Wireless Communication and Localization Systems under Spatial and Temporal Channel Variations

by

Peter Anthony Iannucci

B.S., Massachusetts Institute of Technology (2011)
S.M., Massachusetts Institute of Technology (2013)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2019

© Peter Anthony Iannucci, MMXIX. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Author . . . . . .
Department of Electrical Engineering and Computer Science
January 31, 2019

Certified by . . . . . .
Hari Balakrishnan
Professor of Computer Science
Thesis Supervisor

Accepted by . . . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Wireless Communication and Localization Systems under Spatial and Temporal Channel Variations

by

Peter Anthony Iannucci

Submitted to the Department of Electrical Engineering and Computer Science
on January 31, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Wireless signals inevitably vary in time and space. The three chapters of this dissertation revolve around the exploitation of signal variations. This line of work has yielded new link-layer protocols for rateless codes on half-duplex additive white Gaussian noise channels; a new abstraction for short-range mobile-to-mobile and mobile-to-infrastructure "room-area" networks that adhere to the spatial boundaries of human conversation; a reduced-complexity tone reservation algorithm for optimizing signals to avoid amplifier non-linearities; and new tools for the study of physical-layer privacy and anonymity in wireless systems.

Thesis Supervisor: Hari Balakrishnan
Title: Professor of Computer Science

Acknowledgments

Portions of Chapter 2 are based on joint work with Jonathan Perry, Hari Balakrishnan, and Devavrat Shah, which appeared in the Annual International Conference on Mobile Computing and Networking, MobiCom 2012.

Portions of Chapter 3 are based on joint work with Ravi Netravali, Ameesh Goyal, and Hari Balakrishnan, which appeared in the 14th ACM Workshop on Hot Topics in Networks, HotNets-XIV 2015.

Portions of Chapter 4 are based on joint work with Hari Balakrishnan which appeared in ACM WiSec 2017.

The integration of this work would not have been possible without the guidance of my thesis committee, Profs. Hari Balakrishnan, Muriel Médard, and Gerald Sussman.

I also extend my deepest gratitude to Joel Emer, Bob Iannucci, Dina Katabi, Mehrdad Khani, Alexandre Megretski, Jameson Nash, Yury Polyanskiy, Hariharan Rahul, Madars Virza, Keith Winstein, and many others for lessons, thoughtful comments, and helpful conversations over the years.

This work was supported in part by the industrial partners of the MIT Center for Wireless Networks and Mobile Computing (Wireless@MIT) and the National Science Foundation.

In memory of Professor David Staelin, who helped me to bridge my studies in physics, my training as an amateur radio operator, and my work in electrical engineering and computer science.

`73 DE K6PAI`

Contents

1 Introduction **19**

2 Rateless Link-Layer Protocol **29**

2.1 Terminology and Background 30

2.1.1 Wireless Communications, Coding, and Capacity 30

2.1.2 Rateless Coding from a Systems Perspective 40

2.2 Problem 45

2.3 Marginalized-τ model 46

2.4 Notation 48

2.5 Related work 48

2.5.1 Stop-Only Feedback 49

2.5.2 RLNC and Marginalized-τ 50

2.5.3 Scheduling 50

2.6 Construction and Implementation 51

2.6.1 Truncating the Infinite Recursion 54

2.6.2 Pseudocode 60

2.6.3 Complexity 60

2.7 Learning the Decoding CDF 62

2.8 Packing Packets and Block ACKs 65

2.9 Precise Stop Feedback 67

2.10 Evaluation 69

2.10.1 Baselines for Comparison 70

2.10.2 Baseline comparison to ARQ, Try-after-n HARQ 71

2.10.3 Slow-fading Rayleigh channel 71
2.10.4 Efficiency 72
2.10.5 How well does learning work? 73
2.10.6 Streaming with a low-latency requirement 73
2.10.7 Streaming on a fast-fading channel 74

3 Acoustic Room-Area Networks 77

3.1 Overview 77
3.2 The Case for Room-Area Networks 78
3.3 Related Work 81
3.3.1 Existing Approximations to RAN 81
3.3.2 RAN PHY Candidates 82
3.4 Architecture of Blurt 83
3.4.1 Network Stack 83
3.5 Evaluation of Blurt 92
3.5.1 Experimental Setup 92
3.5.2 Validation of Selected Parameters 93
3.5.3 Reliability versus Range 97
3.5.4 Audibility 97
3.5.5 Hardware Compatibility 99
3.5.6 Robust to Noise 103
3.5.7 Attenuation at Room Boundaries 103
3.5.8 CPU Usage 105
3.6 Applications of Blurt 105
3.6.1 Guest Access for Wi-Fi Networks 105
3.6.2 Rate Limits and Traffic Prioritization 107
3.6.3 BlurtFactor: Two-Factor Authentication with Blurt 107
3.6.4 Building Automation 108
3.7 PAR Minimization with ADMM 108
3.7.1 PAR in a Software Modem 111

3.7.2 Algorithms for Tone Reservation 113
3.7.3 Complexity Model 115
3.7.4 Construction and Convergence 117
3.7.5 Continuous Time 122
3.7.6 Discrete Time 124
3.7.7 Proximal Mappings and the Alternating Direction Method of Multipliers 125
3.7.8 Final Algorithm 136
3.7.9 Convergence Properties 136
3.7.10 Accelerating the Haircut-Step 140
3.7.11 Complexity of the Project-Step 146
3.7.12 Step-size Schedule Optimization 146
3.7.13 Evaluation 147

4 Challenges to PHY Anonymity for Wi-Fi 151

4.1 Overview 152
4.2 PHY De-anonymizer 154
4.2.1 Setting 154
4.2.2 Prior Fingerprinting Schemes 154
4.2.3 MIMO Notation 155
4.2.4 Counting Degrees of Freedom 156
4.2.5 Selecting Degrees of Freedom 156
4.2.6 Robust Eigenspace Signal Tracking (REST) 159
4.2.7 Steps towards Quantitative Evaluation 161
4.3 MMSEE: Rational Gaussian Matrix Moments 162
4.3.1 Neumann-Isserlis Estimation 164
4.4 Maximum-likelihood Estimator 169
4.5 Anonymizing Countermeasures 170
4.5.1 Hiding in a Crowd 170
4.5.2 Antenna-Switching 170

4.6 Limits on Countermeasures 172
4.6.1 Nyquist Bound 173
4.7 Random-Field Bound on Effective Antennas 174
4.7.1 Approach 177
4.7.2 Fields from Currents 178
4.7.3 Random Field Current Model 182
4.7.4 Correlation of Electric Fields 188
4.7.5 Correlation Spectrum 190
4.8 Conclusion 202

A Computer Algebra Code **219**

B Norm Bounds for the Electric Two-Point Function **223**

List of Figures

1-1 Schematic of the decoding CDF. After some minimum number of symbols, below which decoding is not possible, the CDF rises gradually to 1. Under better signal conditions, fewer symbols (units of redundancy) are required, and the CDF shifts to the left and becomes sharper. . . 22

1-2 Proportion of 2.4 GHz radio wave energy accessible to a small cube-shaped transmitter or receiver, as a function of size. The energy which falls in each of a number of low-dimensional subspaces, labeled by ℓ, is indicated on the vertical axis. Any antennas beyond the $\sim$72 accounted for by the subspaces labeled $\ell = 0 \ldots 5$ will have a negligible impact. For a device under 6 mm, the number of effective antennas is only 3. . 26

2-1 Capacity versus average SNR for selected SISO fading models. 35

2-2 In HARQ with incremental redundancy, block decoding errors result in negative acknowledgments and retransmission attempts. The decoder uses all the redundancy available so far. 41

2-3 Symbols from many packets are aggregated into a frame in order to amortize the cost of the inter-frame spaces (in 802.11n, SIFS, 16 μs) and preambles/headers. 66

2-4 With spinal codes and $n_f = 20$. At low, medium, and high SNR, respectively, RateMore reduces overhead by 3.8×, 2.9×, and 2.6× relative to ARQ and 2.8×, 3.6×, and 5.4× relative to Try-after-n HARQ. 71

2-5 Learning performance of RateMore with spinal codes on a Rayleigh fading AWGN channel. Noise power is 15 dB below average faded signal power, giving an equivalent (in terms of capacity) stationary SNR of 12.8 dB. Doppler velocity is 10 m/s. The learning parameter α can be set to a very aggressive value of .8 for an aggregate throughput within 1.6% of the known-SNR "ideal adaptation" throughput. With $n_f = 10$. 72

2-6 Detail of Figure 2-5. 72

2-7 Even though n is random, RateMore makes good guesses and thus wastes very little time sending unnecessary symbols or waiting for unnecessary feedback. 73

2-8 Gaussian learning performance on a stationary channel with spinal codes. Mean and variance are smoothed by an exponentially-weighted moving average with parameter $\alpha = .99$. Steady-state performance is comparable to performance when CDFs are fully known. 74

2-9 RateMore is agnostic to the details of a specific rateless code, and achieves low overhead in latency-critical environments. n_f has been normalized according to the largest number of parallel streams compatible with a given bandwidth-latency product. Cases are shown for a minimum-quality Skype voice call and for an HD video call. For an ideal user experience, latency should be less than 100 ms. Strider and Raptor require relatively large packets, which prevents the use of aggregation on the voice call – in fact, a single packet contains more than 100 ms of audio, as shown in red. In order to meet the latency target at the expense of channel occupation, packets could be fired off partly empty. 75

2-10 RateMore overhead and channel occupation on fast-fading Rayleigh channels. n_f has been normalized as in Figure 2-9 75

3-1 Layers in the proposed RAN stack. 83

3-2 Response curves of a typical desktop loudspeaker and a laptop microphone, as measured against a CM-130 sound pressure level meter. . . 89

3-3 A Blurt datagram, before and after passing through an acoustic channel. From left to right, the short training sequences (which appear in the frequency domain as 12 broadly-spaced spikes), the long training sequences, and the data symbols. 90

3-4 Power-spectral density from a laptop source as observed by sweeping a cellphone microphone through a 180° arc. Deep nulls indicate frequencies at which acoustic waves from the two speakers interfered destructively. Horizontal axis is approximate. 91

3-5 Power-spectral density from a laptop source as observed by sweeping a cellphone microphone through a 180° arc. Shifting the waveform on one stereo channel by a few milliseconds is enough to largely obscure the pattern of nulls, which are now too closely spaced to resolve. . . . 92

3-6 Heat-map of packet detector accuracy vs. repetitions of the OFDM training sequence (in simulation). 95

3-7 Blurt normalizes the output waveform to the 95th percentile of amplitude. An additional amplitude reduction of 50% increases short range performance at the cost of long range performance. 96

3-8 Percentage of successful receptions achieved by Blurt, Quietnet, and Ultrasonic Networking as a function of distance between sender and receiver. Our perferred, inaudible configuration is shown with 95% confidence intervals. 98

3-9 Decibels relative to the threshold of audibility for different laptop volume settings. 99

3-10 Percentage of successful receptions as a function of distance between sender and receiver when using commodity mobile phones and tablets as Blurt senders and laptops as Blurt receivers. 101

3-11 Percentage of sucessful receptions as a function of distance between sender and receiver when using commodity mobile phones and tablets as Blurt receivers and laptops as Blurt senders. 102

3-12 Percentage of successful receptions as a function of distance between sender and receiver when using Blurt in the presence of characteristic ambient noise for three typical settings. 104

3-13 The results of one experiment on the stateful audio envelope limiter in a MacBook Pro laptop. 112

3-14 Power spectral density for model 802.11n HT20 OFDM waveform (N=64, dense QAM on 56 tones, 20 MHz bandwidth) before and after optimizing all unused tones but the DC subcarrier to minimize the peaks. The out-of-band radiation for optimal Tone Reservation is greatest at the edges of the original passband. Convex optimization using discrete epigraphical form (3.9) with Ipopt Interior Point Method software package [1, 2]. If an OFDM receiver is inspecting the subcarrier amplitudes at the band edges to track the carrier-frequency offset, such an "optimal" Tone Reservation scheme could cause the receiver to detune. 114

3-15 Tail of the power distribution before and after optimizing. Optimum computed by unlimited-iteration interior point optimization using Ipopt. 148

3-16 Tail of the PAR distribution before and after optimizing. 149

3-17 Power spectrum before and after optimizing. Note how the true optimum is very far in Euclidean distance from the original symbol. . . . 150

4-1 REST fingerprints of 370 packets (left-to-right, top-to-bottom) from a 5-second trace in a busy office building. The receiver bandwidth is 40 MHz, and the channel is divided into 128 subcarriers. Each complex 128×6×6 fingerprint is shown as a 128-row color bitmap, with hue and saturation representing phase and amplitude. Of the 377 packets in the full trace, 345 are decodable, and 286 include the transmitter's MAC address. A number of visually distinct sender fingerprints can be identified by their recurring patterns and colors. Quantitative clustering results on these fingerprints are future work. 162

4-2 With $\rho_A = 1$, $v \in \{-1, 1\}$, shows how trails of packets from a user's device may be concealed (top), marginally clear (middle), or abundantly clear (bottom) depending on the time τ between transmissions. . . . 171

4-3 For a physically-small transmitter or receiver, the map between electrical currents and radio waves (electric fields) is *low rank*, meaning that it acts as a bottleneck on the number of degrees-of-freedom we can access. 175

4-4 Proportion of 2.4 GHz radiated energy in each subspace $\mathcal{G}_\ell$ for a small cube-shaped ($a_1 = a_2 = a_3 > 0$) transmitter or receiver, as a function of size a_3. The thickening of the lines at the right edge of the plot represents the difference between the Taylor series at orders 22 and 24. 194

4-5 Proportion of 2.4 GHz radiated energy in each subspace $\mathcal{G}_\ell$ for a small slab-shaped ($a_1 \to 0$, $a_2 = a_3 > 0$) transmitter or receiver, as a function of size a_3. 195

4-6 Proportion of 2.4 GHz radiated energy in each subspace $\mathcal{G}_\ell$ for a small stick-shaped ($a_1 = a_2 \to 0$, $a_3 > 0$) transmitter or receiver, as a function of size a_3. 196

4-7 Sub-division of spherical polygons. The vertices (blue) at one level of refinement are used to construct new vertices (green), which are connected to form the triangles of the next level of refinement. The non-triangular case (right) appears only at the first level of refinement. 197

4-8 Convergence of two-point correlation spectrum estimates at four levels of refinement, corresponding to square matrices of dimension 96, 540, 2160, and 8640. Non-negligible modes have a relative power no less than $-50\,\mathrm{dB}$. Using the results for $\mathcal{H}_3$ as ground truth, the relative and absolute errors in the estimated eigenvalues are shown in the lower plots. 203
4-9 Convergence of two-point correlation eigenvector estimates for the first eigenvector. 204
4-10 Convergence of two-point correlation eigenvector estimates for the second eigenvector. 205
4-11 Convergence of two-point correlation eigenvector estimates for the third eigenvector. 206
4-12 Convergence of two-point correlation eigenvector estimates for the fourth eigenvector. 207

List of Tables

1.1 De-anonymization methods and countermeasures at multiple layers of the network stack. 27

2.1 Special cases of variable-length feedback codes with termination, following Polyanskiy et al [3]. VL=variable length; VLF=VL with feedback; FV=fixed-to-variable; ARQ=automatic repeat request. 37

2.2 Summary of experimental results of Chapter 2. 69

3.1 Selected applications of room-area networking 80

3.2 Applications of RAN. RAN fully captures the concepts expressed by the first group. The second group represents applications which stand to gain in usability from the RAN abstraction. 80

3.3 Packet error rate (PER) versus range between Blurt senders and receivers (2014 MacBook Pro laptops). Center frequency 19 kHz, bandwidth 3 kHz. 90

3.4 Design parameters for adapting 802.11a PHY to the acoustic channel. 91

3.5 Comparison of attenuation for Wi-Fi radio and for ultrasound when closing some typical office doors between the sender and receiver. . . 103

3.6 Complexity of peak-to-average ratio optimization schemes. Quantities for Tellado and Cioffi [4] are exact; quantities for Aggarwal and Meng [5] are lower bounds. See Definition 3.7.1 for detailed interpretation. . . 116

3.7 Phases of the evaluation of the proximal mapping of $\lambda\|\cdot\|_\infty$ with complexity estimates. N is the oversampled OFDM order. See Definition 3.7.1 for detailed interpretation. 141

3.8 Complexity of the project(v) proximal mapping. See Definition 3.7.1 for detailed interpretation. 146

4.1 Results of sub-division procedure. 198

Chapter 1

Introduction

Wireless signals have a fundamental tendency to vary in time and space. We can predict that the signals will vary, but we cannot, in general, predict how they will vary. This property creates significant challenges for wireless systems. This dissertation will present a body of work that revolves around the unpredictability of signal variations in wireless communication and localization systems.

When the strength of a wireless signal varies over time, this is usually an indication that something is physically moving—either the sender of the signal, or the receiver, or some object in the space in between: foliage, vehicular traffic, pedestrians, *et cetera*.

When the strength of a wireless signal varies over space, this is usually an indication that signals are reaching the receiver via more than one path. Wireless signals are waves, and as a consequence, waves following more than one path from sender to receiver may add, or they may subtract, depending on the distances involved. If they subtract, there may be little or no signal left for the receiver to pick up.

Naturally, motion and multi-path propagation are far from uncommon in wireless systems with mobile participants.

The high-level goal of wireless communications is to faithfully deliver information between devices. Temporal signal variation may interfere with that goal. Chapter 2 presents a protocol that assists wireless communications systems in adapting to temporal variations. The key intuition of this chapter is the remarkable effectiveness of an abstraction that we term the decoding distribution, or decoding CDF (cumulative

distribution function).

Within the space of communication systems, we have identified an unmet need for mobile-to-mobile and mobile-to-infrastructure networks that combine communication with implicit localization at the granularity of a single room. Chapter 3 makes the case for acoustic Room Area Networks, and provides a reference implementation. In addition, this chapter develops a new application of proximal gradient convex optimization techniques to low-complexity crest-factor minimization in multi-carrier modulation (§3.7).

In wireless localization, the goal is to establish an accurate relationship between signals and places. Signal variation with space may interfere with that goal. Chapter 4 explores the question of when localization systems designed to tolerate spatial variations can, and cannot, be fooled by a determined adversary. The key intuition in this chapter is that a necessary condition to reliably mask or un-mask a signal is to be the party with the largest number of antennas spread out over the largest space. The chapter additionally presents a robust channel fingerprinting mechanism that aids a passive listener in deciding whether two transmissions come from the same source. Finally, the chapter presents bounds on the effective number of spatial degrees of freedom accessible to a spatially-compact transmitter or receiver.

Chapter 2 and the Rateless Link-Layer Protocol

When a user—say, Alice—wishes to access the Internet from her wireless device, the browser on her phone assembles a packet: a discrete chunk of data suitable for independent delivery. This packet represents a request to load Alice's favorite web-page. The phone prepares to emit a wireless signal bearing this packet.

In the past, Alice's device might have sent the data out in a single, one-way burst. But let us suppose that Alice has a recent LTE phone. Rather than a single one-way burst, this device relies on a back-and-forth negotiation that divides each packet across several bursts. Each time the device finishes a burst, it interrogates the other side to determine whether the packet has been successfully received. If the answer is no, then the device prepares to send another burst.

This format is typical because, today, the vast majority of wireless devices cannot send and receive at the same time. Even when Alice is on a call, her phone is never sending and receiving at the same instant: instead, it switches back and forth between the two modes many times per second in order to maintain the illusion of a continuous, two-way connection.

This approach is not without costs. Each switching event takes time, and the total time spent waiting for feedback may become substantial. RateMore, originally presented in MobiCom 2012 [6] and elaborated in Chapter 2, is our system for controlling these costs.

The RateMore algorithm calculates the costs and benefits of different strategies for using wireless signals on devices that cannot transmit and receive simultaneously. The algorithm uses these costs and benefits to determine exactly how long each burst should be, so that Alice's phone will waste as little time as possible waiting for feedback. In particular, RateMore is designed to make this determination in concert with a data protection scheme called a *rateless code*. The rateless code represents a way to protect data from radio interference or noise, by introducing redundancy into the message; as much redundancy as may turn out to be necessary.

RateMore takes as its input a simple model of the wireless physical layer. This model, which we call a decoding CDF, abstracts away the details of the rateless code. It also represents everything we need to know about the current strength of the wireless signal.

The RateMore algorithm then calculates the costs and benefits of different strategies for sending and listening. Using the decoding CDF, RateMore computes the best possible rule, or schedule, for how much redundancy should be sent over the channel, and for when the sender should pause to listen for a feedback signal.

In addition to presenting the 2012 work in greater detail, this dissertation extends RateMore to the case where feedback packets may be lost.

A decoding CDF is a map that takes as input a certain number of symbols, and produces as output the probability that a particular decoder, in some particular conditions, will decode successfully. This decoding CDF is an empirical quantity that

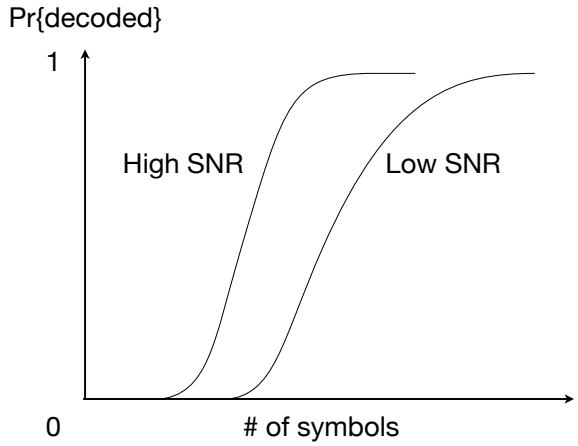


Figure 1-1: Schematic of the decoding CDF. After some minimum number of symbols, below which decoding is not possible, the CDF rises gradually to 1. Under better signal conditions, fewer symbols (units of redundancy) are required, and the CDF shifts to the left and becomes sharper.

we can draw on a graph, with input symbols along the horizontal axis, and output probability on the vertical axis (Figure 1-1.

In general, the exact CDF depends on many factors: the specific rateless code; the instantaneous condition of the wireless channel; the details of any approximations or heuristics employed by the hardware designers who created the decoder; and in some cases even the power and computation budget of the receiver. Estimating and tracking these parameters on-the-fly would be costly and error-prone. However, once the decoding CDF is determined, none of these other factors are needed any longer: the decoding CDF is a sufficient statistic for transmission scheduling on the half-duplex link.

The decoding CDF abstracts away the details of the rateless code. RateMore does not need to know how the code is structured, or how the decoder works; these things can be treated as black boxes. All RateMore needs to know is the cost of feedback, the probability of a lost feedback packet, and how likely decoding is to succeed.

Moreover, the decoding CDF is directly observable: with each successful decoding, the RateMore sender learns a random sample from the distribution, and uses this observation to update its model of the decoding CDF. Due to this property of direct observability, RateMore can side-step the issue of estimating all the individual factors

that affect decoding, and it can rapidly and accurately track fading signal conditions. Adopting the decoding CDF as an abstraction barrier between rateless link-layer protocols and rateless codes leads to tractable throughput-optimization problems.

The decoding CDF is sufficient: it tells RateMore everything it wants to know to schedule transmissions under a given cost of feedback.

The decoding CDF is succinct: it consists of a few thousand parameters at most, and may be approximated with as few as two parameters under on-line estimation.

The decoding CDF is robust: the two-parameter Gaussian approximation results in minimal loss of throughput.

The decoding CDF is directly observable: every time decoding succeeds, RateMore learns how many symbols were required for decoding.

Most of all, the decoding CDF works for any rateless code on the additive white Gaussian noise channel.

RateMore allows Alice's phone to use the wireless channel efficiently, saving time and power, and freeing up network resources for other users. RateMore determines when a rateless sender should stop to ask the receiver whether the message has been decoded successfully. It takes advantage of feedback without incurring huge costs because it takes the costs into account in the optimization. RateMore can do this because the decoding CDF tells it how likely it is to receive negative feedback under any possible strategy.

The decoding CDF is sufficient, succinct, observable, robust, and code-agnostic. It can be readily estimated on-line, and tracks tightly during fading. Chapter 2 details the definition and application of this abstraction.

Chapter 3 and Room Area Networks

Alice lives in the city, where social life is built around *rooms*: spaces where a few people can gather to share an activity or a conversation. Where walls are present, they naturally separate conversations and delineate rooms. Even where walls are absent, as in a restaurant or an outdoor area, many people choose to speak with such a volume that conversations remain private to a small space: a table or booth, a park bench, a

vehicle, or simply an "earshot". We term all of these spaces rooms.

When Alice brings her friends into a room, she is immediately able to network with them on a personal level, subject to whatever degree of privacy is afforded by the particular room. Alice's mobile device, however, shares no special or private link with those of her friends. Room-Area Networks (RAN) represent connectivity between devices that mirrors social communication between people.

RAN enables a broad range of novel location-based policies. RAN-enabled mobile devices pave the way for application developers and users to achieve mobile-to-mobile goals like "share this confidential PowerPoint with everyone in the board-room, and no-one outside this room", and mobile-to-infrastructure goals like "turn up the volume on the television set in front of me" or "turn down the temperature in this room". RAN-enabled spaces provide a natural way for infrastructural engineers to provide prioritized quality-of-service to wireless users in particular physical zones. Perhaps if Steve Jobs's on-stage demonstration of the iPhone 4 had occured in a RAN-enabled conference center, he would not have needed to stop and wait for the audience to shut off their wireless devices so that he could use it on the stage.

We distinguish RAN from indoor localization systems by virtue of its infrastructure-optional nature, and from radio systems by its location-selectivity. We provide a proof-of-concept implementation of RAN as an acoustic software modem, called Blurt. The RAN is exposed to applications as an IPv6 tunnel with link-local addressing.

In the process of characterizing the performance of Blurt, we encountered the well-known challenge of high *peak-to-average power ratio* (PAR; variously also called PAPR, PMEPR, or crest factor) in multi-carrier modulation. We present a new application of the alternating direction method of multipliers to this convex optimization problem, achieving a significant reduction in complexity relative to prior schemes.

Chapter 4 and Wireless Anonymity

Society faces significant stresses concerning anonymous behavior on the Internet. On one hand, anonymity may protect dissidents and whistleblowers from reprisal. On the other hand, it may also shield criminals and predators from justice. As public policy

evolves, technical standards will face pressure to follow, and wireless standards are no exception.

Current Wi-Fi technology makes no effort to perform surveillance, but it also makes little effort to prevent it. Journalist Russell Brandom, on the subject of wireless anonymity, once wrote, "in theory, a person with the right equipment [can] always trace the bark back to the dog". Is this so? Is it necessarily the case that using a wireless network makes an individual identifiable?

In our work, we have explored a number of measures, countermeasures, and counter-countermeasures for tracking moving sources of wireless signals using physical-layer information. Our idea was to design the most powerful de-anonymizer we could conceive, and then to characterize its strengths and weaknesses. The key take-away from this work is that tracking is almost always possible if the tracker has more antennas than the device being tracked, and inversely, tracking is almost always impossible if the device being tracked has more antennas than the tracker. It turns out, though, that access to more physical space is also a critical factor. If the device being tracked is physically very compact, then its many antennas cannot behave independently, and the effective number of antennas is lesser, allowing tracking. In the same way, if the device performing the tracking is physically very compact, then its large number of antennas will not avail (Figure 1-2).

Let us put a face on this problem by introducing Alice's aunt, Betsy. Betsy is a very cautious individual, because she belongs to a persecuted minority in her home country. Betsy has a simple question for Alice: Dangerous people are looking for me. If I use my laptop to connect to a wireless network, am I putting myself and my family in danger?

We will call Betsy the sender, and her adversary the listener. Betsy wishes to push traffic through the network; and her adversary wants to passively eavesdrop on all wireless signals, tie each signal back to a particular sender, and ultimately to learn the identities of the senders.

The key question is, when can two transmissions be tied back to the same device? That is, can one tell, from any given pair of transmissions, whether they have the

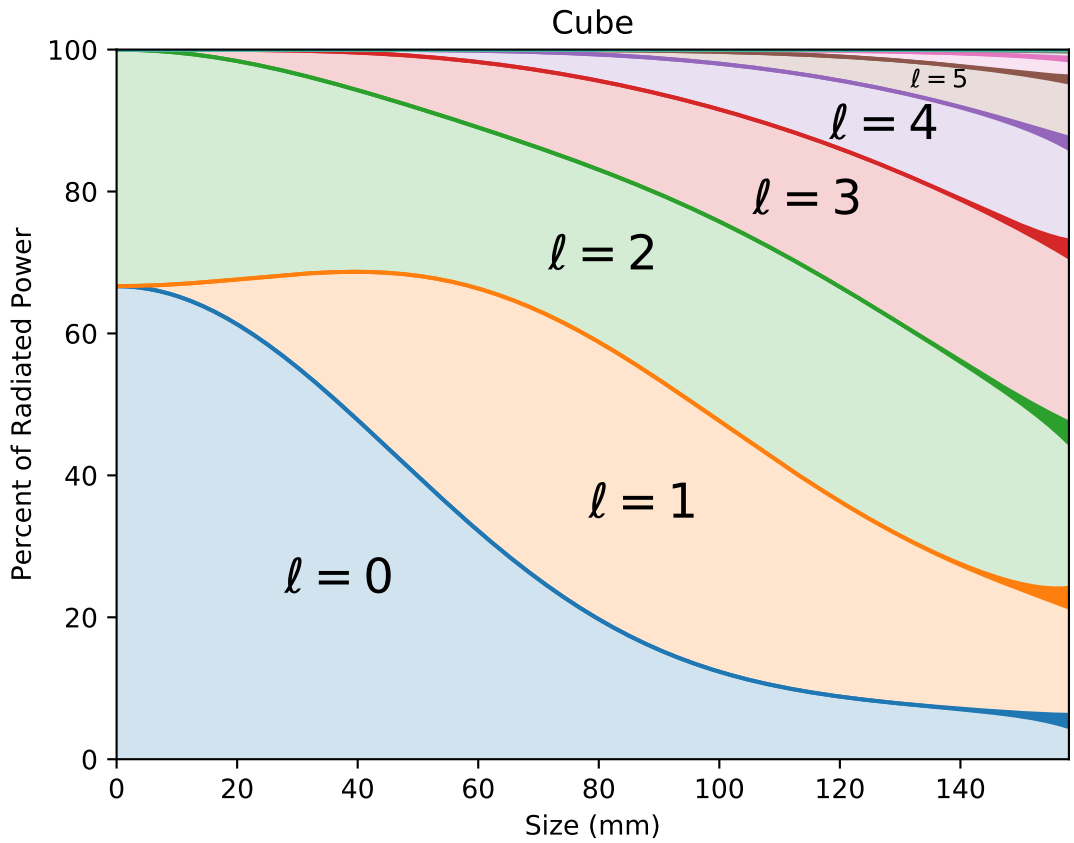


Figure 1-2: Proportion of 2.4 GHz radio wave energy accessible to a small cube-shaped transmitter or receiver, as a function of size. The energy which falls in each of a number of low-dimensional subspaces, labeled by ℓ, is indicated on the vertical axis. Any antennas beyond the $\sim$72 accounted for by the subspaces labeled $\ell = 0 \ldots 5$ will have a negligible impact. For a device under 6 mm, the number of effective antennas is only 3.

same sender? We will call such a procedure a "same-sender test".

This may not be the obvious question to ask. In the end, the question is whether a person can participate in a wireless network without leaking their identity. How does this relate to the non-existence of a "same-sender test"?

If a listener can tie two transmissions together, and then it can tie those to a third, and so on, then it can build up a chain of transmissions, all tied back to the same sender. If the listener then finds a way to de-anonymize any one of these transmissions, it de-anonymizes all of them. This chain is as strong as its weakest link, and information leaks are quite common (Table 1.1).

	Same-sender Test	Countermeasure
Link Layer	MAC sniffing	Randomization
	Traffic analysis	Mix and cover
Physical Layer	Radiometric signatures	Calibration
	Channel State Information (CSI) fingerprinting	Ventriloquism

Table 1.1: De-anonymization methods and countermeasures at multiple layers of the network stack.

Suppose the sender and listener are connected to a secure Wi-Fi network with WPA2 Enterprise. Because of the way such a system works today, the listener can decrypt some of the sender's messages. In particular, they can decrypt broadcast and multicast transmissions, including ARP, Windows file sharing, multicast DNS, and Dropbox file synchronization, leaking a great deal of information useful for de-anonymizing the sender, most notably Betsy's link-layer (MAC) address.

It's not enough to patch every application and secure every protocol if the listener can de-anonymize a chain of transmissions. Eventually, the pattern of behavior revealed by this chain of transmissions will uniquely identify the sender. So to achieve anonymity, the sender will have to stop the listener from building such a chain in the first place. That is, the sender's goal is to defeat every possible "same-sender test".

Today, Wi-Fi makes the listener's job easy: besides MAC addresses, in 2016, Matte et al [7] built a same-sender test using packet timings. Others have shown that imperfections in radio hardware generate unique "radiometric" signatures for each sender.

In each case—MAC sniffing, traffic analysis, radiometric signatures—the technical issues seem tractable. MAC address randomization techniques have been proposed and deployed. There are network protocols, like Vuvuzela by van den Hooff et al, that are resistant to traffic analysis. As for radiometric signatures, modern radios are flexible enough to correct for almost any imaginable non-ideality in software, with enough effort.

However, there are other ways to associate packets that depend only on the nature of radio waves—on the way that they spread out and interact with the environment.

How would the strongest imaginable attacker exploit these physical layer properties? We explored this question in a paper in WiSec 2017 [8]. Chapter 4 presents this work in depth.

The radio information is contained in an object called the channel state information, or CSI. Our first contribution is a strategy for the listener. The listener wants to tell whether two transmissions are from the same sender, and it does this by extracting a hash of the sender's identity from the channel state information using our fingerprinting procedure, which we call REST.

Our second contribution is a strategy for the sender. The idea is to dynamically change the shape of the sender's antenna without having any moving parts. Each configuration is used for a single transmission. By reprogramming an array of switches, the sender can access an exponentially large number of antenna shapes, all occupying the same space. This allows a sender with a small number of antennas to hide from a listener running REST with a larger number of antennas—which would not be possible otherwise.

Our third contribution in Chapter 4 is a theoretical bound on how many antennas the sender can emulate in this way. What we found is that a large smartphone can access, at most, a few dozen distinct antenna patterns in the 2.4 GHz band. With greater numbers, the antennas become indistinguishable to the listener, and they can be tracked. Figure 1-2 depicts this situation graphically: for devices of smartphone dimensions (right) or smaller (left), the vast majority of transmitted or received radio energy falls into a small number of finite-dimensional subspaces, represented by the colored bands. The number of effective antennas in such a device is no more than (roughly) $2\ell_{\max}^2$.

