit coordination* between the transmitter and receiver designers [13].

The goal of achieving an efficient end-to-end system design applies to both the construction costs and operational costs of the system. Depending on the circumstances of the designers, one of these cost types may be considered more critical than the other, but both should be considered. The capital cost of construction may be a high barrier to implementing the system, particularly to a civilisation that is at a lower level of technological development. Alternatively, if the system is intended to operate over very long timeframes, the operational cost of the transmitter (which involves high energy consumption) may be the dominant concern. Minimising the transmitter's operational cost translates to minimising the required energy per transmitted information bit. As shown in Section 2.4, this leads to the conclusion that beacon signals can be expected to be wideband. But, as we will explore in this chapter, investing greater capital to build a transmitter that can direct its power in a more focussed way can also reduce operational cost. There is a design trade-off between capital and operational costs, but it will always be advantageous to employ a signalling method that is maximally energy-efficient.

As mentioned in Section 4.1, the method for signalling over interstellar distances that is generally held to be the most energy-efficient is the use of electromagnetic wave propagation. Assuming this approach, achieving efficient interstellar communications may be simplified to satisfying the following three requirements:

1. Operate within a region of the electromagnetic spectrum in which signals propagate with minimal losses and distortions through the interstellar medium (ISM);

2. Operate within a region of the electromagnetic spectrum in which the signal energy can be cost-effectively directed towards the target receiver(s); and

3. Utilise a signalling (i.e. modulation) method that allows a receiver to extract the embedded information from the signal at the lowest possible S/N (i.e. requiring the minimum energy per information bit).

The third of these requirements was discussed in Chapter 2. In this chapter we focus on the first two requirements; those related to the choice of preferred frequency band.

4.3 "Benford Beacons"

The first detailed consideration of efficiency/cost as a driver for the design of interstellar beacons was presented in two landmark papers by J., G. and D. Benford [28] [30]. Their work represented a very clear example of exploring *implicit coordination*, although they did not use that terminology. They analysed the power levels needed to achieve interstellar signalling and the cost implications of building and operating practical beacon transmitters. They pointed out the trade-off between construction and operational costs, i.e. that building an antenna system with higher gain (i.e. that is more directional) will reduce the amount of input power needed to achieve a given Effective Isotropic Radiated Power (EIRP) – the power transmitted in the direction of the target recipient. Furthermore, they showed that the cost of the transmitter should be split roughly equally between the antenna system and the input power source. This leads naturally to the conclusion that highly directional antennas should be utilised, which implies a narrow beam-width. If the beacon-builder then wishes to send their signal to multiple targets in different regions of space, either multiple beams should be generated or a single beam that can be switched in pointing direction. In the latter case, a given target would only "see" the beam for a limited 'dwell time' at a certain repetition rate - analogous to a sweeping lighthouse beam. From this the Benfords suggest that SETI should

expect any interstellar beacon signal to be transient in its nature, rather than constantly observable. This is a logical conclusion in the case of a single beam that has its pointing direction switched. However, with modern array type telescopes, such as the Square Kilometre Array (SKA), it is typical to generate many simultaneous beams of different pointing directions, and ever higher numbers of beams can be expected in the future as the cost of beam-forming electronics decreases. In the multi-beam scenario, a single beacon transmitter antenna system could potentially *constantly* illuminate a large number of target recipients. The implication is that interstellar beacons may or may not appear as transient signals.

Another important aspect of the Benfords' cost analysis is its implications for the preferred frequency of operation for beacons. They pointed out that, since the EIRP of an antenna is frequency dependent, the cost of achieving a given EIRP is not constant across all frequencies. For a given antenna area, the antenna gain is inversely proportional to the square of the operating wavelength, hence the EIRP is proportional to the square of the operating frequency. This suggests there would be a major benefit to operating an interstellar beacon towards the higher end of the microwave window, i.e. towards 10 GHz. However, working against this is antenna cost. The cost per unit area of an antenna will generally increase with frequency as a result of the increased mechanical accuracy and smoothness required to maintain performance as the wavelength of the signal decreases. The Benfords note that, on Earth, it is commonly assumed that antenna cost can be modelled as proportional to f^x, where f is frequency and x is of the order ⅓. However, they did not attempt a quantitative analysis of how this impacts the choice of optimum frequency, which is the aim of the modelling work presented in Section 4.5.

4.4 Fundamental limits

4.4.1 The Cosmic Microwave Background and receiver noise

If we accept a goal of good beacon system design is to operate with maximum energy efficiency, then it is important to understand the fundamental physical limits to what is achievable. On the transmitting side, what matters is maximising EIRP for a given level of input power – which we have just seen is dependent on the choice of frequency. On the receiving side, what matters is maximising the S/N of the receiver – which we know is also frequency dependent, as illustrated by the plots in Section 4.1. S/N is impacted by both sky noise (which cannot be controlled by the receiver designers) and receiver noise, which consists of both thermal noise (which is a function of the quality of the receiver implementation, and which therefore can be controlled) and shot noise (which is a function of frequency – see Section 4.4.2). In the limit, a very high performance receiver solution could virtually eliminate receiver thermal noise. Here on Earth it is already typical for radio telescope front-ends to perform with noise temperatures of a few tens of degrees K. With improved cooling, there is reason to expect future receivers to perform at a few degrees K, or even below 1 K. Reaching ever lower temperatures becomes increasingly more difficult; a case of diminishing returns. Crucially, however, it is not necessary to push too close to absolute zero. Within the microwave window, the fundamental limit on total noise is determined by the CMB at a temperature of 2.725 K. Once the receiver noise has been reduced below this, the overall temperature will asymptote towards 2.725 K. This is the minimum level of noise that the receiver will "see", regardless of how technically advanced its design.

Another important characteristic of the CMB is its isotropy[35]. This means the same noise limit applies to all directions in the sky. Perhaps less obvious is the implication for receive antenna design. Any antenna, whether isotropic or directional, will always collect the same amount of noise power from the CMB, since the power is summed over all directions in its radiation pattern. There is no way to design a microwave antenna that avoids or attenuates the CMB.

4.4.2 Shot noise and the quantum limit to receiver noise

All systems for detecting electromagnetic radiation experience what is known as *shot noise* (also known as *spontaneous emission noise*). This results from the inevitability of having to count multiple photons[36] to make a reliable detection of a signal [48]. The number needed for detection can be reduced by advanced detector design, but can never be less than one. Each time an individual quantum of energy is detected, this causes a step change in the count, which can be viewed as a form of quantisation noise. In the limit, termed the 'quantum limit', this noise is equal to the energy of one quantum at the frequency of the signal, i.e. h*f*, where h is the Planck constant.

For frequencies below 10 GHz, shot noise can be ignored because its contribution to receiver noise is well below that due to the CMB. However, as the frequency increases above 10 GHz, shot noise starts to become a factor, firstly because the CMB noise begins to decline, and secondly because the quantum limit increases proportionally with frequency. At approximately 40 GHz the noise contributions due to the CMB and the quantum limit on shot noise are equal (a fact discussed further in Section 4.6). As seen in Figure 4-1, at ~40 GHz there

[35] The level of anisotropy in the CMB that is postulated to account for the matter distribution of the Universe is on a miniscule scale and can be ignored for the purposes of S/N calculations.

[36] Or equivalently, extract multiple quanta of energy from the signal field.

is a change in curvature of the total effective receiver noise temperature plot; it is no longer flat, but begins to ascend linearly with frequency. This is an important frequency dependence that must be taken into account when identifying the optimum transmission frequency for interstellar beacons.

4.5 Modelling end-to-end system cost

4.5.1 Modelling assumptions

A credible cost model for an end-to-end interstellar beacon system must be based on fundamental assumptions that are defensible and that, as much as possible, avoid anthropocentrism. To this end, the following four assumptions are posited as the starting point for the analysis that follows:

Assumption 1: ***Cost is a concern to beacon builders.*** Capital costs to build a beacon transmitter and the ongoing cost of supplying radio frequency power are significant for galactic-scale beacons. This will lead beacon builders towards designs that achieve their objective at minimal cost and resource usage.

Assumption 2: ***Transmitter antenna cost per unit of aperture area increases with frequency.*** In the case of radio frequency dish-type antennas, cost per unit of effective aperture area (the 'cost coefficient') increases with frequency mainly due to the more demanding build tolerances needed to maintain good antenna efficiency as the wavelength reduces. Other antenna technologies, such as large dipole arrays of the type discussed by Scheffer [52], also suffer an increasing cost coefficient with increasing frequency due to the larger number of antenna elements per unit area and the associated higher cost of the electronics to perform phasing of the elements. The precise relationship

between the cost coefficient and frequency for a given beacon builder will be determined by the specific technology they employ. However, the physics of radio frequency antenna design would suggest that the cost coefficient would always increase with frequency.

Assumption 3: ***Beacon transmitters and target receivers are not confined to planetary surfaces***. Allowing the beacon transmitter and target receivers to be space-based avoids concerns about atmospheric absorption. It is not unreasonable to believe mankind will be capable of constructing radio telescopes of large aperture outside of Earth's atmosphere (e.g. at a Lagrange Point or far side of the Moon) within the next 100 years. The period during which we will be limited to surface-based receivers is very short on cosmic timescales, so this era is likely to be ignored by extraterrestrial beacon builders.

Assumption 4: ***The thermal noise of target receivers can be assumed to be negligible.*** Beacon builders should assume target receivers would operate with near-ideal performance in terms of their system noise temperature. Earth's radio frequency technology is advancing at a rapid pace and system noise temperatures a few degrees above absolute zero should be achievable within the next 100 years, especially for space-based receivers where thermal noise can be readily controlled. Therefore it seems reasonable to assume that other non-thermal noise sources will dominate and determine the achievable sensitivity performance of target receivers.

4.5.2 End-to-end beacon system

We consider an end-to-end interstellar beacon system consisting of a transmitter and a single receiver, separated by interstellar space, as depicted in Figure 4-3. On this figure are labelled

the key performance parameters of the model. For current purposes, we are not interested in absolute values, but only in how these parameters vary as a function of frequency.

At the left side of Figure 4-3 is a transmitter antenna of effective aperture area A_T, which is driven by a signal source of power P_T. Here we assume a 'reflector antenna' as is commonly used on Earth for transmitting signals over large distances because of its high directivity. For this type of antenna, standard antenna theory [53] [54] tells us the gain, G_T, compared to an isotropic antenna, is proportional to (f^2.A_T), and the EIRP in the boresight direction (towards the receiver) is proportional to the power multiplied by the gain[37], i.e. (f^2.A_T.P_T).

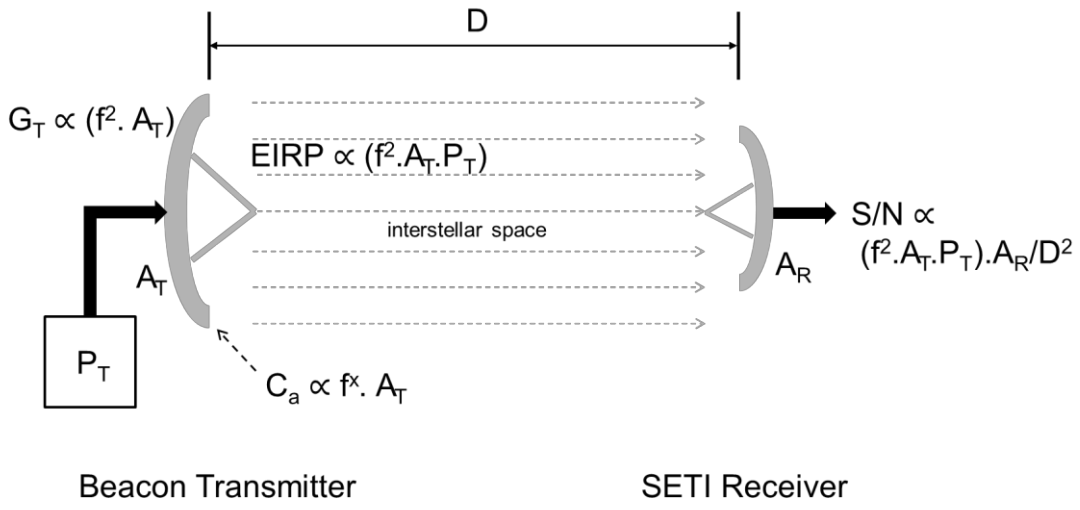

Figure 4-3: Illustrative end-to-end interstellar beacon system, showing the proportionality relationships of key parameters.

The transmitted signal propagates a distance D through interstellar space to the receive antenna at right, which has an effective aperture area A_R. The receiver will have a certain

[37] Other antenna types, including arrays, provide equivalent gain and EIRP relationships as a function of frequency and effective aperture area. Hence the same proportionalities apply generally across a wide range of antenna types.

inherent noise temperature (due to sky noise, and receiver thermal and shot noise), which we can ignore for present purposes[38]. Theory tells us that the signal level at the output of the receiver is proportional to the transmitter EIRP and A_R, and inversely proportional to D^2, i.e. $S \propto \frac{f^2.P_T.A_T.A_R}{D^2}$. For given values of P_T, A_T, A_R and D it is seen that the received signal strength **S scales with f^2**.

4.5.3 Receiver noise as a function of frequency

Adopting Assumption 4 from Section 4.5.1, we assume that receiver thermal noise is insignificant compared to sky or shot noise. Adopting Assumption 3 from Section 4.5.1, we should refer to the free-space noise characteristics shown in Figure 4-1 for the total equivalent thermal noise density as a function of *f*. We can ignore galactic noise because we are only interested in frequencies above 1 GHz. The fundamental limit on total noise density is therefore the sum of the CMB noise density and the quantum limit on shot noise.

The CMB noise density, ψ_b, can be derived from the Planck equation for the spectral radiance of a black-body radiator at temperature *T* (the power emitted per unit area of the black-body, per unit solid angle that the radiation is measured over, per unit frequency) [55]:

$$\boldsymbol{B(f,T) = \frac{2hf^3}{c^2}.\left(\frac{1}{e^{\left(\frac{hf}{kT}\right)}-1}\right)} \qquad (15)$$

in units of $Wm^{-2}st^{-1}Hz^{-1}$, and where *k* is the Boltzmann constant ($1.38x10^{-23}$ JK^{-1}), *h* is the Planck constant (6.63×10^{-34} Js) and c is the speed of light ($3.00x10^8$ ms^{-1}).

[38] For present purposes we can also ignore the various impairments experienced during propagation through the ISM, including dispersion, scattering and scintillation.

As mentioned in Section 4.4.1, the CMB is isotropic so the power density received by an antenna (of any radiation pattern) can be calculated by multiplying Equation (15) by the number of steradians in a sphere (4π) and the effective aperture area of an isotropic antenna, which is given by $A_e = \frac{\lambda^2}{4\pi} = \frac{c^2}{4\pi f^2}$ [54]. In each of two orthogonal polarisations, the antenna will capture half of the total power. Hence the CMB noise density for a single polarisation is given by

$$\boldsymbol{\psi_b = 4\pi.\left(\frac{c^2}{4\pi f^2}\right).\left(\frac{1}{2}\right).\frac{2hf^3}{c^2}.\left(\frac{1}{e^{\left(\frac{hf}{kT_b}\right)}-1}\right)}$$

$$= \frac{\mathrm{h}\boldsymbol{f}}{\boldsymbol{e^{\left(\frac{\mathrm{h}f}{kT_b}\right)}-1}} \qquad (16)$$

where T_b = 2.725 K is the temperature of the CMB. This matches the expression given in [18] and [48], where it was provided without derivation.

The noise density of the quantum limit on shot noise is simply [48]

$$\boldsymbol{\psi_{QL} = hf} \qquad (17)$$

The total noise density, ψ_T, is therefore

$$\boldsymbol{\psi_T = hf + \frac{\mathrm{h}f}{e^{\left(\frac{\mathrm{h}f}{kT_b}\right)}-1}} \qquad (18)$$

which is plotted in Figure 4-4 along with the individual CMB and shot noise components. In this figure the noise densities are expressed as equivalent thermal noise temperatures, obtained by dividing the noise densities by *k*.

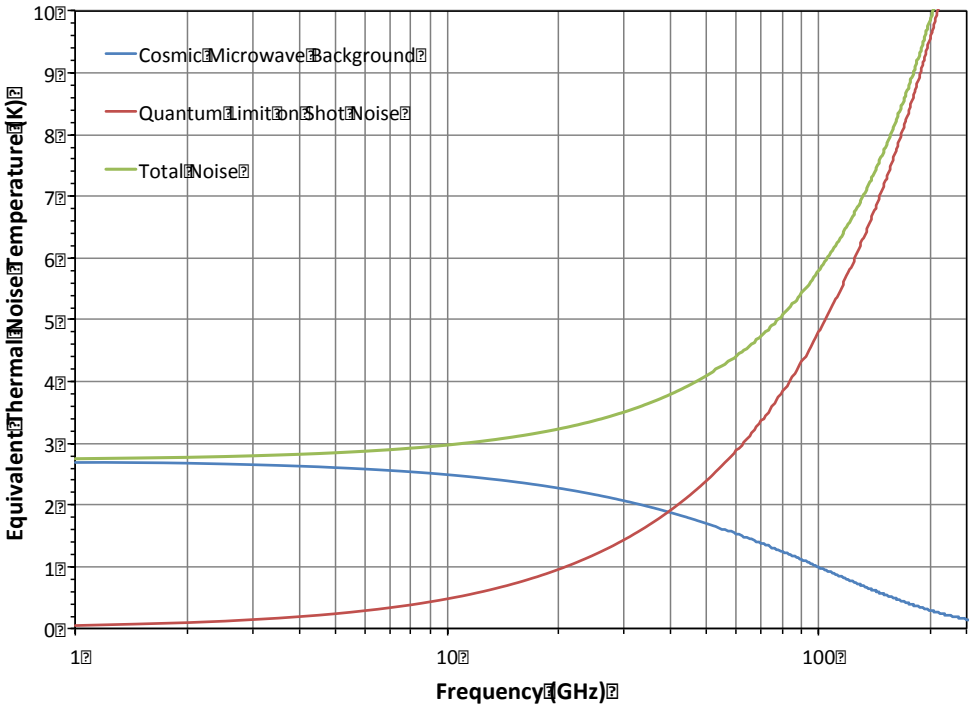

Figure 4-4: Noise densities (expressed as equivalent thermal noise temperatures) as a function of frequency for the CMB and the quantum limit on shot noise. Also shown is their sum, ψ_T, which represents the lower limit on total free-space noise density[39].

It is seen from Figure 4-4 that ψ_b is relatively flat out to approximately 10 GHz when it starts to fall away. At approximately 40 GHz, ψ_b and ψ_{QL} cross and shot noise begins to become the dominant noise source. In this region ψ_T begins to asymptotically approach ψ_{QL} with increasing frequency.

We know from Chapter 2 that there is a minimum threshold S/N that is required for reliable communication of information over the end-to-end 'link'. To achieve this S/N, we will need a sufficient antenna size, A_R, to "capture" sufficient signal in relation to the noise density at the frequency of operation. We note the following:

[39] It can be shown that the first derivative with respect to frequency of ψ_T is positive for all $f > 0$ (and any $T_b > 0$). This means ψ_T is a monotonically increasing function with increasing f.

- The required signal strength at the receiver to achieve the target S/N is nearly **flat with *f* in the CMB region**; and

- The required signal strength at the receiver to achieve the target S/N **scales linearly with *f* in the quantum region**.

Across the whole frequency range, we can apply a common rule: that the required signal strength at the receiver to achieve the target S/N must **track the total noise density** (the sum of CMB and shot noise), as given by Equation (18) and plotted in Figure 4-4.

4.5.4 Antenna cost as a function of frequency

As already explained in Section 4.5.1 Assumption 2, antenna cost per unit of aperture area will increase with frequency. A simple and common way to model the cost is to assume a cost coefficient that is proportional to f^x where f is the centre frequency of the transmitted signal and x is some non-zero exponent. A typical assumption for current Earth technology is an exponent of ~1/3 for reflector antennas at radio frequencies [28]. It is reasonable to assume the exponent for an advanced extraterrestrial beacon builder will be similar or lower, suggesting that exponents from 0.2 to 0.4 would be a reasonable range to consider.

As indicated in Figure 4-3, the cost of an antenna, **C_a, is proportional to (f^x.A)**, where A is the area of the antenna in question, either transmit or receive.

4.5.5 Power system cost as a function of frequency

The antenna cost coefficient increases with frequency, as discussed in Section 4.5.4. In addition, we suggest the signal source power generator system at the transmitter should also be modelled with a frequency-dependent cost-coefficient. Our experience of constructing high-power radio sources on Earth has shown that the output efficiency of such sources tends to decrease with increasing frequency [28]. This occurs due to the physics involved in the high-

power radio generating technologies employed, which include magnetrons, klystrons and travelling-wave tube amplifiers. It is not known conclusively whether this is fundamental to all high power radio generating technologies, or if there may be other technologies, not yet devised on Earth, that avoid this decline in efficiency with frequency. For the present modelling exercise, we make the assumption that such a frequency dependence is present. We further assume that the total transmitter power of the beacon will be generated using a large number (N) of sub-units, and that, for a given size and cost, the output power of each sub-unit will typically decline proportionally to $\frac{1}{f^2}$ [28]. Therefore, the number of sub-units required to maintain the same total power is proportional to f^2. For example, for every doubling of frequency, the required number of sub-units will increase by a factor of 4. To a first order approximation, this would imply that cost is proportional to f^2. However, as pointed out by Benford [56], it is important to consider economies of scale when assessing the cost of manufacturing multiple common sub-units.

As explained in [56], economies of scale can be represented by a *learning curve factor*, which here we will denote *F* (rather than *f*, to avoid confusion with frequency). Depending on the type of product and the technological capabilities of the manufacturer, *F* will typically fall within the range $0.7 < F < 1$.

From [56], the cost of manufacturing N sub-units is as follows:

$$C_N = C_1 . N^{\left(1+\frac{\log F}{\log 2}\right)}$$

Here C_1 is the cost of a single sub-unit, which varies depending on the nature of the product manufactured. For a given product (such as a particular design of radio power sub-unit), it is a constant. The total manufacturing cost for N sub-units will be less than NC_1 whenever $F < 1$.

Assuming N_1 sub-units are required at f = 1 GHz, and since N scales with f^2, we have

$$C_N(f) = C_1.(N_1.f^2)^{\left(1+\frac{\log F}{\log 2}\right)}$$

$$= C_1.N_1^{\left(1+\frac{\log F}{\log 2}\right)}.f^{2\left(1+\frac{\log F}{\log 2}\right)}$$

where $C_N(f)$ is the total cost to manufacture all required sub-units at frequency f. To obtain a cost-coefficient representing the cost per unit of total output power, we normalise to f = 1 GHz by setting $C_N(1) = 1$. Hence the cost-coefficient as a function of frequency is

$$CC(f) = f^{2\left(1+\frac{\log F}{\log 2}\right)}$$

$$= f^z \quad \text{where } z = 2\left(1+\frac{\log F}{\log 2}\right)$$

On Earth it is common to assume a learning curve factor of 0.85 [56]. We may expect a typical extraterrestrial beacon builder to have more sophisticated manufacturing technology than ourselves, and hence a somewhat lower factor would apply. Conveniently, z = 1 corresponds to $F = \frac{1}{\sqrt{2}} = 0.71$, which seems a suitable choice for the nominal case. To assess the sensitivity of the cost modelling to z, we consider z values in the range 0.8 to 1.2, corresponding to F values in the range 0.66 to 0.76.

In summary, the capital cost for a radio power generating system of output power P is modelled as being **proportional to (f^z.P)**, where z is the cost-coefficient exponent which takes values in the range 0.8 to 1.2.

4.5.6 Total system cost as a function of frequency

Model Parameters

In developing the cost model for the complete end-to-end system, we define the following parameters:

- f – the operating frequency in GHz of the beacon transmitter;
- P_T – the output level of the transmitter radio frequency power source;
- C_{op} – the operational power cost of the transmitter, which is proportional to P_T;
- $CC_{ta}(f)$ – the cost-coefficient as a function of frequency for the construction cost of the transmitter antenna system (cost per unit of antenna aperture area), which is modelled as f^x where x is the cost-coefficient exponent, as discussed in Section 4.5.4;
- C_{ta} – the total construction cost of the transmitter antenna system, which is proportional to $(CC_{ta}.A_T)$ where A_T is the aperture area of the transmit antenna;
- $CC_{ra}(f)$ – the cost-coefficient as a function of frequency for the construction cost of the receiver antenna system (cost per unit of antenna aperture area), which is modelled as f^y where y is the cost-coefficient exponent, as discussed in Section 4.5.4;
- C_{ra} – the total construction cost of the receiver antenna system, which is proportional to $(CC_{ra}.A_R)$ where A_R is the aperture area of the receive antenna;
- $CC_{tp}(f)$ – the cost-coefficient as a function of frequency for the construction cost of the transmitter power system (cost per unit of transmitter power), modelled as f^z where z is the cost-coefficient exponent, as discussed in Section 4.5.5;

- C_{tp} – the total construction cost of the transmitter power system, which is proportional to ($CC_{tp}.P_T$);

- EIRP – the transmitter's effective isotropic radiated power in the direction of the receiver, which is proportional to ($f^2.P_T.A_T$), as explained in Section 4.5.2;

- P_R – the received power at the output of the receiver antenna, which is proportional to (EIRP.A_R), as explained in Section 4.5.2;

- S/N(f) – the target S/N at the output of the receiver antenna, given by ($P_R/\psi_T(f)$), where $\psi_T(f)$ is the total receiver noise density as a function of frequency, as described in Section 4.5.3;

- C_T – the total system cost.