Chapter 2

Rateless Link-Layer Protocol

This chapter considers the use of evolved incremental redundancy schemes known as rateless codes to accommodate fading channel conditions in the additive white Gaussian noise (AWGN) model of point-to-point wireless communication systems. For these codes to reach their full potential on standard half-duplex channels, changes are needed to the link layer. This chapter presents RateMore, a rateless link protocol (RLP) to meet this need. RateMore finds the optimum trade-off between retransmissions and up-front forward-error-correction (FEC) risk aversion using a dynamic programming approach.

The main contributions of this section are

1. A relaxation, the *marginalized-τ model* (§2.3), of the RLP problem for systems in the class referred to by Polyanskiy et al [3] as stop-feedback VLF codes for AWGN channels (definitions provided in §2.1).

2. A dynamic programming (DP) algorithm that solves this relaxed problem, with the additional capability of scheduling over lost ACKs. (§2.6.2)

3. Proof of ε-optimality for the DP in poly time, under the marginalized-τ model (§2.6.1). The DP has time complexity $\mathcal{O}(\omega^2)$ and space complexity $\mathcal{O}(\omega)$, where $\omega = \mathcal{O}(\bar{\tau} + \frac{n_f}{p_f}\frac{\sigma^2}{\varepsilon})$ is a concentration bound on the tail of τ. (§2.6.3)

4. Construction of a family of efficient truncation heuristics for the DP. (§2.6.1)

5. A family of heuristics for online tuning of the DP algorithm which perform well in simulation. (§2.7)

6. A scheme that reduces the header size requirement for code-block aggregation of M homogeneous-sized packets from $\mathcal{O}(M)$ to a single bit. (§2.8)

7. Significant refinements to the tuning procedure to improve stability, bias, and rate of convergence. (§2.9)

8. Evaluation of the proposed algorithms in simulation over multiple rateless codes on stationary and fading AWGN channels. (§2.10)

2.1 Terminology and Background

In this section, we sketch out some of what is known from a theoretical and a systems perspective about physical layer wireless communications on fading channels. On the theoretical side, we will highlight a series of channel models of increasing realism. We will attempt to reveal the reasoning that has led to the increasingly widespread deployment of what are, in effect, variable-length forward error correction (FEC) codes. From there, we will segue into the systems that have been designed to meet these needs: automatic repeat requests (ARQ), hybrid ARQ with FEC, hybrid ARQ with incremental redundancy (IR), and finally rateless codes. It is these last objects, rateless codes, that serve as the foundation for the work of this chapter.

2.1.1 Wireless Communications, Coding, and Capacity

The discrete-time AWGN point-to-point channel is a stochastic system

$$y^k = hx^k + n^k,$$

where $x^k \in \mathbb{C}$ is an input signal, $n^k \in \mathbb{C}$ is the cumulative effect of all sources of additive disturbance, $h \in \mathbb{C}$ is the cumulative effect of all sources of multiplicative disturbance, and $y^k \in \mathbb{C}$ is the output signal. n^k is taken to be zero-mean circular complex Gaussian

with power σ_{NN}^2: that is, its real and imaginary parts are independent identically distributed normal variates with mean zero and variance $\sigma_{NN}^2/2$. The AWGN model assumes that the sequence $\{n^k\}$ is jointly Gaussian, i.i.d., and independent of h and $\{x^\ell\}$. The signal-to-noise ratio (SNR) is defined as $\mathbb{E}[|x|^2]/\sigma_{NN}^2$, and is typically given in logarithmic units (decibels). As a convention, and without loss of generality, we set $\mathbb{E}[|x|^2] = 1$; this is equivalent to rescaling n^k.

Neither the transmitter nor the receiver has *a priori* knowledge of n^k. If h is known to the receiver, the model is called *coherent*; otherwise, it is called *incoherent*. Knowledge of h is called channel-state information (CSI). The existence of a scalar value for h is a form of *narrow-band assumption*, but it can often be justified by decomposing a wide-band channel into a number of independent, orthogonal narrow-band channels. We will need h later; for now, let it be 1.

The dimensionless unit of discrete time is a second-Hertz, or a *channel use*. The number of channel uses (Nyquist samples) occupied by a signal is equal to the product of its (epsilon-) duration and its (epsilon-) bandwidth. The *AWGN channel coding problem* is to transmit data at the highest rate possible rate R bits-per-channel-use (i.e. bits/s/Hz) such that the probability of any undetected errors at the receiver (e.g. due to the disruptive influence of n^k) can be made negligible. We will denote the sampling bandwidth by B.

A *fixed-rate code* for the AWGN channel is an encoder $E\colon \{0,1\}^n \to \mathbb{C}^k$ and a decoder $D\colon \mathbb{C}^k \to \{0,1\}^n$. Here, k is the *code-word length*, and n is the *block length*. The error rate of this code is defined as $\varepsilon = \mathbb{P}_{\{n^k\}}(D(E(d) + n) \neq d)$.

Shannon's results [9, 10] on the feasibility of this problem establish: First, that there exists a probability distribution $\mu_{\mathcal{X}}$ (in this case, i.i.d. circular complex Gaussian) such that if E_n is drawn at random from ${\mu_{\mathcal{X}}}^{2^n}$ and $D_n(y) = \arg\min_{x\in\{0,1\}^n} \|y - E_n(x)\|_2^2$, then in the limit $n \to \infty$, the pair (E_n, D_n) achieves an asymptotically zero error rate, provided that

$$R = n/k < C = \log_2(1 + \mathrm{SNR}),$$

where C is called the capacity of the channel. And second, that no code for the AWGN

channel achieves a higher rate without a self-defeating increase in error rate.

Complication 1: finite time, finite error rate

This so-called random codebook construction is not computationally feasible for large n, much less infinite n. Furthermore, real-world communications consists of messages, or blocks, of finite length, which must be delivered in a finite time: we would not prefer to take $n \to \infty$ even if it were possible. It is also neither possible nor necessary to entirely eliminate decoding errors: for better or for worse, real-world networks perform additional error detection steps above the physical layer.

A result of Polyanskiy et al [11] establishes that at a finite block length n, and a finite error rate $\varepsilon > 0$, communication is possible at a rate

$$R^*(n, \varepsilon) = C - \frac{1}{\sqrt{n}} \Phi^{-1}(1-\varepsilon)\sqrt{V} + \mathcal{O}((1/n)\log n) \tag{2.1}$$

$$\approx C - \frac{1}{\sqrt{n}} \Phi^{-1}(1-\varepsilon)\sqrt{\frac{\log_2^2 e}{2}\left(1 - \frac{1}{(1+\mathrm{SNR})^2}\right)}, \tag{2.2}$$

where Φ is the standard normal CDF, and V is the AWGN channel dispersion.

Complication 2: fading

Wireless propagation is subject to *path delay* and *path loss*. If a transmitter and receiver are in relative motion, the path delay varies with time, a condition known as Doppler shift. Doppler shift is observed at the receiver as carrier- and sample-frequency offsets (CFO, SFO). In addition to these delay variations, path loss, and hence received signal power, varies with the inverse-square law and any scattering or absorption of the wireless signal.

CFO and SFO are present even in the absence of motion, due to the the accuracy and stability limitations of low-power/low-cost local oscillators. Similarly, received signal amplitude may vary in the absence of endpoint motion due to temperature- and voltage-dependent amplifier performance, or variation in the environment. For these reasons, feedback-control loops for tracking delay and signal envelope are a

component of every practical asynchronous digital wireless system. These tracking loops compensate for Doppler shift and path loss within a wide range.

When wireless propagation follows more than one path (multipath), the delay and loss along each path vary—to some degree—independently of each other. When the receiver measures the signal, it will see the superposition of the waves arriving along each path. Multipath propagation is overwhelmingly the common case for wireless communications devices in indoor or urban environments.

The combination of multipath and motion gives rise to *fading*. As the paths' relative delays (and hence phases) vary, the channel gain h^k will follow a complicated trajectory in the complex plane. (While the path losses also vary, it is the interference between components of differing phase that leads to large and rapid variations in received signal envelope). When the channel gain h^k passes close to zero for a period of time, we say that the channel has experienced a *deep fade*.

The full effect of fading is not only a multiplicative disturbance. Under the Jakes model of Rayleigh fading, the effect of fading on the power-spectral density (PSD) of the input signal $\{x^k\}$ is to spread the energy at any given frequency ω out over a finite band $\omega(1 \pm \beta)$, where β is the ratio of receiver velocity to wave velocity. If the PSD of the input signal has support only in a narrow band centered on ω_0, and $\beta \ll 1$, then this spreading may be approximated by a convolution in the frequency domain, or a multiplication in the time domain, of the input process $\{x^k\}$ by a stationary fading process $\{h^k\}$. The discrete-time channel needs to be oversampled, compared to the bandwidth of $\{x^k\}$, in order for the product sequence $\{h^k x^k\}$ to be Nyquist sampled.

The fading channel therefore has two natural discrete-time-scales: the code-word length k, and the coherence length (inverse Doppler spread) B/f_d. The capacity of this channel in the $n, k \to \infty$ limit depends critically on whether we take this limit with $k \ll B/f_d$ or the inverse. In the first case (quasi-static fading) the $\varepsilon = 0$ capacity

is zero: no code can eliminate the possibility of *outages* during deep fades, because there is no lower bound on how deep a fade can be, and a deep fade lasts for an entire block. Non-zero rates are possible only for $\varepsilon > 0$.

A more realistic model of the channel takes the fading to be quasi-static over a fixed time-scale L, rather than a timescale of one code-word. This can relax the problem substantially both at finite and at zero error rates: if the fading process has suitable ergodic properties, then averaging over many intervals L gives a code resilience against deep fades. If $n, k \to \infty$ with L constant (the fast-fading limit), so that a single code-word is many times longer than the coherence length, and the fading process $\{h^k\}$ is ergodic in the marginal distribution of h^k, then non-zero rates are possible with $\varepsilon = 0$.

After the fast-fading assumption and the ergodic assumption, the next points of differentiation among fading channel models are the availability of CSI at the receiver (CSIR) and/or the transmitter (CSIT). The theories for +CSIR+CSIT and +CSIR−CSIT are complete and well-known at all SNRs [12, §5.4.5-5.4.6, p. 199-205], though the capacities of these channels as a function of SNR cannot be written in closed form (Figure 2-1). For −CSIR−CSIT, the capacity is known asymptotically for low [13] and high [14, 15] SNR. The high-SNR limits differ critically based on whether the fading process $\{h^k\}$ is block-correlated or finite-entropy; in the latter case, the high-SNR capacity may drop as far as $\log\log \mathrm{SNR} - 1 - \gamma$, with γ the Euler-Mascheroni constant. In either case, if the coherence time of the channel becomes as short as one symbol, the channel is said to be overspread. In this regime, the log and log-log terms may go to zero, leading to a finite capacity even in the high-SNR limit.

These inefficient regimes can, in some cases, be dodged by oversampling [16]. Spatial diversity may also play a critical role [17]. Extensions to MIMO [18] and frequency correlation [19, 20, 21] are also available. The exact capacity results can be quite sensitive to the details of the channel model [22].

The approach taken in real-world systems has been to make one fundamental change to the problem set-up, and one fundamental change to the solution space. The change to the problem is to allow feedback from the receiver to the transmitter. Stated

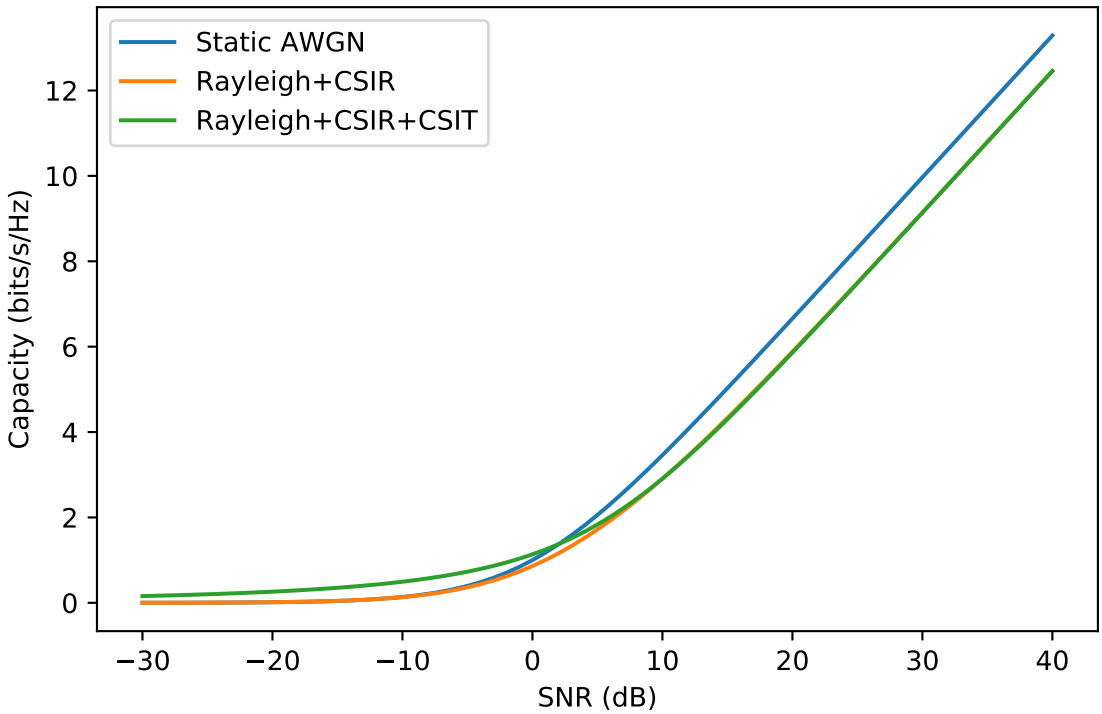


Figure 2-1: Capacity versus average SNR for selected SISO fading models.

most generally, and without regard for the practicality of such a notion, this feedback could be the entire sequence $\{y^k\}$, fed back synchronously with the transmission $\{x^k\}$. The change to the solution space is to re-consider the assumption of a fixed ratio n/k of block length to channel uses.

Variable-length coding with feedback

Polyanskiy et al [3] offers the following characterization of this altered coding problem, where we have omitted randomization and specialized to the (non-fading) AWGN channel.

Definition 2.1.1 (Variable-length feedback code with termination (VLFT)) [3, Definition 2, p. 4]

An $(\ell, M, \varepsilon) \in \mathbb{R}_{>0} \times \mathbb{Z}_{>0} \times [0, 1]$ *deterministic VLFT code for the AWGN channel is defined by:*

1. *A sequence of encoders* $f_k \colon \{1, \ldots, M\} \times \mathbb{C}^{k-1} \to \mathbb{C}$, $k \geq 1$, *defining channel*

inputs

$$x^k = f_k(W, y_{k-1}),$$

where $W \in \{1, \ldots, M\}$ *is the equiprobable message, and* y_{k-1} *is a prefix of the channel output sequence* $\{y^k\}$.

2. *A sequence of decoders* $g_k \colon \mathbb{C}^k \to \{1, \ldots, M\}$ *providing the best estimate of* W *at time* k.

3. *A non-negative integer-valued random variable* τ*, a stopping time for the filtration* $\mathcal{G}_k = \sigma\{y^1, y^2, \ldots, y^k\}$*, which satisfies*

$$\mathbb{E}[\tau] \leq \ell$$

The final decision $\hat{W}$ *is computed at the time instant* τ*:*

$$\hat{W} = g_\tau(y_\tau)$$

and must satisfy

$$\mathbb{P}[\hat{W} \neq W] \leq \varepsilon$$

Polyanskiy et al go on to identify the following special cases of this definition, characterized by requiring the encoding and termination (i.e. f_k and τ) to depend only on certain variables:

In terms of the capacity C of traditional fixed-length AWGN codes, the ε-capacity of VL, VLF, and VLFT codes is $C/(1-\varepsilon)$. In the non-asymptotic (finite block-length) regime for AWGN, Polyanskiy et al show [3, Thms. 2, 9, 11] that the highest achievable

System	Encoding	Termination	Interpretation
	depends only on		
VLF	k, W, y	k, y	Error detection at receiver
stop-feedback	k, W	k, y	Error detection at receiver + Encoding is non-adaptive
VL	k, W	k, W	Ragged code-book
FV	k, W	$(g_k(y_k) = W)$	Oracle decoder
ARQ	$(k \bmod K), W$	$(k = 0 \bmod K) \wedge (g_k(y_k) = W)$	Keep trying fixed-rate code until decoding succeeds

Table 2.1: Special cases of variable-length feedback codes with termination, following Polyanskiy et al [3]. VL=variable length; VLF=VL with feedback; FV=fixed-to-variable; ARQ=automatic repeat request.

rate $R^*(\ell, \varepsilon)$ satisfies

$$\frac{C}{1-\varepsilon} - (1/\ell)\log \ell + \mathcal{O}(1/\ell) \leq R^*_{\text{VLF}}(\ell, \varepsilon) \leq \frac{C}{1-\varepsilon} + \mathcal{O}(1/\ell) \tag{2.3}$$

$$R^*_{\text{VLF}}(\ell, \varepsilon) \leq R^*_{\text{VLFT}}(\ell, \varepsilon) \leq \frac{C + (1/\ell)\log \ell}{1-\varepsilon} + \mathcal{O}(1/\ell) \tag{2.4}$$

$$R^*_{\text{VLF}}(\ell, 0) = 0 \tag{2.5}$$

$$C + \mathcal{O}(1/\ell) \leq R^*_{\text{VLFT}}(\ell, 0) \leq C + (1/\ell)\log \ell + \mathcal{O}(1/\ell) \tag{2.6}$$

where (2.3), (2.4) hold for $0 < \varepsilon < 1$.

Note that (2.5) does not hold for general discrete memoryless channels as the other results do; it holds only in the case where Burnashev's C_1 parameter, the maximum K-L divergence between any pair of conditional output distributions, is finite, as it is for AWGN channels. If it is infinite, as on erasure channels, then a result of Burnashev [23] gives instead that

$$C - \mathcal{O}(\sqrt{(1/\ell)\log \ell}) \leq R^*_{\text{VLF}}(\ell, 0) \tag{2.7}$$

One intuition for these results—particularly the possibility of rates $R^* > C$—is to articulate exactly how VLFT codes may "cheat" relative to the traditional coding abstraction.

Note first that the difference between VLF and VLFT codes is that in VLF codes, the termination time τ is not allowed to depend on W. Put another way, this means that in VLFT codes, the decoder may have access to an oracle that indicates whether a given decoding $g_k(y_k)$ is correct. If we suppose that the sequence of hypotheses eventually converges to the correct W, then knowing when to terminate decoding is tantamount to knowing τ, the stopping time upper-bounded in expectation by ℓ. So, relative to the traditional channel coding problem, the VLFT abstraction communicates τ without spending channel uses. The entropy of this extra information is upper bounded by $(\log \ell + \log e)$, as can be shown by maximizing the entropy of a positive-integer-valued random variable over the space of distributions subject to a mean constraint to get a geometric distribution with parameter $p = 1/\ell$. The extra information therefore contributes to the maximum rate by at most $(1/\ell) \log \ell + \mathcal{O}(1/\ell)$.

The second way that VLFT codes can cheat is to first drive errors to zero, and then to selectively abandon a fraction ε of blocks by spending zero time transmitting them. This may not seem like cheating, because driving errors to zero seems like a very costly first step if we carry over our expectations from the fixed-length abstraction. However, the VLFT abstraction has the property that its zero-error capacity, even in the non-asymptotic regime, is the full Shannon capacity C of the channel. So the first step is free; then the second step buys a rate increase of $1/(1-\varepsilon)$ by decreasing the mean termination time. (note that this intuitive argument does not explain why VLF codes, and indeed VL codes, achieve the same $C/(1-\varepsilon)$ capacity; VLF codes certainly do not have the free-first-step property per (2.5)).

Comparing the achievability (lower) bound of (2.3) to (2.1), we find that the VLF and VLFT models have a third benefit: they can avoid the $\mathcal{O}(n^{-1/2})$ maximum-rate penalty suffered by fixed-rate AWGN codes due to channel dispersion. This is advantageous, but not revolutionary; for a TCP ACK with Wi-Fi headers ($n = 576$ bits) with SNR in the range 5–20 dB and a maximum error rate of 10%, the dispersion penalty is only 0.8–2.6% of the Shannon capacity, increasing to 2.0–6.2% at a maximum error rate of 10^{-3}.

We will argue that the most significant benefit of these new models of channel

coding is none of the three rate-boosts mentioned so far, but instead the opportunity they afford to adapt to slow-fading channels. Before we do, however, we must address the elephant in the room: the utter impracticality of general channel-output feedback.

Even if it were possible to instantly and accurately convey the channel outputs $\{y^k\}$ back to the transmitter, we could surely find some better use to which to put such an infinite-capacity back-channel. The question, then, is how much feedback is truly needed. A remarkable result of Polyanskiy et al is that general channel-output feedback buys us next to nothing beyond what we get from stop-feedback. In particular, the lower bound in (2.3), which bounds both R^*_{VLF} and R^*_{VLFT}, is established under the stop-feedback assumption. In what follows, we will only consider what Polyanskiy et al call stop-feedback and fixed-to-variable (FV) codes (which include ARQ).

Our terminology will differ slightly. Polyanskiy et al define stop-feedback codes as a special case of VLF codes with feedback-agnostic encoding, and they define FV codes as a special case of VLFT codes with the same property. The difference is that VLF codes cannot rely on a decoding oracle, while FV codes must rely on a decoding oracle. The decoding oracle, we recall, is a predicate that indicates at the receiver whether $g_k(y_k) = W$ holds for a particular k. In the presence of a decoding oracle, the termination time τ is simply the greatest lower bound on k such that the predicate accepts the decoding.

Polyanskiy et al go on to highlight FV codes as the prototype for rateless or fountain codes, suggesting that the decoder of such codes "can get a highly reliable estimate of τ autonomously", referring to Verdú and Shamai [24, Remark 6], which suggests a hashing-based approach. One presumes that the meaning is to introduce error detection into the message, effectively increasing the minimum distance of the codebook and allowing termination to be determined by a threshold on the relative likelihood of the second-most-likely candidate decoding. (This machinery is not needed on an erasure channel, where counting arguments will determine whether decoding is possible after a certain number of un-erased symbols have been received.)

We point out, however, that such an error detection scheme is precisely what stop-feedback codes are required to use, and precisely what FV codes are required

not to use. We therefore regard the association of rateless or fountain codes with FV codes as a source of potential confusion. An intuitive distinction might be to treat stop-feedback VLF codes with vanishing ε as a practical relaxation of FV codes, which are zero-error but require a non-existent oracle.

To further muddy the waters, we observe that in the absence of general channel-output feedback, FV codes require the use of stop-feedback to indicate to the transmitter that the oracle has accepted the decoding. That is, stop-feedback is both required and sufficient in both cases (stop-feedback VLF codes, FV codes). Because stop-feedback is therefore not a useful term for distinguishing the two cases of interest in this work, we will instead call the two cases error-detection termination (EDT) and oracle termination (OT).

2.1.2 Rateless Coding from a Systems Perspective

A reliable system must offer error correction. There are two main tools for error correction: backward error correction (BEC) and forward error correction (FEC). BEC attempts to correct errors after they are detected by introducing retransmissions; FEC attempts to correct errors before they are detected by introducing enough redundancy to accommodate worst-case expectations of channel conditions. FEC is health insurance: you pay a constant up-front cost of additional time and bandwidth on the channel in order to mitigate the cost of a low-probability error condition. BEC is a trip to the hospital: you only pay in the low-probability case, but you pay a high price.

FEC and BEC are not mutually exclusive; their combination is referred to as hybrid automatic repeat requests (HARQ). The presence of BEC allows the FEC scheme to be less conservative in its choice of transmission rate; and the presence of FEC allows the BEC mechanism to be invoked less frequently. A further refinement, incremental redundancy (Figure 2-2), modifies FEC in two ways: first, the decoder takes as input not only one transmission attempt, but all attempted transmissions of this block so far; and second, the encoder designs its retransmissions to be decoded in combination, rather than to be decodable in isolation.

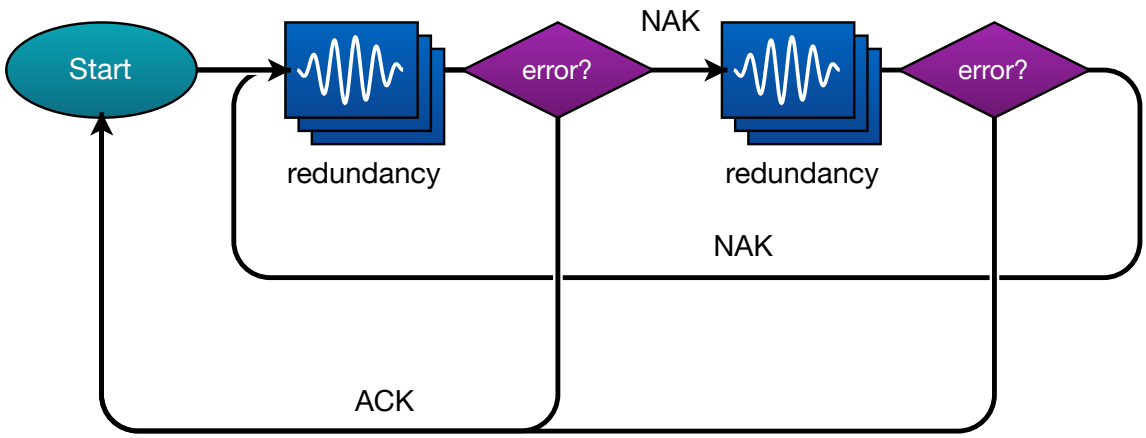


Figure 2-2: In HARQ with incremental redundancy, block decoding errors result in negative acknowledgments and retransmission attempts. The decoder uses all the redundancy available so far.

If, as in Wi-Fi, the duration of a transmission is flexible while the message (block) size is fixed by upper layers, incremental redundancy means that retransmissions can be much shorter, since they need provide only as much additional redundancy as the receiver requires. This frees channel resources for other users. If, as in LTE, the time and bandwidth of one scheduling quantum is fixed, so that the retransmission has to be as long as the original transmission, then the benefit of incremental redundancy is that the chance of successful decoding after the second attempt is higher than would be achieved from two independent attempts.

Finally, a reliable system must offer applications a degree of control over the error recovery process. For instance, in LTE voice communications, it takes 8 ms for a mobile phone to solicit a retransmission of a block that failed to decode. However, the latency budget for voice traffic only allows the radio subsystem to introduce 20 ms of end-to-end latency [25]. It is preferable to drop the failed block and allow the vocoder to infill the missing data, rather than to wait. On the other hand, TCP-based applications on the same LTE link might prefer that the link-layer retransmission mechanism keep trying for some time before giving up [26, 27]. By reasoning explicitly about bandwidth and latency costs, rateless link protocols allow applications to impose policies at a fine level of granularity.

Ratelessness

What is ratelessness? As above, we say that a system transmits at a rate $R = n/k$, in bits per channel usage, if it sends a block of n bits over a channel using a time-bandwidth product of k channel uses. We say that a system is fixed-rate if a decision must be made in advance of transmission as to what rate to transmit at, and we say that the system is rateless if no such decision needs to be made.

The advantage of postponing the rate decision is that we do not need to guess the future fading of the channel, removing a source of regret—that is, in the event that our prediction is wrong, we might waste time transmitting at too high a rate, so that the receiver cannot possibly decode the block, or at too low a rate, so that the receiver gets more redundancy than it needs.

From a theoretical perspective, ratelessness and feedback are not new concepts, and the characterizations given in §2.1.1 are more than sufficient. From a systems perspective, however, there seems to have been some disagreement over the term.

The definition of ratelessness in terms of not needing to make a decision is problematic: it fails to distinguish between lacking good options and having no regrets, and it fails to specify what, precisely, is being allowed to vary, and what is being held constant. In the systems literature, rateless codes are, invariably, fixed-to-variable codes (fixed number of bits per block; variable number of channel uses) for the AWGN channel and its fading variants. Such a scheme can be viewed as a family of block codes of different rates such that the code-words of the high-rate codes are prefixes of the code-words of the low-rate codes.

The central systems concept that the definition is trying to reach is this: we wish to eliminate regret by eliminating up-front decision-making without sacrificing performance. This extends beyond rate decisions to other factors that impact decoding performance. We call this larger class *abstemious codes*, because they avoid making choices that they might regret later.

We are not aware of any wireless system that entirely avoids the need for regrettable decisions in advance of transmission. For instance, existing rateless systems [28, 29]

still require up-front decisions of the number of MIMO streams and the allocation of power across sub-carriers. Both decisions impact optimal performance to a degree quantified by the water-filling algorithm [12, §5.3, p. 183].

In 2011–2012, the SIGCOMM and IEEE Trans. Info. Theory communities discussed ratelessness in the form of proposals from Perry et al [28], Gudipati et al [29], and Erez et al [30]. The distinguishing feature of these systems was that they obviated two particular up-front decisions: the length k of an encoded block, and the constellation order (e.g. 16-QAM). These systems were called rateless to capture both of these salient properties, but in fact the two properties are more-or-less orthogonal.

We offer an alternative terminology which addresses some of these headaches.

A system is *strongly abstemious* with respect to a parameter iff the system never suffers ex-post regret over this parameter—that is, a situation never arises where the receiver or sender wishes retroactively that it had set the parameter differently.

A system is *weakly abstemious* with respect to a parameter if the system may suffer regrets about design limitations affecting the parameter, but not about about runtime decisions.

System ratelessness is the property of fixed-to-variable codes which allows them to be abstemious with respect to rate (by varying codeword length, with block size held fixed).

Constellation universality is the property of certain codes that allows them to be abstemious with respect to constellation order.

One way a system can be weakly abstemious is if the abstemious property breaks in edge cases. Spinal codes [28] are constellation universal, but only up to a finite rate limit determined by the branching factor of the coding tree. Because this parameter has performance implications for the decoder, it must be set to a finite value, and this determines the maximum rate achievable by the system at runtime. The system may very well discover at runtime that channel conditions are unexpectedly favorable, so that the design limitation on performance is regrettable. However, aside from this edge case, spinal codes do not suffer from regret over the choice of constellation.

Another way a system can be weakly abstemious is if the property is constrained

by design to take on a discrete set of values, and at runtime some intermediate value is found to be preferable to those in the discrete set. For instance, in LTE HARQ, after an initial transmission at some rate R, the system can use retransmissions to amend its choice to $R/2$ or $R/3$, but not to some value in-between, even if we can determine that such a value would be optimal. We say that LTE HARQ is weakly abstemious with respect to rate, or weakly rateless.

The construction of spinal codes [28] is illustrative of practical rateless coding. To encode, a few message bits at a time are hashed into a state variable, forming a chain of cumulative hashes. Each hash is used to seed one pseudo-random number generator (PRNG), forming a chain of PRNGs. Next, each PRNG is invoked in round-robin order to pick, from a pre-set constellation, one symbol to transmit. Thus, any small or large quantity of redundancy can be produced from a given message, simply by continuing to invoke the PRNGs until decoding is successful. The resulting code-words are as random as possible, subject to a constraint that makes decoding feasible: any two messages with a common prefix generate hash sequences which share a common prefix, and hence result in identically-seeded PRNGs up to the length of the prefix. This causes a fraction of the transmitted symbols for the two messages to be identical, with the fraction being in proportion to the length of the common prefix.

The decoder exploits this property that common prefixes result in common symbols among related code-words. It performs a pruned tree search over all possible messages, computing hash chains as it proceeds. The pruning criterion is to keep a constant number of message candidates alive at each search depth, according to the similarity of the signal that would have been transmitted for a hypothetical message to the signal that was actually received.

2.2 Problem

Rateless systems like spinal codes [28], Strider [29], and LTE hybrid automatic repeat requests (HARQ) remove the need to make an up-front decision of the coded length of a packet, and hence the rate in bits per channel usage. Instead, they assume that the transmitter will send incremental redundancy until the receiver reports success. At that time, the system will move on to the next block. This concept represents a fundamental change to the model of communication adopted by traditional fixed-rate FEC schemes. Rather than describing a fire-and-forget transmission which either decodes or does not, as in a fixed-rate code, rateless codes expect to have a fine-grained stream of yes-no feedback from the receiver to the sender. In the absence of full duplex capability in typical radios, this feedback stream poses a distinct challenge to the protocol designer.

The question then arises, on an ordinary half-duplex link, when should the sender pause transmitting redundancy to solicit feedback from the receiver? If the sender chooses to pause only after sending a great deal of redundancy, then it can be confident that the receiver has decoded the message; however, it can also anticipate that, on average, more redundancy will have been sent than is strictly needed. Sending unneeded redundancy slows the system down. On the other hand, if the sender chooses to pause early, then the receiver has most likely not decoded the message, and the sender will, with high probability, needlessly incur an expensive feedback cycle.

In order for rateless codes to meet their potential in eliminating codeword length as a source of regret on half-duplex links, a novel link-layer protocol is needed that can balance these costs.

This is some delicious irony. Though we have pursued system ratelessness in the name of avoiding making up-front decisions about rate—that is, decisions about how much redundancy to send for a given block—we have now come full circle, and we are discussing a system that adapts rateless codes to half-duplex channels by making a decision up-front about how much redundancy to send before pausing for feedback.

Nevertheless, the need is real: rateless codes are fish out of water on a half-duplex

link. Without a rateless link protocol to schedule retransmissions optimally, their potential is wasted. The system presented in §2.6 meets this need. This rateless link protocol, which we call RateMore, learns the probability distribution of the number of channel uses required to decode a packet, and uses this learned distribution in a dynamic programming strategy to produce an optimal plan of when to pause for feedback. Fading-channel simulations show that this protocol reduces wasted channel resources by 2.6× to 3.9× compared to the design of Wi-Fi and by 2.8× to 5.4× compared to the design of LTE HARQ. Using information from feedback, RateMore achieves 98% of the throughput of an omniscient scheme that knows the channel's fading behavior in advance.

2.3 Marginalized-τ model

RateMore seeks to address the problem of scheduling rateless error-correction codes (EDT or OT) over half-duplex quasi-static point-to-point AWGN channels using only stop-feedback.

In general, τ depends on the channel outputs, and through them on both the message W being transmitted and the realization of the noise $\{n_k\}$ (which in turn depends on the state of the channel). In the case of oracle termination, τ is $\inf\{k \mid g_k(y_k) = W\}$. In either case, we only have stop-feedback, so the transmitter does not know the channel outputs. The greatest amount of knowledge the transmitter can have about τ comes from conditioning on W, conditioning on the channel state, conditioning on the most recent negative acknowledgement, if any, and marginalizing over the unknown channel outputs.