Weighting factors

The following three weighting factors are defined, the combinations of which are used to specify different modelling scenarios:

- u – a multiplication factor applied to C_{op} to reflect the degree of concern the beacon builder has regarding operating cost versus capital construction costs. The choice of u < 1 results in operational cost having less influence than construction costs, while u > 1 results in operational cost having more influence than construction costs. A very small u (e.g. 0.01) can be chosen to essentially eliminate operational cost as a concern, while a very large u (e.g. 100) can be chosen to essentially eliminate construction cost as a concern.

- v – the ratio of the construction costs for the transmit antenna and transmit radio power system. It was shown by Benford et al. [28] that the total construction cost for

a beacon transmitter is minimised when the capital cost components for the power source and transmit antenna are equal, i.e. v = 1. Whilst the implemented cost model was capable of using different values of v, for simplicity a value of 1 was used for all the results presented in this chapter.

- w – a cost cap for the construction cost of the receiver system. A small value of w represents the scenario where *transmitter* construction costs dominate (i.e. the receiver cost is not a significant influence on the decision-making of the builder of the beacon transmitter). A large value of w represents the scenario where the builder of the beacon transmitter is primarily concerned with how to minimise the cost burden on the builder of the receiver.

Other modelling decisions

- The same cost-coefficient exponents were used for both the transmit and receive antennas, i.e. x = y. The implemented model was capable of applying different exponents for the transmit and receive cases. However, it was found that using different values did not alter the findings of the modelling exercise[40], so for simplicity, a common exponent value was used.

- Since we are concerned only with the relative variation of the component and total costs as a function of frequency, the various cost-coefficients, $CC_{ta}(f)$, $CC_{ra}(f)$, and $CC_{tp}(f)$, were normalised to value 1 at the lowest frequency considered (1 GHz). They

[40] Using different x and y exponents alters the ratio of A_T to A_R, but if the average of x and y is kept the same, the total cost will be unchanged. Therefore, exploring the effect of larger or smaller exponents can be performed equivalently by varying the average exponent value, or by varying a single common exponent value.

were then scaled with frequency (in GHz) as described in Section 4.5.4 and Section 4.5.5. The target S/N at 1 GHz was also normalised to value 1, and scaled at all other frequencies according to $\psi_T(f)$.

Cost equations

Establishing the overall system cost at a given frequency involved solving the following set of equations:

1. $\mathrm{A_R} = \frac{\mathrm{w}}{\mathrm{CC_{ra}}(f)}$
2. $\mathrm{C_{ta}} = \mathrm{v}.\,\mathrm{C_{tp}}$
3. $f^2.\,\mathrm{P_T}.\,\mathrm{A_T}.\,\mathrm{A_R} = \frac{S}{N}(f)$
4. $\mathrm{C_T} = \mathrm{C_{op}} + \mathrm{C_{tp}} + \mathrm{C_{ta}} + \mathrm{C_{ra}}$

$$= \mathrm{u}.\,\mathrm{P_T} + \mathrm{CC_{tp}}(f).\,\mathrm{P_T} + \mathrm{CC_{ta}}(f).\,\mathrm{A_T} + \mathrm{w}$$

$$= \mathrm{u}.\,\mathrm{P_T} + (1+\mathrm{v}).\,\mathrm{CC_{tp}}(f).\,\mathrm{P_T} + \mathrm{w}$$

$$= \Big(\mathrm{u} + (1+\mathrm{v}).\,\mathrm{CC_{tp}}(f)\Big).\,\mathrm{P_T} + \mathrm{w}$$

The second and third equations together produce a unique solution for P_T. A_T can then be established from the second equation. Once C_T has been calculated for all frequencies across the range of interest, the values are normalised by dividing each value by the minimum value found, thus forcing the minimum of all cost curves to a common value of 1 (we are not interested in absolute costs).

4.5.7 Alternative cost scenarios

Using the weighting factors described in Section 4.5.6, a range of different scenarios can be explored, to determine the implications of varying the balance between (i) operational cost

versus capital cost, and (ii) transmitter cost versus receiver cost[41]. For selected scenarios, the sensitivity to variations in cost-coefficient exponents x, y and z was also investigated. Table 4-1 summarises the scenarios modelled.

Table 4-1: Summary of modelled cost scenarios.

Scenario ID	Scenario Description	u	v	w	x	y	z
A	**Cost-coefficient sensitivity:** low operational cost, high Tx:Rx investment ratio	0.01	1	0.01	0.2, 0.3, 0.4	=x	0.8, 1.0, 1.2
B	**Cost-coefficient sensitivity:** high operational cost, high Tx:Rx investment ratio	100	1	0.01	0.2, 0.3, 0.4	=x	0.8, 1.0, 1.2
C	**Cost-coefficient sensitivity:** low operational cost, low Tx:Rx investment ratio	0.01	1	2	0.2, 0.3, 0.4	=x	0.8, 1.0, 1.2
D	**Rx investment sensitivity:** low operational cost, variable Tx:Rx investment ratio	0.01	1	0.01, 1, 2	0.3	=x	1.0
E	**Rx investment sensitivity:** high operational cost, variable Tx:Rx investment ratio	100	1	0.01, 1, 2	0.3	=x	1.0
F	**Operational cost sensitivity:** mid Tx:Rx investment ratio, variable operational cost	0.01, 1, 10, 30, 100	1	1	0.3	=x	1.0

Tx = transmitter
Rx = receiver

Note that, since v =1 in all cases, varying x and y by ±0.1 will have an identical effect to varying z by ±0.2. For example, using [x=0.2, y=0.2, z=1] will give the same result as [x=0.3, y=0.3, z=0.8], as is shown in Figure 4-5.

[41] As explained in Section 4.5.6, we employ v = 1 for all modelling scenarios, meaning the transmitter capital costs for the power system and antenna system are equated (known to result in the minimum total capital cost). Varying the value of v was found to have a negligible effect on the modelling outcomes.

Scenario A: Cost-coefficient sensitivity: low operational cost, high Tx:Rx investment ratio

This scenario represents the case where the operational cost of the beacon transmitter is considered less of a concern to the transmitter builder than capital costs, and they take on considerably more of the capital cost burden than the receiver. The cost of building the receiver is assumed to be low and it will have little impact on determining the optimum transmitter frequency. Figure 4-5 presents the modelling results expressed as normalised total system cost (C_T) versus frequency.

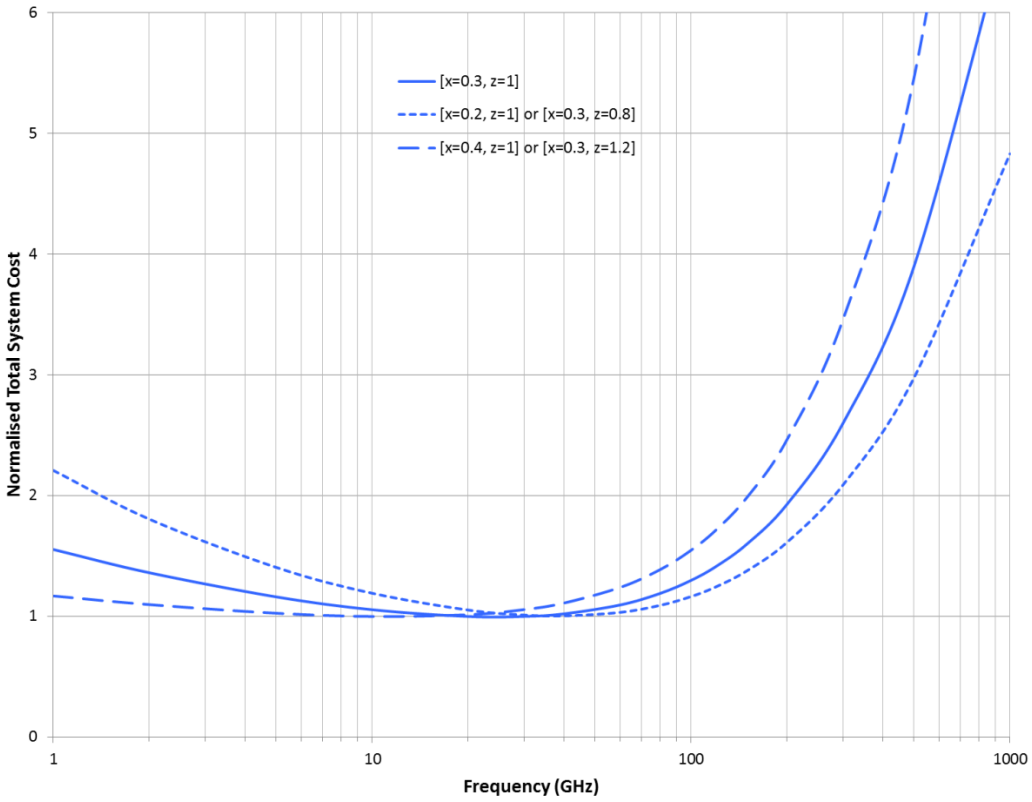

Figure 4-5: Normalised total system cost versus frequency for **Scenario A** [u = 0.01, w = 0.01], with different combinations of the x (= y) and z cost-coefficient parameters.

Regardless of the combination of cost-coefficient parameter values, the curves for total cost all display a convex shape, with a minimum at a similar frequency of ~30 GHz. The different cost-coefficient values only alter the slopes of the curves either side of this minimum.

The observed shape can be explained as follows: starting at the left, the cost is found to decrease as frequency increases, driven by the increasing transmitter EIRP with frequency, while the required S/N remains relatively constant in the CMB region. However, the cost of antenna and power systems is increasing with frequency, and when the required S/N begins to increase due to shot noise, the two effects in conjunction overwhelm the effect of increasing EIRP to produce a total cost that begins increasing above ~30 GHz.

When transmitter capital costs are assumed to dominate over transmitter operational cost, the model suggests there exists an optimum choice of transmitter frequency in the vicinity of 30 GHz.

Scenario B: Cost-coefficient sensitivity: high operational cost, high Tx:Rx investment ratio

This scenario represents the case where the operational cost of the beacon transmitter is considered to be of greater concern than the capital construction costs, while the transmitter builder is still taking on the bulk of the capital cost burden. As with Scenario A, the cost of building the receiver is assumed to be low and it will have little impact on determining the optimum transmitter frequency. Figure 4-6 presents the modelling results for C_T versus frequency.

As with Scenario A, different combinations of cost-coefficient parameter values have a minor effect, and the curves also exhibit a convex shape. A major difference, however, is that the optimum frequency has increased to ~90 GHz. This shift can be explained by the fact that cost is dominated by the transmitter power P_T, so reducing this is more important than constraining the capital costs. The increasing EIRP with frequency continues to drive P_T down, and it is at a much higher frequency that the capital costs and increased S/N requirement combine to overwhelm the effect of increasing EIRP.

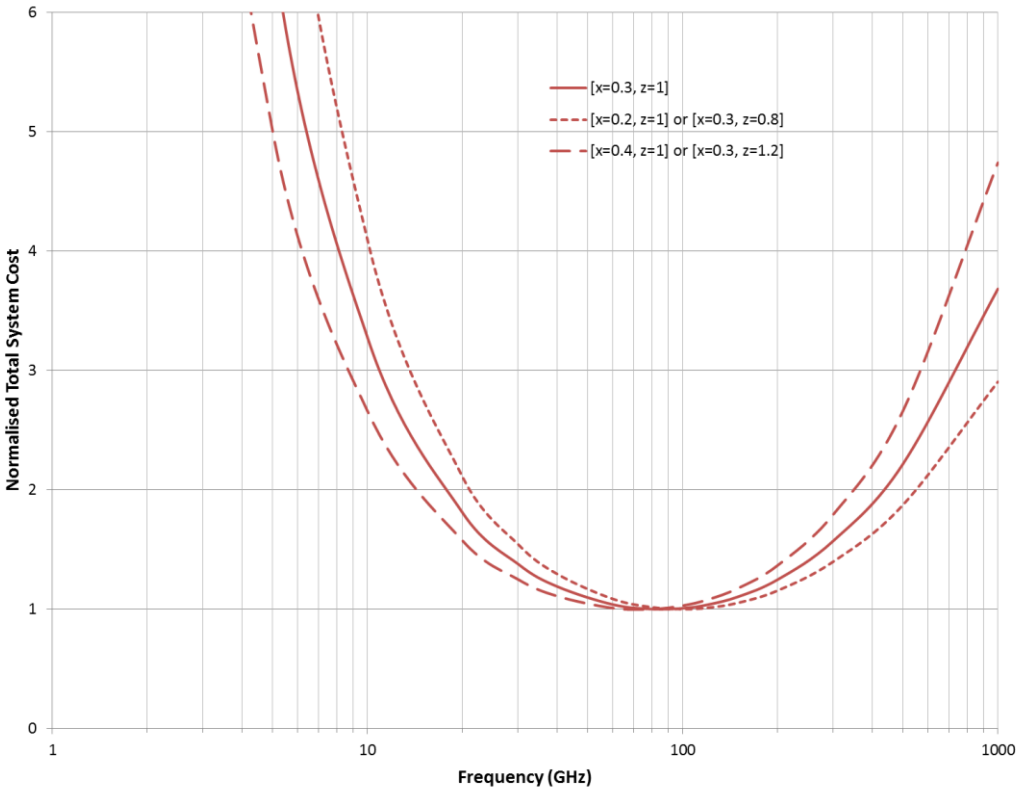

Figure 4-6: Normalised total system cost versus frequency for **Scenario B** [u = 100, w = 0.01], with different combinations of the x (= y) and z cost-coefficient parameters.

When transmitter operational cost is assumed to dominate over transmitter capital costs, the model suggests there exists an optimum choice of transmitter frequency in the vicinity of 90 GHz.

Scenario C: Cost-coefficient sensitivity: low operational cost, low Tx:Rx investment ratio

This scenario represents the case where the operational cost of the beacon transmitter is considered less of a concern to the builder of the transmitter than capital costs, and more of the capital cost burden is taken by the receiver builder. Figure 4-7 presents the modelling results for C_T versus frequency.

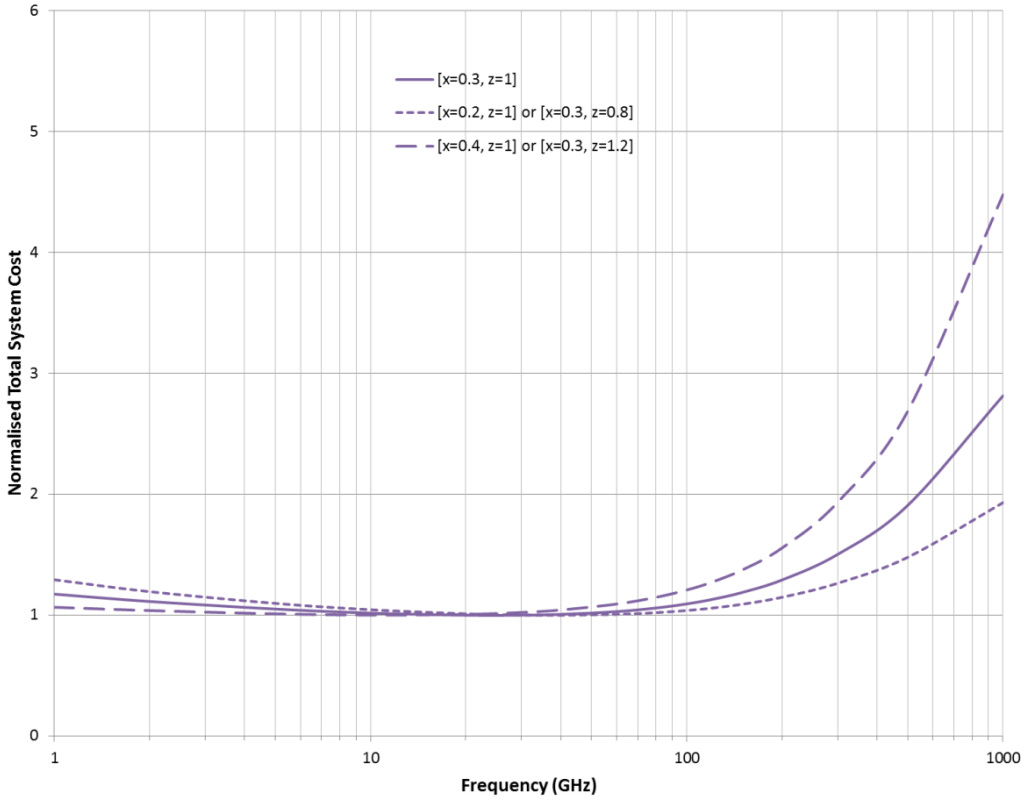

Figure 4-7: Normalised total system cost versus frequency for **Scenario C** [u = 0.01, w = 2], with different combinations of the x (= y) and z cost-coefficient parameters.

The result for Scenario C is very similar to Scenario A, giving the same optimum frequency but producing "flattened" curves. It is apparent that the ratio of operational to capital costs (i.e. u) has more influence on the optimum frequency than the extent to which receiver cost is taken into account (i.e. w). When receiver cost is more of a concern, this adds another substantial component to C_T; a flat value of w across the frequency range. After normalisation, where all C_T values are divided by the minimum value (which is now higher than Scenario A), the same shaped curves are produced, but they exhibit a reduced degree of variation with frequency.

Scenario D: Rx investment sensitivity: low operational cost, variable Tx:Rx investment ratio

This scenario explores the effect of varying the level of concern for receiver cost, as controlled by the parameter w, while regarding transmitter operational cost as a lower concern than the

transmitter capital costs [u = 0.01]. Three values of parameter w were used: 0.01, 1 and 2, representing low, mid and high concern for receiver cost, respectively. This allows a direct comparison to be made of the results from Scenario A [w = 0.01] and Scenario C [w = 2], with the additional case of [w = 1] also included to better observe trends. Figure 4-8 presents the modelling results for C_T versus frequency.

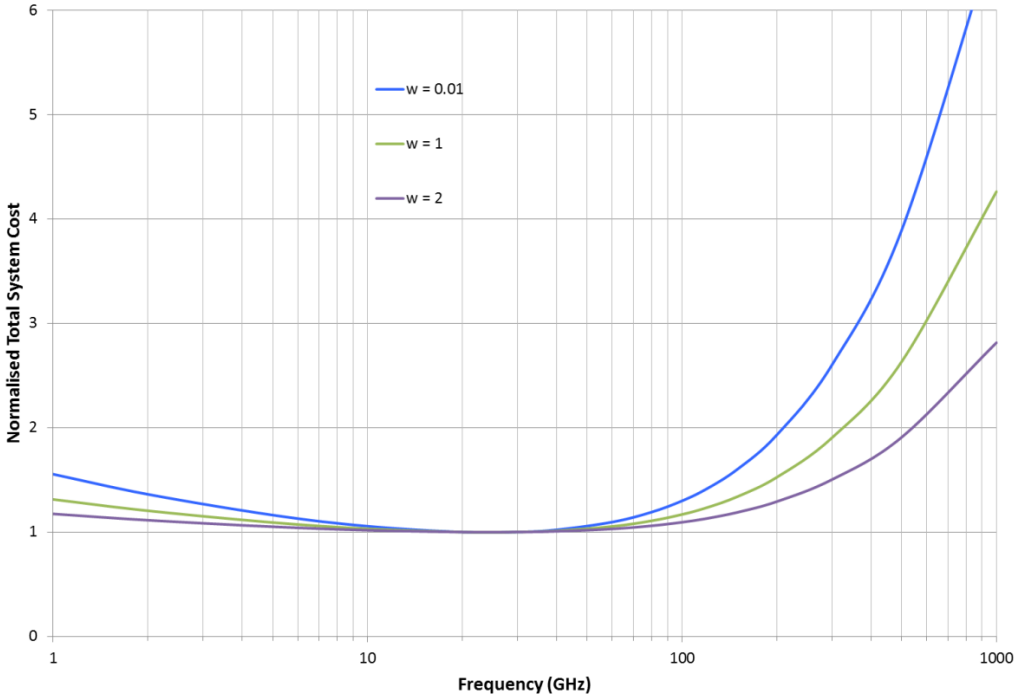

Figure 4-8: Normalised total system cost versus frequency for **Scenario D** [u = 0.01]**.** In this scenario capital cost is more of a concern to the transmitter than operational cost, and different levels of receiver cost have been assumed: [w = 0.01] (low receiver cost), [w = 1] (mid receiver cost) and [w = 2] (high receiver cost).

Figure 4-8 shows clearly the influence of varying the level of receiver cost concern. In all cases the curves have the same shape and indicate the same optimum frequency of ~30 GHz. The only change is a flattening effect as receiver cost increases. As noted previously, this is explained by the normalisation process dividing all C_T values by a larger minimum value when receiver cost is higher.

The significance of this result is that the decision on optimum frequency to be made by the transmitter designer is not influenced by the assumption they make about the level of investment made by the receiver. Whether it is the transmitter or receiver that is to carry most of the cost burden of the end-to-end system, the same optimum frequency is suggested.

Scenario E: Rx investment sensitivity: high operational cost, variable Tx:Rx investment ratio

This scenario is similar to Scenario D but here the assumption is that the transmitter's operational cost is a higher concern than its capital costs [u = 100]. Three values of parameter w were used: 0.01, 1 and 2, representing low, mid and high concern for receiver cost, respectively. This allows a direct comparison to be made of the result from Scenario B [w = 0.01] with the additional cases of [w = 1] and [w = 2]. Figure 4-9 presents the modelling results for C_T versus frequency.

Figure 4-9 reinforces the conclusion from Scenario D – that the degree of concern for receiver cost has no influence on the optimum frequency. However, in this case where transmitter operating cost is the major concern, we find the optimum frequency has increased to ~90 GHz, as was found in Scenario B.

Scenario F: Operational cost sensitivity: mid Tx:Rx investment ratio, variable operational cost

This scenario assumes a balanced partitioning of costs between the transmitter and receiver [w = 1] and explores the effect of varying the level of concern for transmitter operational cost, as controlled by the parameter u. Five values of u were used: 0.01, 1, 10, 30 and 100, spanning a range from low to high concern for operational cost. Figure 4-10 presents the modelling results for C_T versus frequency.

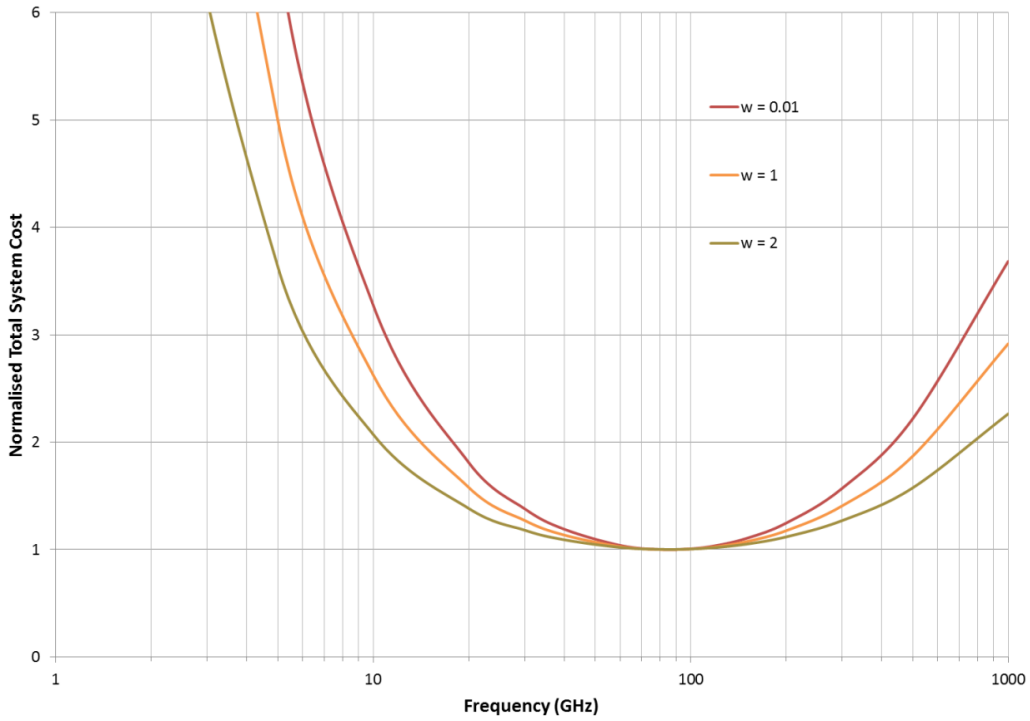

Figure 4-9: Normalised total system cost versus frequency for **Scenario E** [u = 100]**.** In this scenario operational cost is more of a concern to the transmitter than capital cost, and different levels of receiver cost have been assumed: [w = 0.01] (low receiver cost), [w = 1] (mid receiver cost) and [w = 2] (high receiver cost).

Figure 4-10 clearly shows that the optimum frequency is highly dependent on the assumed level of concern for transmitter operational cost. When operational cost is of low concern, this results in a lower optimum frequency of ~30 GHz. As the level of concern for operational cost increases, the optimum frequency is pushed higher, reaching ~90 GHz when operational cost is dominant [u = 100].

Arguably the black curve [u = 10] in Figure 4-10 represents the most balanced scenario considered throughout the whole modelling exercise. It assumes a balance between transmitter and receiver capital costs, and a mid-range level of concern for operational costs. This scenario results in an optimum frequency of ~40 GHz, which is interesting in light of the discussion to come in Section 4.6.