Such a family of conditional distributions has far too many degrees of freedom to represent in any naïve way. Moreover, it depends on implementation details of the rateless decoder. In the real world, decoders are large circuits that are power and latency bottlenecks in mobile devices, and they are subject to an extreme degree of proprietary micro-architectual optimization, including the introduction of sub-optimal approximations. In addition, an advantage of certain rateless codes is that

the decoder has free parameters (in spinal codes, the number of hypotheses alive at each step) that can be chosen to trade off rate against computational complexity. Any scheduler thata assumes that it can perfectly model the decoder to determine when to terminate a transmission then becomes *prima facie* fragile. We therefore prefer to minimize RateMore's dependence on the details of the decoder. If possible, we'd like to treat decoder optimality and/or effort as part of the channel state, and rely only on statistical observations of the combined channel-decoder gestalt to drive scheduling.

In particular, if we can eliminate the dependence of τ on W, then the conditional distribution of τ given the channel state, marginalized over the channel outputs, encodes everything the transmitter can know about τ.

Under what conditions might τ be independent of W? Clearly, if this property exists, it is fragile: knocking out a handful of nearest-neighbor messages to some message W', i.e. removing them from the code, will hardly affect the rate at all, but it will allow a candidate decoding of W' to pass any given confidence threshold in less time. Perhaps if the code-book formed a lattice over some module whose base ring has a cyclic additive group; but this is certainly not true on the AWGN channel.

In the asymptotic regime of error detection termination where only the correct message and its nearest neighbor(s) are plausible candidates, the termination time can be bounded using the relative rates of growth of the nearest-neighbor distance and the noise radius with k. Unfortunately, the identity of the nearest neighbor to W may vary with k, and the running of the nearest-neighbor distance with k may certainly vary with W. For instance, truncated to one value of k, a given message may lie far out on the tail of the distribution of code-word norms $\|x_k\|_2$, and its nearest neighbor may be relatively distant; while for a later time k' its code-word norm may have regressed towards the mean.

Fully justifying the assumption that τ is independent of W is beyond the scope of this work. We nevertheless introduce an approximation, modeling τ as if it were a random variable independent of W whose distribution reflects the indirectly-observable channel and decoder state. In essence, our model for τ is a random variable independent of everything else, distributed according to the true marginal distribution of τ over

messages and the transmitter's beliefs about channel state, decoder optimality, and decoder effort.

Within this marginalized-τ model, and assuming only stop-feedback, RateMore is essentially optimal. That is, RateMore can be used to find ε-optimal transmission schedules for any $\varepsilon > 0$, where a transmission schedule is a sequence {number of symbols sent before n^{th} pause for feedback}$_n$, and optimality is with respect to time-averaged rate $\mathbb{E}[\tau R]/\mathbb{E}[\tau] = n/\mathbb{E}[\tau]$.

2.4 Notation

Relaxing the problem to the marginalized-τ model with stop-feedback allows us to represent the transmitter's knowledge about τ by a vector

$$p_i \triangleq \mathbb{P}(\tau \leq i \,|\, \text{current channel and decoder beliefs}) \tag{2.8}$$

From this point on, we will drop the conditional notation. We will also need, for $k \geq i$,

$$\begin{aligned}\mathbb{P}(\tau > k \,|\, \tau > i) &= \mathbb{P}(\tau > k \wedge \tau > i)/\mathbb{P}(\tau > i) \\ &= \mathbb{P}(\tau > k)/\mathbb{P}(\tau > i) \\ &= (1 - p_k)/(1 - p_i)\end{aligned}$$

and

$$\begin{aligned}n_f &= \text{time cost in channel usages to pause for feedback} \\ p_f &= \mathbb{P}(\text{feedback decoded successfully})\end{aligned}$$

2.5 Related work

Lucani et al [31] solve the related problem of scheduling random linear network codes (RLNC) over half-duplex point-to-point and multi-cast links. RLNC are rateless

packet-erasure-correction codes: at each time k, the channel will either deliver a random linear combination (over a finite field of order q) of M blocks, or an erasure. Decoding succeeds as soon as the receiver has M linearly-independent combinations.

2.5.1 Stop-Only Feedback

Their system explicitly depends on non-stop-only feedback. Stop-only feedback would seem to be a particularly poor fit to RLNC over the erasure channel, where the true conditional distribution of τ at each time k is highly sensitive to the number of erasures in y_k: in this case, it is advantageous to feed back e.g. the number of erasures.

Polyanskiy's results indicated that stop-only feedback was essentially sufficient: for $0 < \varepsilon < 1$, the achievability bounds for VLF and VLFT codes use stop-only feedback, and for $\varepsilon = 0$, the achievability bound for VLFT uses stop-only feedback. So why does the packet erasure channel seem to require non-stop-only feedback? Is stop-only feedback a significant limitation of RateMore after all?

A few observations. First, we should point out that the packet erasure channel has $C_1 = \infty$, since the K-L divergence between the conditional output distributions for distinct messages is infinite. So the null zero-error capacity result (2.5) for VLF, and hence EDT, does not apply. Instead, the lower bound (2.7) applies.

Next, we observe that there is no difference between oracle termination and error detection on the packet erasure channel. Decoding is algebraic; it either succeeds or fails, and if it succeeds, then the message is correct. Thus, RLNC over the packet erasure channel is a zero-error VLF or VLFT code. The capacity of the packet erasure channel for n-bit packets is $(1 - p_e)n$, where p_e is the probability of an erasure. Then the best zero-error VLFT rate is [3, Thm. 11]

$$(1 - p_e)n + \mathcal{O}(1/\ell) \leq R^*_{\mathrm{VLFT}}(\ell, 0) \leq (1 - p_e)n + (1/\ell)\log \ell + \mathcal{O}(1/\ell)$$

where $\ell = M/(1 - p_e)$, and the achievability bound uses stop-only feedback.

The discussion in Lucani et al leads us to believe that the system is envisioned to code over groups of $M = 5$ to 10 packets at a time, each of $n = 10000$ bits, though this

is of course not a limitation. Substituting these values suggests that, unless the hidden constant factors are very large, the leading term dominates by three to four orders of magnitude. This would seem to indicate that it is possible to design a rateless alternative to RLNC that works over the packet erasure channel, achieves zero error and essentially the best possible rate, and only requires stop-feedback. In view of this, the restriction of RateMore to stop-only feedback is not a significant disadvantage.

2.5.2 RLNC and Marginalized-τ

While RLNC codes over the packet erasure channel do not have the stop-only feedback property, they do have the property that the termination time τ is independent of the message W. The termination time only depends on the realization of the random linear combinations and erasures of packets and ACKs. So a restriction to the marginalized-τ model is not sub-optimal in the setting of Lucani et al.

2.5.3 Scheduling

Because the time-cost of feedback is large and more-or-less independent of bandwidth, both schemes amortize the cost of feedback over a large number of coded bits. In RateMore, this is by aggregating incremental redundancy for a number of in-flight ratelessly-coded blocks, which are scheduled and decoded independently. In Lucani et al, RLNC is applied over a block of M (~ 5 to 10) packets.

In order to physically realize a packet erasure channel like the one in Lucani et al on an AWGN wireless link, one must introduce an underlying error-correcting block code operating in the finite-ε regime. For Lucani et al, if the data bandwidth is low (e.g. voice over IP), then the queues will never have a depth greater than one, and hence $M = 1$. In this case [32, Eq. 13], the scheme is Type-I HARQ with repetition coding and no combining. This is sub-optimal with respect to the inner wireless channel: one should prefer to code for the inner channel with incremental redundancy, so that at the very least a retransmission attempt may take advantage of the time and energy spent on prior attempts to reach a lower error rate.

Alternatively, one should prefer to modulate the ε of the underlying AWGN block code towards a lower value when aggregation is not possible, since we are going to have to perform error recovery anyway, and reducing ε before the first transmission attempt is less costly than incurring a re-transmission.

In fact, these alternatives describe RateMore precisely. If the amortized cost of feedback grows because the number of in-flight coded blocks falls, RateMore schedules its retransmissions more conservatively, allocating a larger amount of redundancy to each block, and achieving a lower effective rate with a correspondingly lower rate of decode failures on the first attempt. If an attempt fails, RateMore's inner rateless AWGN code permits all subsequent attempts to take full advantage of incremental redundancy.

Put another way, the quantum of scheduling for the scheme of Lucani et al is one degree-of-freedom (DoF), which corresponds to an interval of transmission time as long as one entire (FEC-coded but not network-coded) packet. The quantum of scheduling for RateMore can be as small as a single channel use, so that the effective coding rate varies smoothly. This allows RateMore to instantly and accurately rebalance the amortized cost of feedback in the event of varying queue depth or low offered load.

2.6 Construction and Implementation

We seek a schedule of pauses that minimizes the expected time-to-completion (ETTC) for a given block using a given rateless code. We assume that n_f and p_f are given, and we suppose, for now, that the distribution parameters p_i (2.8) are known.

Now let

$$
\begin{aligned}
t_{ijk}\colon \text{ETTC} &\left| \begin{array}{l} i \geq 0 \text{ symbols NAKed, } j \geq i \text{ symbols sent,} \\ \text{next pause after } k \geq 0 \text{ more symbols,} \\ \text{remaining pauses optimal} \end{array} \right. \\
&= k + n_f + p_f \frac{1 - p_{j+k}}{1 - p_i} t^*_{j+k,j+k} + (1 - p_f) t^*_{i,j+k} \\
t^*_{ij}\colon \text{ETTC} &\left| \begin{array}{l} i \geq 0 \text{ symbols NAKed, } j \geq i \text{ symbols sent,} \\ \text{remaining pauses optimal} \end{array} \right. \\
&= \min_{k \geq 0} t_{ijk}
\end{aligned}
$$

If we define an auxiliary quantity, $s_{ijk} = j + t_{ijk}$, then we may observe that s depends only on i and $j + k$, and not on j or k independently. Furthermore, the $\arg\min_k$ of s_{ijk} is the same as that of t_{ijk}. So we can rewrite the problem in terms of a two-index quantity s_{ij}, defined as

$$
\begin{aligned}
s_{ij}\colon \text{ETTC + symbols sent so far} &\left| \begin{array}{l} i \text{ symbols NAKed,} \\ \text{next pause after } j \text{ total symbols sent,} \\ \text{remaining pauses optimal} \end{array} \right. \\
&= j + n_f + p_f \frac{1 - p_j}{1 - p_i} t^*_{jj} + (1 - p_f) t^*_{ij} \\
t^*_{ij} &= \min_{k \geq 0} s_{i,j+k} - j \\
&= \min\{s_{ij} - j, t^*_{i,j+1} + 1\}
\end{aligned}
$$

Now s_{ij}, $j \geq i \geq 0$ depends on t^*_{jj} and t^*_{ij}, and t^*_{ij}, $j \geq i \geq 0$ depends on s_{ij}. We can generally perform the computation of these quantities in the order

1. For all i, largest to smallest; then
2. For all j, largest to smallest; then
3. Compute s_{ij}, then t^*_{ij}.

However, there is a circular dependence between t^*_{ij} and s_{ij}. This corresponds to the

possibility that sending 0 additional symbols before pausing for feedback might be optimal, which occurs iff

$$
\begin{aligned}
i = j :& \\
& n_f = 0 \\
& s_{ii} = i + t^*_{ii} \\
& s_{ii} \leq i + 1 + t^*_{i,i+1} \\
i < j :& \\
& \frac{n_f}{p_f} \leq 1 + t^*_{i,j+1} - \frac{1-p_j}{1-p_i} t^*_{jj}
\end{aligned}
$$

The first case ($i = j$ and $n_f = 0$) is uninteresting because if feedback is free, then there is no need to schedule pauses, because any schedule is optimal. We therefore assume that $n_f > 0$. Now we can break the cyclic dependency, in the case where it matters (i.e. the min takes the left branch), by solving for s_{ij}, to get

$$
\begin{aligned}
s_{ij} &= \begin{cases} j + \frac{n_f}{p_f} + \frac{1-p_j}{1-p_i} t^*_{jj} & (i < j) \wedge (\frac{n_f}{p_f} + \frac{1-p_j}{1-p_i} t^*_{jj} \leq 1 + t^*_{i,j+1}) \\ j + n_f + p_f \frac{1-p_j}{1-p_i} t^*_{jj} + (1-p_f) t^*_{ij} & \text{otherwise} \end{cases} \\
t^*_{ij} &= \min\{s_{ij} - j, t^*_{i,j+1} + 1\}
\end{aligned}
$$

The value of s_{ij} does not matter if it is not chosen as the minimum. So we can rewrite this as

$$
\begin{aligned}
& t^*_{ij} := t^*_{i,j+1} + 1 \\
& \text{If } (i < j) \text{ then} \\
& \qquad t^*_{ij} := \min\left\{ t^*_{ij}, \frac{n_f}{p_f} + \frac{1-p_j}{1-p_i} t^*_{jj} \right\}
\end{aligned}
$$

The body of the if statement is a no-op if $i = j$, so it can be executed unconditionally. Additionally, as is typical in dynamic programs, we use an auxiliary back-tracking array to store the $\arg\min$ that goes along with each min. Defining min over tuples

lexicographically,

$$\begin{aligned} (t^*,\, k^*)_{ij} &:= (t^*_{i,j+1} + 1,\, k^*_{i,j+1} + 1) \\ (t^*,\, k^*)_{ij} &:= \min\left\{ (t^*,\, k^*)_{ij}\,,\, \left(\tfrac{n_f}{p_f} + \tfrac{1-p_j}{1-p_i} t^*_{jj},\, 0\right) \right\} \end{aligned} \tag{2.9}$$

(n.b. these assignments are sequential because t^*_{jj} may refer to the result of the first statement)

We perform the computation of these quantities in the order

1. For all i, largest to smallest; then

2. For all j, largest to smallest.

The computation terminates when we reach the point (i, j) with i the actual number of symbols NAKed and j the actual number of symbols sent. k^*_{ij} then indicates the optimum number of symbols to send before pausing.

2.6.1 Truncating the Infinite Recursion

Let q^* represent the optimal schedule under marginal-τ cumulative distribution function (τ-CDF) $\{q_i\}$, and let p^* represent the optimal schedule under τ-CDF $\{p_i\}$.

Let $\mathrm{ETTC}_p[q^*]$ represent the expected time to completion of the strategy q^* under τ-CDF $\{p_i\}$.

We would like to establish that truncating the infinite recursion of the previous section—that is, starting from finite i, j—introduces a departure from optimality that is bounded.

One way to effect this truncation is to set $p_\omega = 1$ for some finite ω to get a new τ-CDF $\{q_i\}$. The dynamic program on this new CDF takes finite time. Our goal, then, is to bound the gap to optimality of the q-optimal strategy q^* when used with the true τ-CDF $\{p_i\}$. Theorem 2.6.1 gives the desired bound. We will require the following lemma.

Lemma 2.6.1 *If a non-negative random variable τ has finite mean $\bar{\tau}$ and variance*

σ^2, and $\omega > \bar{\tau}$, then

$$\mathbb{E}[\tau \mid \tau > \omega] \leq \frac{2\omega\sigma^2}{(\omega - \bar{\tau})^2 \, \mathbb{P}(\tau > \omega)} \tag{2.10}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\tau \mid \tau > \omega] &= \sum_{k>\omega} k \, \mathbb{P}(\tau = k \mid \tau > \omega) \\
&= \frac{1}{\mathbb{P}(\tau > \omega)} \sum_{k>\omega} k \, \mathbb{P}(\tau = k) \\
&= \frac{1}{\mathbb{P}(\tau > \omega)} \sum_{k>\omega} \left(k \, \mathbb{P}(\tau > k-1) - k \, \mathbb{P}(\tau > k)\right)
\end{aligned}$$

Since we started with a finite quantity, this is a convergent series. If we change the associativity by telescoping the sum, we will get a new convergent series with the same limit provided that the limit of the sequence $\{k \, \mathbb{P}(\tau > k-1), k \, \mathbb{P}(\tau > k), \ldots\}_{k=\omega+1}^{\infty}$ exists and is zero. In this case, the limit does exist and is zero, because if the tail of the CDF does not go to zero at least as fast as $1/k$, then the variance does not exist, which is a contradiction. So we can write

$$\begin{aligned}
\mathbb{E}[\tau \mid \tau > \omega] &= \frac{1}{\mathbb{P}(\tau > \omega)} \left((\omega + 1) \, \mathbb{P}(\tau > \omega) + \sum_{k>\omega} \left(-k \, \mathbb{P}(\tau > k) + (k+1) \, \mathbb{P}(\tau > k)\right) \right) \\
&= \frac{1}{\mathbb{P}(\tau > \omega)} \left((\omega + 1) \, \mathbb{P}(\tau > \omega) + \sum_{k>\omega} \mathbb{P}(\tau > k) \right) \\
&= \frac{1}{\mathbb{P}(\tau > \omega)} \left((\omega + 1) \, \mathbb{P}(\tau \geq \omega + 1) + \sum_{k>\omega} \mathbb{P}(\tau \geq k + 1) \right)
\end{aligned}$$

By the one-tailed Chebyshev's inequality, for $k \geq \omega > \bar{\tau}$,

$$\mathbb{P}(\tau \geq k) \leq \frac{\sigma^2}{\sigma^2 + (k - \bar{\tau})^2}$$

So we have

$$
\begin{aligned}
\mathbb{E}[\tau \mid \tau > \omega] &\leq \frac{1}{\mathbb{P}(\tau > \omega)} \left((\omega + 1) \frac{\sigma^2}{\sigma^2 + (\omega + 1 - \bar{\tau})^2} + \right. \\
&\qquad \left. \sum_{k \geq \omega + 1} \frac{\sigma^2}{\sigma^2 + (k + 1 - \bar{\tau})^2} \right) \\
&\leq \frac{\sigma^2}{\mathbb{P}(\tau > \omega)} \left((\omega + 1) \frac{1}{\sigma^2 + (\omega - \bar{\tau})^2} + \right. \\
&\qquad \left. \int_{\omega+1}^{\infty} \frac{dk}{(k - \bar{\tau})^2} \right) \\
&\leq \frac{\sigma^2}{\mathbb{P}(\tau > \omega)} \left(\frac{\omega + 1}{(\omega - \bar{\tau})^2} + \frac{1}{\omega - \bar{\tau}} \right) \\
&\leq \frac{\sigma^2}{\mathbb{P}(\tau > \omega)} \frac{2\omega + 1 - \bar{\tau}}{(\omega - \bar{\tau})^2} \\
&\leq \frac{2\omega\sigma^2}{(\omega - \bar{\tau})^2 \, \mathbb{P}(\tau > \omega)}
\end{aligned}
$$

■

Theorem 2.6.1 *Suppose that two cumulative distributions $\{p_i\}$, $\{q_i\}$ satisfy*

$$
\begin{aligned}
\forall i \leq \omega \qquad & q_i = p_i \\
\forall i \qquad & q_i \geq p_i \\
& \mathrm{Var}_p\{\tau\} < \infty
\end{aligned}
$$

Then for any $\varepsilon > 0$,

$$
\mathrm{ETTC}_p[q^*] - \mathrm{ETTC}_p[p^*] \leq \varepsilon
$$

provided that

$$
\omega \geq 2 \left(\mathbb{E}_p[\tau] + \left(1 + \frac{n_f}{p_f} \right) \frac{\mathrm{Var}_p\{\tau\}}{\varepsilon} \right).
$$

Proof. We write the gap to optimality as

$$\begin{aligned}
\text{gap} &= \text{ETTC}_p[q^*] - \text{ETTC}_p[p^*] \\
&= \text{ETTC}_p[q^* \,|\, \tau \le \omega\,]\,\mathbb{P}(\tau \le \omega) + \text{ETTC}_p[q^* \,|\, \tau > \omega\,]\,\mathbb{P}(\tau > \omega) \quad - \text{ETTC}_p[p^*] \\
&= \text{ETTC}_q[q^* \,|\, \tau \le \omega]\,p_\omega + \text{ETTC}_p[q^* \,|\, \tau > \omega\,]\,(1 - p_\omega) \quad - \text{ETTC}_p[p^*] \\
&\le \text{ETTC}_q[q^*] + \text{ETTC}_p[q^* \,|\, \tau > \omega\,]\,(1 - p_\omega) \quad - \text{ETTC}_p[p^*]
\end{aligned}$$

The first term is the q-ETTC of the q-optimal strategy. Since $q_i \ge p_i$ for all i, this is bounded above by the p-ETTC of the p-optimal strategy: under the q distribution, τ only terminates earlier, so the optimum strategy cannot be slower.

The second term is bounded above by the worst possible p-strategy: pausing for feedback after every symbol, and repeating the pause for feedback until the ACK decodes successfully. The expected time spent per symbol under this strategy is $(1 + n_f/p_f)$, and the expected number of symbols is $\mathbb{E}[\tau]$; since τ is independent of whether ACKs decode successfully, we can multiply these expectations together and apply Lemma 2.6.1 to obtain

$$\begin{aligned}
\text{gap} &\le \left(1 + \frac{n_f}{p_f}\right)(1 - p_\omega)\mathbb{E}_p[\tau \,|\, \tau > \omega\,] \\
&\le \left(1 + \frac{n_f}{p_f}\right)(1 - p_\omega)\frac{2\omega {\sigma_p}^2}{(\omega - \bar{\tau}_p)^2(1 - p_\omega)} \\
&\le \left(1 + \frac{n_f}{p_f}\right)\frac{2\omega {\sigma_p}^2}{(\omega - \bar{\tau}_p)^2}
\end{aligned}$$

Now if we take $\omega = a\bar{\tau}_p$, then

$$\begin{aligned}
\text{gap} &\le \left(1 + \frac{n_f}{p_f}\right)\frac{{\sigma_p}^2}{\bar{\tau}_p}\frac{2a}{(a-1)^2} \\
&\le \left(1 + \frac{n_f}{p_f}\right)\frac{{\sigma_p}^2}{\bar{\tau}_p}\frac{2}{a-2} \text{ for } a > 2.
\end{aligned}$$

So we can make the gap as small as we like by a suitable choice of a. For instance, if

$$a \geq 2\left(1 + \left(1 + \frac{n_f}{p_f}\right)\frac{{\sigma_p}^2}{\bar{\tau}_p}\frac{1}{\varepsilon}\right)$$

then the gap to optimality is no larger than ε. This is equivalent to the choice

$$\omega \geq 2\left(\bar{\tau}_p + \left(1 + \frac{n_f}{p_f}\right)\frac{{\sigma_p}^2}{\varepsilon}\right)$$

■

Theorem 2.6.1 implies that truncating the dynamic program to a recursion depth of $\omega = \mathcal{O}(1/\varepsilon)$ is sufficient to obtain a ε-optimal schedule for the un-truncated problem. In fact, any alteration of the CDF tail is acceptable, provided that two conditions are met:

1. The CDF only increases at each point, and
2. The CDF is unaltered up to argument ω.

Soft truncation

One disadvantage of hard truncation ($q_i = 1$ for $i > \omega$) is that it does not give us any sensible schedule to follow in the unlikely event that $\tau > \omega$. An alternative heursitic is to replace the tail of the true τ-CDF with a memoryless distribution, so that the dynamic program has a fixed point:

$$\begin{aligned}
\mathbb{P}(\tau = \omega + 1 + i \,|\, \tau > \omega) &= (1-\alpha)\alpha^i \\
\forall k > \omega. \quad q_k &= \sum_{i=0}^{k} \mathbb{P}(\tau = i) \\
&= p_\omega + (1 - p_\omega)\sum_{i=\omega+1}^{k} \mathbb{P}(\tau = i \,|\, \tau > \omega) \\
&= p_\omega + (1 - p_\omega)(1 - \alpha^{k-\omega}) \\
&= 1 - (1 - p_\omega)\alpha^{k-\omega} \\
\forall \omega < i \leq j. \quad \frac{1 - q_j}{1 - q_i} &= \alpha^{j-i}
\end{aligned}$$

We next solve the dynamic program for its fixed point in the form of two arrays $(t^*, k^*)_{i-j}$:

$$
\begin{aligned}
t^*_{ij} &= k^*_{ij} + \frac{n_f}{p_f} + \frac{1-q_{j+k^*_{ij}}}{1-q_i} t^*_{j+k^*_{ij}, j+k^*_{ij}} \\
k^*_{ij} &= \arg\min_{k\geq 0} \left(k + n_f + p_f \frac{1-q_{j+k}}{1-q_i} t^*_{j+k,j+k} + (1-p_f) t^*_{i,j+k} \right)
\end{aligned}
$$

$$
\forall \omega < i \leq j. \quad \begin{cases}
t^*_{j-i} &= k^*_{j-i} + \frac{n_f}{p_f} + \frac{1-q_{j+k^*_{j-i}}}{1-q_i} t^*_0 \\
&= k^*_{j-i} + \frac{n_f}{p_f} + \alpha^{j-i+k^*_{j-i}} t^*_0 \\
k^*_{j-i} &= \arg\min_{k\geq 0} \left(k + n_f + p_f \frac{1-q_{j+k}}{1-q_i} t^*_0 + (1-p_f) t^*_{j-i+k} \right) \\
&= \arg\min_{k\geq 0} \left(k + p_f \alpha^{j-i+k} t^*_0 + (1-p_f) t^*_{j-i+k} \right)
\end{cases}
$$

$$
\forall i \geq 0. \quad \begin{cases}
t^*_i &= k^*_i + \frac{n_f}{p_f} + \alpha^{i+k^*_i} t^*_0 \\
k^*_i &= \arg\min_{k\geq 0} \left(k + p_f \alpha^{i+k} t^*_0 + (1-p_f) t^*_{i+k} \right)
\end{cases}
$$

Relaxing the integer minimization to a continuous minimization,

$$
\forall i \geq 0. \quad \begin{cases}
t^*_i &= k^*_i + \frac{n_f}{p_f} + \alpha^{i+k^*_i} t^*_0, & \text{and either} \\
0 &= 1 + p_f \alpha^{i+k^*_i} t^*_0 \log \alpha + (1-p_f)\, \partial_k t^*_{i+k} \Big|_{k=k^*_i} & \text{or} \\
k^*_i &= 0
\end{cases}
$$

For the special case $p_f = 1$, this has the solution

$$
\begin{aligned}
t^*_0 &= k^*_0 + n_f + \frac{1}{\log \frac{1}{\alpha}} & t^*_i &= \max\{0, t^*_0 - i\} \\
k^*_0 &= \log_{1/\alpha} \left((k^*_0 + n_f) \log \frac{1}{\alpha} + 1 \right) & k^*_i &= \max\{0, k^*_0 - i\}
\end{aligned}
$$

This can be evaluated with a few iterations of

$$
\begin{aligned}
A &:= 1 + n_f \log \frac{1}{\alpha} \\
B^0 &:= 1 \\
B^{k+1} &:= \log A + \log(1 + B^k) \\
k^*_0 &:= \frac{AB}{\log \frac{1}{\alpha}}
\end{aligned}
$$

In the $p_f < 1$ case, if we take $\partial_k t^*_{i+k}\Big|_{k=k^*_i} \approx -1$ based on our experience with $p_f = 1$, then we get

$$t^*_0 = k^*_0 + \frac{n_f}{p_f} + \frac{1}{\log \frac{1}{\alpha}} \qquad t^*_i = \max\{0, t^*_0 - i\}$$
$$k^*_0 = \log_{1/\alpha}\left(\left(k^*_0 + \frac{n_f}{p_f}\right)\log\frac{1}{\alpha} + 1\right) \qquad k^*_i = \max\{0, k^*_0 - i\},$$

which can be evaluated using the same iteration as before under the substitution $n_f \to \frac{n_f}{p_f}$ in the definition of A.

When the dynamic program (2.9) evaluates t^*,k^* for $(i,j) \preccurlyeq \omega$, any accesses to variables defined outside of this domain can be substituted as $t^*_{ij} \to t^*_{j-i}$ and $k^*_{ij} \to k^*_{j-i}$ using the results of the fixed-point iteration. Similarly, if the protocol requires access to these variables in the unlikely event that $\tau > \omega$, it can perform the same substitutions.

2.6.2 Pseudocode

See Algorithm 1. There is a subtlety having to do with the proper behavior when $\hat{\jmath} > \omega \geq \hat{\imath}$, i.e. the DP's schedule is truncated because $\hat{\jmath}$ is large, not because $\hat{\imath}$ is large. In this case, based on our ε-optimality results, we are free to fall back on the "worst-case" behavior: to demand as many pauses for acknowledgement as are required until an acknowledgement is decoded, and we end up in the $\hat{\jmath} = \hat{\imath} > \omega$ regime where termination applies.

2.6.3 Complexity

Since each subproblem (i,j) relies on subproblems (j,j) and $(i,j+1)$, we need linear storage to keep the diagonal solutions (j,j) alive for the entire problem, plus constant storage to keep $(i,j+1)$ alive from the previous iteration. So the time complexity is $\mathcal{O}(\omega^2)$, and the space complexity is $\mathcal{O}(\omega)$, where $\omega = \mathcal{O}(\bar{\tau} + \frac{n_f}{p_f}\frac{\sigma^2}{\varepsilon})$.

Algorithm 1: RateMore scheduler.

Input : n_f time cost of a pause for feedback
p_f probability that a feedback packet is decoded successfully
p_i marginal-τ probability that $\tau \leq i$
ω truncation point of the dynamic program
$\hat{\imath}$ number of symbols NAKed
$\hat{\jmath}$ number of symbols sent
t_i^* one-index ETTC in the event of truncation
k_i^* one-index symbols-before-next-pause in the event of truncation

Output : $k_{\hat{\imath}\hat{\jmath}}^*$ optimal number of symbols to send before pausing for feedback

```
if î > ω then
    return k*_{ĵ−î}
else if ĵ > ω then
    return 0
else
    for i ← ω to î do
        t_right ← t*_{ω−i+1}
        k_right ← k*_{ω−i+1}
        for j ← ω to ĵ do
            # (t*_ij, k*_ij) ← (t*_{i,j+1} + 1, k*_{i,j+1} + 1)
            # (t*_ij, k*_ij) ← min{ (t*, k*)_ij , ( n_f/p_f + (1−p_j)/(1−p_i) t*_jj, 0) }
            if i = j then
                t*_ij ← t_right + 1
                k*_ij ← k_right + 1
            else
                t_pause ← n_f/p_f + (1−p_j)/(1−p_i) t*_jj
                if t_pause < t_right + 1 then
                    t*_ij ← t_pause
                    k*_ij ← 0
                else
                    t*_ij ← t_right + 1
                    k*_ij ← k_right + 1
            t_right ← t*_ij
            k_right ← k*_ij
    return k*_îĵ
```

2.7 Learning the Decoding CDF

The marginal τ-CDF is the primary input to the dynamic programming algorithm, specifying the probability that a given number of symbols will be required by the receiver for successful decoding. These probabilities depend on current channel conditions. When the channel can be characterized by a single parameter, we can obtain CDFs for different values of that parameter using off-line simulations or experiments. For example, in the case of the additive white Gaussian noise channel (AWGN) we could obtain CDFs for a range of signal-to-noise ratio (SNR) values. In practice, however, wireless radios operate in complex environments, and we expect the channel to have too many parameters for such an approach to be practical; we expect these parameters to vary unpredictably over time; and we expect that off-line simulations will differ significantly from the actual implemented system, particularly when encoding and decoding systems are implemented by a variety of hardware vendors.

A robust alternative to this approach is to *learn* the CDF on the fly: that is, to estimate the CDF based on the recent history of number of symbols required for successful decoding at the receiver. The sender always consults the strategy derived from the most recent CDF estimate. This form of online learning directly accommodates variations in channel conditions due to fading and mobility.

The most general empirical distribution for the probability of successful decoding after any number number of symbols is the multinomial distribution. Thus, a very general Bayesian approach would be to learn this multinomial distribution beginning from a Dirichlet prior. This entails maintaining a histogram over the number of symbols required for decoding so far. While straightforward, a model with such a large state-space leads to slow learning and slow adaptation to variations in the channel.

Ideally, we would like to maintain a parametric distribution with a minimal number of parameters as our surrogate for the actual CDF. With this in mind, we propose two approaches: (a) Gaussian approximation, and (b) Constant-plus-Geometric approximation. From these, we pick the Gaussian approximation, for the

reasons explained below.

Gaussian approximation Our inspection of the decoding CDFs indicates that the Gaussian distribution should be a reasonable approximation at low SNR. Maximum-likelihood (ML) estimation of a Gaussian distribution requires nothing more than computing the empirical mean and variance of the number of symbols needed for successful decoding, which can be accomplished with a few accumulators. In the face of time-varying conditions, the empirical mean and variance can be *filtered* using a moving average or a similar scheme.

We selected Algorithm 2 for learning a Gaussian CDF. The algorithm keeps exponentially-weighted moving averages (EWMAs) to estimate the mean, $\bar{\tau}$, and variance, σ^2; in addition, it keeps an EWMA of the number of samples it has seen. The EWMA parameter α ranges from 0 (no memory) to 1 (unlimited memory). The averages track the input with a time constant of $1/\ln(1/\alpha)$.

This scheme has three advantages over a traditional EWMA of the form $y \leftarrow (1-\alpha)x + \alpha y$. First, the start-up transient dies out more quickly, because we weight our initial conditions less heavily. Second, the estimator's behavior is well-defined for $\alpha = 1$: inputs are retained forever, and we recover the ML estimator. And third, the estimator framed in this way offers a natural way to gradually reweight and forget old input without having the predicted mean and variance regress towards zero during an idle period.

We present results in §2.10 showing that under the Gaussian approximation, performance of RateMore when the true decoding CDF is not available differs by only a few percent from performance when the true CDF is available. At low SNR, the Gaussian approximation with a memory parameter of $\alpha = 0.99$ performs within 1% of the known-CDF case. We also show that very aggressive learning rates of $\alpha = 0.80$ perform very well on simulated fading channels.

Constant-plus-Geometric approximation The Constant-plus-Geometric approximation treats the random variable τ as $\tau = c + g$, where c is a constant and g is

Algorithm 2: Gaussian learning with EWMA for the marginal distribution of τ. Long-term performance is not sensitive to initial values.

```
1  def init():                                  # initialize the state machine
2      samples ← 1
3      sum     ← 100                                          # for instance
4      sumsq   ← sum^2 + 10^2                                 # for instance
5  def learn(sample, α):                      # update state with a new sample
6      samples ← samples · α + 1
7      sum     ← sum     · α + sample
8      sumsq   ← sumsq   · α + sample^2
9  def forget(idle_time, α):                # update state with an idle period
10     samples ← samples · α^idle_time
11     sum     ← sum     · α^idle_time
12     sumsq   ← sumsq   · α^idle_time
13 def get_ccdf(x):                    # evaluate estimated complementary CDF
14     τ̄   = sum/samples
15     σ^2 = sumsq/samples − τ̄^2
16     return normal_ccdf(τ̄, σ^2, x)
```

a geometric random variable with parameter α. Like the Gaussian, this family of distributions has two parameters, c and α. We considered this approximation because of the analogy with the presence of *error exponents* in most good codes. When a code has an error exponent, the probability of a decoding error after transmitting τ symbols scales as $\exp(-\gamma n)$ for some constant γ as long as τ is large enough to push the rate below the channel capacity. This minimal τ corresponds to the constant summand c, and the exponential drop in error probability with increasing τ corresponds to the geometric drop in probability of unsuccessful decoding with increasing g.

Additionally, if this approximation is good, then the resulting dynamic programming strategy is essentially the same as that of the fixed point computed above. Only the first transmission is modified, and it is changed by the addition of c additional symbols before the first pause.

Unfortunately, despite this elegance, this distribution does not have a convenient maximum-likelihood estimator. The most likely value of the constant c is upper bounded by the min of τ observed over all decoding attempts. In practice, this will

lead the dynamic programming algorithm to under-estimate the number of symbols necessary for decoding. For this reason, we performed our experiments using the Gaussian approximation.

2.8 Packing Packets and Block ACKs

Rateless codes, particularly in situations of high bandwidth and high tolerance for latency, may benefit from a method to pack symbols belonging to multiple distinct network-layer packets into the same link-layer frame, and using a "block ACK" scheme (as in 802.11e/n) to amortize the cost of sending feedback about the decoding state of each packet. Packing packets into a single frame is a useful amortization not only when the packets are small in size, but also when only a few more symbols are required to be transmitted to successfully decode the packet. As link rates increase, it is likely that any code (whether rateless or using incremental redundancy) that decodes using symbols from previous frame transmissions will benefit from packing multiple distinct packets into the same frame and using block ACKs. In RateMore, this amortization is accomplished by dividing the cost of feedback n_f among all active packets.

(Unfortunately, systems and coding terminology conflict in this case. A block ACK is a *block of ACKs*, not the ACK for a (code) block.)

While fixed-rate codes would also benefit from reduced feedback time, RateMore intelligently trades some of this reduction to obtain more rounds of feedback on each packet. Such a fine-grained feedback schedule decreases the risk that many symbols will be sent beyond those necessary for decoding.

If per-packet latency were not a concern, we could pack as many packets as we wished into one frame, making n_f arbitrarily small. RateMore's scheduling algorithm assumes that n_f is exogenously given by latency requirements, and that it has already been amortized over multiple in-flight packets packed into one frame (Figure 2-3).

We can use a trick to reduce the cost of packing and unpacking symbols from multiple packets (coded blocks) into a single transmitted frame. The straightforward, robust approach is to encode the number of symbols included in a transmitted frame

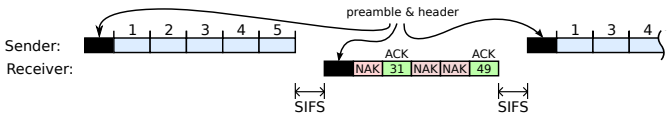


Figure 2-3: Symbols from many packets are aggregated into a frame in order to amortize the cost of the inter-frame spaces (in 802.11n, SIFS, 16 μs) and preambles/headers.

from each in-flight packet into a header using a low-rate block code. The trick is to recognize that the sender's decoding CDF is learned from the receiver's feedback, and hence both the sender and receiver can agree on the exact CDF and the exact feedback schedule currently in use (supposing that ω and n_f/p_f are agreed upon in advance). Then all the receiver requires in the frame header is a single bit indicating whether the most recent block ACK was decoded successfully, and a length field for the entire transmission.

This bit allows it to determine, for each in-flight packet that it currently knows about, what the sender's $\hat{\imath}$ and $\hat{\jmath}$ would have been. The receiver also knows that for any in-flight packets that it did not previously know about, $\hat{\imath} = \hat{\jmath} = 0$. If the most recent block ACK was decoded, and it contained any positive acknowledgement bits, then the receiver can update its view of the state machine of Algorithm 2. Either way, it uses Algorithm 2 to compute the τ-CDF $\{p_i\}$ used by the transmitter, and it runs Algorithm 1 to compute k^*_{ij} for all in-flight packets, known or unknown. This only requires a single pass through the dynamic program with $\hat{\imath} = \hat{\jmath} = 0$, provided that the algorithm is modified to return the entire array k^*_{ij} rather than a single value.

Using k^*_{ij}, the receiver can now determine how many symbols from each known in-flight packet are present in the received frame. Any remaining symbols then correspond to new in-flight packets, and are divided into groups of k^*_{00} symbols. Now that the receiver knows how many symbols are present from each packet, it can de-multiplex them to the appropriate decode buffers. Note that, if we wish, we can assign symbols to packets in any order known to both parties; for instance, round-robin ordering helps to ensure that all of the packets experience the same fading environment, while concatenative ordering (as in Figure 2-3) may improve latency by allowing the decoder to start working on e.g. Packet 1 before the end of the transmission.

The first benefit of this trick is that we do not need to explicitly communicate the number of symbols per in-flight packet, because both parties agree on the transmission schedule. The receiver can even compute a tentative updated schedule in parallel with sending a block ACK. Whether the receiver will use or discard this updated schedule depends on the header bit from the sender indicating whether the block ACK decoded successfully.

The second benefit of the trick over the straightforward scheme is that header decoding is considerably simpler, and hence the latency from frame preamble until the packet decoder(s) can start working is minimized.

Whether these benefits are worth the added complexity remains to be seen. The regimes in which the trick is most useful and least costly are when the number of in-flight blocks is large, and when $\bar{\tau}$ and σ^2 are small, respectively.

2.9 Precise Stop Feedback

The rateless codes we are considering require only stop-feedback, and hence a single bit per in-flight packet in a block acknowledgement. In the VLF model, one such bit arrives per transmitted signal. Over the course of k transmitted symbols, the transmitter would collect k feedback bits. On our half-duplex channel, up to this point, the receiver has only made one decoding attempt per packet per pause for feedback, generating a single feedback bit rather than k bits.

We can tighten this gap substantially by noting that once the decoder declares success, the feedback bits will be one, forever. So the number of distinct k-bit sequences a VLF sender can observe over the course of k transmitted bits is only $\log_2 k$, and thus the maximum amount of feedback information that the RLP receiver can generate is $\log_2 k$ bits per in-flight packet per pause for feedback.

Without this "extra" feedback, the learning scheme of Algorithm 2, applied naïvely, may suffer some biases. Based on its observations, it will generate a CDF, which will be used to generate a transmission schedule, which will determine at what times feedback is generated, which will determine what values the "sample" of τ can take at

the next update of Algorithm 2.

A first defense against this problem is to feed Algorithm 2 not the sample of τ implied by the time of the first positive acknowledgement, but instead a random pick from the old distribution for τ, conditioned to lie between the time of the last NAK and the first positive ACK. If the estimated distribution is actually correct, then over multiple packets these picks will be unbiased samples from the same distribution. Thus the algorithm with $\alpha = 1$ has a fixed point which is the ML estimate of the true distribution.

A second defense is to allocate a few more bits per packet in the block ACK. The exact number needed is logarithmic in the number of decode attempts the receiver is willing to make per packet per pause for feedback. That is, the receiver might choose to perform extra decodings, after the first successful one, using fewer symbols. Suppose that 100 symbols were available at the time it generated the last NAK, 50 more symbols have arrived, and it is time to generate another block ACK. The receiver would first make an attempt to decode using all 150 symbols. If this fails, the receiver can move on to the next packet. If the decoding attempt succeeds, the receiver might forward the packet to the upper layers, and then, in parallel, try to decode the same packet again using only 125 symbols.

Why would the receiver do extra work after decoding succeeds? If the transmitter gets a more accurate estimate of the decoding CDF, then the system as a whole will achieve a higher throughput. If the right stars align, this might be a worthwhile trade-off to the receiver. For instance, perhaps the receiver is a base-station, and via channel reciprocity the uplink τ estimates are useful for determining the downlink τ distribution. In that case, the base-station could burn power making extra decoding attempts in order for the system to achieve higher spectral efficiency on both channels.

To provide the most information to the transmitter using a limited number of decoding attempts, the receiver should make its attempts using numbers of symbols chosen by binary search, not on τ, but on the CDF of τ. If the distribution is smooth over scales of k^* and nearly linear, then there will not be much difference. Because the transmitter and receiver agree on the decoding CDF, the transmitter can perform

the reverse entropy coding to determine τ from the available feedback bits.

A special bit-pattern for a given packet within a block ACK—for instance, all ones—is reserved to indicate a NAK.

2.10 Evaluation

Fig., §	**Experiment**	**Result**
2-4, §2.10.2	RateMore vs. ARQ, Try-after-n HARQ	Overhead reduced by up to 5.4×; throughput increased by up to 26.6%.
2-5, 2-6, §2.10.3	Learning under slow Rayleigh fading	Throughput only 1.6% below case of known CDFs and SNR.
2-7, §2.10.4	RateMore efficiency for various n_f	Better than 88% efficiency even when $n_f = 100$: time is well spent.
2-8, §2.10.5	Impact of learning on throughput	RateMore's learning has an impact of only 0.25%-6%.
2-9, §2.10.6	Streaming with a latency requirement	Block ACK scheduling and code agnosticism enables efficient streaming.
2-10, §2.10.7	Streaming on a fast-fading channel	Less than 15% channel occupancy even at low bandwidth.

Table 2.2: Summary of experimental results of Chapter 2.

We have evaluated RateMore in simulation on stationary, slow-, and fast-fading AWGN channels. Our principal results are summarized in Table 2.2. Some of our results involve the following metrics:

Definition 2.10.1 (Efficiency) *The efficiency of a transmission strategy is the fraction of its ETTC that is strictly necessary for reliable delivery of a stream of packets.*

We make two observations:

1. For any zero-error link-layer protocol, it is essential that the sender successfully obtain feedback at least once for each packet, so the sender can progress to other packets. This imposes a time cost of $\frac{n_f}{p_f}$ in expectation.

2. For any zero-error link-layer protocol over a rateless code, it is essential that the sender transmits, on average, at least $\bar{\tau}$ symbols for each packet. If the

transmitter sends less than this number of symbols, then some positive fraction of packets must be lost.

Thus, we define

Definition 2.10.2 (Efficiency) *The efficiency of a transmission strategy s is*

$$\eta = \frac{\bar{\tau} + \frac{n_f}{p_f}}{\mathrm{ETTC}[s]}$$

We further take

Definition 2.10.3 (Overhead) *The overhead of a transmission strategy s is the portion of its ETTC which is not included in its efficiency:*

$$\text{overhead} = 1 - \eta$$

2.10.1 Baselines for Comparison

For comparison, we consider two alternative feedback schedules: ARQ and Try-after-n HARQ. With ARQ, the sender picks some number of symbols to transmit, then waits for feedback in the form of an ACK. If it receives no ACK, it drops the packet and starts over. This scheme effectively chooses a rated version of the rateless code, and does not incorporate incremental redundancy. In principle, ARQ could choose a different rate for each SNR value. However, existing ARQ systems with rate adaptation, like 802.11, do not have such a large number of rates. 802.11n with one spatial stream has only eight. We therefore require ARQ to pick eight rates to cover the SNR range from -5 to 35 dB. These eight rates are selected to minimize ETTC over all SNRs. For spinal codes, this gives the set $R \in \{0.28, 0.52, 0.9, 1.6, 2.5, 3.6, 4.7, 6.0\}$ bits/s/Hz.

For Try-after-n HARQ, we choose a family of eight parameter values, as before, and sent n additional symbols of incremental redundancy before each pause for feedback.

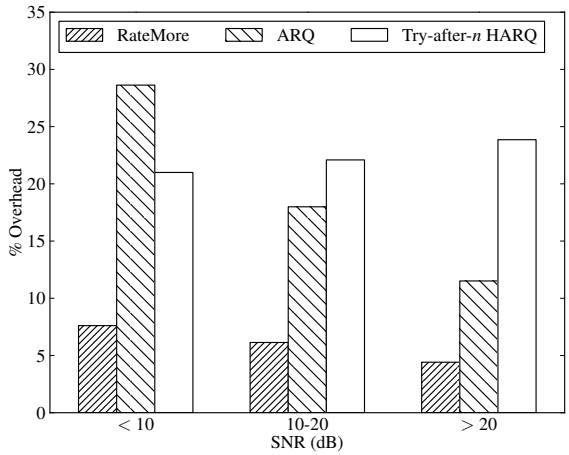


Figure 2-4: With spinal codes and $n_f = 20$. At low, medium, and high SNR, respectively, RateMore reduces overhead by 3.8×, 2.9×, and 2.6× relative to ARQ and 2.8×, 3.6×, and 5.4× relative to Try-after-n HARQ.

2.10.2 Baseline comparison to ARQ, Try-after-n HARQ

Figure 2-4 shows the overhead of RateMore compared to ARQ and "Try-after-n" HARQ running atop spinal codes with a feedback value $n_f = 20$. These experiments are using a stationary channel simulator that introduces Gaussian noise of different variances. We run 20 trials of 100 packets at each SNR.

Across a range of SNRs, the reduction in overhead is between 2.6× and 3.8× relative to ARQ, and between 2.8× and 5.4× relative to HARQ. These reductions are significant; they translate to link-layer throughput improvements up to 26%.

2.10.3 Slow-fading Rayleigh channel

Using our simulated Rayleigh fading coefficients for a Doppler velocity of 10 meters per second, we compared the throughput of RateMore running with known CDFs and SNR against the throughput of RateMore with Gaussian CDF learning. Figure 2-5 shows a typical trace of 100 milliseconds of adaptation on a fading channel. The average SNR is 15 dB. We found that long integration times (e.g. learning parameter α close to 1) were not necessary to obtain good performance. In fact, with $\alpha = 0.8$, corresponding to an exponential time constant of only 4.5 packets, we found aggregate throughput with learning to be within 1.6% of throughput under known CDFs and

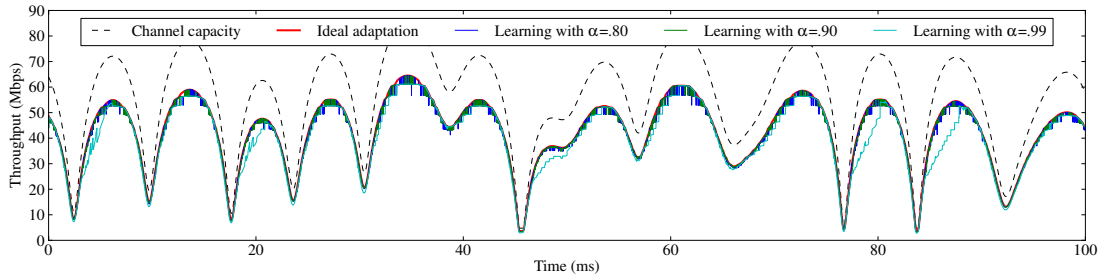


Figure 2-5: Learning performance of RateMore with spinal codes on a Rayleigh fading AWGN channel. Noise power is 15 dB below average faded signal power, giving an equivalent (in terms of capacity) stationary SNR of 12.8 dB. Doppler velocity is 10 m/s. The learning parameter α can be set to a very aggressive value of .8 for an aggregate throughput within 1.6% of the known-SNR "ideal adaptation" throughput. With $n_f = 10$.

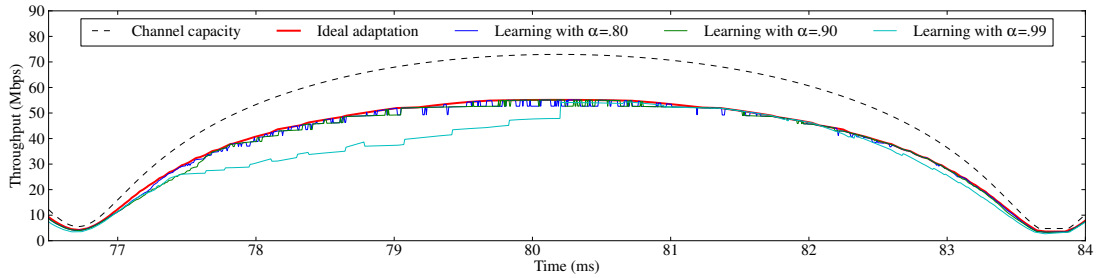


Figure 2-6: Detail of Figure 2-5.

SNR. Figure 2-6 is a detail of Figure 2-5.

2.10.4 Efficiency

Figure 2-7 shows the efficiency of RateMore across a range of SNRs for different values of the feedback cost, n_f. These graphs are for spinal codes (the other codes show similar results). We have grouped the data according to "low", "medium", and "high" SNRs for illustration. In general, efficiency is extremely high, always over 95% at medium and high SNRs, and over 88% for low SNR values, even when n_f is as high as 100. The conclusion is that RateMore produces a good transmission schedule across different channel conditions.

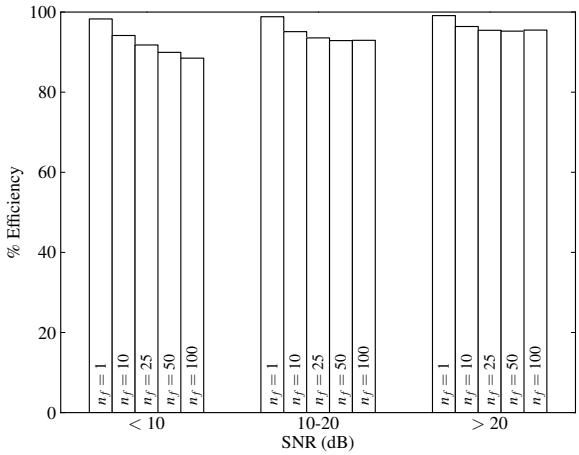


Figure 2-7: Even though n is random, RateMore makes good guesses and thus wastes very little time sending unnecessary symbols or waiting for unnecessary feedback.

2.10.5 How well does learning work?

Figure 2-8 shows the results of experiments that evaluate how much we lose in our *learning* the decoding CDF compared to *knowing* the true CDF. We include the comparison for spinal codes, but the results are similar for other codes. At low and medium SNRs, the cost of learning, i.e., how much we lose, is a negligible 2%. At higher SNR values, the cost is always less than 6%, which is still tolerable.

The conclusion is that our Gaussian approximation of the decoding CDF, which is a simple two-parameter fit (mean and variance) works extremely well, as does the simple filtering method to estimate the mean and variance.

2.10.6 Streaming with a low-latency requirement

For spinal codes, Strider, and Raptor, we compared overhead and channel occupation fraction using RateMore for streaming multimedia designed to emulate a Skype voice and HD video call (Figure 2-9). The main constraint here is latency: we require that the number of packets in flight, times the size of a packet, should represent no more than 100 ms of audio or audio and video. The Strider and Raptor codes we used have single-packet sizes larger than this limit for the low-rate voice call. In order to meet the latency requirements, these codes would be forced to transmit partly-empty

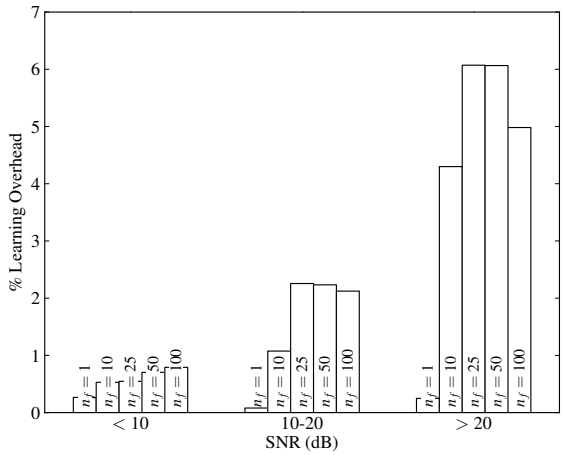


Figure 2-8: Gaussian learning performance on a stationary channel with spinal codes. Mean and variance are smoothed by an exponentially-weighted moving average with parameter $\alpha = .99$. Steady-state performance is comparable to performance when CDFs are fully known.

packets. Because RateMore is agnostic to the details of a specific rateless code, it enables a direct comparison of different rateless coding schemes with different packet sizes, including for instance the effects of relatively smaller packet sizes in spinal codes.

2.10.7 Streaming on a fast-fading channel

When run on CDFs generated for a fast-fading Rayleigh channel, RateMore still maintains low overhead and channel occupation (Figure 2-10).

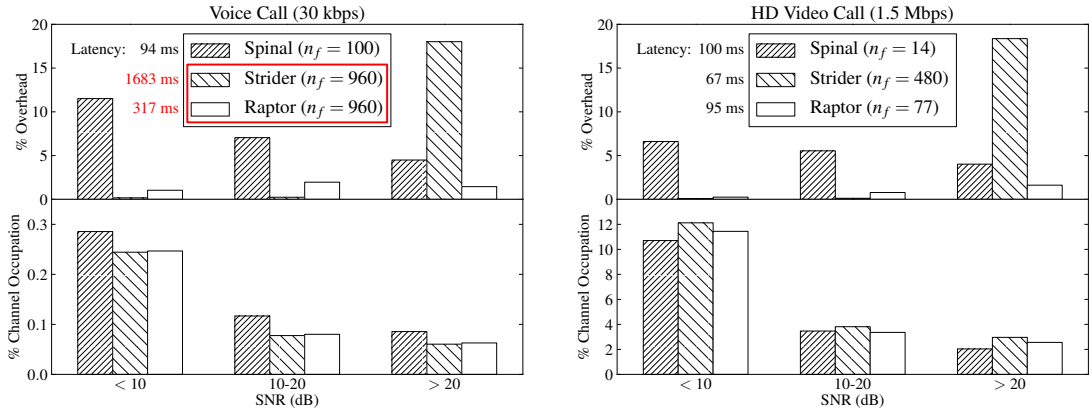


Figure 2-9: RateMore is agnostic to the details of a specific rateless code, and achieves low overhead in latency-critical environments. n_f has been normalized according to the largest number of parallel streams compatible with a given bandwidth-latency product. Cases are shown for a minimum-quality Skype voice call and for an HD video call. For an ideal user experience, latency should be less than 100 ms. Strider and Raptor require relatively large packets, which prevents the use of aggregation on the voice call – in fact, a single packet contains more than 100 ms of audio, as shown in red. In order to meet the latency target at the expense of channel occupation, packets could be fired off partly empty.

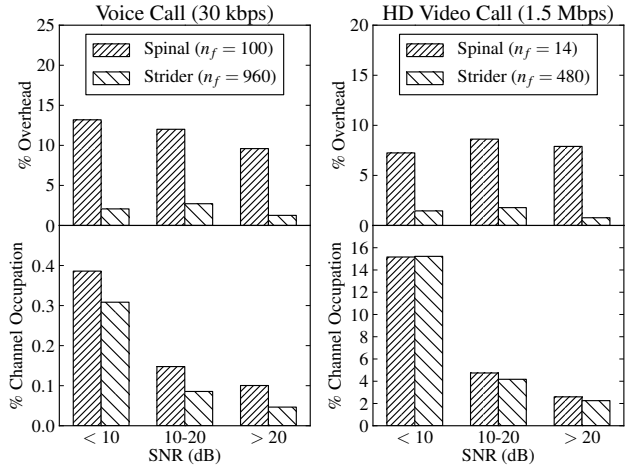


Figure 2-10: RateMore overhead and channel occupation on fast-fading Rayleigh channels. n_f has been normalized as in Figure 2-9

Chapter 3

Acoustic Room-Area Networks

3.1 Overview

This chapter explores the unmet need for a new class of mobile-to-mobile and mobile-to-infrastructure wireless networks. These are zero-configuration, backward compatible systems that enable infrastructure-optional discovery and localization with room granularity. The contributions of this chapter are:

1. Definition of the problem: the abstraction of room-area networks (RAN) §3.2
2. Differentiation of the problem:
   (a) How related systems fail to meet this abstraction, and
   (b) How RAN would make existing applications simpler to design and more capable, and how it enables new applications (§3.3)
3. A solution to the problem: acoustic RAN
   (a) The architecture of Blurt (§3.4), a real-time acoustic software modem that implements the RAN abstraction
   (b) Experimental evaluation of Blurt (§3.5)
   (c) Prototypical applications of Blurt (§3.6)

In addition, this chapter explores the peak-to-average ratio (PAR) problem in orthogonal frequency-division multiplexing (OFDM), which arises in Blurt as it does in other OFDM systems. The PAR problem represents how the limitations of the analog front-end (AFE; particularly, the linear amplifier) impact the range, reliability, and power consumption of OFDM transmitters. The PAR problem raises two well-studied questions: how can the transmitted signal be modified to accommodate the AFE, or, how can the AFE and the transmitted signal best be co-designed? The exploration in this chapter is limited to the first question, and focuses on the search for algorithms with low (additional) circuit complexity compared to an unmodified OFDM transmitter. It contributes

4. A new application of the alternating-direction method of multipliers (ADMM) algorithm for convex optimization to the PAR minimization problem:
   (a) Construction and convergence properties §3.7.4
   (b) Offline step-size schedule optimization §3.7.12
   (c) Arithmetic complexity in the serial and parallel models; estimates of hardware implementation complexity and latency; complexity comparison with other optimization-based solutions. §3.7.3, §3.7.10, §3.7.11
   (d) Evaluation; PAR comparison with other optimization-based solutions. §3.7.13

3.2 The Case for Room-Area Networks

Social interaction happens in rooms: spaces where a few people can gather to share an activity or a conversation. Besides an office, a conference room, or a living room, this usage of the term room extends to a car or bus, a picnic table, a group on the sidewalk, or two seats at a bar. A room is a conversation space; it could equally (if awkwardly) be called an earshot.

When a number of conversants enter a room, they automatically form a human multicast network. If they carry mobile devices, however, those devices do not automatically form a network. While there exist ad-hoc wireless neighbor discovery systems

based on Bluetooth, Apple Wireless Direct Link, and near-field communications, these each travel either too far or not far enough to reliably indicate whether another device is in the same room.

What differentiates human communication from typical wireless channels? One does not leave the room; the other does. In fact, architectural materials and construction techniques are designed, in part, with the express intention of stopping human visible and audible communication at the walls of the room, because otherwise those communications would be disruptive to other tenants. On the one hand, our wireless radio links are designed with the express intention to penetrate walls, in order to reduce capital expenditure on base stations and access points.

This is not to say that communication is impossible to confine to a room with radio. However, in order to multicast exclusively within a room, we would have to first validate the presence of each reachable node, and negotiate a shared secret with the nodes that we trust. We could then accomplish room access control cryptographically. The radio link could contribute to the validation phase through ranging and attenuation measurements, though this might prove challenging if some parties were uncooperative or if a detailed floor-plan were unavailable.

Ultimately, the visibility semantics of radio are quite different from the media of human communication. An attempt to make one match the others is likely to be an uphill battle—and one which we sidestep by the observation that acoustic communications automatically has the property that audibility implies room co-location. Visible and infrared communications share this property, but both suffer from strong geometrical shadowing (connectivity is line-of-sight only) and sensitivity to sunlight.

We began to identify a number of applications that rely on some notion of zero-configuration networking at the room level (Tables 3.1, 3.2). The common factor among these apps appears to be the need to impose location-based policies. A typical mobile-to-mobile or ad-hoc objective might be, "share this confidential PowerPoint with everyone in the room, and nobody outside the room", "list the other people in the room playing Scrabble", or "turn up the volume on the TV in front of me".

Location-based Policy	Mobile-to-Mobile / Ad-Hoc	Mobile-to-Infrastructure
Access Control	Share files/links/contacts	Guest access, rate limits, traffic priority for Wi-Fi
Discovery	Multiplayer gaming	
Remote Control	TV, appliance control	Smart building

Table 3.1: Selected applications of room-area networking

Domain	Application	
Multimedia	Waves by Wham City Lights	[33]
Retail & Marketing	shopBeacon by Shopkick	[34]
—"—	Dash Button by Amazon	
Home Theater	Chromecast Guest Mode by Google	[35]
Wi-Fi Configuration	Secure guest access	§3.6.1
—"—	Rate limits and traffic prioritization	§3.6.2
Building Automation	Room detection for comfort control	§3.6.4
Contact Management	Bump by Bump Technologies	[36]
Building Automation	Comfy / BOSS by UC Berkeley	[37, 38]
Gaming	GameKit Session by Apple	[39]
File/Link Sharing	Google Tone by Google	[40]
—"—	Chirp by Animal Systems	[41]
—"—	AirDrop by Apple	[42]
—"—	Android Beam by Google	[43]

Table 3.2: Applications of RAN. RAN fully captures the concepts expressed by the first group. The second group represents applications which stand to gain in usability from the RAN abstraction.

A mobile-to-infrastructure policy goal might be "grant scheduling priority to Wi-Fi nodes on the demo stage over those in the audience", "authenticate the guest in room 446 to the Wi-Fi", or "adjust the temperature in this room".

A number of organizations have found the need to re-invent this particular wheel, including Bump Technologies, Microsoft Research, Wham! City Lights, Chirp Share by Asio, Apple, Google, and Amazon. Each solution works for a single app or platform. Some also require internet connectivity on each device to facilitate discovery. We see a need for a universal solution that provides both mobile-to-mobile and mobile-to-infrastructure room-area networking without the need for an internet connection or a map.

3.3 Related Work

3.3.1 Existing Approximations to RAN

Apple's Multi-peer Connectivity Framework [44] enables discovery of and communication with nearby mobile devices using Wi-Fi and Bluetooth. It supports unicast, multicast (up to a limit of 8 peers), and service discovery. However, it does not adhere to room boundaries due to RF leakage through walls. In addition, it uses a proprietary protocol and support is limited to Apple devices. Multi-peer Connectivity optionally uses DTLS to provide security, while our RAN abstraction requires applications to implement their own security.

Sonicnet.js [45] uses audio to enable ultrasonic communication between devices. The library provides applications with a socket-like API. Sonicnet naturally adheres to physical room boundaries, but does not provide service discovery and imposes restrictions on transmitted data due to digital signal processing limitations.

Indoor Localization. Spatial coordinates alone are not sufficient to determine in which room a node is located. It would be possible to augment indoor localization systems [46, 47, 48] with up-to-date floorplan information to build a RAN. One advantage of the acoustic medium is that it does not require floorplan information to

respect room boundaries; in fact, mobile nodes can set up a RAN without any support from infrastructure.

3.3.2 RAN PHY Candidates

Wi-Fi suffers from long association times. This phenomenon has been noted in the opportunistic vehicular communication literature for many years [49], and experience indicates that it has not improved in practice. Hadaller et al. [50] report Wi-Fi setup delays of 13.1 ± 12.3 seconds. Blurt can deliver a large number of packets in this time, and may even be useful for accelerating the Wi-Fi association process.

Bluetooth has further usability issues when secure pairing is required, involving waiting for device discovery and manually copying a PIN code. A number of schemes for improving this situation have been considered previously by Uzun et al. [51]. RANs do not have inherent security guarantees or overhead; these may be added with layering.

Technologies using radio frequency (RF) links are not well-suited for a RAN because it is hard to restrict their communication range to the confines of a physical room. Radio signals usually penetrate walls in offices and living spaces to provide good coverage [52].

Near-field communication (NFC) addresses the pairing problem in a zero-configuration way by requiring close physical proximity (a few centimeters) between nodes. This precludes NFC from being useful for a room-area network.

A RAN using infrared (IR) [53] might seem well-suited because it does not penetrate walls, but IR is highly directional, suffers from dead spots, and may not work in rooms with solar illumination. IR has all but disappeared from mobile devices today.

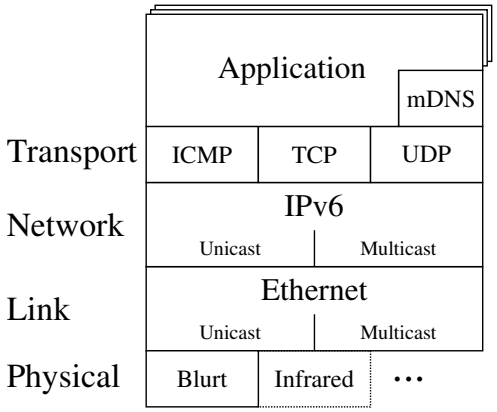


Figure 3-1: Layers in the proposed RAN stack.

3.4 Architecture of Blurt

3.4.1 Network Stack

Our RAN stack is depicted in Fig. 3-1. Other physical layers besides Blurt are possible, including infrared (§3.3.2).

Link Layer

A RAN link is a shared, possibly low-bandwidth Ethernet segment providing best-effort delivery. Each node applies carrier-sense multiple access (CSMA) and exponential-weighted random back-off.

Ideally, we will provision globally unique MAC addresses to RAN-equipped devices. If this is too onerous or expensive, one alternative is for each node to randomly choose a MAC address from the locally-administered unicast range. There are 2^{46} such addresses, giving a negligible probability of collision in a room containing less than 10000 devices.

At the link layer, unicast traffic is addressed directly to the destination node. Multicast relies on the broadcast nature of the medium to reach interested nodes. The software stack at the receiver is responsible for keeping track of which unicast and multicast addresses the interface will accept.

Network layer

Because the entire RAN is one shared segment at the link layer, nothing is to be gained from maintaining a distinction between MAC and IP addresses. If it is feasible to remove this distinction, then we will have no need for ARP, saving time for unicast connection establishment.[1] We will then also have no need for DHCP, since IP addresses will be computed from unique MAC addresses. Avoiding DHCP is highly desirable, since designating a DHCP server introduces an undesirable asymmetry between nodes in the RAN and raises thorny questions about what happens when the DHCP server leaves the room.

There are two reasons why we cannot achieve this goal with IPv4. First, IPv4 has only 65536 link-local addresses in the 169.254.0.0/16 range, which is small enough to lead to collisions in a room with just a few hundred nodes (see the Birthday Paradox). Second, for any IPv4 address outside the link-local range, we cannot be certain that those addresses are not routable on some RAN clients' other network interfaces, raising substantial interoperability issues.

IPv6 link-local addresses have no such problem. There is a canonical and invertible mapping from MAC addresses (that is, 48-bit IEEE 802 addresses) to 128-bit IPv6 link-local addresses [54]. In addition, the IPv6 socket programming interface [55] defines, in the `sockaddr_in6` struct, a field (`sin6_scope_id`) for the purpose of identifying the interface for link-locally addressed connections. This means that client applications can directly address services running on other nodes by MAC address and port. Multicast group addresses at the network layer are also mapped canonically (but non-invertibly) to multicast addresses at the link layer.

Unfortunately, the use of scope IDs (also known as zone IDs) in URLs containing explicit link-local IPv6 addresses, e.g.

```
http://[fe80::%utun1]/
```

[1] Because our acoustic PHY is approximately 2000 times slower than Wi-Fi, the saved round trip time is on the order of 400 ms.

is not supported by major browsers. The proposed standard for this usage, RFC6874 [56], has been a subject of disagreement. The disagreement, in part, is due to leaky abstractions: the OS- and machine-specific contents of the network interface namespace are incorporated into the IPv6 namespace; the IPv6 namespace is incorporated into the URL namespace; and now we have Resource Locators that are not, by any stretch, Uniform. Under certain conditions, like redirections from HTTP to HTTPS, it may become necessary for the server to know by which URL the client has accessed it so that the server can embed that URL into the HTML response. Thus, the private network interface namespace becomes public—an outcome unacceptable to certain parties.