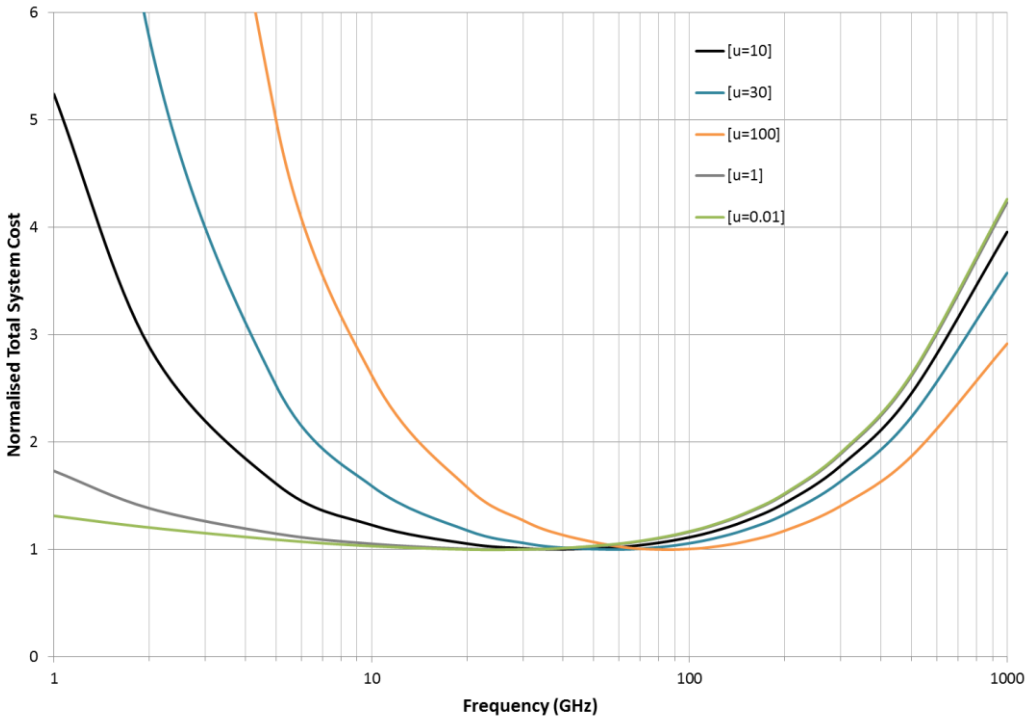

Figure 4-10: Normalised total system cost versus frequency for **Scenario F** [w = 1]**.** In this scenario the transmitter and receiver capital costs are balanced and different levels of transmitter operational cost have been assumed, from [u = 0.01] (low operational cost) to [w = 100] (high operational cost).

It is difficult to say anything concrete about the expected ratio between the operational and capital cost concerns for an extraterrestrial beacon builder. This ratio will depend on many factors, including the beacon builder's level of technological sophistication, the planned operational lifetime of the system, and various economic and political factors. Therefore, it is appropriate to consider the entire band of ~30 to ~90 GHz for the possible locations of interstellar beacons.

Consolidation of model results

We have seen that the cost model produces a range of conclusions for the optimum frequency, depending on parameter assumptions. It is instructive to combine all the results on a single plot for comparison, which has been done with Figure 4-11.

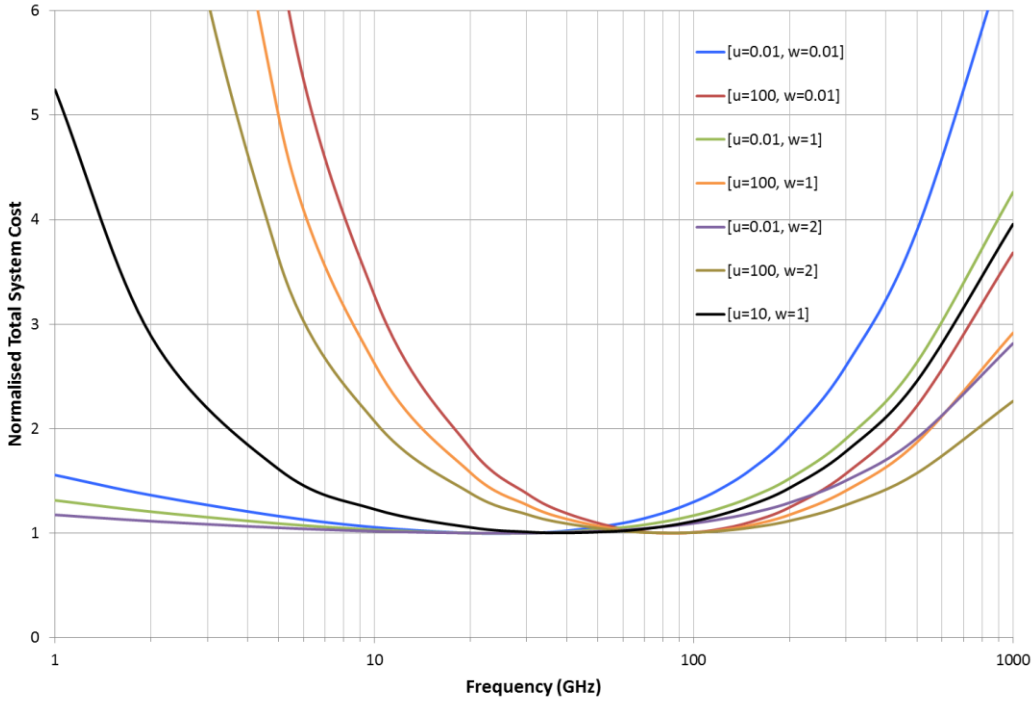

Figure 4-11: Normalised total system cost versus frequency for multiple diverse scenarios.

Although there is a spread of optimum frequencies suggested by the various curves plotted in Figure 4-11, the *envelope* of all the curves is worth considering. In the vicinity of 55 GHz, it is seen that **all** the curves are close to their minimum cost; no more than ~10% above their lowest value. This is interesting because a beacon builder could choose this frequency and be certain of operating close to maximum efficiency, regardless of what assumptions are made about operational cost versus capital cost, or what level of investment will be made by the target receiver. This same "safe bet" conclusion could be reached from cost modelling performed by the transmitter builder and multiple receiver builders. If all parties were to come to a similar conclusion, then this represents a form of implicit coordination. Designing the beacon to operate at a frequency in the vicinity of 55 GHz could make a good choice, as it is a frequency that should be prioritised by the SETI programmes at each target receiver, based on the implicit coordination argument. Here on Earth, it therefore seems reasonable to suggest that SETI should give an increased search priority to the region around 55 GHz.

The interpretation of the modelling results presented so far is suggestive that operating a beacon below 30 GHz would be inefficient and therefore would represent an unlikely choice by any beacon builder. Since the overwhelming majority of past SETI searches have been conducted below 10 GHz, it is worth analysing this conclusion more closely. Referring to Figure 4-11, there is a grouping of curves (those taken from Figure 4-8) that only weakly constrain the optimum frequency to be above 10 GHz. This is scenario D, where capital costs are much more of a concern than operational cost. Such scenarios are certainly conceivable, for example, if energy costs for a civilisation are low compared to material costs, or politico-economic factors make up-front construction costs the primary barrier to establishing a beacon, or the planned operational lifetime for the beacon is short so that operating costs represent a small fraction of total lifetime costs. From one viewpoint, the weak constraint emerging in scenario D may let past searches below 10 GHz "off the hook" to an extent. From another viewpoint, choosing to search below 10 GHz can now be seen to represent an implicit assumption that most civilisations will have capital cost as their dominant concern. This might reasonably be criticised as anthropocentric or, at least, unnecessarily restrictive. Removing this assumption brings all the different modelling scenarios into play, and strengthens the argument for an increased emphasis on frequencies above 10 GHz in future searches.

4.6 The "CMB-QL-intersection" natural SETI frequency

When developing the cost model employed in this chapter, it became apparent that the transition from the CMB-dominated noise region to the shot-noise-dominated noise region has a special significance in determining the optimum choice of beacon frequency. Essentially, it is the combination of this vertex point on the total noise density (ψ_T) curve (shown in Figure 4-4) and the increasing antenna and power system cost coefficients that drives this optimum. The vertex of the ψ_T curve occurs at the frequency where the CMB noise density (ψ_b) and the quantum limit on shot noise (ψ_{QL}) intersect. This frequency is completely determined by

astrophysical parameters, i.e. the temperature of the CMB and the fundamental quanta of electromagnetic energy. As such, it becomes interesting as a possible "natural frequency" for SETI, since it is a unique frequency that any extraterrestrial beacon builder would be able to calculate unambiguously, and which they in turn can assume will be known by all target receiving civilisations. The fact that it lies within the optimum beacon band of ~30 to ~90 GHz makes it particularly interesting.

We propose the "CMB-QL-intersection" natural frequency for interstellar beacons, defined as ***the frequency at which the equivalent thermal noise density due to the CMB equals the equivalent thermal noise density due to shot noise at the quantum limit***[42].

This "new" natural frequency was proposed by the author in a presentation to the Astrobiology Science Conference in 2012 [57]. Later it was pointed out it had already been suggested in a 1973 letter to Nature by Drake and Sagan [58]. They too had been motivated by the universality of this frequency that is "determined simultaneously by quantum mechanics and cosmology". Although they define their natural frequency as "the intersection of two fundamental sources of noise", surprisingly they calculate something different: the frequency at which the equivalent thermal temperature of the quantum limit equals the temperature of the CMB. Based on a CMB temperature of 2.7 K, they calculated their natural frequency to be 56 GHz. However, at 56 GHz the equivalent thermal noise temperature of the CMB is actually significantly lower than the CMB temperature of 2.7 K. Drake and Sagan appear to have made the assumption that the equivalent thermal noise due to the CMB remains flat (and equal to

[42] Interestingly, a different natural frequency (and corresponding wavelength) will be found depending on whether the noise density is defined to be *per-unit-frequency* or *per-unit-wavelength*. Here we choose the per-unit-frequency definition.

the CMB temperature) throughout this part of the spectrum. This is equivalent to assuming the 'Rayleigh-Jeans law' as an approximation to the CMB's black-body radiation. However, this approximation begins to fail around 10 GHz, above which the more precise 'Planck's law' expression for black-body radiation should be used, as given in Equation (15) of Section 4.5.3. The more appropriate value for the natural frequency – and one consistent with how Drake and Sagan define it in the text of [58] – is the frequency at which the *equivalent thermal noise of the CMB equals the equivalent thermal noise of the quantum limit*. This is the frequency at which the CMB noise density and quantum limit curves intersect on Figure 4-4. So, while this author was not the first to propose this natural frequency, the contribution here is the publication of what is arguably a more appropriate value, as calculated below.

The value of the "CMB-QL-intersection" natural frequency, $f_{CMB\text{-}QL}$, is found by equating Equation (16) with Equation (17). Equality is achieved when the denominator of Equation (16) is unity, i.e.

$$\left(e^{\left(\frac{\mathrm{h}f_{CMB-QL}}{kT_b}\right)}-1\right)=1$$

Hence

$$\frac{\mathrm{h}f_{CMB-QL}}{kT_b}=ln2$$

and

$$f_{CMB-QL}=\left(\frac{kln2}{\mathrm{h}}\right)T_b \qquad (19)$$

For the present-day CMB temperature of T_b = 2.725 K, this results in a frequency of

$$f_{CMB-QL}=39.3568\ \mathrm{GHz} \qquad (20)$$

At this point it is worth noting that T_b is not static but has continually decreased since the big bang to its present-day value of 2.725 K. This means that the value of $f_{CMB\text{-}QL}$ has changed over

cosmic time. However, as seen in Figure 4-12, the current rate of change of T_b is low, so both it and $f_{CMB\text{-}QL}$ can be considered static over timescales of millions of years. Certainly over the timeframes involved in communicating within our Galaxy (tens of thousands of years), the present-day value of $f_{CMB\text{-}QL}$ = 39.3568 GHz can be treated as a constant.

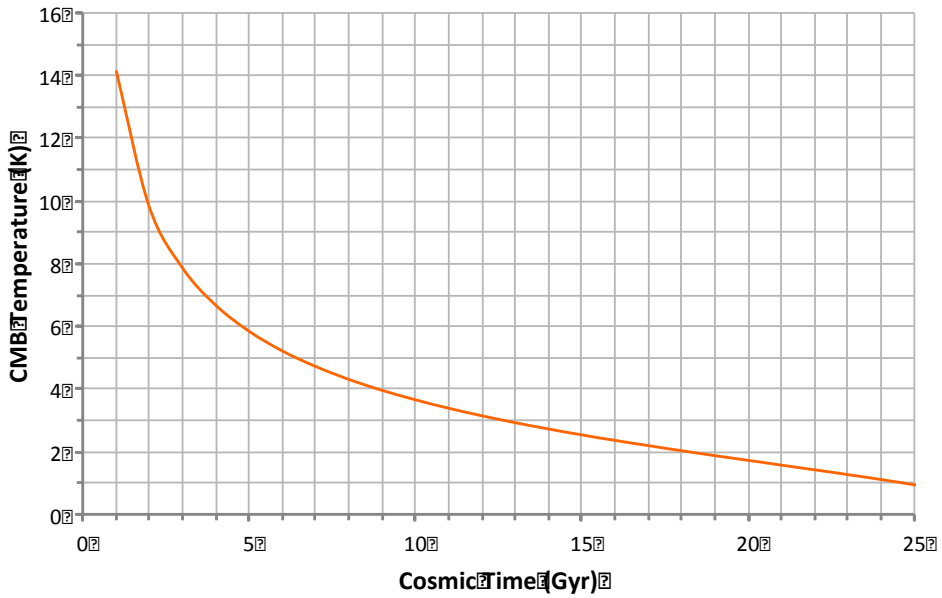

Figure 4-12: The variation in the CMB temperature, T_b, as a function of cosmic time. The present time is ~14 Gyr.

From Equation (19) it is seen that the intersection of ψ_b and ψ_{QL} always occurs at a constant multiplicative factor of T_b. This is because the shape of the ψ_b curve does not change with T_b, only the frequency at which it begins to decline (a result of the peak frequency in a black-body's spectral radiance being proportional to its temperature). The quantum limit is a linear function of frequency, so the frequency for any given shot noise equivalent temperature is also proportional to that temperature. Since both functions scale with temperature, so too does their intersection point. Furthermore, the shape of the ψ_T curve will not change with T_b; it will always monotonically increase with f, as noted in Section 4.5.3, regardless of the value of T_b (i.e. for all time epochs).

4.7 Conclusions and implications for SETI

In conjunction with natural physical laws and characteristics of the Galaxy, it has been shown that the four starting assumptions of Section 4.5.1 taken together can provide guidance as to the optimum frequency band for interstellar beacons.

The key observation from end-to-end modelling is that EIRP increases with the square of frequency, which more than compensates for the linear increase in shot noise with frequency. At first sight this might suggest ever higher frequencies should be preferred. However, the physics and practicalities of constructing radio frequency power sources and antennas results in increasing costs per unit power output and aperture area with increasing frequency. Taken together, there is found to be a minimum in the overall system cost that occurs at a frequency between ~30 GHz and ~90 GHz, depending on the specific technology costs and cost partitioning assumptions made. However, the modelling exercise conducted here demonstrates that there is a particular region of the spectrum (in the vicinity of 55 GHz) that offers near-optimal performance over a wide range of technology and costing assumptions. It is postulated that this makes this part of the spectrum an attractive choice for the transmitter designer who must decide on a transmission frequency without knowledge of the assumptions made by those constructing the target receivers. The existence of an astrophysically-derived natural SETI frequency at ~40 GHz provides another possible choice for priority attention.

If extraterrestrial beacon builders in other parts of the Galaxy were to reach similar conclusions, then the implication for Earth-based SETI would be to focus on searching for beacons in the range 30 to 90 GHz, with priority given to regions in the vicinity of 40 GHz and 55 GHz. It should not be a concern for SETI that our current level of terrestrial technology does not support all of the assumptions made in reaching this conclusion – specifically assumptions 3 and 4 of Section 4.5.1. Neither should the close proximity to the atmospheric O_2 absorption

band around 60 GHz effect our judgement. What matters is the choice we expect to have been made by extraterrestrial beacon builders, who will not be influenced by our anthropocentric concerns.

Ultimately the best sensitivity to signals in the range 30 to 90 GHz will require telescopes above Earth's atmosphere (e.g. located in space or on the far side of the Moon). However, despite the performance limitations of Earth-based telescopes, this author advocates placing increased emphasis on searching in this region of the spectrum, on the basis that there is a defensible scientific argument for believing signals are more likely to be present in this band. Put another way… where is the logic in expending resources searching for beacons in bands where it is known to be inefficient to operate a beacon? SETI to date has tended to follow Dyson's Dictum and search those bands where we have the best search capability – which has meant predominantly the microwave window between 1 and 10 GHz. No beacons have yet been discovered in that part of the spectrum, and, in light of the findings of this chapter, this should not be surprising. If we know we never passed by a particular street-light, why look for our keys under it?

5 Conclusions and recommendations

The discovery space for SETI is vast and multi-dimensional. At the present time, there are insufficient technical and financial resources to allow a comprehensive search covering all the sky, all the time, across all parts of the electromagnetic spectrum, with sensitivity to all types of signal waveforms. In order to make best use of the available resources it is necessary to make choices in regards to prioritising regions of the discovery space. With few exceptions, the traditional choices for SETI have been to intermittently target nearby star systems within the terrestrial microwave window, looking for narrowband tones. This represents a tiny fraction of the total discovery space. Furthermore, these choices have been influenced by anthropocentric biases and have not always been underpinned by impartial scientific reasoning.

This thesis has attempted to provide a more rigorous consideration of appropriate SETI priorities by suggesting constraints on the discovery space that derive from laws of physics and information theory, and our growing knowledge of the Universe through astronomy and cosmology. While it is impossible for an anthropoidal author to completely eliminate anthropocentric logic, every effort has been made to marginalise such thinking and focus on those factors that are likely to be universal (or at least plausibly universal).

The key conclusions from this exercise are summarised as follows:

- The number of potential targets for SETI is likely to be many orders of magnitude higher if the search volume is extended beyond our local stellar neighbourhood to include the inner Galaxy and other nearby galaxies, since there is no robust evidence to indicate our local neighbourhood is preferential for the development of intelligent life;

- Due to the physics of radio propagation and limits to telescope sensitivities on Earth, discovery of radio emissions from extraterrestrial civilisations more than a few hundred light years distance from Earth is likely to be prohibitively difficult unless those emissions have been intentionally transmitted to us, i.e., there are civilisations deliberately attempting to communicate with us by way of interstellar beacon transmissions;

- It is reasonable to assume that interstellar beacon transmissions will contain embedded information, and information theory dictates that communicating with high power efficiency requires the utilisation of wideband signal formats;

- It is reasonable to assume that the waveforms employed in interstellar beacon transmissions will have been designed for efficient propagation through the ISM/IGM, and for straightforward, unambiguous discovery by their intended target recipients;

- Even when designed for ease of discovery, the detection of wideband information-bearing signals at low S/N is extremely challenging and new techniques are required – an example of which is the "SWAC" algorithm presented in this thesis;

- Transmission in the range 30 to 90 GHz would appear to represent an attractive choice for the designer of an interstellar beacon, based on efficiency considerations and taking account of the propagation characteristics of the ISM/IGM.

As a consequence of these conclusions, the following recommendations are suggested to increase SETI's future chances of success:

- Increase the emphasis on observing the inner Galaxy and other nearby galaxies, searching for signals that are beamed intentionally towards Earth;

- Increase the emphasis on observing at frequencies above the terrestrial microwave window, specifically the range 30 to 90 GHz;

- Alongside narrowband signal searches, include wideband signal searches that are sensitive to the broadest possible range of waveform types. Increased resources should be directed toward conceiving, optimising and validating new wideband detection algorithms tailored to the discovery of low S/N signals of extraterrestrial origin.

References

[1] L. J. Rothschild and R. L. Mancinelli, "Life in extreme environments," *Nature,* vol. 409, pp. 1092-1101, 2001.

[2] R. Cavicchioli, "Extremophiles and the Search for Extraterrestrial Life," *Astrobiology,* vol. 2, no. 3, pp. 281-292, 2002.

[3] M. Perryman, "The History of Exoplanet Detection," *Astrobiology,* vol. 12, no. 10, pp. 928-939, 2012.

[4] "The Extrasolar Planets Encyclopaedia," established February 1995 by Jean Schneider, CNRS/LUTH - Paris Observatory, [Online]. Available: http://exoplanet.eu/.

[5] G. Cocconi and P. Morrison, "Searching for Interstellar Communications," *Nature,* vol. 184, no. 4690, pp. 844-846, 1959.

[6] F. D. Drake, "Project Ozma," *Physics Today,* vol. 14, no. 140, 1961.

[7] J. Tarter, "The Search for Extraterrestrial Intelligence (SETI)," *Annual Review of Astronomy and Astrophysics,* vol. 39, pp. 511-548, 2001.

[8] H. P. Shuch, Searching for extraterrestrial intelligence - SETI past, present, and future, Berlin: Springer, 2011.

[9] "SETI Institute," [Online]. Available: www.seti.org.

[10] A. Siemion, J. Benford, J. Cheng-Jin, J. Chennamangalam, J. Cordes, H. Falcke, S. Garrington, M. Garrett, L. Gurvits, M. Hoare, E. J. Korpela, J. Lazio, D. Messerschmitt, I. Morrison, T. O'Brien, Z. Paragi, A. Penny, L. Spitler, J. Tarter and D. Werthimer, "Searching for Extraterrestrial Intelligence with the Square Kilometre Array," in *Advancing Astrophysics with the Square Kilometre Array*, SKA Organisation, 2015.

[11] F. Dyson, “Let's look for life in the outer solar system,” ted.com, February 2003. [Online]. Available: www.ted.com/talks/freeman_dyson_says_let_s_look_for_life_in_the_outer_solar_system/transcript?language=en. [Accessed 8 October 2015].

[12] G. Moore, “Progress in Digital Integrated Electronics,” *IEDM Tech Digest,* pp. 11-13, 1975.

[13] D. G. Messerschmitt and I. S. Morrison, “Design of Interstellar Digital Communication Links: Some Insights from Communication Engineering'',” *Acta Astronautica,* vol. 78, pp. 80-89, 2012.

[14] I. S. Morrison and M. G. Gowanlock, “Extending Galactic Habitable Zone Modeling to Include the Emergence of Intelligent Life,” *Astrobiology,* vol. 15, no. 8, pp. 683-696, 2015.

[15] A. Loeb and M. Zaldarriaga, “Eavesdropping on radio broadcasts from galactic civilizations with upcoming observatories for redshifted 21 cm radiation,” *Journal of Cosmology and Astroparticle Physics,* vol. 2007, 2207.

[16] D. H. Forgan and R. C. Nichol, “A failure of serendipity: the Square Kilometre Array will struggle to eavesdrop on human-like extraterrestrial intelligence,” *International Journal of Astrobiology,* vol. 10, pp. 77-81, 2011.

[17] J. Billingham and J. Benford, “Costs and Difficulties of Interstellar 'Messaging' and the Need for International Debate on Potential Risks,” *JBIS,* vol. 67, pp. 17-23, 2014.

[18] B. M. Oliver (ed.), “Project Cyclops: A Design Study of a System for Detecting Extraterrestrial Intelligent Life,” Stanford/NASA/Ames Research Center Summer Faculty Program in Engineering Systems Design, Technical Report, 1971.

[19] R. J. Cohen, G. Downs, R. Emerson, M. Grimm, S. Gulkis, G. Stevens and J. Tarter, "Narrow polarized components in the OH 1612-MHz maser emission from supergiant OH-IR sources," *Monthly Notices of the Royal Astronomical Society,* vol. 225, pp. 491-498, 1987.

[20] E. O. Brigham, The Fast Fourier Transform, New York: Prentice-Hall, 2002.

[21] D. G. Messerschmitt, "Design for minimum energy in interstellar communication," *Acta Astronautica,* vol. 107, p. 20–39, 2015.

[22] J. Benford and D. Benford, "Power Beaming Leakage Radiation as a SETI Observable," *The Astrphysical Journal,* vol. 825, no. 2, 2016.

[23] G. S. Shostak, "SETI at wider bandwidths, in: G. Seth Shostak (Ed.), Progress in the Search for Extraterrestrial Life," *ASP Conference Series,* vol. 74, p. 74, 1995.

[24] D. G. Messerschmitt, "Interstellar communication: The case for spread spectrum," *Acta Astronautica,* vol. 81, no. 1, p. 227–238, 2012.

[25] D. Torrieri, Principles of spread-spectrum communication systems, 2nd ed., Springer, 2015.

[26] P. F. Clancy, "Some advantages of wide over narrow band signals in the search for extraterrestrial intelligence/SETI/," *J. Br. Interplanet. Soc.,* vol. 33, p. 391–395, 1980.

[27] H. W. Jones, "Optimum signal modulation for interstellar communication," in *Astronomical Society of the Pacific Conference Series*, 1995.

[28] J. Benford, G. Benford and D. Benford, "Messaging with cost-optimized interstellar beacons," *Astrobiology,* vol. 10, no. 5, p. 475–490, 2010.

[29] J. Benford, D. Benford and G. Benford, "Building And Searching For Cost-Optimized Interstellar Beacons," in *Communication with Extraterrestrial Intelligence*, D. A. Vakoch, Ed., New York, SUNY Univ. Press, 2011, pp. 279-306.

[30] G. Benford, J. Benford and D. Benford, “Searching for cost-optimized interstellar beacons,” *Astrobiology,* vol. 10, no. 5, p. 491–498, 2010.

[31] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal,* vol. 27, no. 3, pp. 379-423, 1948.