As a work-around, links into the RAN (for instance via printed QR codes) can rely on multicast DNS via URLs of the form

`http://servicename.local/`

Without a distinction between MAC and IP addresses, the IPv6 header is almost entirely redundant for nearly all unicast and multicast traffic. Opportunistic header compression is not a new idea [57], but it works particularly well here: of the 40-octet IPv6 header, we only need to keep one octet in the common case (see §3.4.1). It is important to note that the compression we use is *opportunistic*: that is, so long as we retain the ability to distinguish between compressed and uncompressed packets, our algorithm is permitted to reject some inputs entirely, forcing them to be sent uncompressed.

Transport Layer

We briefly consider the performance of TCP on a RAN. Unlike Wi-Fi systems, RANs do not provide reliable delivery at the link layer. Furthermore, for link-local traffic, RAN transport protocols do not need to tolerate packet drops at routers. In this case, the only two sources of loss are stochastic losses at the physical layer, such as collisions or decoding failures, and transmission time-outs.

TCP is designed to interpret lost packets as a sign of congestion in the network and to respond by backing off its congestion window size. However, in the presence of stochastic losses, TCP reacts rather poorly [58]. Many solutions are possible.

One possibility is that if collisions can be eliminated (hypothetically, via perfect carrier sense), then the TCP sender does not need to sense congestion from ACKs at all. This is not to say that the TCP sender will not encounter congestion – fairness is still an issue – but rather that the signaling mechanism of detecting dropped packets should no longer be used to control the congestion window size, and instead the link layer should be polled to determine whether queued packets are reaching the medium or whether the medium is 100% utilized. Only then should the sender back off.

Application Layer

Many RAN applications are well served by blind UDP multi-cast; for instance, an indiscriminate file- or link-sharing app might broadcast information to every instance of the same app in the room, or a RAN-enabled "We Have Wi-Fi" sign in a café could broadcast log-in information. These applications require no special system support, except for a canonical mechanism to identify which network interface is the RAN interface. Once the interface is identified, these applications bind a UDP socket to it and call `sendto` or `recvfrom`.

Other applications, like the match-making lobby in a multi-player game, can be divided into two phases: a discovery phase that keeps the UI up-to-date with a list of nearby nodes, and a narrow-cast phase that shares real-time game data. Consequently, fully RAN-equipped devices (as opposed to small, embedded nodes supporting only blind multi-cast) should run an mDNS responder as a system service. Applications access this service via a file-domain socket or other (platform specific) standard inter-process communication mechanism.

An additional service that RAN nodes may provide is connection offloading, which they can advertise via mDNS. The idea is that some RAN physical layers, like BLE, IR, or the acoustic reference PHY, may not be suitable for high-bandwidth or low-latency streaming data. In those cases, the devices may negotiate (over the RAN) a connection

over standard Bluetooth, Wi-Fi direct, a mutually-visible access point, WAN unicast via NAT traversal, or, in the worst case, via a mutually-accessible WAN cloud proxy. Fall-back to a cloud proxy would require infrastructural investment beyond what is provided by RAN. Ideally, this case would be highly unlikely due to the prevalence of Wi-Fi and Bluetooth capabilities.

Application programming interface

A RAN client application sees the RAN in two ways. First, the RAN appears as a standard network interface with only a link-local IPv6 address. The application can bind sockets to this interface as normal, and accept or create connections to other nodes if it knows their addresses. Second, the RAN system service provides the following IPCs:

1. Discover RAN interface ID for use with `sockaddr_in6.sin6_scope_id`
2. Issue mDNS query with timeout
3. Start/stop broadcasting a set of mDNS records

Header Compression

We considered two methods for header compression: a bespoke method, and a standard method.

In the bespoke method, we compress if, at the sender, we detect a series of headers (beginning from the LLC header) satisfying the following tests:

1. DSAP = `0xaa` (SNAP)
2. SSAP = `0xaa` (SNAP)
3. Control = 3 (Unnumbered information)
4. Protocol ID = 0 (Ethertype)
5. Ethertype = `0x86DD` (IPv6)

---

1. Version = 6 (IPv6)
2. Next Header = 6, 17, or 58 (TCP, UDP, or ICMPv6)

3. Source Address is link-local[2]
4. Destination Address is link-local[2]

We anticipate that these conditions will hold for the vast majority of application traffic.

If any test fails, we transmit a zero octet followed by the unmodified headers. If all tests pass, then we know that the LLC and network headers can be reconstructed from the MAC header. This is because the source and destination MAC addresses are either unicast, in which case the corresponding IPv6 address can be derived via the canonical mapping, or multicast, in which case the low 32 bits of the IPv6 multicast group can be extracted from the canonical mapping. We have already arranged (footnote 2) to only compress in the (common) case where the first 96 bits of any IPv6 multicast addresses are known, so by this reasoning we can completely reconstruct the source and destination IP addresses at the receiver if compression is used.

If the tests pass, we transmit an octet with the value 6, 17, or 58 (based on the Next Header field of the IPv6 header) followed by the TCP, UDP, or ICMPv6 header and payload.

In the standard method, we used a combined fragmentation and compression layer for IPv6 traffic following RFC4944 [59] and RFC6282 [57]. Fragmentation is required because IPv6 requires that the link support a maximum transmissible unit (MTU) of 1280 octets or greater [60, §5]. Acoustic RAN operates in the narrow range between short time-scales over which very little data can be modulated, and long time scales over which the channel is non-stationary. Fragmentation allows the use of smaller and more reliable physical-layer MTUs.

Reference Physical Layer

Our reference PHY, Blurt, is implemented as an open-source cross-platform library in Python and C++ [61]. For a proven physical layer (PHY) solution, we turned to the 802.11a OFDM scheme with convolutional coding for forward error correction and a

[2]If address is multicast, we also require that it matches the prefix `FF02::/96`, which includes mDNS and many other common multicast groups.

32-bit cyclic redundancy check (CRC) for error detection. This choice of modulation was primarily for expedience; however, it is justified by the high spectral efficiency of OFDM and the long impulse response of the indoor acoustic channel.

We choose a center frequency of 19 kHz, at the upper edge of human hearing (usually given as a ball-park figure of 20 kHz). The response curves of some typical audio hardware are shown in Figure 3-2. We could have chosen a lower center frequency for more sensitivity, at the cost of a certain shrillness. We use more sub-carriers than Dhwani [62] (64 versus 35) over less bandwidth (2.4 kHz without guard band, versus 6 kHz) so that we tolerate longer multi-path delays. An exemplary Blurt datagram is shown in Figure 3-3.

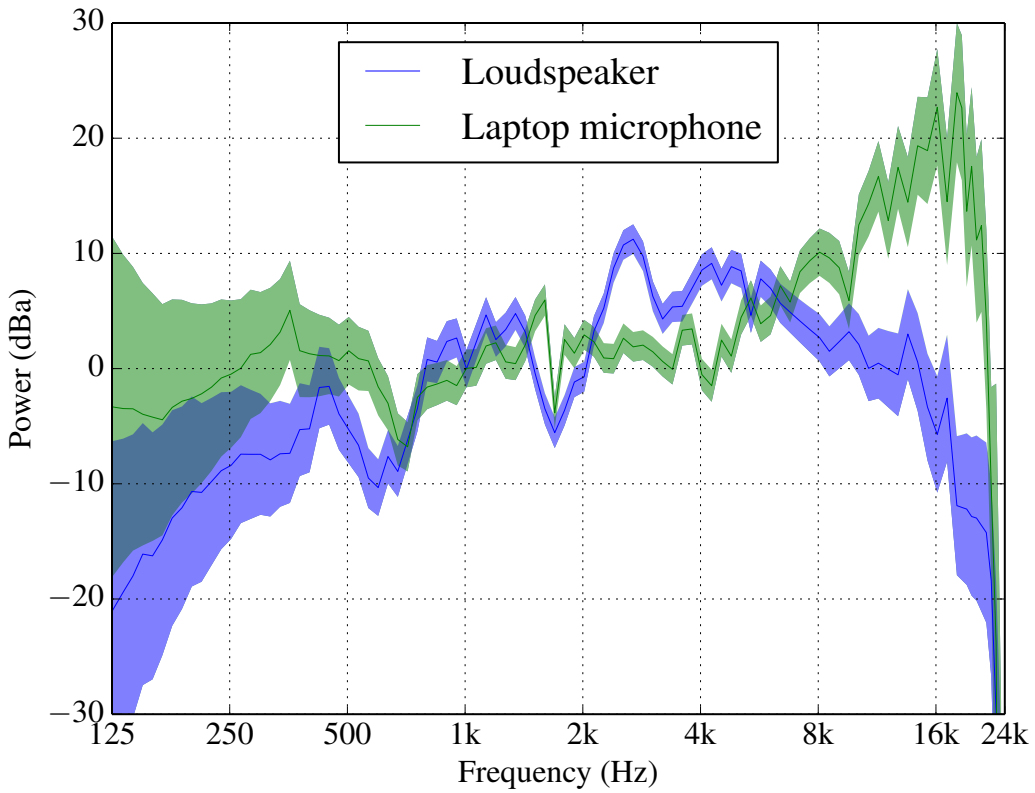
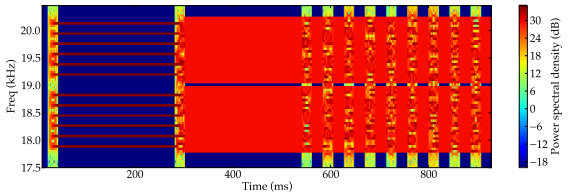


Figure 3-2: Response curves of a typical desktop loudspeaker and a laptop microphone, as measured against a CM-130 sound pressure level meter.

Parameters. The main design parameters in our adaptation of the 802.11a PHY to audio are summarized in Table 3.4.

(a) Before passing through the channel



(b) After passing through the channel

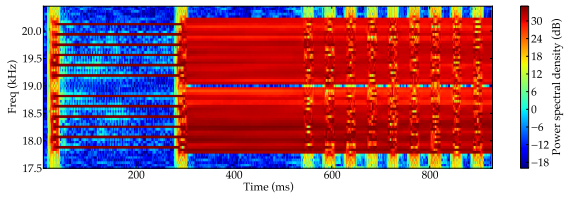


Figure 3-3: A Blurt datagram, before and after passing through an acoustic channel. From left to right, the short training sequences (which appear in the frequency domain as 12 broadly-spaced spikes), the long training sequences, and the data symbols.

Range (m)	0.5	1	2	4	8	12
PER	3%	12%	11%	19%	22%	84%

Table 3.3: Packet error rate (PER) versus range between Blurt senders and receivers (2014 MacBook Pro laptops). Center frequency 19 kHz, bandwidth 3 kHz.

The stereo shift is worth mentioning. Because of the short wavelength of ultrasound, the radiation pattern of a pair of stereo laptop speakers exhibits deep frequency-selective nulls at certain angles (Figure 3-4). By applying cyclic delay diversity—shifting the OFDM symbols transmitted on one speaker relative to the other by half a symbol period, cyclically—we can obtain a much more uniform antenna pattern (Figure 3-5). To be sure, this procedure does not eliminate nulls; it simply makes them so numerous and so closely-spaced in frequency that they are no longer a major problem.

Caveats. The acoustic reference PHY throughput ranges from 900 bps in its most robust configuration to 8100 bps in its fastest configuration, with RTTs of 408±4 ms. Off-loading traffic onto RF technologies after service discovery through the RAN will be desirable.

Not every mobile device has an adequate audio passband to support Blurt's inaudible ultrasonic operation. Table 3.3 shows results from an experiment using

Parameter	Significance	802.11a/g	Blurt
Center frequency	Must be within pass-band of speaker, microphone; should be above limit of human hearing	2.4-5.7 GHz	19 kHz
Bandwidth	Same as above; directly improves throughput	20 MHz	3 kHz
# Sub-carriers	Must be short relative to coherence time of channel but long relative to echo time; impacts PAPR	64	64
Cyclic prefix length	Echo/multi-path tolerance, but cuts into throughput	800 ns	21.3 ms
Training seq. reps.	Determines packet detection and channel estimation accuracy; affects per-packet overhead	2	6
Power	Improves signal at receiver but worsens audibility	< 100 mW	~ 500 mW
Stereo shift	Prevent unintentional beam-forming	0-200 ns	5 ms

Table 3.4: Design parameters for adapting 802.11a PHY to the acoustic channel.

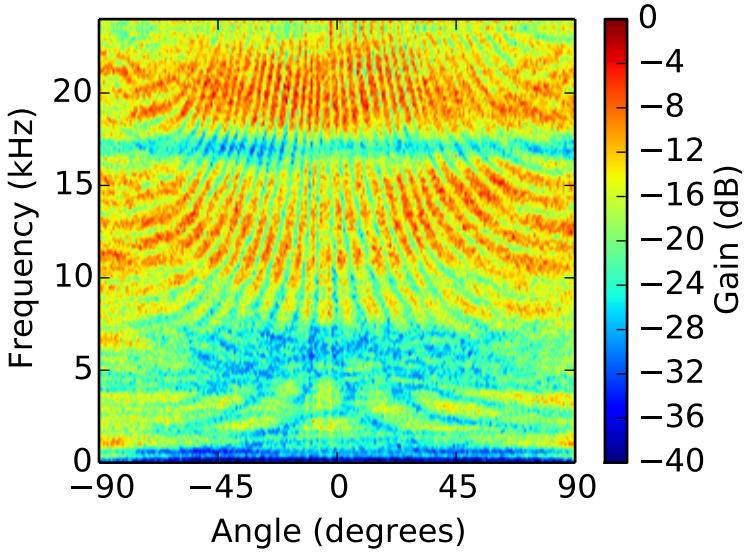


Figure 3-4: Power-spectral density from a laptop source as observed by sweeping a cellphone microphone through a 180° arc. Deep nulls indicate frequencies at which acoustic waves from the two speakers interfered destructively. Horizontal axis is approximate.

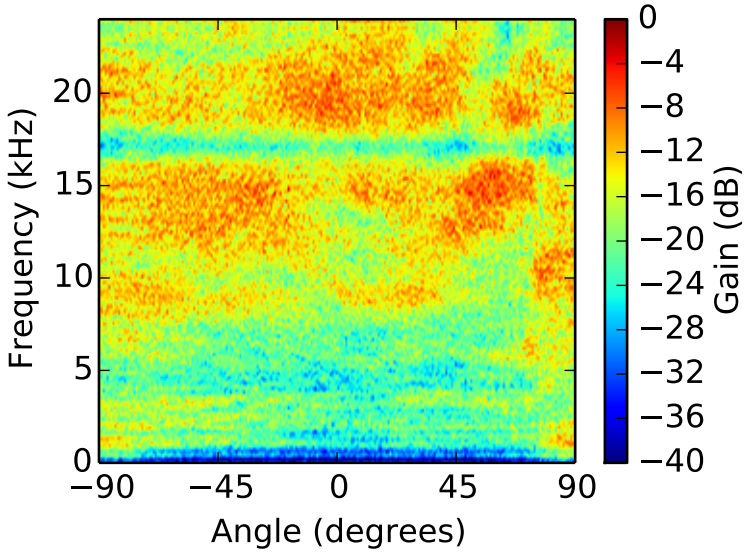


Figure 3-5: Power-spectral density from a laptop source as observed by sweeping a cellphone microphone through a 180° arc. Shifting the waveform on one stereo channel by a few milliseconds is enough to largely obscure the pattern of nulls, which are now too closely spaced to resolve.

MacBook Pro laptops as Blurt senders and receivers. As expected, packet error rates (PER) increase with distance. Blurt successfully delivers packets at ranges up to 12 m.

3.5 Evaluation of Blurt

In addition to presenting data validating our choices of certain key parameters of Blurt, in this section, we evalute Blurt's performance on several figures of merit: reliability in transporting datagrams over long distances; inaudibility; compatibility with existing mobile devices; robustness to common sources of noise; attenuation at room boundaries, affecting conformance to the desired room-area abstraction; and CPU requirements.

3.5.1 Experimental Setup

We carried out laptop-to-laptop, laptop-to-mobile, and mobile-to-laptop experiments in a residential hallway and a basement corridor. The hallway was 1.5 m wide, with sheetrock and steel doors every 5.5 m, wall-to-wall carpeting, and a hanging ceiling.

The basement corridor measured 142 m×4.4 m, with painted cinderblock walls, a polished cement floor, and a large number of insulated ducts, conduits, and pipes running over a series of joists elevated 2.4 m off the floor and spaced by 2-5 m. The hard ceiling was 4.4 m above the floor. One end of the corridor was sheetrocked; our experiments were performed 14.6 m from this end wall.

Blurt was configured to adjust its volume so that 95% of samples did not clip at the sender. We used a pair of desktop loudspeakers to inject controlled noise into the room, allowing us to emulate a variety of (otherwise uncontrollable) acoustic environments, such as an urban intersection or a crowded restaurant. We calibrated the volume of this injected noise using a CHECKMATE CM-130 sound pressure level (SPL) meter.

Our Python implementation of Blurt works online and in real time, and runs indefinitely on any platform supported by Python and NumPy. For certain experiments, however, particularly on mobile phones, we found it expedient to collect raw audio recordings and run the Blurt decoder offline. Our offline experiments use the exact same encoder and decoder as our online experiments, and have identical performance. The only difference is that the offline decoder retrieves audio samples from a WAV file rather than from the operating system's audio API.

For each experiment, we transmitted 30 or more 20-byte packets over distances ranging from 0.5 m to 12 m, and recorded the number of successful receptions. We tested two Apple MacBook laptops—the larger Pro model, and the ultralight Air model—and a number of mobile phones and tablets from various manufacturers.

3.5.2 Validation of Selected Parameters

Blurt's center frequency and bandwidth are adjustable. We measured packet success rates using different center frequencies at a sender-receiver separation of 4 m and a bandwidth of 3 kHz.

Center Frequency	Packet Success Rate
17 kHz	100%
19 kHz	90%
21 kHz	0%

At 21 kHz, Blurt is unable to successfully send and receive packets, most likely because the laptop's speakers do not respond over the entire range of 21 ± 1.5 kHz. The marginal difference in success rate between 17 kHz and 19 kHz is offset by the greater audibility of the 17 kHz-centered signal. Except as noted, the rest of our experiments are performed with a center frequency of 19 kHz.

Packet success rates also vary with bandwidth. We measured packet success rates at a 4 m separation for various choices of bandwidth.

Bandwidth	Packet Success Rate
6 kHz	0%
3 kHz	90%
1.5 kHz	100%

No packets were successfully delivered at 6 kHz bandwidth, again most likely because the laptop's speakers do not respond over the entire range of 19 ± 3 kHz. The marginal difference in success rate between 3 kHz and 1.5 kHz is offset by the twofold reduction in throughput at the narrower bandwidth. Except as noted, the rest of our experiments are performed with a bandwidth of 3 kHz.

In Figure 3-6, we find that packet detection accuracy at low SNR increases rapidly with the number of training sequence repetitions, up to 4 or so. We used 6 repetitions for our experiments.

Figure 3-7 presents data collected from the basement corridor. Normalizing the output waveform to a clipping rate of 5% achieves longer range, while halving this amplitude gives higher success rates at short range, possibly due to clipping at the receiver. We used the larger of the two amplitude choices.

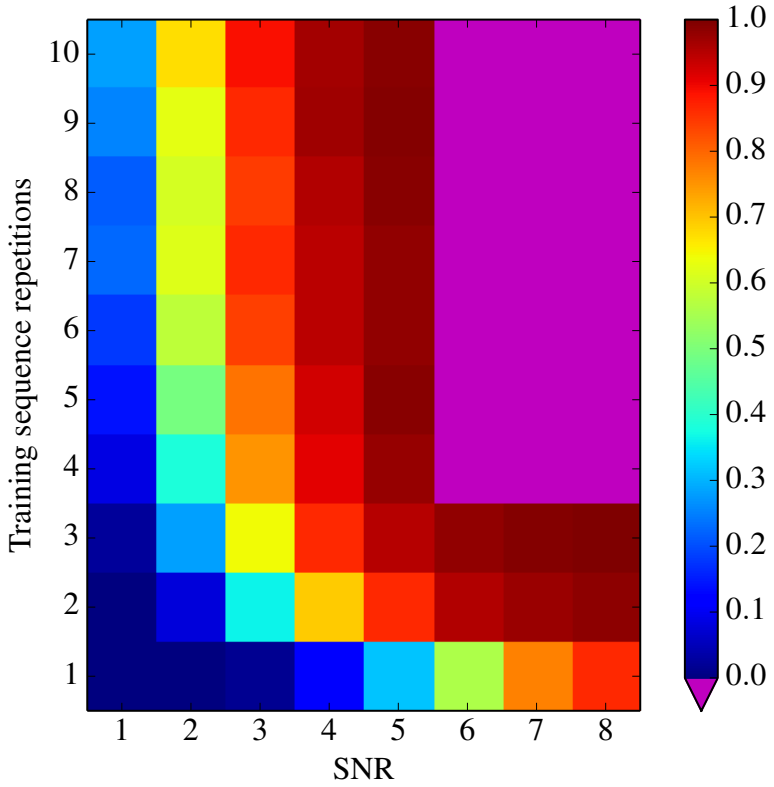


Figure 3-6: Heat-map of packet detector accuracy vs. repetitions of the OFDM training sequence (in simulation).

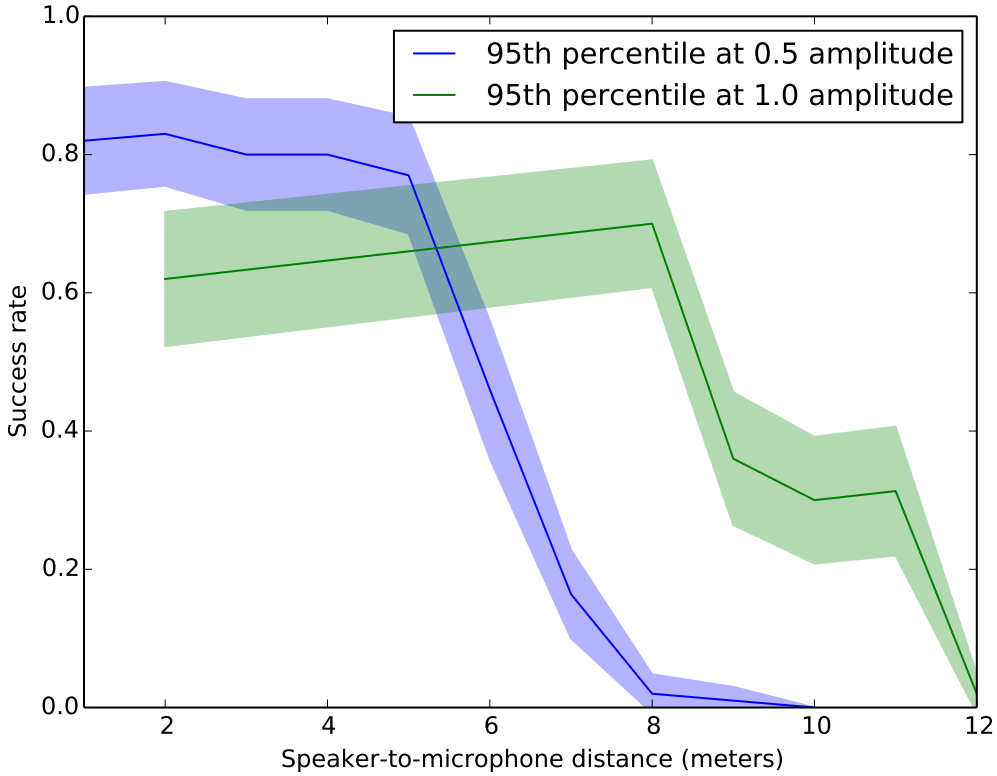


Figure 3-7: Blurt normalizes the output waveform to the 95th percentile of amplitude. An additional amplitude reduction of 50% increases short range performance at the cost of long range performance.

3.5.3 Reliability versus Range

In the residential hallway, we measured the communication range achieved by Blurt, Quietnet [63], and Ultrasonic Networking [64, 65] with COTS laptops (MacBook Pro, MacBook Air) as the sender and receiver (Figure 3-8). For Blurt, we tested one audible and one inaudible volume. Blurt achieves packet success rates of 93% at 8 m and 33% at 12 m at the audible volume. The better performance of the lower volume at distances below 4 m may be a consequence of clipping at the receiver; the lower volume achieves 97% decodability at 0.5 m. The performance at 8 m substantially exceeds the performance observed at 8 m in Figure 3-7. Quietnet and Ultrasonic Networking use near-ultrasonic frequencies around 19 kHz, for data transmission. We modified each to transmit packets of the same size as Blurt (20 bytes). Except at very short ranges, Blurt outperforms both Quietnet and Ultrasonic Networking by more than 2×. Under ideal signal conditions, Blurt successfully decoded a number of packets at 20 m.

Quietnet's failure beyond 1 m and Ultrasonic Networking's failure beyond 4 m may be due to their lack of echo cancellation, which is not needed over very short distances with a strong direct path. Ultrasonic Networking's Gaussian frequency-shift keying implementation substantially outperforms Quietnet's amplitude modulation decoder.

Dhwani [62] does not have a public source code release; however, it is is meant for NFC applications and thus is designed to only send data over distances up to 20 cm. Blurt has no such design criterion or limitation.

3.5.4 Audibility

Under the same conditions as §3.5.3, we relate sound intensity to audibility at each distance by measuring how many decibels of attenuation or gain are needed to bring the signal to the threshold of audibility for two human experimenters. Northern et al. [66] have tabulated absolute thresholds of hearing for various age and gender groups at various frequencies. A median listener in any age group above 20-29 years would be expected to find all of these trials completely inaudible, as per Northern's Fig. 1,

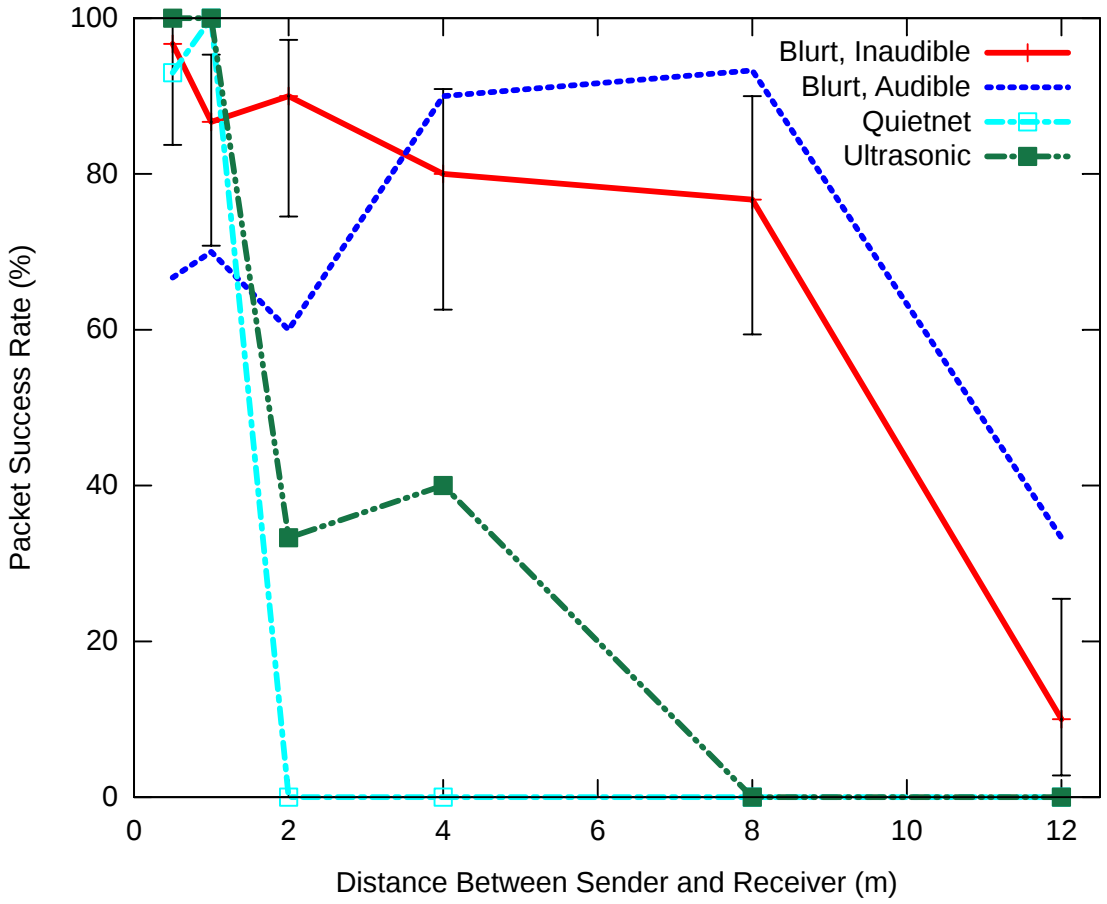


Figure 3-8: Percentage of successful receptions achieved by Blurt, Quietnet, and Ultrasonic Networking as a function of distance between sender and receiver. Our perferred, inaudible configuration is shown with 95% confidence intervals.

showing a 10 dB gap between age groups around 18 kHz; our highest volume setting did not reach 10 dB above the threshold of audibility.

Our results are shown in Figure 3-9. At 8 m, laptop speaker volumes upwards of 80% yield audible Blurt signals, while volumes below 65% are inaudible.

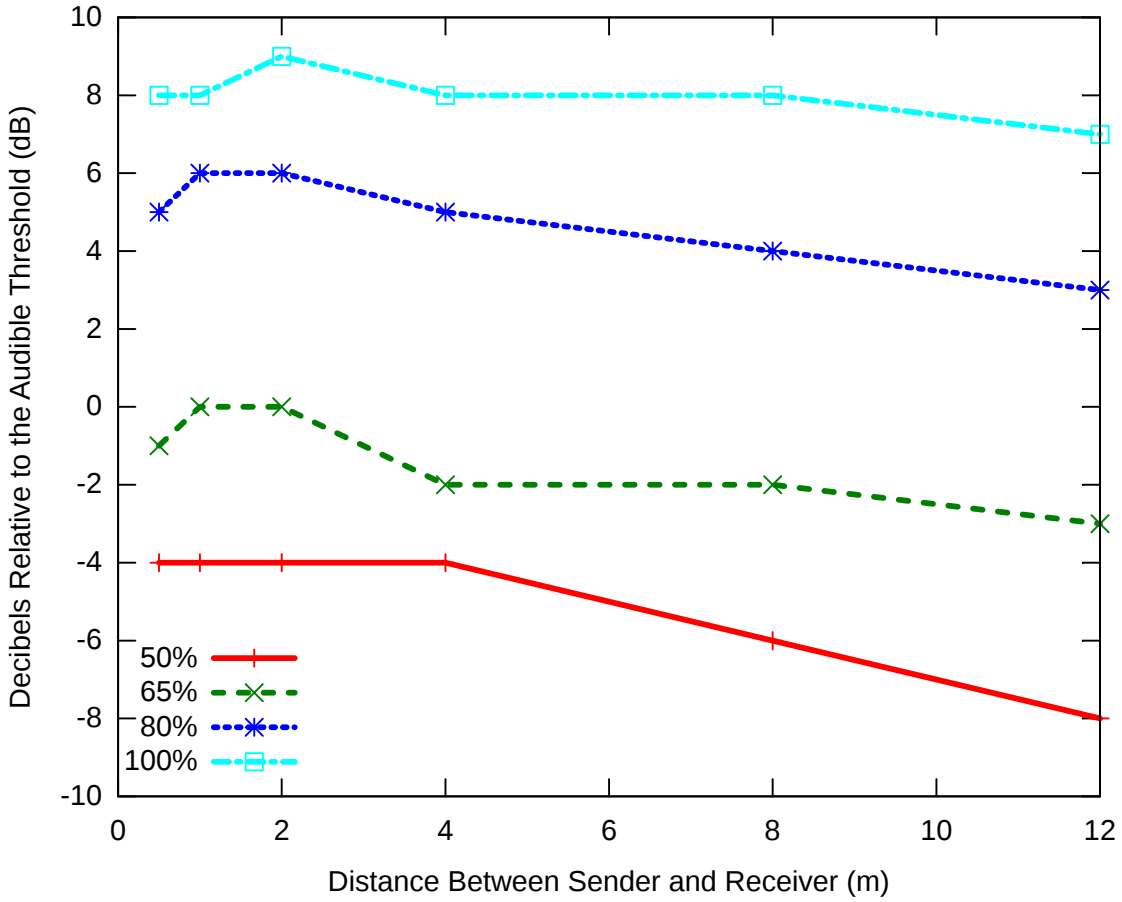


Figure 3-9: Decibels relative to the threshold of audibility for different laptop volume settings.

3.5.5 Hardware Compatibility

We evaluate Blurt's performance on commodity mobile devices as many of Blurt's potential applications pertain to mobile communication. For this evaluation, we use a MacBook Air laptop and one of the following devices:

1. Apple iPhone 5 (iOS 7)

2. Apple iPad (iOS 7)
3. Samsung Galaxy Tab 2 (Android 4)
4. HTC EVO 4 (Android 3)

We selected these devices as they incorporate three leading mobile device manufacturers, Apple, Samsung, and HTC, and two popular mobile operating systems, iOS and Android.

To use each device as a transmitter, we first create an audio bitstream file to represent Blurt's transmission. We use Blurt's *wav-out* flag to output a WAV file pertaining to a 30 packet transmission. Then, for each device, we play this WAV file using the built-in speakers. The Blurt receiver detects packets normally and returns the number of successful packet receptions.

Figure 3-10 shows, per transmitter device, Blurt's packet success rate as a function of distance between sender and receiver. As shown, Blurt is able to achieve reliable transmission of packets at distances up to 8 m. Apple's iPhone 5 yields a 33% packet success rate at 8 m and as high as a 97% packet success rate at 2 m. Similarly, Samsung's Galaxy Tab 2 has successful communication for 83% of transmitted packets at 1 m and 10% at 2 m.

To evaluate each mobile device as a Blurt receiver, we use each device's microphone to record Blurt transmissions from our laptop as WAV files. We then use Blurt's *wav-in* flag to process our recorded WAV files and note the percentage of successful packet receptions. Figure 3-11 shows, per receiver device, Blurt's packet success rate as a function of distance between sender and receiver. Several of the mobile devices we considered support Blurt packet reception at distances up to 12 m away from our Blurt sender. However, we do notice large fluctuations in packet success rate, even for the same device across distances. For example, Apple's iPad was 54% more successful at 4 m than at 2 m separation. Other distance modifications, however, resulted in milder changes to the performance of Blurt.

We note that some mobile devices, including HTC's EVO 4, are unable to use Blurt as their speakers and coding are not yet capable. However, we anticipate future models to include the necessary hardware and software capabilities for Blurt.

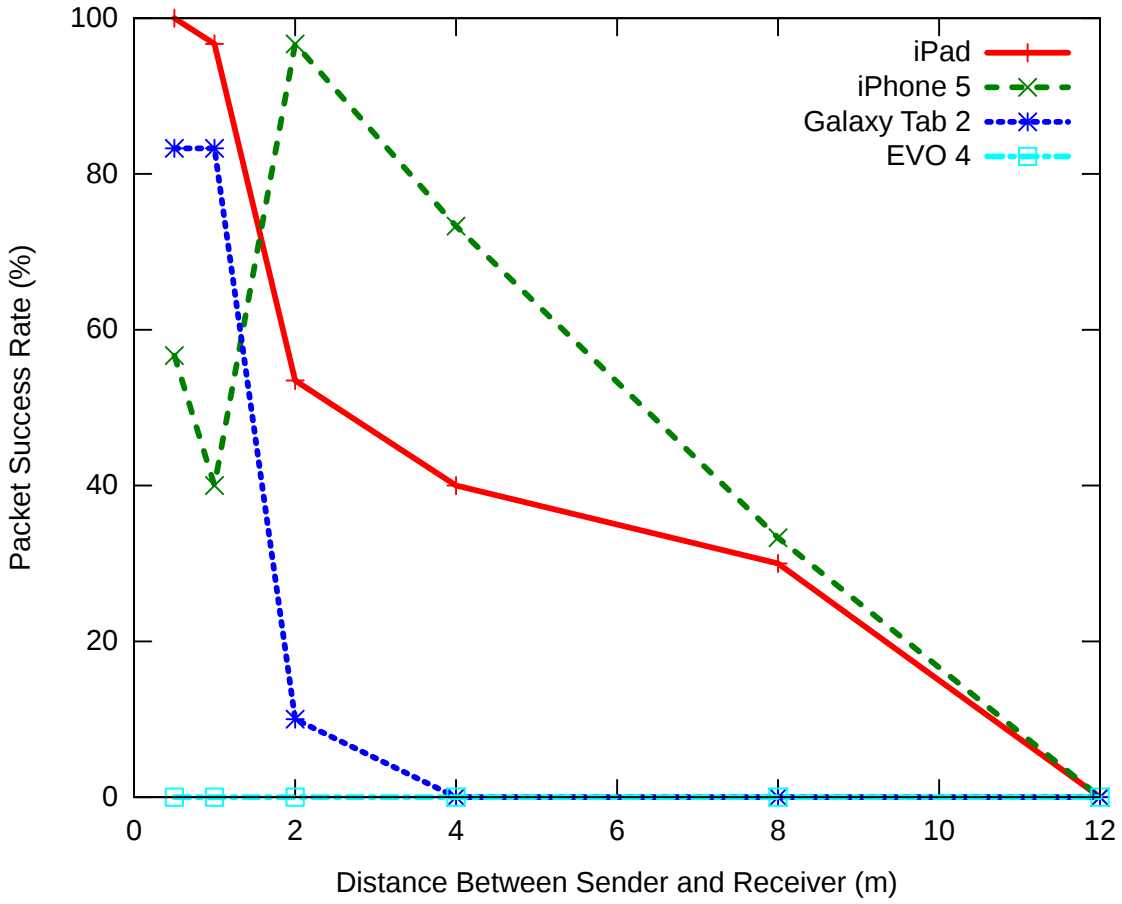


Figure 3-10: Percentage of successful receptions as a function of distance between sender and receiver when using commodity mobile phones and tablets as Blurt senders and laptops as Blurt receivers.

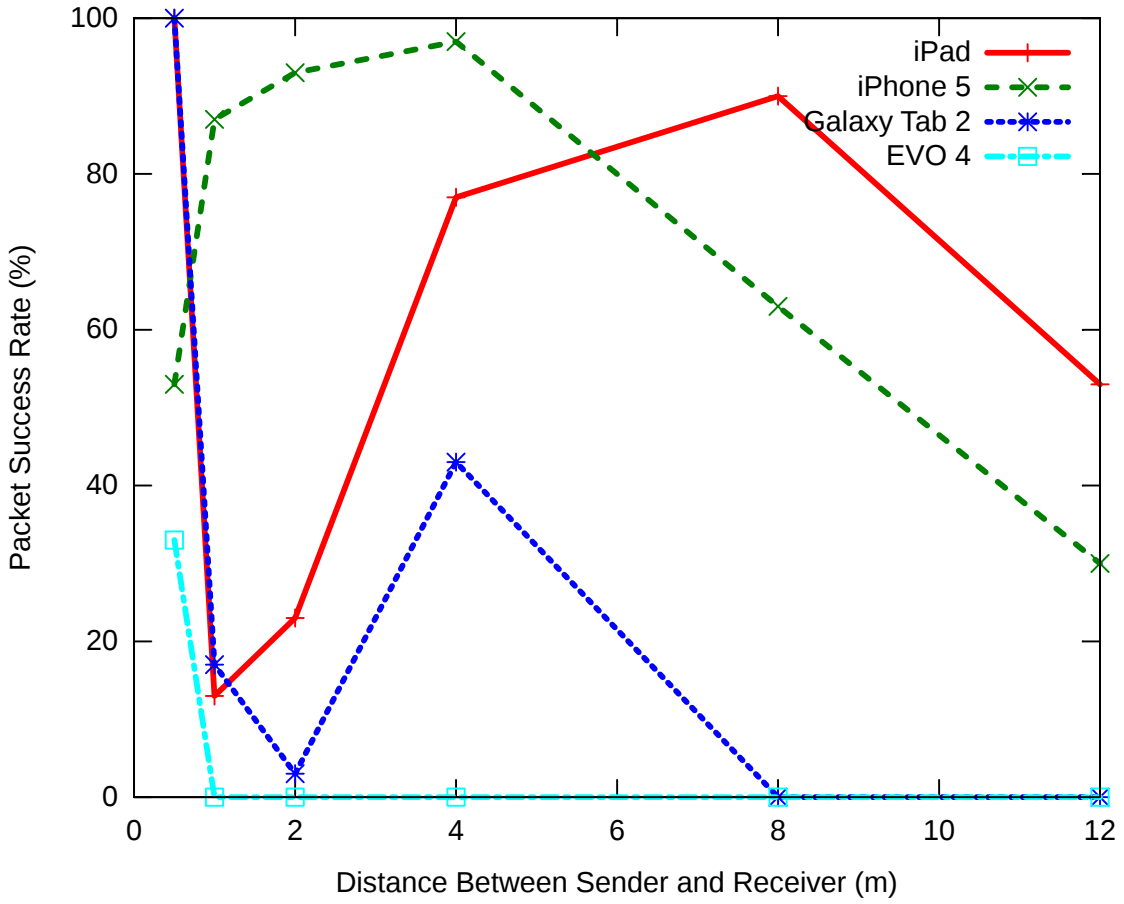


Figure 3-11: Percentage of sucessful receptions as a function of distance between sender and receiver when using commodity mobile phones and tablets as Blurt receivers and laptops as Blurt senders.

3.5.6 Robust to Noise

To evaluate the effect of ambient noise on Blurt's performance, we consider three different cases:

1. Busy intersection (65 dBA)
2. Crowded restaurant (65 dBA)
3. Office (62 dBA)

For each scenario, we recorded 90 seconds of ambient noise. To evaluate Blurt with each noise, we played back each recording during Blurt's transmission, while ensuring that the noise level from the speakers matched that seen while recording.

Figure 3-12 shows, per noise type, the packet success rate when using Blurt to transmit packets at different distances. At a distance of 4 m, Blurt was able to achieve packet success rates of 93%, 70%, and 53% for the busy intersection, crowded restaurant, and office noise types, respectively. At the same distance, with no ambient noise recording to interfere, Blurt was able to achieve a packet success rate of 90%, showing an average success reduction of 17% in the presence of noise. This mild performance drop illustrates Blurt's resiliency to noise.

For comparison, we evaluate Quietnet with each recorded noise. Quietnet did not have any successful packet receptions at each of the considered distances (0.5 m to 12 m).

3.5.7 Attenuation at Room Boundaries

	Glass Door	Wooden Door
Blurt	36 dB	40 dB
Wi-Fi	1.5 dB	2 dB

Table 3.5: Comparison of attenuation for Wi-Fi radio and for ultrasound when closing some typical office doors between the sender and receiver.

We compare Blurt's confinement to a single room (and hence suitability as a RAN) to Wi-Fi.

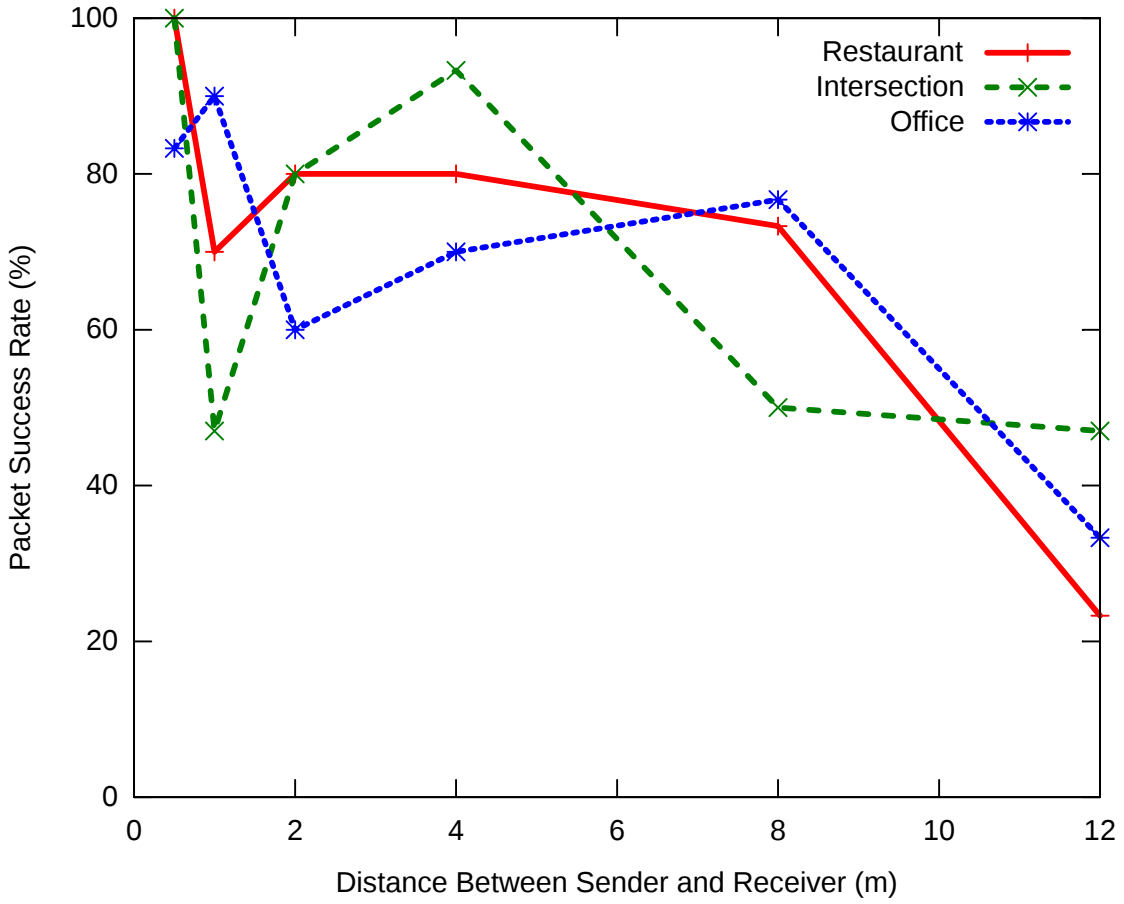


Figure 3-12: Percentage of successful receptions as a function of distance between sender and receiver when using Blurt in the presence of characteristic ambient noise for three typical settings.

We first measure the attenuation experienced by Wi-Fi when inside a room and outside a room, behind a closed door. We consider two rooms; one with a wooden door and one with a glass door. Attenuation for Wi-Fi is measured as difference in RSSI seen inside the room and outside the room. We then measure the attenuation of audio frequencies in Blurt's range of operation in the same conditions. Table 3.5 shows the results from these experiments. Blurt's attenuation at the room boundary exceeded Wi-Fi's by more than 4000×.

3.5.8 CPU Usage

We measured the CPU requirements of Blurt for real-time operation on an Intel Core i7 processor by integrating the instantaneous frequency of the CPU while encoding and decoding 100 packets as quickly as possible. This was necessary because the CPU dynamically scales its clock frequency. Using the Python implementation of Blurt, we obtained CPU usage of 36.1 ± 0.6 MHz for the encoder and 111.2 ± 2.0 MHz for the decoder. Our C++ implementation requires only 18 MHz for encode. The C++ decoder is not ready.

3.6 Applications of Blurt

In this section we describe some applications of room-area networks and how they map into our proposed system architecture.

3.6.1 Guest Access for Wi-Fi Networks

Blurt can be used to establish an authorization scheme where only users in a certain room are able to connect to a given Wi-Fi network, for instance in a company or inside a café, by making the access point (AP) RAN-capable.

Insecure guest access. The user connects to an open Wi-Fi network, and using standard methods is diverted to a captive web portal. The portal challenges the user's device to prove that it is in-room by communicating via the RAN: it includes a tag

`<script src="http://auth.BSSID.local/NONCE.js" onerror="retry()" onload="done()">`, causing the user's browser to issue an mDNS query over the RAN. The AP answers this query with an AAAA record pointing to its link-local address. The user's browser then uses this address to open a socket via the RAN, and receives an empty JavaScript file. Now the AP knows that the user is in the room, since the user's browser contacted it via the RAN and provided the correct nonce. The web portal can tell whether the resource was loaded using HTTP header `Access-Control-Allow-Origin` and `<script>` attribute `crossorigin` , and might retry. Note that the user's browser is not trusted to report success. Once the resource request succeeds, the user's browser is redirected to the URL she originally tried to access before reaching the captive portal. All these steps are performed without the user typing in any passwords.

WPA2-encrypted secure guest access. In fact, we can use RAN to provide WPA2-encrypted guest access to an enterprise Wi-Fi network. Today, this would require each guest to be given a Wi-Fi password or personal certificate that then must be entered manually, a somewhat cumbersome task on smartphones.

This time, the AP sends periodic RAN multicast datagrams containing the BSSID and a WPA2 password. A native application running on the user's mobile device listens to this multicast group and displays each advertisement it receives. Once the user selects a network from the list, the application causes the user's device to associate with the corresponding AP. In the process, the device and the AP negotiate a pairwise transient key (PTK) over e.g. EAP-TTLS. The user's device now goes through the in-room challenge detailed above for the case of insecure guest access, with the native application replacing the web browser.

We'd like to give the user confidence that she has not connected to a rogue access point. The in-room challenge nonce is served over TLS with forward secrecy. The server certificate will include the AP's BSSID. The user can now be certain that the TLS connection terminates at a legitimate access point. But a rogue AP could be tunneling the TLS connection to a legitimate AP; we need to bind the TLS connection to the Wi-Fi association. So the AP uses its copy of the PTK to encrypt a known

plain-text, and sends the resulting cipher-text to the native application through the TLS connection. Now the application can validate the AP to which it is associated by comparing the BSSID, certificate, and pairwise transient key.

The name-spacing of the TLS certificates bears some discussion. One approach is to have the root certificate signed by a RAN certificate authority (CA), which would have to be trusted. The domain name of the AP would then be of the form BSSID.ran.campus.edu (or company.com). If organizations don't want to reveal their BSSIDs to us, another level of indirection to hide that would suffice.

3.6.2 Rate Limits and Traffic Prioritization

As we have seen, RAN lets us challenge a device to confirm its location. We can now prioritize traffic from certain rooms (e.g., important meeting rooms), rate-limit traffic from certain rooms, and provide guarantees for traffic from certain rooms. When the user interacts with the captive portal, her location is tabulated with her MAC address on the access point, and used to assign her traffic to a QoS class for routing. To update this information as the user moves, we arrange for her mobile device to advertise an mDNS record containing its Wi-Fi MAC address. Any time the AP would like to update its location mappings, it issues a query over the RAN in each room to learn the set of user devices in that location.

3.6.3 BlurtFactor: Two-Factor Authentication with Blurt

Two-factor authentication augments the security of password authentication by requiring each authentication attempt to involve a unique challenge-response transaction with a cryptographic token. This token (which may be a physical key-fob or a smartphone app) can be lost or stolen and subsequently revoked without a security concern, since the passphrase must be supplied simultaneously with using the token. YubiKey [67] is an example of such a token. A YubiKey device is inserted into a computer's USB port, where it simulates a keyboard. When the user wishes to log in, he or she taps the button on the YubiKey and it generates, encrypts, and "types

out" a unique one-time password (OTP). This password contains the user's identity, a timestamp, and an incrementing counter. At the server end of the authentication connection, the OTP is decrypted and checked for replay attacks. Once the OTP has been used, it is no longer valid and reveals nothing to an attacker who learns what it was. Subsequent OTPs generated by the same Yubikey will be completely different.

We have implemented an iOS application that uses the Blurt library to send YubiKey-compatible OTPs over audio. A companion application running on a laptop or desktop types out the OTP when it is received from the smartphone. Any eavesdropper attempting to steal and use the OTP must race the legitimate user to present it to the server. A curious property of the acoustic channel is that an attacker several feet away from the legitimate user has a handicap of several milliseconds due to the propagation delay of sound.

3.6.4 Building Automation

Recently, Dawson-Haggerty et al. proposed a building operating system called BOSS. RAN provides a useful and interesting way to enforce space-based access control and authorization to control the buildings actuators because it provides evidence of the room in which the device currently is running. Each room of a building could send periodic RAN multicast datagrams containing a cookie that is unique to the room and changes, say, every five minutes. Then the user could supply this cookie to a building application accessible via Wi-Fi to prove their recent location. The same idea can be extended to controlling other devices in a home, campus, or company.

3.7 PAR Minimization with ADMM

A peculiarity of OFDM is that the distribution of the outputs of the encoder may have a much larger support than the distribution of the inputs. This is quantified by the peak-to-average ratio (PAR) of the signal's instantaneous power. This quantity, or its square root, are also known as crest factor, peak-to-mean envelope ratio (PMEPR), or peak-to-average power ratio (PAPR).

In a typical OFDM transmitter, the complex baseband signal is produced by inverse FFT, cyclic prefix insertion, and windowing. This pair of digital fixed-point signals (in-phase, I, and quadrature, Q) feed a pair of digital-to-analog converters, the outputs of which are mixed with a quadrature local oscillator to form the real baseband signal. This signal is then filtered and amplified.

The high PAR of OFDM is problematic because DACs and power amplifiers must be provisioned for the worst-case scenario of a large output spike; otherwise, the signal will suffer saturation non-linearities when the output stage cannot faithfully generate the waveform computed by the encoder.

Saturation non-linearities are particularly harmful for OFDM. When the desired waveform includes a peak of amplitude A, and the circuitry can only produce a peak of amplitude B, the difference between the desired signal and the generated signal is an impulse of amplitude A-B. This difference signal has a large magnitude on every sub-carrier at the receiver, causing a burst of bit errors. Depending on bit interleaving and the FEC scheme, this condition may not be recoverable.

Infrequent, large-amplitude spikes force the system designer to either use a more expensive power amplifier, or to back off the average transmit power so that peaks are manageable. PAR is one reason why high-order OFDM designs are more favorable for applications with a few expensive transmitters and a large number of inexpensive receivers, like DVB-T or LTE.

Aside: PAR as a cost driver The phenomenon of amplifier de-rating under multi-carrier input predates OFDM by many years. In the now-famous Wall of Sound public address system designed and used for the band The Grateful Dead, the bass guitarist's performance could be recorded separately for each string, and directed through four separate power amplifiers and speaker arrays [68, p. 144–145] [69, p. 455] [70].

To give a concrete example of this problem for RF systems, consider that Wi-Fi adapters typically have a maximum transmit power of 20 dBm using

OFDM, while cellular phones typically have a maximum transmit power of 30 dBm using a different modulation technology (e.g. SC-FDMA). For an apples-to-apples comparison, consider a bandwidth of 20 MHz, in-spec for both technologies. Suppose that we wish to build a mobile phone with the same ultimate Wi-Fi range as its cellular range.

For an average radiated power of 30 dBm and a worst-case voltage standing-wave ratio (VSWR) of 3, the amplifier may be required to produce an output voltage as large as $\sqrt{2\,\mathrm{VSWR}\,Z_0 P_{\mathrm{radiated}}} \approx 17.3\ \mathrm{V_{rms}}$. As often as 1% of the time, a given 4 µs window of the 802.11n HT20 OFDM waveform will have a PAR of at least 9.7 dB, requiring the amplifier to produce brief (circa 50 ns) excursions of ±52.7 V.

Suppose we want 20 dB of SINR, because our operating point is at, say, 10 dB of SNR at the limit of our range, and we want no more than half a decibel of SNR degradation at the receiver due to transmitter non-ideality. Then we require the IM_3 intermodulation product to be below 10 dBm. During 1%-level peak events (which occur 2500 times per second) we can estimate the symbol-averaged IM_3 power as

$$\frac{(\langle P_{\mathrm{tx}}\rangle 10^{9.7/10})^3 (50\,\mathrm{ns}/4\,\mu\mathrm{s})}{{\mathrm{OIP}_3}^2} \leq 10\,\mathrm{mW}, \text{ so we require}$$

$$\mathrm{OIP}_3 \geq 45\,\mathrm{dBm}$$

Furthermore, we require a 1 dB saturation power of at least 41 dBm. There is no device that meets these requirements simultaneously on the 2.4 GHz to 6 GHz bands in the Mouser Electronics or Digi-Key catalogues as of November 2018. The nearest facsimiles are devices in the $400 to $600 range that consume 13 W to 35 W even when no signal is present. These would be intolerable thermal, electrical, and cost parameters for integration into a mobile phone. To make such a design work, a PAR reduction scheme and/or a digital compensation scheme for amplifier nonlinearity would be required.

3.7.1 PAR in a Software Modem

For the ultrasonic software modem, the operating system nominally requires us to produce a waveform clipped to $[-1, +1]$. If we use OFDM to tolerate multi-path, then we need to be prepared for the possibility of clipping.

Solution 1: Allow the operating system to perform clipping. This solution is unlikely to behave in a consistent way across operating systems, or even hardware revisions.

On MacOS 10.13.2, a series of experiments allow us to characterize the audio subsystem's fallback behavior when an application produces a waveform outside the $[-1, +1]$ range. All of the experiments involve combinations of silence, 440 Hz tones, and rectangular pulses of height 10000 and duration 1 ms. The measurements reported below were taken from a software spectrum analyzer running on a second computer, connected to a microphone near the transmitting speaker. Signal amplitude parameters are given in linear units; measurements are in decibels.

First, the behavior is stateful. If we produce a 1 s tone at amplitude 1.0, followed by a pulse, and then a 30 s tone at amplitude 1.0, the observed audio gain drops sharply after the pulse, and then ramps back up.

Second, the behavior is global rather than per-application: if we produce a tone from one application and a pulse from another, the audio gain of the tone drops and then recovers, as before.

This behavior is consistent with an audio processing system that mixes (adds) unclipped samples from each client application, then applies a "limiter" to keep the peak magnitude below 1.0. The limiter's attack behavior is quite rapid, in the range $10\,\mu\text{s} < \tau \leq 1\,\text{ms}$, based on its response to short and long pulses.

We also observe that produce a tone with a rising linear amplitude ramp (amplitude ≥ 1) gives a steady tone, while the same signal with a falling ramp gives a pop followed by a decrescendo. That is, the limiter tracks increases in signal envelope eagerly, but it tracks decreases in signal envelope with caution.

These observations do not explain the appearance of odd harmonics in the steady-

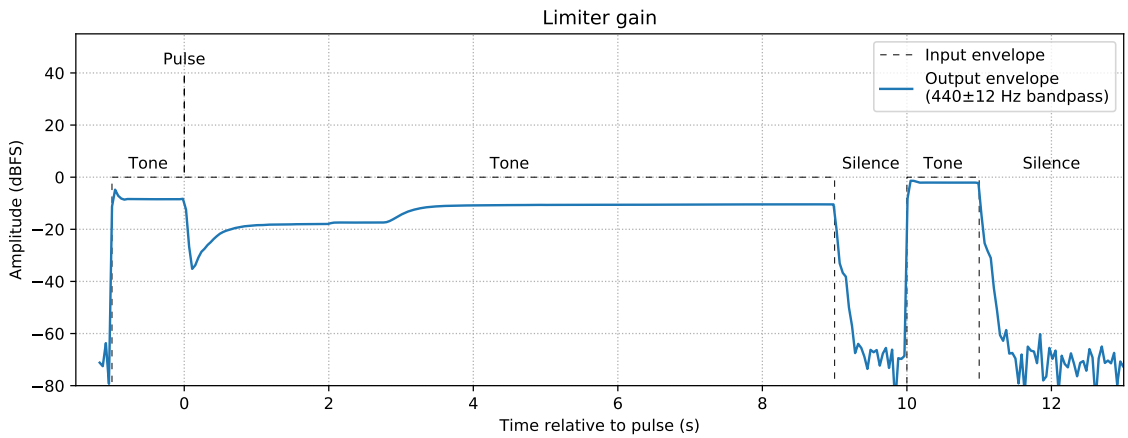


Figure 3-13: The results of one experiment on the stateful audio envelope limiter in a MacBook Pro laptop.

state audio power spectrum: evidence of saturation or clipping, which the limiter could, in principle, entirely prevent. The problem of measuring these harmonics is complicated by second-order nonlinearities in the audio front-end; these can be rendered unobservably small by setting the system volume control to $-20\,\mathrm{dB}$ or lower. Note that the order of operations appears to be mix, limit, clip, quantize, and finally apply the system volume control.

We do not find discernable third harmonic power for tones of amplitude 100 or less. The lowest amplitude that gives a third harmonic distinguishable from the receiver noise floor is 102. This suggests that the limiter's gain is confined to the range $[-40\,\mathrm{dB}, 0\,\mathrm{dB}]$.

Finally, the limiter seems to have a multi-stage release. The initial release seems to be complete after one second, with the gain rising to $-9.6\,\mathrm{dB}$ of its pre-pulse value. After $2.8\,\mathrm{s}$ from the pulse, the limiter gain ramps again up to $-2.1\,\mathrm{dB}$, where it remains. There also appear to be edge cases where the limiter shows positive gain of $6.4\,\mathrm{dB}$ (Figure 3-13).

Solution 2: Static reduction in transmit power. We can examine the distribution of PAR, and set the transmit power level low enough to achieve a given target error rate. This option has the advantages of being deterministic and adding no

latency, but the power back-off limits range and throughput.

Solution 3: Non-causal gain control. In software, it is not too expensive to encode the entire packet in advance, measure its peak amplitude, and normalize the entire waveform before sending it to the operating system. This approach costs a small amount of additional memory, and a small amount of additional encoding latency, but eliminates crests as a source of errors.

Solution 4: Modify the signal. Techniques abound for modifying the transmitted OFDM signal to ameliorate the PAR problem [71]. Many, like Tone Injection [4], Constant-Envelope OFDM [72], Constant-Modulus OFDM [73], Single-Carrier Frequency Division Multiple Access [74, 75], bit-reserving schemes [76, 77], and error-correction coding/selected mapping schemes [78, 79], are explicitly non-backwards-compatible with standard OFDM receivers. Others, like Active Constellation Extension [80], are intended to be backwards-compatible, but may cause problems for receivers that track fading coefficients or frequency offsets using data subcarrier residuals. This work focuses on two backwards-compatible techniques: Tone Reservation (TR) [4] and optimized in-band distortion [5], sometimes called Clipping and Filtering (CF) or error vector magnitude (EVM) optimization. The error vector is defined as the difference between the data subcarriers' intended values and their transmitted values.

3.7.2 Algorithms for Tone Reservation

Tone Reservation will, if permitted, generate a spectrum of out-of-band radiation (Figure 3-14). On a radio channel, a regulatory "spectral mask" must generally be imposed as a constraint; on the ultrasonic channel, the only spectral mask that matters is the threshold of human annoyance. Both TR and EVM techniques have an additional side effect: thanks to the nonlinear processing performed by the transmitter, it is no longer straightforward to implement an efficient Maximum Likelihood (ML) receiver, though there has been work on this problem [81].

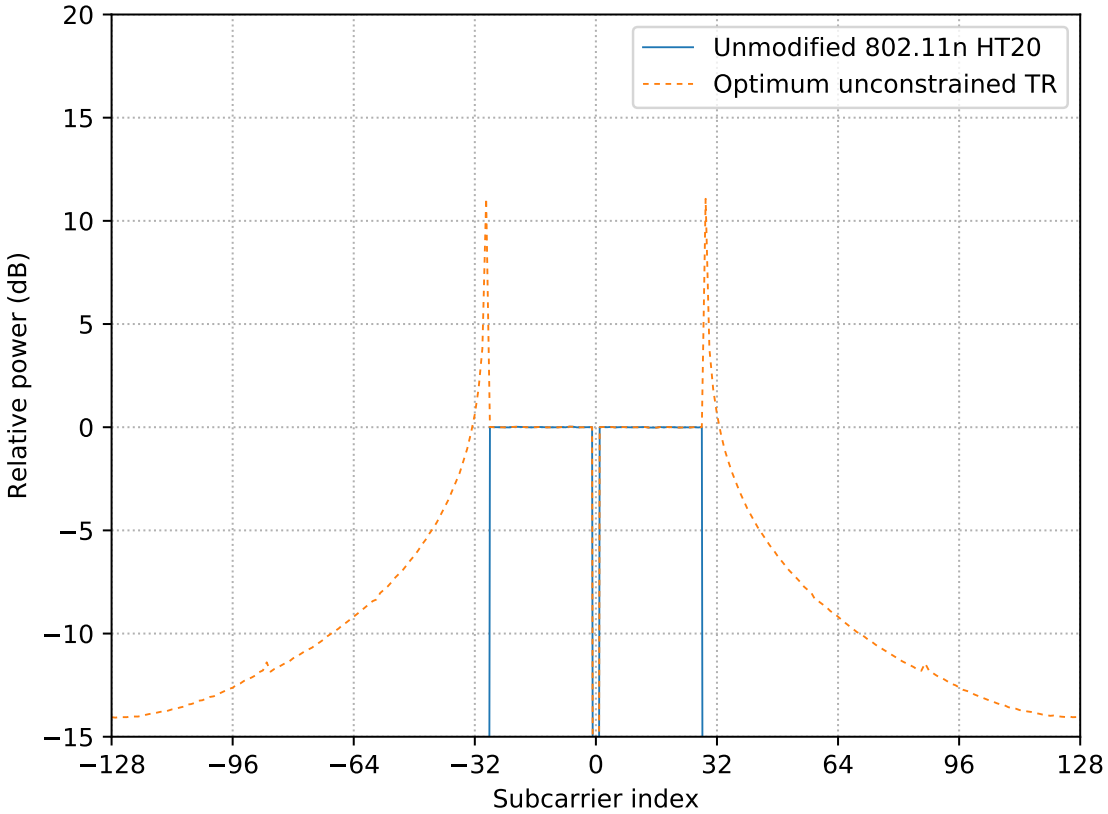


Figure 3-14: Power spectral density for model 802.11n HT20 OFDM waveform (N=64, dense QAM on 56 tones, 20 MHz bandwidth) before and after optimizing all unused tones but the DC subcarrier to minimize the peaks. The out-of-band radiation for optimal Tone Reservation is greatest at the edges of the original passband. Convex optimization using discrete epigraphical form (3.9) with Ipopt Interior Point Method software package [1, 2]. If an OFDM receiver is inspecting the subcarrier amplitudes at the band edges to track the carrier-frequency offset, such an "optimal" Tone Reservation scheme could cause the receiver to detune.

Convex optimization provides a well-known and quite general solution to the problem of reducing PAR via TR and EVM optimization [5], and offers a great deal of freedom to trade off PAR against error-vector magnitude and out-of-band radiation. However, these schemes lead to a significant increase in transmitter complexity.

In addition to asymptotics, we will consider concrete finite complexity numbers for a hypothetical 802.11n HT20 transmitter using 4× oversampling. The baseline complexity of this transmitter is one $N = 256$-point FFT every 4 µs.

3.7.3 Complexity Model

Definition 3.7.1 (Complexity model) *Sequential, parallel, and arithmetic circuit complexity figures are given in a model where fused multiply-add, compare-and-swap, mux-2, divide, square root, and inverse square root are one operation, each. Sequential and parallel complexity may be given as asymptotics; circuit complexity is non-asymptotic. Four figures are reported: sequential time complexity $t_\perp$, parallel time complexity $t_\parallel$, circuit size complexity $s_\circ$, and circuit depth complexity $t_\circ$. Parallel complexity is in the PRAM model. Arithmetic circuits have a maximum fan-in of 3 (for the FMA and mux-2 operations) and unlimited fan-out. Trivial combinatoric wiring that can be expressed with a single logic gate may be absorbed into the cost of the downstream operations. The problem size parameter N is the number of samples per OFDM symbol, not including the cyclic prefix.*

Pipelining notwithstanding, the latency of a hardware implementation will scale like $t_\circ$. A hardware design may have fewer functional units than $s_\circ$ by re-using units from cycle to cycle. However, the total number of usages of functional units per OFDM symbol will scale like $s_\circ$, and the number of OFDM symbols-per-second is fixed by the standard. Thus, the power consumed by a hardware implementation will scale like $s_\circ$. If, in addition, the design operates with a duty cycle of 100% (i.e. the transmitter is clocked just fast enough to meet its real-time requirements), then the area of silicon required scales like $s_\circ$.

We can now quantify the cost of the schemes under comparison (Table 3.6).

Scheme	$t_\perp$ (Sequential)	$t_\parallel$ (Parallel)	$s_\circ$ (Circuit size)	$t_\circ$ (Depth)	
Baseline	$3N \log N$	$3 \log N$	$3N \log N$	$3 \log N$	
	Added cost per iteration				Iters
[4]	$8N + 2$	$\log N + 5$	$\frac{N}{2} \log N + 8N + 2$	$2 \log N + 5$	40
[5]	$(4/3)N^3 + 12N \log N + 2N^2$	$12 \log N + 14N - 4$	$(4/3)N^3 + \mathcal{O}(N^2)$	$12 \log N + 14N - 4$	2
§3.7.8	$6N \log N + 16N$	$7 \log N + 11$	$6N \log N + 16N$	$7 \log N + 11$	5
	Total cost with $N = 256$				
Baseline	6144	24	6144	24	
[4]	88144	544	129104	864	
[5]	45056683	7376	45580288	7376	
§3.7.8	88064	359	88064	359	

Table 3.6: Complexity of peak-to-average ratio optimization schemes. Quantities for Tellado and Cioffi [4] are exact; quantities for Aggarwal and Meng [5] are lower bounds. See Definition 3.7.1 for detailed interpretation.

One main functional block is the N-point complex FFT. The baseline design given in Table 3.6 consists entirely of a single such block. The specific implementation used for the calculation is the radix-4 design given by Karner et al [82], which those authors have optimized in terms of real fused multiply-add (FMA) functional units.