[32] J. G. Proakis, Digital Communication, McGraw-Hill, 2001.

[33] D. Tse and P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.

[34] I. S. Morrison, “Detection of Antipodal Signalling and its Application to Wideband SETI,” *Acta Astronautica,* vol. 78, pp. 90-98, 2012.

[35] I. Jacobs, “Comparison of M-ary Modulation Systems,” *Bell System Technical Journal,* vol. 46, no. 5, p. 843–864, 1967.

[36] P. A. Fridman, “SETI: The transmission rate of radio communication and the signal's detection,” *Acta Astronautica,* vol. 69, pp. 777-787, 2011.

[37] “Normal Distribution,” Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Normal_distribution. [Accessed 22 April 2016].

[38] G. M. Nita and D. E. Gary, “The generalized spectral kurtosis estimator,” *Monthly Notices of the Royal Astronomical Society: Letters,* vol. 406, no. 1, p. 60–64, 2010.

[39] W. A. Gardner, “The spectral correlation theory of cyclostationary time-series,” *Journal of Signal Processing,* vol. 11, no. 1, 1986.

[40] W. A. Gardner, “Exploitation of spectral redundancy in cyclostationary signals,” *IEEE Signal Processing Magazine,* vol. 8, no. 2, 1991.

[41] F. D. Drake, “Chapter IX: The Radio Search for Intelligent Extraterrestrial Life,” in *Current Aspects of Exobiology*, Pergamon Press, 1965.

[42] G. R. Harp, R. F. Ackermann, S. K. Blair, J. Arbunich, P. R. Backus and J. Tarter, “A new class of SETI beacons that contain information,” in *Communication with Extraterrestrial Intelligence*, State University of New York Press, 2011.

[43] D. R. Lorimer and M. Kramer, Handbook of Pulsar Astronomy, Cambridge University Press, 2005.

[44] C. Maccone, Deep Space Flight and Communications: Exploiting the Sun as a Gravitational Lens, Praxis Publishing, 2009.

[45] G. Burel, “Detection Of Spread Spectrum Transmissions Using Fluctuations Of Correlation Estimators,” in *Proc. IEEE Intelligent Signal Processing and Communication Systems*, Hawaii, USA, 2000.

[46] J. M. Cordes, T. J. W. Lazio and C. Sagan, “Scintillation-induced intermittency in SETI,” *The Astrophysical Journal,* vol. 487, no. 2, pp. 782-808, 1997.

[47] M. G. Gowanlock, D. R. Patton and S. McConnell, “A model of habitability within the Milky Way Galaxy,” *Astrobiology,* vol. 11, pp. 855-873, 2011.

[48] B. M. Oliver, “The Rationale for a Preferred Frequency Band: The Water Hole,” in *SP-419 SETI: The Search for Extraterrestrial Intelligence*, NASA Scientific and Technical Information Office, 1977.

[49] B. M. Oliver, “Rationale for the water hole,” *Acta Astronautica,* vol. 6, pp. 71-79, 1979.

[50] D. G. Blair and M. G. Zadnik, “A List of Possible Interstellar Communication Channel Frequencies for SETI,” *Astronomy and Astrophysics,* vol. 278, pp. 669-672, 1993.

[51] L. Gindilis, V. Davydov and V. Strelnitski, “New "Magic" Frequencies for SETI,” in *Third Decennial US-USSR Conference on SETI*, 1993.

[52] L. K. Scheffer, “A scheme for a high-power, low-cost transmitter for deep space applications,” *Radio Science,* vol. 40, no. RS5012, 2005.

[53] J. D. Kraus, Antennas, 2nd ed., New York: McGraw-Hill, 1988.

[54] H. J. Visser, Antenna theory and applications, John Wiley & Sons, 2012.

[55] M. Planck, The theory of heat radiation, Philadelphia: P. Blakiston's Son & Co., 1914.

[56] J. Benford, "Starship Sails Propelled by Cost-Optimized Directed Energy," *JBIS,* vol. 66, pp. 85-95, 2013.

[57] I. S. Morrison, "Interstellar Beacons Should Transmit at 50 GHz," in *Astrobiology Science Conference (AbSciCon)*, Atlanta, Georgia, USA, 2012.

[58] F. D. Drake and C. Sagan, "Interstellar Radio Communication and the Frequency Selection Problem," *Nature,* vol. 245, pp. 257-258, 1973.

[59] "Variance," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Variance. [Accessed 20 August 2016].

[60] D. Divsalar and M. Simon, "Multiple Symbol Differential Detection of MPSK," *IEEE Transactions on Communications,* vol. 38, no. 3, pp. 300-308, 1990.

Appendix A

D.G. Messerschmitt and I.S. Morrison, “Design of Interstellar Digital Communication Links: Some Insights from Communication Engineering”, Acta Astronautica, vol. 78, pp. 80-89, 2012.

Published paper has been removed due to copyright restrictions

Appendix B

I.S. Morrison, “Detection of Antipodal Signalling and its Application to Wideband SETI”, Acta Astronautica, vol. 78, pp. 90-98, 2012.

Published paper has been removed due to copyright restrictions

Appendix C

Analytical derivation of detector sensitivities

The sensitivity of a detector for can be characterised by its output S/N, S/N_{out}. In terms of evaluating detection miss and false alarm probabilities, the appropriate definition is given by the following expression:

$$\boldsymbol{S/N_{\text{out}}} = \frac{\left(\mathbf{E}[\boldsymbol{D}_{\text{signal+noise}}] - \mathbf{E}[\boldsymbol{D}_{\text{noise}}]\right)^2}{\mathbf{Var}[\boldsymbol{D}_{\text{noise}}]} \quad (21)$$

where D is the detector output metric, E[x] is the expectation of x and Var[x] is the variance of x. It is appropriate to describe the ratio of Equation (21) as an S/N because it takes the form of power over variance, which is consistent with how the S/N of the detector input is defined, as we will see below. Note that the formulation of Equation (21) is only strictly valid in the low input S/N regime – which we presume to be the case for the SETI discovery scenario. When the input S/N is small, Var[$D_{\text{signal+noise}}$] ≈ Var[D_{noise}], so we can use Var[D_{noise}] consistently in the denominator of Equation (21), applying it to both the miss and false alarm cases.

When the detector metric, D, exhibits Gaussian statistics (which we show later in this appendix to be a good approximation for the scenarios under consideration), it can be shown that the detection miss and false alarm probabilities are completely determined by the number of standard deviations between the detector output threshold and the expected detector outputs when the target signal is, respectively, present and not present. That is, the required miss and false alarm probabilities will be achieved if the difference between E[$D_{\text{signal+noise}}$] and E[D_{signal}] equals or exceeds the appropriate multiple of standard deviations.

As an example, consider the case where the desired miss and false alarm probabilities are both to be a maximum of 10^{-3}. To achieve this, ($E[D_{signal+noise}]$ - $E[D_{signal}]$) needs to exceed 6.2 standard deviations, with the detector threshold set mid-way between the two expected values[43]. Squaring this figure provides the corresponding S/N, so in this example the required S/N_{out} is ~38 (or ~16 dB).

It is important to recognise that the detector sensitivity is in general signal dependent because $E[D_{signal}]$ may vary depending on the signalling alphabet. It may also be data-dependent, i.e. dependent on the specific data pattern with which the signal was modulated during the measurement interval. For the purposes of deriving and comparing the sensitivity of different detection algorithms, we will assume in all cases a signal waveform where $E[D_{signal}]$ is time-invariant and known (ignoring channel impairments). For this purpose we choose the binary antipodal spread-spectrum modulation described in Section 2.8.2, which represents a best-case scenario. The sensitivity for other signal classes may be below that of binary antipodal modulation, but it is important to establish upper bounds on achievable performance – and for this purpose binary antipodal modulation is appropriate.

We begin by assuming a transmitted signal $s(t)$ that is sampled at rate W and where:

- Each sample has the same signal amplitude, $\pm s$, and power s^2;
- The total energy in one symbol is: $E_s = s^2.T_s$, where T_s is the symbol period.

[43] The Q-function, $Q(x)$, provides the probability that a standard normal random variable will obtain a value greater than x. Since $Q(3.1) \approx 10^{-3}$, the detection threshold should be set at 3.1 standard deviations above the mean detector output when noise only is present for a 10^{-3} false alarm probability. The expected detector output when signal and noise are present will need to be a further 3.1 standard deviations above this threshold to achieve a 10^{-3} miss probability. Hence, in this example, $E[D_{signal+noise}]$ needs to equal or exceed 6.2 standard deviations above $E[D_{noise}]$.

Assume that $s(t)$ is combined with Gaussian noise $n(t)$ that is white across bandwidth W and has variance σ^2, i.e. the noise power in each sample is σ^2. We take ($s(t)$ + $n(t)$) as the input to the detector.

- The *per-sample* input S/N is given by $S/N_{\mathrm{in}} = \frac{s^2}{\sigma^2}$.
- $\tau = WT_s$ is the number of samples per modulation symbol.
- The *per-symbol* input S/N is given by $\frac{E_s}{N_0} = \frac{s^2 T_s}{\frac{\sigma^2}{W}} = WT_s\left(\frac{s^2}{\sigma^2}\right) = \tau\left(\frac{s^2}{\sigma^2}\right)$.
- We define M as the number of input symbols processed by the detector over an observation time of $T_{obs} = MT_s$.

Without loss of generality we can set s = 1, which will simplify the analysis. We can now use $E_s = T_s$, $S/N_{\mathrm{in}} = \frac{1}{\sigma^2}$ and $\frac{E_s}{N_0} = \frac{\tau}{\sigma^2}$.

We wish to determine how S/N_{out} varies as a function of E_s/N_0 and M for given assumptions for W and T_s. This involves formulating expressions for each of the following terms, and then combining them according to Equation (21):

- $E[D_{signal+noise}]$,
- $E[D_{noise}]$, and
- $Var[D_{noise}]$.

We consider the following detector types, as described in Sections 2.10.1, and 2.10.3 and 2.11:

1. Matched filter – data-aided
2. Matched filter – data-blind (ABS variant)
3. Matched filter – data-blind (SQR variant)
4. Energy detector
5. SWAC – basic (ABS variant)
6. SWAC – basic (SQR variant)

Table C-1 presents the analysis stages in separate rows for each detector type. This tabular arrangement allows the commonalities and points of difference for the different detectors to be seen. The last column in the table provides the final formulae for S/N_{out} as a function of E_s/N_0, W and T_s.

One strong point of commonality between detector types is that they all involve correlating samples of $s(t)$ with samples of another waveform $u(t)$. In the case of matched filtering, $u(t) = h(t)$, the originally transmitted waveform. In the case of energy detection, $u(t) = s(t)$, the received waveform (i.e. a squaring operation). In the case of basic SWAC, $u(t) = s(t+T_s)$, a one-symbol delayed version of the received waveform. We define random variable X to be the result of correlating one-symbol spans of $s(t)$ and $u(t)$, defined as:

$$X = Re\left[\sum_{k=k_0+(n-1)\tau}^{k=k_0+n\tau}(s_k.\bar{u}_k)\right].$$

where $\bar{u}_k$ is the complex conjugate of sample u_k and *Re*[ ] takes only the real component. We assume here the best performance case, which is when phase coherence between symbols has been maintained, i.e. phase shifts between symbols have been tracked and removed, as explained in Section 2.11.1. This ensures the signal component of the complex inter-symbol correlation score will be real, allowing us to ignore the imaginary component. Since the noise power is shared equally between the real and imaginary components, taking only the real component results in a halving of the variance of X, which ultimately improves the detection sensitivity by 3 dB.

Throughout the derivations in Table C-1, we develop expressions for the mean and variance of X, and operations that have been performed on X (specifically $|X|$ or X^2). In all cases, X consists of the sum of τ terms, each with a normal-product distribution (assuming Gaussian noise). According to the central limit theorem, the distribution of X will be approximately

Gaussian with mean and variance equal to τ times the mean and variance of individual sample correlations. In the general case, each individual sample correlation will involve the multiplication of the terms (s_1+N_1) and (s_2+N_2), where N_1 and N_2 are Gaussian noise samples that may be equal or independent, depending on the detector type. For the binary antipodal alphabet considered here, s_k take (normalised) values of ±1 (one of these values for the duration of each symbol, randomly selected for each symbol). In general, the mean of X is obtained as:

$$\mathrm{E}[X_{s+n}] = \tau.(\mathrm{E}[(s_1 + N_1)(s_2 + N_2)])$$

$$= \tau.(\mathrm{E}[s_1.s_2] + \mathrm{E}[s_1.N_2] + \mathrm{E}[s_2.N_1] + \mathrm{E}[N_1.N_2])$$

The first term in brackets is always equal to +1 (if $s_1 = s_2$) or -1 (if $s_1 \neq s_2$). The second and third terms will always be zero, since E[N_1] = E[N_2] = 0. The fourth term will be σ^2 when $N_1 = N_2$, or zero when N_1 and N_2 are independent. In the case of matched filtering, N_2 is effectively set to zero, in which case the fourth term will also be zero.