Two prior TR optimization schemes are highlighted in the table: Tellado and Cioffi [4], and Aggarwal and Meng [5].

Tellado and Cioffi proceed by primal subgradient descent, a natural choice for its low-complexity iterations. Each step consists of finding a global maximum of amplitude in the time domain, and then permuting and scaling a pre-computed step direction. This low complexity is, however, offset by the need for tens of iterations.

Aggarwal and Meng offer an optimized Interior Point Method that requires just two iterations, each requiring four FFTs of size N and a positive definite linear system solve of size $2N$. Counting only FFTs and linear solves—reported as the dominant costs—allows us to place a lower bound on the complexity of each iteration. Note, however, that the circuit depth estimate may be pessimistic by as much as 3% at $N = 256$ due to the unknown mutual parallelizability of certain blocks.

Cutting-plane method. If we consider a hypothetical alternative design using the Analytic Center Cutting-Plane Method (ACCPM) [83, 84, 85, 86], the complexity per iteration is dominated by linear solves costing $\mathcal{O}(N^3)$ in the Newton method re-centering procedure. Evaluation of the oracle to generate separating hyperplanes for the TR or EVM problems requires evaluating one FFT of size N, $2N$ complex magnitudes, two log-depth N-point maximums to find the most-violated constraint and the highest peak, and the generation of a complex exponential sequence $\{e^{2\pi jnm/N}\}$ by look-up (serial, PRAM) or FFT (circuit). The number of iterations required by this algorithm is bounded by $\mathcal{O}(N^2/\varepsilon^2)$ [85], assuming an ε-accurate solution is sufficient. Alternatively, if we have a bound R on the half-side-length of a cubical region in which the optimum may be found (Lemma 3.7.3), and we are satisfied with a $1/R$-accurate solution, then the number of iterations is $\mathcal{O}(N \log^2 R)$. Constant factors notwithstanding, this bound on iterations is unlikely to be lower than 2 for a typical instance of size $N = 256$; and so, from a complexity standpoint, the ACCPM seems not to dominate the $\mathcal{O}(N^3)$ scheme of Aggarwal and Meng.

Alternating Direction Method of Multipliers (ADMM). The following section details a new application of ADMM to this optimization problem. Variants of this new scheme compare favorably in circuit size and depth. Additionally, these same variants have the property that the computation consists almost entirely of FFTs, allowing implementors to reuse existing hardware blocks or block designs.

3.7.4 Construction and Convergence

Because it would worsen the challenge of channel estimation and tracking at the receiver, we will eschew varying the transmit gain from symbol to symbol. Instead, we seek an efficient algorithm that will tamp down the peaks of each symbol as much as possible; we will then normalize the entire packet to its largest peak. If we perform this optimization on each symbol in turn, we can terminate our iterations early once the current symbol's peaks are no worse than the worst optimized symbol so far.

With this plan in mind, it is clear that the quantity we need to optimize is not

the peak-to-average ratio of a symbol, but simply the peak instantaneous power. In complex baseband notation, the "instantaneous power" is defined as the non-reactive (i.e. dissipated) power averaged over one cycle of the carrier:

$$\begin{aligned}\langle \Re\{v(t)e^{j\omega_c t}\}\Re\{i(t)e^{j\omega_c t}\}\rangle &= (1/2)\Re\{v(t)i^*(t)\} \\ &= (1/2)|v(t)|^2/Z_0\end{aligned}$$

From this point forward, we will drop the constant factor $(1/2Z_0)$ and write the power simply as $|v(t)|^2$. The complex baseband voltage of an OFDM waveform of bandwidth B is, within one symbol,

$$v(t) = \frac{1}{N}\sum_{k=-\lfloor\frac{N}{2}\rfloor}^{\lfloor\frac{N-1}{2}\rfloor} e^{2\pi jktB/N} X_k \tag{3.1}$$

leading to the problem statement

$$\begin{aligned}&\min_{X\in\mathcal{S}}\max_{t}\left|\frac{1}{N}\sum_{k=-\lfloor\frac{N}{2}\rfloor}^{\lfloor\frac{N-1}{2}\rfloor} e^{2\pi jktB/N} X_k\right|^2 \\ =&\min_{X\in\mathcal{S}}\max_{\theta\in[0,2\pi)}\left|\frac{1}{N}\sum_{k=-\lfloor\frac{N}{2}\rfloor}^{\lfloor\frac{N-1}{2}\rfloor} e^{jk\theta} X_k\right|^2 \\ =&\min_{X\in\mathcal{S}}\|v_X(\theta)\|_\infty^2\end{aligned} \tag{3.2}$$

for some set $\mathcal{S}$ of acceptable modified signal vectors. We will refer to this form (3.2) as the *squared normal continuous form* of the problem: the objective is a squared max-norm over a continuous variable θ. If the norm were not squared, we would refer to the problem formulation as the linear normal continuous form.

Constraint sets. Meaningful assignments for $\mathcal{S}$ include the affine set $\{X|(\forall i \in \mathcal{D} : X_i = d_i)\}$, which constrains a subset of subcarriers to carry data values d_i, while leaving the remainder free to vary. This is the classical Tone Reservation problem. Another is $\{X|(\forall i : |X_i - d_i| \leq c_i)\}$, which allows us to bound the components of

the error vector as well as the amount of out-of-band radiation on the reserved tones. Letting $A = \sum_{i \in \mathcal{D}} e_i {e_i}^\dagger$, we can notate the first set as $\{AX = d\}$, and the second set as $\{|X - d| \preccurlyeq c\}$. Note that the second set is also convex: it is an intersection of convex sets.

Epigraphical form. An alternative *squared epigraphical continuous form* with a linear objective is given by

$$\left.\begin{array}{ll} \inf & \eta \\ \text{s.t.} & |v_X(\theta)|^2 - \eta \leq 0 \quad \forall \theta \in [0, 2\pi) \\ & X \in \mathcal{S} \end{array}\right\} \tag{3.3}$$

Squaring serves the purpose of making the constraints differentiable.

Closed convexity.

Lemma 3.7.1 *The normal* (3.2) *and epigraphical* (3.3) *forms have closed convex objectives over* $\mathbb{R}^{2N}$ *and* $\mathbb{R}^{2N+1}$*, respectively.*

Proof. The normal objective is a closed convex function because its epigraph is closed and convex; the epigraph is a closed and convex set because it is an intersection of (uncountably many) epigraphs of closed convex functions.

The epigraphical objective is a closed convex function because it is linear.

■

Proposition 3.7.1 *Either of the two constraint sets* $\mathcal{S}$ *is convex and closed, and the inequality constraints of the epigraphical form can be written in terms of closed convex functions* $h_\theta(X) \leq 0$.

Proposition 3.7.2 *The normal objective* (3.2) *is locally Lipschitz and locally bounded [87, Theorem 3.3.1, p. 64], and subdifferentiable on* $\mathbb{R}^{2N}$ *[87, Proposition 3.4.1, p. 17].*

Proposition 3.7.3 *Both objectives are bounded below by zero on the constraint set, and both are finite for any finite point, so both are proper convex functions on the constraint set provided that it is non-empty.*

Equivalence.

Lemma 3.7.2 *Any optimum of the squared epigraphical form is an optimum of the squared normal form, provided that $\mathcal{S}$ is convex.*

Proof. The feasible set of the squared epigraphical form is the epigraph of the objective of the squared normal form over its feasible set. That is, a point (η, X) is feasible in the epigraphical form iff X is feasible in the normal form and $\eta \geq \|v_X(\theta)\|_\infty^2$. So a feasible point in the epigraphical form gives an upper bound on the optimum of the normal form, and an optimal point of the epigraphical form infimizes the objective of the normal form over its feasible set. Now the objective is bounded from below (by zero) and the problem is convex, so the infimum is attained.

■

Proposition 3.7.4 *The nonnegativity of the norm implies a bijection between the optimal sets of the squared and linear forms of the PAR problem, and similarly for the feasible sets.*

Strong duality. The normal form has no inequality constraints, and its equality constraints are affine. The affine hull of the domain of its objective and constraint functions is the full space $\mathbb{R}^{2N}$, and hence so is its relative interior. The feasible set is nonempty (e.g. set all free variables to zero). So the Slater condition holds for the normal form.

The epigraphical form has affine equality constraints. The affine hull of the domain of its objective and constraint functions is the full space $\mathbb{R}^{2N+1}$, and hence so is its relative interior. Given a feasible point in the normal form, we can construct a point for which no inequality constraint binds in the epigraphical form by letting $\eta = \|v(\theta)\|_\infty + \epsilon$. So the Slater condition holds for the epigraphical form.

Both forms satisfy the Slater condition, so both exhibit strong duality. In addition, the objective of each form is bounded below by 0.

Boundedness, existence. We can bound the primal variables to aid in applying convergence proofs later.

Lemma 3.7.3 *If an optimum point exists for the problem in either the squared normal or epigraphical forms, with any constraint set $\mathcal{S}$, it lies in the closed, bounded set*

$$\mathcal{X} = \{\|X\|_\infty \leq R\}$$

for the normal form, or

$$\{0 \leq \eta \leq R^2/N^2\} \times \mathcal{X}$$

for the epigraphical form, with $R = N\|v_d(\theta)\|_\infty$. Furthermore, $\mathcal{X} \subset \mathcal{X}_{poly}$, where $\mathcal{X}_{poly}$ is the closed, bounded, polyhedral set

$$\{\|\Re X\|_\infty \leq R \cap \|\Im X\|_\infty \leq R\}$$

Proof. First, by the equivalence of the optima of the squared normal and epigraphical forms, if we can establish the bound for the epigraphical form, then it follows for the normal form. So we will consider only the epigraphical form.

Suppose (η, X) is an optimum, with $\eta < 0$. But for any feasible point, $0 \leq |v_X(\theta)|^2 \leq \eta$ for any i. Contra. $0 \leq \eta$.

Suppose (η, X) is an optimum, with $\eta > \|v_X(\theta)\|_\infty^2$. Then $(\|v_X(\theta)\|_\infty^2, X)$ is feasible and has a better objective. Contra. $0 \leq \eta = \|v_X(\theta)\|_\infty^2$.

Suppose (η, X) is an optimum, with $\eta > \|v_d(\theta)\|_\infty^2$. Then $(\|v_d(\theta)\|_\infty^2, d)$ is feasible and has a better objective. Contra. $0 \leq \eta = \|v_X(\theta)\|_\infty^2 \leq \|v_d(\theta)\|_\infty^2$.

In the inequality $\|v_X(\theta)\|_\infty^2 \leq \|v_d(\theta)\|_\infty^2$, the maximization of $|v_X(\theta)|$ over a continuous set $\theta \in [0, 2\pi)$ is lower bounded by a maximization over any subset. In particular,

take the subset $\frac{N\theta}{2\pi} \in \{0, \ldots, N-1\}$. On this subset, the values of $v(\theta)$ are a permuted IDFT of X_i, so that

$$\begin{aligned}
\|v_X(\theta)\|_\infty^2 &\geq \|\mathrm{IDFT}(X)\|_\infty^2 \\
&\geq N^{-1}\|\mathrm{IDFT}(X)\|_2^2 \\
&= N^{-2}\|X\|_2^2 \\
&\geq N^{-2}|X_i|^2 \text{ for any } i
\end{aligned}$$

So if $|X_i| > N\|v_d(\theta)\|_\infty$, then

$$\|v_X(\theta)\|_\infty^2 > \|v_d(\theta)\|_\infty^2,$$

and $(\|v_X(\theta)\|_\infty^2, X)$ cannot be an optimum.

The inclusion of $\mathcal{X} \subset \mathcal{X}_{\text{poly}}$ follows immediately from bounding the component-wise complex magnitude by the absolute value of the real and imaginary parts.

■

Lemma 3.7.4 *The optimal sets of the* $\{$*squared, linear*$\} \times \{$*normal, epigraphical*$\}$ *forms of the problem are nonempty, compact, and convex provided that* $\mathcal{S}$ *is closed, convex, and non-empty.*

Proof. Follows immediately [88, Theorem 16. p. 44].

3.7.5 Continuous Time

One approach—which is not the one we will ultimately take—is to keep the signal in continuous-time form, representing the instantaneous power as a rational function $0 \leq p_X(\theta) = P(z)/z^{N-1}$ in $z = e^{j\theta}$. The coefficients P_k are Hermitian forms in X. The relative extrema of the instantaneous power are then in one-to-one correspondence with the unit-magnitude roots $z_i^\star$ of $R = zP' - (N-1)P$. There is no need to introduce a constrained optimization or Lagrange multiplier because $p_X(z)$ is analytic away

from $z = 0$, and hence its partial derivative along the unit circle is zero iff its complex derivative is zero.

At a point X such that the maximum of $|v_X(\theta)|^2$ over θ is unique, the max-norm is differentiable. In this case, we can write the objective as

$$f(X) = p_X(\arg z_i^\star(X)) = P(z_i^\star(X))/(z_i^\star(X))^{N-1}$$

Using the chain rule, we can compute $\nabla_X f(X)$ from four quantities: $z_i^\star$, the X-partials of P/z^{N-1} at $z_i^\star$, the z-partials of P/z^{N-1} at $z_i^\star$, and the X-partials of $z_i^\star$. This last quantity can be computed by implicit differentiation:

$$\begin{aligned}
0 &= R(z_i^\star) \\
0 &= (R + dR)(z_i^\star + dz_i^\star) \\
0 &= R(z_i^\star) + R'(z_i^\star)dz_i^\star + dR(z_i^\star) \\
dz_i^\star &= -dR(z_i^\star)/R'(z_i^\star) \\
\nabla_X z_i^\star &= -\sum_k (\nabla_X R_k)(z_i^\star)^k / R'(z_i^\star)
\end{aligned}$$

At points of non-differentiability, the subgradient of the objective is the convex hull of the set of gradients obtained by picking a single extremum from the $\arg\max_\theta$.

One disadvantage of performing gradient descent on the continuous-time objective is that it involves solving a number of large eigenvalue problems to find polynomial roots. Without some further subtlety, this is unlikely to provide an adequate level of performance or numerical stability. The same problem arises in using cutting plane methods on this objective: in order to evaluate the subgradient, we must first locate the extrema by root-finding (though early iterations could be accelerated by testing only the constraints at equally-spaced values of θ using the IDFT). It is conceivable that an active set algorithm could be devised using a continuation method [89], reaching the optimum in a bounded number of steps. Alternatively, there exist interior point methods for Linear Semi-Infinite Programming (LSIP) [90]; posed as an LSIP, the

linear epigraphical continuous form with affine $\mathcal{S}$ reads

$$\left.\begin{array}{ll} \inf & \eta^{1/2} \\ \text{s.t.} & \eta^{1/2} - \Re\left\{e^{-j\phi} v_X(\theta)\right\} \geq 0 \quad \forall(\theta,\phi) \in [0,2\pi)^2 \\ & AX = d \end{array}\right\} \tag{3.4}$$

3.7.6 Discrete Time

Our approach will be to work in discrete time with a sample rate equal to an integer multiple of the bandwidth B. The reason for this over-sampling is because the instantaneous power function is a trigonometric polynomial of order $2N-1$, and it may contain sharp peaks which to a certain extent "fall between the cracks" if we only sample at the Nyquist rate B of the complex baseband voltage. This phenomenon is called peak regrowth, and has been extensively characterized [91]. Our problem is now

$$\min_{X\in\mathcal{S}} \max_{0\leq i<LN} \left| \frac{1}{N} \sum_{k=-\lfloor\frac{N}{2}\rfloor}^{\lfloor\frac{N-1}{2}\rfloor} e^{2\pi jik/LN} X_k \right|, \text{ or}$$
$$\min_{X\in\mathcal{S}} \|F^\dagger X\|_\infty, \tag{3.5}$$

where $F^\dagger$ is an IDFT-like linear operator; if we re-index the components of X by ℓ such that

$$\begin{aligned} k &= (\ell + \lfloor N/2 \rfloor)\%N - \lfloor N/2 \rfloor \\ \ell &= k\%N \\ -\left\lfloor \frac{N}{2} \right\rfloor \leq k &\leq \left\lfloor \frac{N-1}{2} \right\rfloor \\ 0 \leq \ell &\leq N-1 \end{aligned}$$

then we can write $F^\dagger$ as the matrix

$$F^\dagger_{i\ell} = (1/N) e^{2\pi ji((\ell+\lfloor N/2\rfloor)\%N - \lfloor N/2\rfloor)/LN} \tag{3.6}$$

so that for $(m-n)\%L = 0$,

$$(F^\dagger F)_{mn} = \sum_{\ell=0}^{N-1} F^\dagger_{m\ell} {F^\dagger_{n\ell}}^* = (1/N)\delta_{mn}, \tag{3.7}$$

and for $0 \le p, q \le N-1$,

$$(FF^\dagger)_{pq} = \sum_{i=0}^{LN-1} {F^\dagger_{ip}}^* F^\dagger_{iq} = (1/N)\delta_{pq}. \tag{3.8}$$

We can once again write the problem in epigraphical form to get a linear objective and, depending on $\mathcal{S}$, affine equality and/or quadratic inequality constraints:

$$\left.\begin{array}{rll} \inf & \eta & \\ \text{s.t.} & \eta - \left|(F^\dagger X)_i\right|^2 \ge 0 & \forall 0 \le i < N \\ \text{and either} & AX = d & \\ \text{or} & {c_i}^2 - \left|X_i - d_i\right|^2 \ge 0 & \forall 0 \le i < N \end{array}\right\} \tag{3.9}$$

Either form of (3.9) is suitable for use with Interior Point Methods that require twice-differentiable objective and constraint functions. The gradients and Hessians can be computed efficiently using the FFT.

3.7.7 Proximal Mappings and the Alternating Direction Method of Multipliers

This section is provided as background for the algorithm presented in §3.7.8.

Preliminaries

These definitions are standard [92], except insofar as we are working over $\mathbb{C}^n$ viewed as $\mathbb{R}^{2n}$. For notational simplicity, we define the dot product notation for complex vectors to be their dot product over $\mathbb{R}^{2n}$: $u \cdot v \triangleq \Re\{u^\dagger v\}$. The coincidence of the Euclidean norm on the two spaces will also work to our advantage.

Definition 3.7.2 (Epigraph) *The epigraph of a function $f\colon \mathbb{C}^n \to (-\infty, +\infty]$ is a*

subset of $\mathbb{C}^n \times (-\infty, +\infty]$ *given by*

$$\mathrm{epi}(f) = \{(x, y) | y \geq f(x)\}$$

Definition 3.7.3 (Subdifferential) *The subdifferential of a convex function* $f\colon \mathbb{C}^n \to \mathbb{R}$ *is the mapping*

$$\begin{aligned}(\partial f) &: \mathbb{C}^n \to 2^{\mathbb{C}^n} \\ (\partial f)(x) &= \{v \,|\, \forall y.\, f(y) - f(x) \geq v \cdot (y - x)\}\end{aligned}$$

An element $v \in (\partial f)(x) \subseteq \mathbb{C}^n$ *is called a subgradient.*

Definition 3.7.4 (Proximal mapping) *Given a function* $f\colon \mathbb{C}^n \to \mathbb{R}$*, for* $\lambda > 0$*,*

$$\begin{aligned}\mathrm{prox}_{\lambda f} &\colon \mathbb{C}^n \to \mathbb{C}^n \\ \mathrm{prox}_{\lambda f}(v) &= \arg\min_x \left(f(x) + \frac{1}{2\lambda} \|x - v\|_2^2 \right)\end{aligned}$$

The positive scale factor λ is not *per se* part of the definition of the proximal mapping, but it arises so frequently that we will insert it wherever practical.

Definition 3.7.5 (Indicator function) *Given a set* $\mathcal{S} \subseteq \mathbb{C}^n$*,*

$$\begin{aligned}I_{\mathcal{S}} &\colon \mathbb{C}^n \to [0, +\infty] \\ I_{\mathcal{S}}(x) &= \begin{cases} 0 & x \in \mathcal{S} \\ +\infty & o.w. \end{cases}\end{aligned}$$

Proposition 3.7.5 *An indicator function* $I_{\mathcal{S}}$ *is convex and closed iff* $\mathcal{S}$ *is convex and closed.*

Proposition 3.7.6 *An indicator function* $I_{\mathcal{S}}$ *is proper convex iff* $\mathcal{S} \neq \varnothing$*.*

Definition 3.7.6 (Convex conjugate) *Given a function $f\colon \mathbb{C}^n \to \mathbb{R}$,*

$$f^*(y)\colon \mathbb{C}^n \to \mathbb{R}$$
$$f^*(y) = \sup_x (y \cdot x - f(x))$$

Proposition 3.7.7 *The convex conjugate of the scalar multiple λf is given by*

$$(\lambda f)^*(y) = \lambda f^*(y/\lambda)$$

Definition 3.7.7 (Dual norm) *Given a norm $\|\cdot\|\colon \mathbb{C}^n \to [0,\infty)$, its dual norm is*

$$\|\cdot\|_*\colon \mathbb{C}^n \to [0,\infty)$$
$$\|y\|_* = \sup_{\|x\|\leq 1} |y \cdot x| = \sup_{x\neq 0} |y \cdot x|/\|x\|$$

Note that the dual norm so constructed using the dot product over $\mathbb{R}^{2n}$ coincides with the dual norm constructed "natively" over $\mathbb{C}^n$ for any p-norm. This is because $|y \cdot x| = |\Re\{y^\dagger x\}| \leq |y^\dagger x|$, with equality when the phases of the components of y and x align. Since $\arg x_i$ does not affect $\|x\|_p$, the inequality holds with equality for some x in the domain of the supremum, and so $\sup |y \cdot x| = \sup |y^\dagger x|$.

Lemma 3.7.5 *If f is a norm $\|\cdot\|$, then its convex conjugate is the indicator function $I_{\mathcal{B}}$ of the ball $\mathcal{B}_{\|\cdot\|_*}(1) = \{x \mid \|x\|_* \leq 1\}$:*

Proof. Note that $f^*(y) = \sup_x (y \cdot x - \|x\|) \geq 0$, since $x = 0$ gives a supremand of 0. Then consider two cases:

1. If $\|y\|_* \leq 1$, $\forall x\,.\,(y \cdot x - \|x\|) \leq 0$ and hence $f^*(y) = 0$.

2. If $\|y\|_* > 1$, $\exists(\|v\| \leq 1)\,.\,\|v\| \leq 1 < y \cdot v$, and $\lim_{a\to+\infty}(y \cdot (av) - \|av\|) = +\infty = f^*(y)$.

■

Theorem 3.7.1 (Moreau decomposition) *At a point v, the proximal mapping of a function and the proximal mapping of its convex conjugate sum to v:*

$$v = \mathrm{prox}_{\lambda f}(v) + \mathrm{prox}_{(1/\lambda)f^*}(v/\lambda)$$

Proposition 3.7.8 *The proximal mapping of the indicator function of a convex set, evaluated at v, is the nearest point in the set by Euclidean distance:*

$$\mathrm{prox}_{\lambda I_{\mathcal{S}}}(v) = \arg\min_{x \in \mathcal{S}} \frac{1}{2}\|x - v\|_2^2$$

Proposition 3.7.9 *The proximal mapping of a norm $\|\cdot\|$, evaluated at v, is a step toward zero by the projection of v onto a dual norm ball $\mathcal{B}_{\|\cdot\|_*}(\lambda) = \{x \mid \|x\|_* \leq \lambda\}$:*

$$\begin{aligned}
\mathrm{prox}_{\lambda\|\cdot\|}(v) &= v - \mathrm{prox}_{(1/\lambda)I_{\mathcal{B}(1)}}(v/\lambda) \\
&= v - \arg\min_{x \in \mathcal{B}(1)} \frac{1}{2}\|x - v/\lambda\|_2^2 \\
&= v - \arg\min_{x \in \mathcal{B}(\lambda)} \frac{1}{2}\|x - v\|_2^2
\end{aligned}$$

Lemma 3.7.6 *The proximal mapping of the indicator function of the tone reservation set $\{Ax = d\}$ is given by the standard least-squares projection*

$$\mathrm{prox}_{\lambda I_{\{Ax=d\}}}(v) = (\mathbf{1} - A^\dagger(AA^\dagger)^{-1}A)v + A^\dagger(AA^\dagger)^{-1}d$$

Lemma 3.7.7 *The proximal mapping of the indicator function of the bounded-EVM set $\{|x - d| \preccurlyeq c\}$ is given by*

$$\begin{aligned}
\mathrm{prox}_{\lambda I_{\{|x-d| \preccurlyeq c\}}}(v) &= \mathrm{prox}_{\lambda \sum_i I_{\{x_i - d_i \leq c_i\}}}(v) \\
&= \Big[\mathrm{prox}_{\lambda I_{\{|x_i - d_i| \leq c_i\}}}(v_i), \quad 0 \leq i < N\Big] \quad \text{[93, Theorem 6.6, p. 135]} \\
&= d + \Big[\min(|E_i|, c_i)E_i/|E_i|, \quad 0 \leq i < N\Big], \textit{ where} \\
E &= v - d
\end{aligned}$$

Definition 3.7.8 (Subscript-plus operator)

$$a_+ \triangleq \max\{a, 0\}$$

Lemma 3.7.8 *The Euclidean projection of v onto $\mathcal{B}_{\|\cdot\|_1}(\lambda)$, $\lambda > 0$, is the vector with components $(1 - \mu_+/|v_i|)_+ v_i$, where μ is the unique solution of $\lambda = \sum_i (|v_i| - \mu)_+$.*

To prove Lemma 3.7.8, it suffices to consider the nonnegative real case $0 \preccurlyeq v$, because the Euclidean distance function and the set $\mathcal{B}_{\|\cdot\|_1}(\lambda)$ are invariant under phase-rotations of the coordinate axes. Furthermore, we can restrict the domain of the minimization to vectors $0 \preccurlyeq x \in \mathcal{B}_{\|\cdot\|_1}(\lambda)$: any other solution point is (weakly) dominated by its component-wise absolute value, which is also in the ball, and is no farther from v.

So it suffices to show that at the optimum of the quadratic program

$$\left.\begin{array}{rr} \min\limits_x & \frac{1}{2}\|x - v\|_2^2 \\ \text{s.t.} & \sum\limits_i x_i \leq \lambda \\ & 0 \leq x_i \end{array}\right\}, \tag{3.10}$$

$x_i = (v_i - \mu_+)_+$, where μ is the unique solution of $\lambda = \sum_i (v_i - \mu)_+$.

(Note that the RHS of the equation for μ is convex and piecewise linear. Its asymptotic value and slope for large μ are 0, and it grows without bound for negative μ. So for any $\lambda > 0$, the solution is unique.)

We verify the optimum with the KKT conditions, since this problem satisfies the linear constraint qualification. We identify μ_+ as the dual variable for the constraint $\sum_i x_i - \lambda \leq 0$, and we introduce μ_i as the dual variable for the constraint $-x_i \leq 0$. Pulling a rabbit out of a hat, we supply the assignment $\mu_i = (\mu - v_i)_+$, and claim that

this putative optimum with these dual variables will satisfy KKT. We need only prove

Stationarity:	$x_i = v_i + \mu_i - \mu_+$	
Primal feasibility:	$0 \le x_i$	$\sum_i x_i \le \lambda$
Dual feasibility:	$0 \le \mu_i$	$0 \le \mu_+$
Complementary slackness:	$0 = \mu_i x_i$	$0 = \mu_+ \left(\sum_i x_i - \lambda\right)$

Substituting our putative solution, we require

$$
\begin{aligned}
(v_i - \mu_+)_+ &= v_i + (\mu - v_i)_+ - \mu_+ \\
0 &\le (v_i - \mu_+)_+ \\
\sum_i (v_i - \mu_+)_+ &\le \sum_i (v_i - \mu)_+ \\
0 &\le (\mu - v_i)_+ \\
0 &\le \mu_+ \\
0 &= (\mu - v_i)_+ (v_i - \mu_+)_+ \\
0 &= \mu_+ \left(\sum_i (v_i - \mu_+)_+ - \sum_i (v_i - \mu)_+ \right)
\end{aligned}
$$

We then divide the problem into two cases: $\mu \ge 0$, and $\mu < 0$. In either case, straightforward manipulations show that these conditions are met. Lemma 3.7.8 follows.

Lemma 3.7.9 (Haircut mapping) *The proximal mapping of $\|\cdot\|_\infty$, evaluated at v, gives the vector x with components $x_i = (\mu_+/|v_i|)_+ v_i$, where μ is the unique solution of $\lambda = \sum_i (|v_i| - \mu)_+$.*

If we interpret the vector v as a complex discrete-time signal, then $x = \mathrm{prox}_{\lambda\|\cdot\|_\infty}(v)$ is the signal after we shave off $\|v - x\|_1 \le \lambda$ units of hair (amplitude), starting from the highest spikes and working our way inward. Naturally, if λ exceeds the total amount of hair $\|v\|_1$, then we are left with the zero vector, no hair at all.

Augmented Lagrangian Methods

Definition 3.7.9 (Lagrangian dual problem) *Given a convex optimization problem with $f\colon \mathbb{C}^n \to \mathbb{R}$ continuous and affine equality constraints,*

$$\begin{aligned} p^\star &= \min_x f(x) \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

the Lagrangian dual problem is

$$\begin{aligned} d^\star &= \max_y g(y), \text{ where} \\ \Lambda(x,y) &= f(x) + y \cdot (Ax - b) \\ g(y) &= \inf_x \Lambda(x,y) = -f^*(-A^\dagger y) - y \cdot b \end{aligned}$$

and convexity+continuity implies strong duality, $p^\star = d^\star$.

Definition 3.7.10 (Augmented Lagrangian)

$$\Lambda_\rho(x,y) = f(x) + y \cdot (Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2$$

Proposition 3.7.10 *The optima of the augmented and unaugmented Lagrangians are identical, because (grouping the new term into the objective) the KKT conditions for optimality are unchanged: the value and gradient of the new term are identically zero on the feasible set.*

Algorithm 3.7.1 (Projected primal descent) *At each iteration, we update the primal variable by a step in the projected (sub-)gradient direction:*

$$x^{k+1} := x^k - \alpha^{(k)}(\mathbf{1} - A^\dagger(AA^\dagger)^{-1}A)(\partial f)(x^k)$$

Note that for the PAR problem, f is $\|\cdot\|_\infty$, and its subdifferential at $x \neq 0$ is the convex hull of the set $\{(x_i/|x_i|)e_i \mid i \in \arg\max_i |x_i|\}$. If we perform the projected primal descent step using a subgradient $e_i(x_i/|x_i|)$, then this algorithm is identical

to that proposed by Tellado et al [4]: the projection operator $(\mathbf{1} - A^\dagger(AA^\dagger)^{-1}A)$ acts on basis vectors e_i to give the most-impulse-like signal which contains energy only in the subspace of reserved tones. As we have argued, the arithmetic complexity of this iteration is high.

Ideally, we would like to find an algorithm that can take large steps that push down all of the peaks at once in the time domain, while maintaining either our TR or EVM constraints in the frequency domain. One way to *separate* these two concerns is to rewrite our problem as

$$\min_x(\|F^\dagger x\|_\infty + I_{\mathcal{S}}(x)) \tag{3.11}$$

There is no magic here; the reason why this separation is valuable to us is because each of these terms has a proximal mapping which we can evaluate exactly and efficiently: the first is the haircut mapping of Lemma 3.7.9, and the second is one of the projections of Lemmas 3.7.6 and 3.7.7. There exist algorithms, like the Alternating Direction Method of Multipliers, which have quite favorable convergence properties on problems of this (separable) form.

In fact, the claim of the preceding paragraph is not quite correct. The proximal mapping of $(\|\cdot\|_\infty \circ F^\dagger)$ is not quite so nice an object as the haircut mapping $\text{prox}_{\lambda\|\cdot\|_\infty}$. The property needed [93, Theorem 6.15, p. 140] for the composition of the "proximable" function $\|\cdot\|_\infty$ with the linear map $F^\dagger$ to have a well-known solution is $F^\dagger F = \alpha\mathbf{1}$ for some $\alpha > 0$. The property we have is $FF^\dagger = \mathbf{1}/N$.

The solution is to embed the vector $x \in \mathbb{C}^N$ into the larger space $\mathbb{C}^{LN}$; that is, $F^\dagger x$ is still the oversampled discrete-time sequence, but now, x is the full LN-point DFT of this oversampled sequence. This does not change the optimization problem, provided that we constrain all of the tones above the original Nyquist frequency to be zero. For the affine TR variant of the feasible set $\mathcal{S}$, this means that we must grow the matrix A and the vector

d to constrain these new degrees of freedom to be zero. Similarly, for the EVM-optimization feasible set, we must grow the d and c vectors with new components $d_i = c_i = 0$.

With this change, the matrix $F^\dagger$ simplifies to

$$F^\dagger_{i\ell} = (1/N)e^{2\pi j i\ell/LN} \tag{3.12}$$

so that $FF^\dagger = F^\dagger F = (L/N)\mathbf{1}$. Now we obtain [93, Theorem 6.15, p. 140], using the fact that the ∞-norm is a proper closed convex function,

$$\begin{aligned}\operatorname{prox}_{\lambda\|\cdot\|_\infty \circ F^\dagger}(x) &= x + (N/L)F(\operatorname{prox}_{\lambda(L/N)\|\cdot\|_\infty}(F^\dagger x) - F^\dagger x) \\ &= F^{-\dagger}\operatorname{prox}_{\lambda(L/N)\|\cdot\|_\infty}(F^\dagger x)\end{aligned}$$

In words, to evaluate the proximal mapping of the first term in (3.11), we evaluate the haircut mapping on $F^\dagger x$ with a scale parameter of $\lambda(L/N)$, and then transform back into the frequency domain.

The remainder of this section will recall the construction of Alternating Direction Method of Multipliers as presented by Boyd et al [94], highlighting at each step how the intermediate methods fail to meet our needs. Boyd et al's construction reaches the concept of the proximal mapping only in passing, after the construction is complete. Alternative constructions exist which approach ADMM from the perspective of proximal point methods. The approach makes no difference to the final algorithm; let this paragraph serve merely as a "heads-up" that proximal mappings are the jack-in-the-box that emerges after we finish turning the crank.

Looking, then, towards primal-dual methods, we'd like to use information from the dual variables to accelerate our optimization. The dual variables encode the importance of various constraint directions; indeed, in the epigraphical form of the PAR problem, the nonzero dual variables for the inequality constraints correspond directly to the peaks we need to suppress. Our first step on the way to ADMM is dual ascent.

Algorithm 3.7.2 (Dual ascent) [94, (2.2-3), p. 8] *At each iteration, we update the dual variable by a gradient step.*

$$\begin{aligned} x^{k+1} &:= \arg\min_x \Lambda(x, y^k) \\ y^{k+1} &:= y^k - \alpha^{(k)} \partial(-g)(y^k) \\ &:= y^k + \alpha^{(k)}(Ax^{k+1} - b) \end{aligned}$$

One issue with dual ascent for our problem is that if $\|A^\dagger y^k\|_1 > 1$, then $\Lambda(x, y) = \|x\|_\infty + y \cdot (Ax - b)$ is unbounded below in x, and the x-update fails. By passing to the augmented Lagrangian, we can ensure that $\Lambda(x, y)$ is bounded below in x for all y.