When N_1 and N_2 are independent, we use the expression for the variance of a product of independent random variables [59] to obtain a general expression for the variance of X:

$$\mathrm{Var}[X_{s+n}] = \tau.(\mathrm{Var}[(s_1 + N_1)(s_2 + N_2)])$$

$$= \tau.(\mathrm{Var}[s_1 + N_1].\mathrm{Var}[s_2 + N_2] + \mathrm{E}[s_1 + N_1]^2.\mathrm{Var}[s_2 + N_2] + \mathrm{E}[s_2 + N_2]^2.\mathrm{Var}[s_1 + N_1])$$

$$= \tau.(\sigma^2.\sigma^2 + (\mathrm{s}_1)^2.\sigma^2 + (\mathrm{s}_2)^2.\sigma^2)$$

$$= \tau.(\sigma^4 + 2\sigma^2)$$

In the case where we take only the real component of the correlation outputs, this has the effect of halving the variance, in which case:

$$\mathrm{Var}[X_{s+n}] = \tau.\left(\sigma^2 + \frac{\sigma^4}{2}\right)$$

In the case of matched filtering (N_2 = 0), we can see from the general expression above that only the final term is non-zero, and hence we will get (after halving):

$$\mathrm{Var}[X_{s+n}] = \tau.\frac{\sigma^2}{2}$$

When there is no signal present, we can see from the general expression above that we will get either:

$$\mathrm{Var}[X_n] = \frac{\tau\sigma^4}{2} \quad \text{when } N_2 \neq 0, \text{or}$$

$$\mathrm{Var}[X_n] = \frac{\tau\sigma^2}{2} \quad \text{when } N_2 = 0.$$

This common set of general expressions for E[*X*] and Var(*X*) is used to derive all the specific expressions employed in Table C-1 for each particular detector type.

Table C-1: Derivations of detector sensitivity formulae

	$\mathbf{E}[D_{\text{signal+noise}}]$	$\mathbf{E}[D_{\text{noise}}]$	$\mathbf{Var}[D_{\text{noise}}]$	$S/N_{\text{out}} = \frac{\left(\mathbf{E}[D_{\text{signal+noise}}] - \mathbf{E}[D_{\text{noise}}]\right)^2}{\mathbf{Var}[D_{\text{noise}}]}$
Matched Filter – data-aided	$= M.\mathrm{E}[X_{s+n}]$ $= M\tau$	$= M.\mathrm{E}[X_n]$ $= 0$	$= M.\mathrm{Var}[X_n]$ $= \frac{M\tau\sigma^2}{2}$	$= \frac{(M\tau)^2}{\frac{M\tau\sigma^2}{2}}$ $= 2M\left(\frac{\tau}{\sigma^2}\right)$ $= 2M\left(\frac{E_s}{N_0}\right)$
Matched Filter – data blind (ABS)	$= M.\mathrm{E}[\lvert X_{s+n}\rvert]$ (Here the expectation term is the 1st non-central absolute moment of X_{s+n}) $= M\sqrt{\frac{2\mathrm{Var}[X_{s+n}]}{\pi}}.\Gamma(1).{}_1F_1\left(-\frac{1}{2},\frac{1}{2},-\left(\frac{1}{2}\right)\left(\frac{\tau^2}{\mathrm{Var}[X_{s+n}]}\right)\right)$ (where ${}_1F_1$ is confluent hypergeometric function [37].) $= M\sqrt{\frac{\tau\sigma^2}{\pi}}.{}_1F_1\left(-\frac{1}{2},\frac{1}{2},-\frac{\tau}{\sigma^2}\right)$ (Since $\mathrm{Var}[X_n] = \frac{\tau\sigma^2}{2}$ in this case.) $= M\sqrt{\frac{\tau\sigma^2}{\pi}}.{}_1F_1\left(-\frac{1}{2},\frac{1}{2},-\left(\frac{E_s}{N_0}\right)\right)$	$= M.\mathrm{E}[\lvert X_n\rvert]$ (Here the expectation term is the 1st central absolute moment of X_n) $= M\sqrt{\frac{\tau\sigma^2}{2}}.\sqrt{\frac{2}{\pi}}$ $= M\sqrt{\frac{\tau\sigma^2}{\pi}}$	$= M.\mathrm{Var}[\lvert X_n\rvert]$ $= M(\mathrm{E}[X_n^2] - \mathrm{E}[\lvert X_n\rvert]^2)$ $= M\left(\mathrm{Var}[X_n] - \left(\sqrt{\frac{\tau\sigma^2}{\pi}}\right)^2\right)$ $= M\left(\left(\frac{\tau\sigma^2}{2}\right) - \left(\frac{\tau\sigma^2}{\pi}\right)\right)$ $= M\tau\sigma^2\left(\frac{1}{2} - \frac{1}{\pi}\right)$	$= \frac{\left(M\sqrt{\frac{\tau\sigma^2}{\pi}}\left({}_1F_1\left(-\frac{1}{2},\frac{1}{2},-\left(\frac{E_s}{N_0}\right)\right) - 1\right)\right)^2}{M\tau\sigma^2\left(\frac{1}{2} - \frac{1}{\pi}\right)}$ $= \frac{M\left({}_1F_1\left(-\frac{1}{2},\frac{1}{2},-\left(\frac{E_s}{N_0}\right)\right) - 1\right)^2}{\left(\frac{\pi}{2} - 1\right)}$

Matched Filter – data blind (SQR)	$= M.\mathrm{E}[X_{s+n}^2]$ (Here the expectation term is the 2nd non-central moment of X_{s+n}) $= M((\mathrm{E}[X_{s+n}])^2 + \mathrm{Var}[X_{s+n}])$ $= M\left(\tau^2 + \frac{\tau\sigma^2}{2}\right)$ $= M\tau\left(\tau + \frac{\sigma^2}{2}\right)$	$= M.\mathrm{E}[X_n^2]$ $= M.\mathrm{Var}[X_n]$ $= \frac{M\tau\sigma^2}{2}$	$= M.\mathrm{Var}[X_n^2]$ $= M(\mathrm{E}[X_n^4] - \mathrm{E}[X_n^2]^2)$ (Here the first expectation term is the 4th central moment of X_n) $= M(3.\mathrm{Var}[X_n]^2 - \mathrm{Var}[X_n]^2)$ $= M\left(2\left(\frac{\tau\sigma^2}{2}\right)^2\right)$ $= \frac{M\tau^2\sigma^4}{2}$	$= \frac{(M\tau^2)^2}{\frac{M\tau^2\sigma^4}{2}}$ $= 2M\left(\frac{\tau^2}{\sigma^4}\right)$ $= 2M\left(\frac{E_s}{N_0}\right)^2$
Energy Detector	$= M.\mathrm{E}[X_{s+n}^2]$ (Here the expectation term is the 2nd non-central moment of X_{s+n}) $= M((\mathrm{E}[X_{s+n}])^2 + \mathrm{Var}[X_{s+n}])$ $= M(\tau + \tau\sigma^2)$ $= M\tau(1 + \sigma^2)$	$= M.\mathrm{E}[X_n^2]$ $= M.\mathrm{E}[N^2]$ $= M\tau\sigma^2$	$= M\tau.\mathrm{Var}[N^2]$ $= M\tau(\mathrm{E}[N^4] - \mathrm{E}[N^2]^2)$ (Here the first expectation term is the 4th central moment of N) $= M\tau(3.\mathrm{Var}[N]^2 - \mathrm{Var}[N]^2)$ $= M\tau(2(\sigma^2)^2)$ $= 2M\tau\sigma^4$	$= \frac{(M\tau)^2}{2M\tau\sigma^4}$ $= \frac{M}{2\tau}\left(\frac{\tau^2}{\sigma^4}\right)$ $= \frac{M}{2WT_s}\left(\frac{E_s}{N_0}\right)^2$

SWAC – basic (ABS)	$= M.\mathrm{E}[\lvert X_{s+n}\rvert\,]$ (Here the expectation term is the 1st non-central absolute moment of X_{s+n}) $= M\sqrt{\frac{2\mathrm{Var}[X_{s+n}]}{\pi}}.\Gamma(1).{}_1F_1\left(-\frac{1}{2},\frac{1}{2},-\left(\frac{1}{2}\right)\left(\frac{\tau^2}{\mathrm{Var}[X_{s+n}]}\right)\right)$ (where ${}_1F_1$ is confluent hypergeometric function [37].) $= M\sqrt{\frac{\tau(2\sigma^2+\sigma^4)}{\pi}}.{}_1F_1\left(-\frac{1}{2},\frac{1}{2},\frac{-\tau}{(2\sigma^2+\sigma^4)}\right)$ (Since $\mathrm{Var}[X_n]=\tau.\left(\sigma^2+\frac{\sigma^4}{2}\right)$ in this case.) $= M\sqrt{\frac{\tau(2\sigma^2+\sigma^4)}{\pi}}.{}_1F_1\left(-\frac{1}{2},\frac{1}{2},\left(\frac{-1}{2\left(\frac{E_s}{N_0}\right)^{-1}+\tau\left(\frac{E_s}{N_0}\right)^{-2}}\right)\right)$	$= M.\mathrm{E}[\lvert X_n\rvert\,]$ (Here the expectation term is the 1st central absolute moment of X_n) $= M\sqrt{\frac{\tau\sigma^4}{2}}.\sqrt{\frac{2}{\pi}}$ $= M\sqrt{\frac{\tau\sigma^4}{\pi}}$	$= M.\mathrm{Var}[\lvert X_n\rvert]$ $= M(\mathrm{E}[X_n^2]-\mathrm{E}[\lvert X_n\rvert]^2)$ $= M\left(\mathrm{Var}[X_n]-\left(\sqrt{\frac{\tau\sigma^4}{\pi}}\right)^2\right)$ $= M\left(\tau\left(\sigma^2+\frac{\sigma^4}{2}\right)-\left(\frac{\tau\sigma^4}{\pi}\right)\right)$ $= M\tau\left(\sigma^2+\sigma^4\left(\frac{1}{2}-\frac{1}{\pi}\right)\right)$	$= \frac{M\left(A\sqrt{(2\sigma^2+\sigma^4)}\ -\sigma^2\right)^2}{\pi\sigma^2+\left(\frac{\pi}{2}-1\right)\sigma^4}$ $= \frac{M\left(A\sqrt{2\left(\frac{E_s}{N_0}\right)^{-1}+WT_s\left(\frac{E_s}{N_0}\right)^{-2}}-\sqrt{WT_s}\ \left(\frac{E_s}{N_0}\right)^{-1}\right)^2}{\pi\left(\frac{E_s}{N_0}\right)^{-1}+WT_s\left(\frac{\pi}{2}-1\right)\left(\frac{E_s}{N_0}\right)^{-2}}$ where $A = {}_1F_1\left(-\frac{1}{2},\frac{1}{2},\left(\frac{-1}{2\left(\frac{E_s}{N_0}\right)^{-1}+WT_s\left(\frac{E_s}{N_0}\right)^{-2}}\right)\right)$
SWAC – basic (SQR)	$= M.\mathrm{E}[X_{s+n}^2]$ (Here the expectation term is the 2nd non-central moment of X_{s+n}) $= M((\mathrm{E}[X_{s+n}])^2+\mathrm{Var}[X_{s+n}])$ $= M\left(\tau^2+\tau\left(\sigma^2+\frac{\sigma^4}{2}\right)\right)$ $= M\tau\left(\tau+\sigma^2+\frac{\sigma^4}{2}\right)$	$= M.\mathrm{E}[X_n^2]$ $= M.\mathrm{Var}[X_n]$ $= \frac{M\tau\sigma^4}{2}$	$= M.\mathrm{Var}[X_n^2]$ $= M(\mathrm{E}[X_n^4]-\mathrm{E}[X_n^2]^2)$ (Here the first expectation term is the 4th central moment of X_n) $= M(3.\mathrm{Var}[X_n]^2-\mathrm{Var}[X_n]^2)$ $= M\left(2\left(\frac{\tau\sigma^4}{2}\right)^2\right)$ $= \frac{M\tau^2\sigma^8}{2}$	$= \frac{\left(M\tau(\tau+\sigma^2)\right)^2}{\frac{M\tau^2\sigma^8}{2}}$ $= 2M\left(\frac{\tau^2}{\sigma^8}+\frac{2\tau}{\sigma^6}+\frac{1}{\sigma^4}\right)$ $= \frac{2M}{\tau^2}\left(\left(\frac{E_s}{N_0}\right)^4+2\left(\frac{E_s}{N_0}\right)^3+\left(\frac{E_s}{N_0}\right)^2\right)$ $= \frac{2M}{(WT_s)^2}\left(\left(\frac{E_s}{N_0}\right)^4+2\left(\frac{E_s}{N_0}\right)^3+\left(\frac{E_s}{N_0}\right)^2\right)$