Algorithm 3.7.3 (Method of multipliers) [94, (2.7-8), p. 11] *At each iteration, we take a dual ascent step on the augmented Lagrangian, with the penalty parameter ρ equal to the dual variable step size:*

$$\begin{aligned} x^{k+1} &:= \arg\min_x \Lambda_\rho(x, y^k) \\ y^{k+1} &:= y^k + \rho(Ax^{k+1} - b) \end{aligned}$$

The use of the same parameter ρ in the primal and dual updates is special: it maintains dual feasibility of the solution pair at each step.

In order to apply the method of multipliers to the PAR problem, we need to be able to compute the x-update. This requires solving an instance of the convex subproblem

$$\min_x \left(\|x\|_\infty + y \cdot (Ax - b) + (\rho/2)\|Ax - b\|_2^2 \right), \text{ with optimality when}$$

$$-\Re\{A^\dagger(y + \rho(Ax - b))\} \in (\partial\|\cdot\|_\infty)(x)$$

We could stop here, and go down the path of solving this subproblem. However, we would only get a solution to the affine-$\mathcal{S}$ Tone Reservation problem, not the EVM-optimization variant. Instead, we will shift gears. From this point forward, the equality constraints will play a different role. They will no longer be used to

enforce the equality of the data subcarriers with their nominal values. Rather, they will be used to enforce *consensus* between the terms of the PAR problem in separable form (3.11).

In separable form, the feasible set is enforced by an indicator function added to the ∞-norm objective. In consensus form, the two objective terms are evaluated on disjoint variables:

$$\left.\begin{array}{ll} \min\limits_{x,z} & \left(\|F^\dagger x\|_\infty + I_{\mathcal{S}}(z)\right) \\ \text{s.t.} & x = z \end{array}\right\} \tag{3.13}$$

We can now state the alternating direction method of multipliers in terms of the augmented Lagrangian for this problem as follows:

Algorithm 3.7.4 (ADMM) [94, (3.2-4), p. 14]

$$\begin{aligned} \Lambda_\rho(x,z,y) &= f(x) + g(z) + y \cdot (x - z) + (\rho/2)\|x - z\|_2^2 \\ x^{k+1} &:= \arg\min_x \Lambda_\rho(x, z^k, y^k) \\ z^{k+1} &:= \arg\min_z \Lambda_\rho(x^{k+1}, z, y^k) \\ y^{k+1} &:= y^k + \rho(x^{k+1} - z^{k+1}) \end{aligned}$$

We can rewrite Algorithm 3.7.4 in the so-called scaled form to make the role of the proximal mapping manifest. This leads to the equivalent algorithm

Algorithm 3.7.5 (ADMM, scaled form) [94, (3.5-7), p. 15]

$$\begin{aligned} x^{k+1} &:= \arg\min_x \left(f(x) + (1/2\lambda)\|\lambda y^k + x - z^k\|_2^2\right) \equiv \mathrm{prox}_{\lambda f}(z^k - \rho^{-1} y^k) \\ z^{k+1} &:= \arg\min_z \left(g(z) + (1/2\lambda)\|\lambda y^k + x^{k+1} - z\|_2^2\right) \equiv \mathrm{prox}_{\lambda g}(x^{k+1} + \rho^{-1} y^k) \\ y^{k+1} &:= y^k + \rho(x^{k+1} - z^{k+1}) \\ \rho &= 1/\lambda \end{aligned}$$

Now that the primal update proceeds in separate x and z steps, the use of the inverse dual step size $\lambda = 1/\rho$ for the primal updates no longer guarantees that the iterate $(x^{k+1}, z^{k+1}, y^{k+1})$ is dual feasible. Some authors maintain the fixed relationship [94], but it is possible to relax it [95] provided that $0 < \rho \leq \frac{1+\sqrt{5}}{2\lambda}$.

Generically, taking smaller dual steps ρ will delay the convergence of the primal variables to the constraint set, while taking smaller primal steps λ will delay the convergence of the primal variables to the optimal set. For ADMM, small ρ means that consensus will be achieved more slowly.

3.7.8 Final Algorithm

Algorithm 3.7.6 (ADMM PAR)

$$
\begin{aligned}
\text{haircut}(v, \lambda) &\triangleq \text{prox}_{\lambda\|\cdot\|_\infty}(v) \\
\text{project}(v) &\triangleq \arg\min_{x \in \mathcal{S}} \frac{1}{2}\|x - v\|_2^2 \\
x^{k+1} &:= \text{DFT}\left(\text{haircut}(\text{IDFT}(z^k - \rho^{-1} y^k), \lambda/N)\right) \\
z^{k+1} &:= \text{project}(x^{k+1} + \rho^{-1} y^k) \\
y^{k+1} &:= y^k + \rho(x^{k+1} - z^{k+1})
\end{aligned}
$$

Further refinements to the definition of the haircut function follow. If ρ is constant from iteration to iteration, then we can save a few multiplies by re-defining y to be smaller by a factor of ρ. Otherwise, the structure of Algorithm 3.7.6 is complete.

3.7.9 Convergence Properties

The literature contains many general-purpose proofs of convergence for ADMM, each establishing a distinct set of properties. We will draw on a number of these.

Theorem 3.7.2 (Convergence of ADMM) [94, §3.2.1, p. 17] *Suppose that f, g are closed, proper, and convex, and that Λ_0 has a saddle point:*

$$\exists x^\star, z^\star, y^\star. \forall x, z, y. \Lambda_0(x^\star, z^\star, y) \leq \Lambda_0(x^\star, z^\star, y^\star) \leq \Lambda_0(x, z, y^\star)$$

Then the iterates (x^k, z^k, y^k) of ADMM (3.7.5) have the following properties:

1. $x^k - z^k \to 0$
2. $f(x^k) + g(z^k) \to p^\star$
3. $y^k \to y^\star$

We have shown that f and g are closed, proper, and convex (Lemma 3.7.1, Propositions 3.7.3, 3.7.5, and 3.7.6).

Lemma 3.7.10 *Λ_0 has a saddle point.*

Proof. The normal form has a nonempty bounded optimal set, and hence it has a finite primal optimum point $x^\star$ with a finite objective value $p^\star$. The consensus form is equivalent on its feasible set, so $(x^\star, x^\star)$ is in the (primal) optimal set of the consensus form.

By the linear constraint qualification, the subdifferential KKT conditions apply, and hence there exists a $y^\star$ such that

$$\begin{aligned}
&0 \in (\partial_{x,z}\Lambda_0)(x^\star, z^\star, y^\star) \\
&\quad = \begin{bmatrix} (\partial f)(x) + y \\ (\partial g)(z) - y \end{bmatrix} (x^\star, z^\star, y^\star) \\
&\quad \Leftrightarrow \\
&0 \in (\partial f)(x^\star) + y^\star, \\
&0 \in (\partial g)(z^\star) - y^\star
\end{aligned}$$

Then by the definition of the subdifferential,

$$\forall x.\, f(x) - f(x^\star) \geq -y^\star \cdot (x - x^\star)$$
$$\forall z.\, g(z) - g(z^\star) \geq +y^\star \cdot (z - z^\star)$$
$$\forall x, z.\, (f(x) + g(z) + y^\star \cdot (x - z)) \geq (f(x^\star) + g(z^\star) + y^\star \cdot (x^\star - z^\star))$$
$$\forall x, z, y.\, \Lambda_0(x^\star, z^\star, y^\star) \leq \Lambda_0(x, z, y^\star)$$

Additionally, we have

$$0 = x^\star - z^\star$$
$$\forall y.\, y.(x^\star - z^\star) \leq y^\star.(x^\star - z^\star)$$
$$\forall y.\, f(x^\star) + g(z^\star) + y.(x^\star - z^\star) \leq f(x^\star) + g(z^\star) + y^\star.(x^\star - z^\star)$$
$$\forall x, z, y.\, \Lambda_0(x^\star, z^\star, y) \leq \Lambda_0(x^\star, z^\star, y^\star)$$

■

So Theorem 3.7.2 applies to our problem, provided that $\mathcal{S}$ is convex, closed, and nonempty.

Lemma 3.7.11 (Convergence of ADMM PAR) *The iterates x_k converge to the optimal set.*

Proof. We have $f(x^k) + g(z^k) \to p^\star$. Now $g(z) = I_{\mathcal{S}}(z)$ is an indicator function, so its proximal mapping enforces $z^k \in \mathcal{S}$ for $k > 1$. Thus, $g(z^k) = 0$, and so $f(x^k) \to p^\star$.

Now z^k is in the feasible set, and $x^k - z^k \to 0$. so x^k approaches the feasible set.

If x^k simultaneously approaches the feasible set and the optimum objective value, then by the local Lipschitz continuity of the objective, x^k must approach the intersection of the feasible set and the set of points with the optimum objective. This is precisely the optimal set.

■

Theorem 3.7.3 (Convergence of ADMM) [96, Theorem 1, p. 2] *Suppose that f*

and g are real-valued convex functions on a polyhedral domain $\mathcal{X}$, and that the problem

$$\begin{array}{ll} \textit{minimize} & f(x) + g(x) \\ \textit{subject to} & x \in \mathcal{X} \end{array} \tag{3.14}$$

has a solution with optimal objective $p^\star$. Then the iterates (x^k, z^k, y^k) of ADMM (3.7.5) *have the following properties:*

1. $f(x^k) + g(z^k) \to p^\star$
2. $\{(x^k, z^k)\}$ *has a single limit point $(x^\star, x^\star)$; furthermore, $x^\star$ solves* (3.14)

This theorem does not immediately pertain, because the range of g includes $+\infty$. The effective domain of g (the region where it is finite) is $\mathcal{S}$, which may not be polyhedral; so we cannot simply remove the complement of the effective domain from $\mathcal{X}$.

In their text, Mota et al [96] claim that Theorem 3.7.3 "can be straightforwardly adapted to the f and g assumptions made in" Boyd et al [94]. Then there is no problem; we have already satisfied the assumptions of Boyd et al.

Beyond Theorem 3.7.2, Theorem 3.7.3 establishes that the iterates do not "wander around" in the optimal set, but have a limit point. Furthermore, since the limit point is unique, distinct initial conditions will not lead to distinct optima.

It would appear that we cannot apply the convergence theorem of Nishihara et al [97], because it requires one of f or g to be strongly convex and differentiable with Lipschitz gradient.

Theorem 3.7.4 (Convergence of ADMM) [98, Ch. 3, Proposition 4.2, p. 256] *Suppose that f and g are real-valued convex functions on a nonempty polyhedral set $\mathcal{X}$, and that the optimal set of* (3.14) *is nonempty. Then the iterates (x^k, z^k, y^k) of ADMM* (3.7.5) *have the following properties:*

1. *The sequence $\{x^k, z^k, y^k\}$ is bounded.*
2. *Every limit point of x^k is an optimal solution to* (3.14)

The same considerations apply here: because our g is extended-real-valued, the theorem does not immediately apply; because $\mathcal{X}$ is required to be polyhedral, we cannot simply intersect it with the effective domain of g. Here as well, the authors claim in the text that "the subsequent ... convergence result can be formulated for more general convex constraint sets".

Beyond the other two theorems, Theorem 3.7.4 establishes that the sequence of iterates does not become arbitrarily large before it converges.

Theorem 3.7.5 (Convergence of ADMM) [99] *The sequence of iterates converges with worst-case ergodic rate $\mathcal{O}(1/n)$ in the number of iterations.*

3.7.10 Accelerating the Haircut-Step

The haircut function of Lemma 3.7.9, evaluated at a point v, can be separated into three phases. In the first phase, we evaluate $|v_i|$. In the second phase, we compute the ∞-norm μ_+ of the output x such that $\|x - v\|_1 = \min(\lambda, \|v\|_1)$. In the third phase, we use μ_+ and $|v_i|$ to rescale v_i into x_i. The complexity of these steps is detailed in Table 3.7.

Algorithm 3.7.7 (Haircut)

$$
\begin{aligned}
&\text{def } \mu(a, \lambda): \\
&\quad y \quad := \text{concatenate}([0], [\text{sort}(a)]) \\
&\quad u \quad := \text{sum}(a) - \lambda \\
&\quad v_i \quad := \text{cumulativeSum}(y)_i + (LN - i)y_i \qquad \forall 0 \le i < LN \\
&\quad \text{return } \text{interpolate}(u, v, y)
\end{aligned}
$$

def haircut(v, λ) :

Phase I:

$a_i \quad := |v_i| \qquad \forall 0 \leq i < LN$

Phase II:

$\mu_+ \quad := \mu(a, \lambda)$

Phase III:

$x_i \quad := (a_i \neq 0)\,?\,(v_i \min(1, \mu_+/a_i)) \;:\; 0 \qquad \forall 0 \leq i < LN$

return x

Phase	$t_\perp$ (Sequential)	$t_\parallel$ (Parallel)	$s_\circ$ (Circuit size)	$t_\circ$ (Depth)
I	$\mathcal{O}(N)$	$\mathcal{O}(1)$	$3N$	3
II	$\mathcal{O}(N \log N)$	$\mathcal{O}(\log^2 N)$	$\frac{N}{4}\log^2 N + \frac{3N}{4}\log N + 8N + 3$	$\frac{1}{2}\log^2 N + \frac{5}{2}\log N + 5$
III	$\mathcal{O}(N)$	$\mathcal{O}(1)$	$4N$	3
Heuristics	$\mathcal{O}(N)$	$\mathcal{O}(1)$	N	$\log N + 1$

Table 3.7: Phases of the evaluation of the proximal mapping of $\lambda\|\cdot\|_\infty$ with complexity estimates. N is the oversampled OFDM order. See Definition 3.7.1 for detailed interpretation.

The time complexity of the second phase is dominated by sorting, followed by binary search and interpolation. For this calculation, sorting is accomplished using the bitonic sorting network of Batcher [100].

The complexity of the second phase dominates every other phase in every category. In the sequential case, the time complexity of the second phase is comparable to an FFT; but in the parallel case, it is greater by a logarithmic factor. This raises a question: is it necessary to treat λ as the independent variable and μ_+ as the dependent variable? Evaluating the relationship in the opposite direction is far less expensive: given μ_+, run Phases I and III to determine x; then $\lambda = \|v - x\|_1$.

We therefore consider two heuristics for μ_+ which bypass the computation of $\mu_+(\lambda)$.

Both are equivalent to evaluating Phase II with some value of λ dependent on v.

$$\mu_\alpha^{(\max)}(v) = \alpha\|v\|_\infty$$
$$\mu_\alpha^{(\text{mean})}(v) = (1/N)\alpha\|v\|_1$$

We will refer to these as the discounted-max and the inflated-mean, respectively. The cost of either heuristic is shown in the final row of Table 3.7.

From now on, we may refer to $\lambda = N\|v - x\|_1$ as the implicit primal step size, re-introducing the factor of N arising from the IDFT.

There are two potential problems with the implicit primal step size approach. The first is that our convergence results have assumed constant step sizes from iteration to iteration. The second is that λ is needed to compute the argument v to the haircut function, before μ can be chosen.

Problem: unequal step sizes

The first potential problem is that the proofs of convergence have assumed both equal primal and dual step sizes and constant step sizes from iteration to iteration.

Problem: λ needed before μ chosen

The haircut argument $v = \text{IDFT}(z^k - \lambda y^k)$ needs to be evaluated before Phase I can be executed (Algorithm 3.7.6). This problem may be bypassed if we relax the requirement of inverse primal and dual step sizes, so long as we maintain $0 < \rho \leq \frac{1+\sqrt{5}}{2\lambda}$ [95].

Is it straightforward to find an acceptable ρ before executing the haircut function? That is, can we obtain a positive lower bound on this quantity φ/λ (where φ is the golden ratio)? The following lemmas will prove useful.

Lemma 3.7.12 *For any μ,*

$$\|v - x\|_1 = \|v\|_1 - \|x\|_1 \leq \|v\|_1$$

Proof. By the definition of Algorithm 3.7.7, $\arg x_i = \arg v_i$ and $|x_i| = \min(|v_i|, \mu) \leq |v_i|$. The lemma follows.

Lemma 3.7.13 *For v generated by Algorithm 3.7.6,*

$$\|v\|_2 = \sqrt{N}\|z^k - \rho^{-1}y^k\|_2$$

Proof. Follows from Parseval's theorem for the IDFT and Algorithm 3.7.6.

Lemma 3.7.14 *Suppose that for some μ heuristic and some $p \geq 1$, $\forall v. \|v - x\|_1 \leq \beta\|v\|_p$. Then a sufficient condition for convergence is that the dual step size ρ satisfies*

$$\begin{cases} |2a\rho + b| & \leq \sqrt{b^2 - 4ac} \qquad a \neq 0 \\ c & \leq 0 \qquad\qquad\quad a = b = 0 \end{cases}$$

where

$$\begin{aligned} a &= \|z^k\|_2^2 \\ b &= -2z^k \cdot y^k \\ c &= \|y^k\|_2^2 - \frac{\varphi^2}{N^{2\xi}\beta^2} \\ b^2 &\geq 4ac \\ (c < 0) &\vee (b < 0) \\ \xi &= \max(1/p + 1, 3/2) \end{aligned}$$

Proof. A dual step size is convergent [95] if $0 < \rho \leq \frac{\varphi}{N\|v-x\|_1}$. Then the following conditions are sufficient:

$$
\begin{aligned}
& & 0 < \rho &\leq \frac{\varphi}{N\beta\|v\|_p} \\
\Leftarrow & & 0 < \rho &\leq \frac{\varphi}{N^{\max(1/p+1/2,1)}\beta\|v\|_2} \\
& & &= \frac{\varphi}{N^{\xi}\beta\|z^k - \rho^{-1}y^k\|_2} \\
\Leftrightarrow & & 0 < \rho, \quad 0 &\geq \|\rho z^k - y^k\|_2^2 - \frac{\varphi^2}{N^{2\xi}\beta^2} \\
& & &= a\rho^2 + b\rho + c
\end{aligned}
$$

These inequalities are feasible iff $b^2 - 4ac \geq 0$ and $\frac{-b+\sqrt{b^2-4ac}}{2a} > 0$.

Suppose that $b^2 \geq 4ac$. Then the second inequality reads $\sqrt{b^2 - 4ac} > b$. If $b < 0$, the second inequality is satisfied. Otherwise, we square and simplify to get $c < 0$.

The strict inequality $a > 0$ holds provided that the z^k iterate is nonzero. This is guaranteed for $k = 0$ if we initialize $z^0 := d$ and d is nonzero; it is guaranteed for all $k > 0$ provided that $0 \notin \mathcal{S}$ is explicitly infeasible. This is true if, for instance, the data vector is modulated from a QAM constellation and the components of the EVM constraint vector c are small enough to prevent bit-errors at the transmitter. If $a > 0$, then the bound of the lemma follows using the quadratic formula.

In the alternative case where $z^k = 0$, then $a = b = 0$, and the condition of the lemma (which is now independent of ρ) follows.

■

The case $b < 0$, $b^2 \geq 4ac$ corresponds to z^k and y^k nearly parallel: the angle between them satisfies

$$\theta \leq \arcsin(\varphi/(N^{\xi}\beta\|y^k\|_2))$$

The case $c < 0$, $b^2 \geq 4ac$ corresponds to y^k not too large:

$$\|y^k\|_2 < \frac{\varphi}{N^{\xi}\beta}$$

Lemma 3.7.14 with $\beta = 1$, $p = 1$ gives a (possibly empty) range for ρ in which convergence is not conditional on μ.

To apply Lemma 3.7.14 to the discounted-max heuristic, we can use the following lemma:

Lemma 3.7.15 *For the discounted-max heuristic with parameter α,*

$$\|v - x\|_1 \leq N(1-\alpha)\|v\|_\infty$$

Proof. Substituting $\mu_\alpha^{(max)}(v)$ into the definition of Algorithm 3.7.7,

$$\|v - x\|_1 = \sum_i (|v_i| - \alpha\|v\|_\infty)_+$$

Among all vectors v of equal $\|v\|_\infty$, this quantity is greatest when $|v_i| = \|v\|_\infty$ for all i; then

$$\|v - x\|_1 \leq N(1-\alpha)\|v\|_\infty$$

■

To apply Lemma 3.7.14 to the inflated-mean heuristic, we can use the following lemma:

Lemma 3.7.16 *For the inflated-mean heuristic with parameter α,*

$$\|v - x\|_1 \leq (1 - (\alpha/N))\|v\|_1$$

Proof. Substituting $\mu_\alpha^{(mean)}(v)$ into the definition of Algorithm 3.7.7,

$$\|v - x\|_1 = \sum_i (|v_i| - (\alpha/N)\|v\|_1)_+$$

Among all vectors v of equal $\|v\|_1$, this quantity is greatest when v has only a single nonzero component; then

$$\|v - x\|_1 \leq (1 - (\alpha/N))\|v\|_1$$

■

3.7.11 Complexity of the Project-Step

For the affine tone reservation set, evaluating project(v) is trivial; it simply substitutes the values of d_i for the components v_i for all $i \in \mathcal{D}$ (Lemma 3.7.6). For the EVM constraint set, project(v) is given by Lemma 3.7.7. Complexities are given in Table 3.8.

Set	$t_\perp$ (Sequential)	$t_\parallel$ (Parallel)	$s_\circ$ (Circuit size)	$t_\circ$ (Depth)
TR	$\mathcal{O}(N)$	$\mathcal{O}(1)$	0	0
EVM	$\mathcal{O}(N)$	$\mathcal{O}(1)$	$8N$	7

Table 3.8: Complexity of the project(v) proximal mapping. See Definition 3.7.1 for detailed interpretation.

3.7.12 Step-size Schedule Optimization

(For conciseness, we will refer to the step-size parameter as α in what follows, regardless of whether it refers to the non-heuristic step size λ or the heuristic parameter α).

Rather than using the same step-size for each iteration, we might prefer to choose a schedule of step-sizes by an optimization procedure. We begin by fixing an iteration count; in our case, five iterations. We then search for the schedule $\{\alpha_1, \alpha_2, \ldots\}$ that achieves, for instance, the best 99.9^{th}-percentile PAR.

The search begins by testing different constant step sizes $\alpha_i = \alpha$ on a large batch of randomly-generated 802.11n HT20-style OFDM symbols (described below), normalized to unit power per sub-carrier. Using $\{\alpha, \alpha, \ldots\}$ as a starting point, we then perform successive relaxation, changing one iteration at a time, sweeping either back-to-front, front-to-back, or back-and-forth.

For our case, we reach the following optimized schedules:

1. Full haircut function: $\{2.18, 1.85, 1.30, 1.69, 1.52\}$.

2. Discounted-max heuristic: $\{0.30, 0.28, 0.36, 0.68, 0.50\}$.

3.7.13 Evaluation

We ran our algorithm on 80,000 randomly-generated OFDM symbols. Each symbol has the sub-carrier layout of 802.11n HT20, with unit-power "uniform QAM" modulation on each sub-carrier. By uniform QAM, we mean that the real and imaginary parts are i.i.d. uniform over a range $\pm\sqrt{3/2}$. It is well-known that PAR of OFDM waveforms is generally an increasing function with constellation order; these results may be considered to be for the asymptotic case of high constellation order.

Our OFDM symbols are modulated using an FFT size of 256, representing $4\times$ digital over-sampling. Transceivers using digital over-sampling do not have as stringent requirements on analog anti-aliasing filter roll-off as those which do not use digital oversampling. All of the non-data, non-pilot subcarriers, other than the DC subcarrier, are reserved tones. The step-size schedules are as given in the previous section.

Results are shown in Figures 3-15, 3-16, and 3-17. At the 10^{-4} level, our algorithm reduces tail PAR by 3.5 to 3.7 dB, out of a maximum possible improvement of 10.6 dB. At the same level, it reduces tail peak power by 3.0 to 3.2 dB, out of a maximum possible improvement of 4.4 dB.

Why the discrepancy? The problem is that the optimization, as we have stated it, is over a single OFDM symbol at a time. It does us little good to reach a PAR of unity for each symbol in isolation, if the power of each symbol in turn is quite different. While the optimum succeeds in converting each OFDM symbol into a nearly-constant-envelope complex waveform, it is really only necessary to optimize for the worst-case symbol in the entire packet. For that reason, the distribution of peak power may be more apropos than the distribution of PAR for designing an OFDM system.

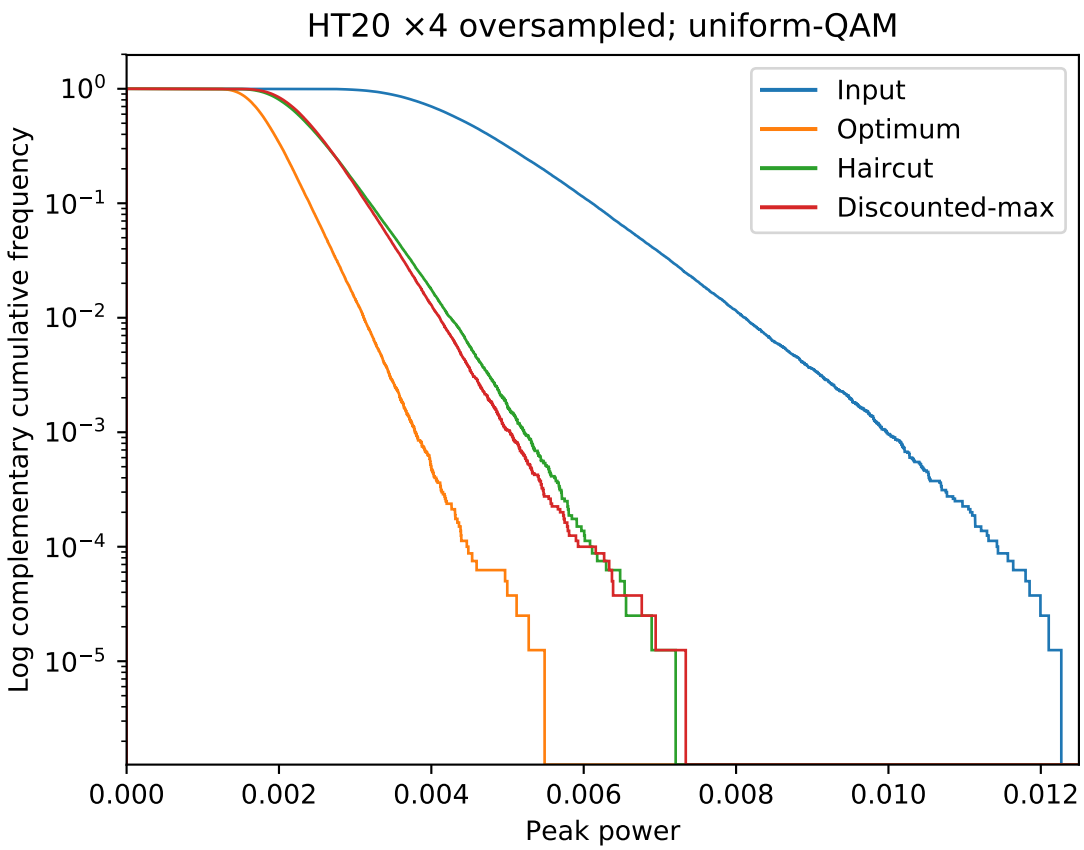


Figure 3-15: Tail of the power distribution before and after optimizing. Optimum computed by unlimited-iteration interior point optimization using Ipopt.

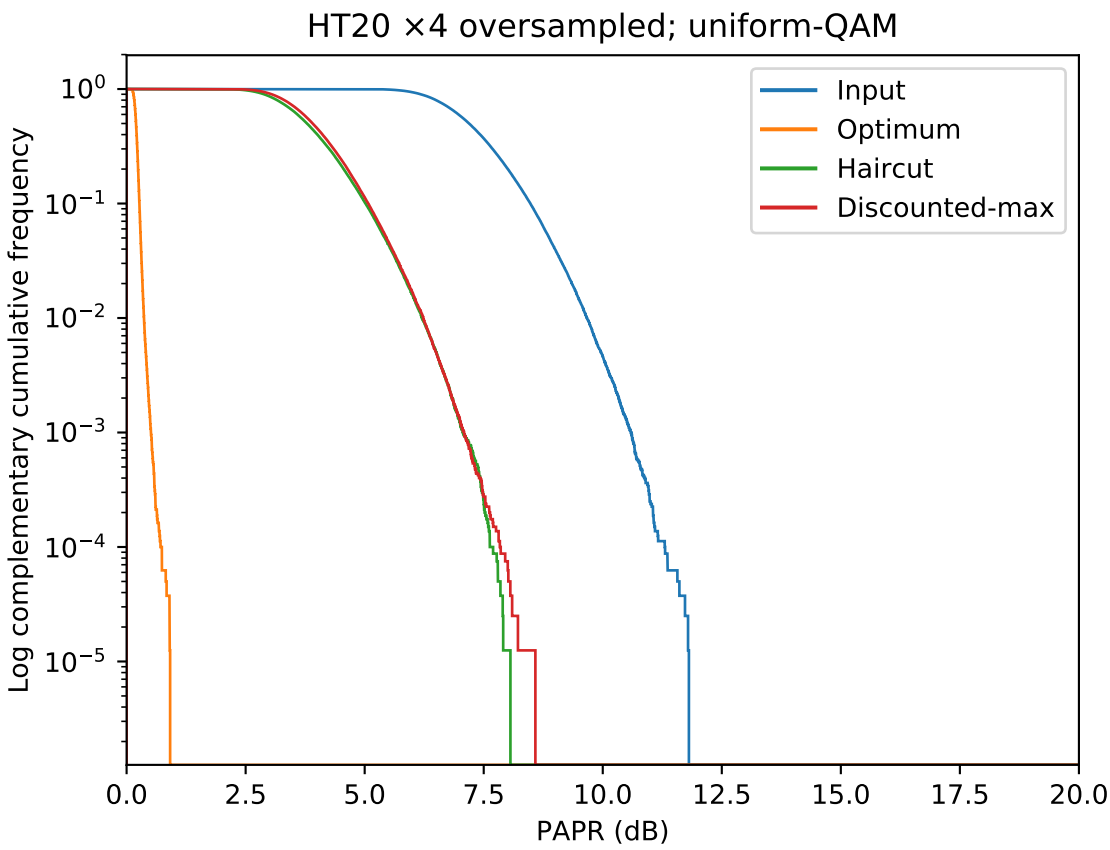


Figure 3-16: Tail of the PAR distribution before and after optimizing.

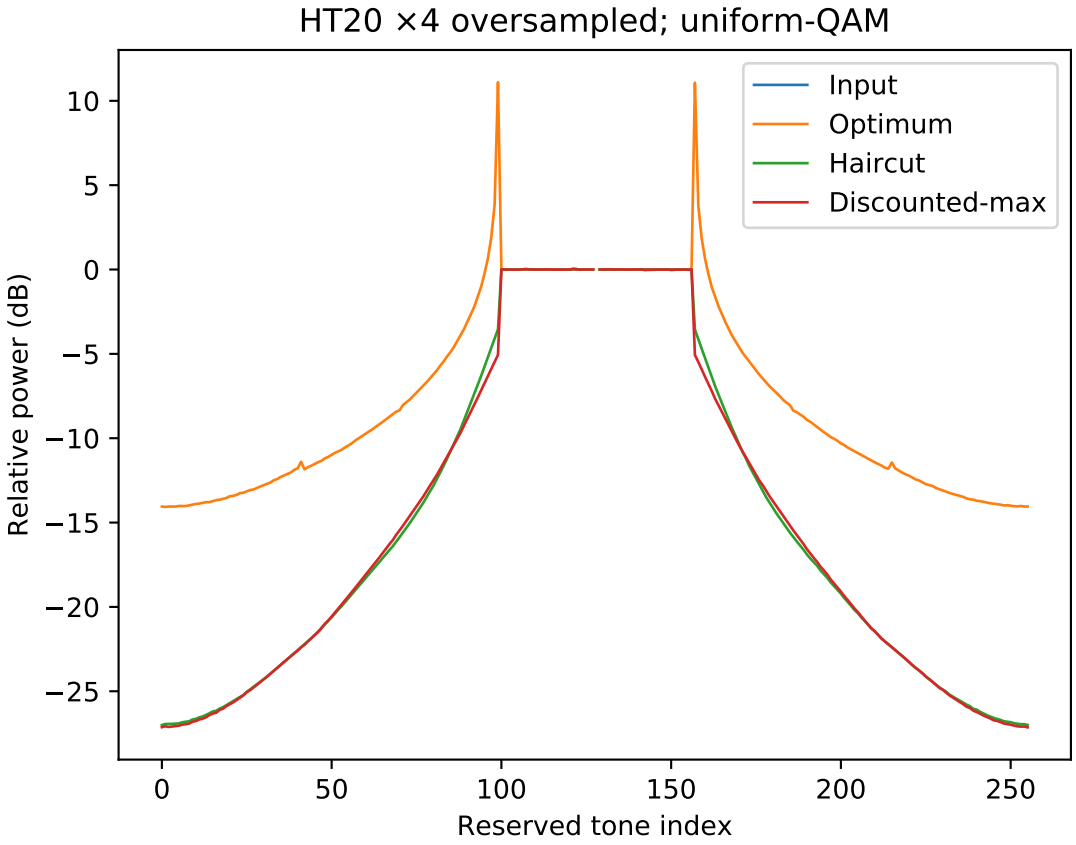


Figure 3-17: Power spectrum before and after optimizing. Note how the true optimum is very far in Euclidean distance from the original symbol.

Chapter 4

Challenges to PHY Anonymity for Wi-Fi

Prior work has shown that unique non-idealities in each wireless device give rise to radiometric fingerprints in the transmitted signal. However, even a device with no such uniqueness would betray channel state information (CSI) at the wireless physical layer. A passive listener can use this CSI to de-anonymize the sender.

To bound the information gained by a powerful de-anonymizer, we develop an array-processing technique, **REST**, that extracts slowly-varying channel fingerprints from a Wi-Fi signal. We argue that REST is universal: for every other channel fingerprinting or angle-of-arrival scheme, either that scheme is too paranoid, leaving information unused; too trusting, relying on degrees of freedom under the transmitter's control; or else it is equivalent to REST.

In response to REST, on the anonymizer's side, we propose a countermeasure to radio-layer tracking based on switchable antennas: a kind of "exponential ventriloquism".

In response to exponential ventriloquism, on the de-anonymizer's side, we derive a theoretical upper bound on the effectiveness of a wide class of antenna- or MIMO-based techniques intended to conceal channel fingerprints.

4.1 Overview

Today's wireless networks do not attempt to anonymize users, so that profiling users for de-anonymization is straightforward at every layer of the protocol stack. We ask, however, who would be the victor in an earnest technological contest: a sender desiring anonymity, or a listener wishing to de-anonymize them? Whether wireless anonymity is socially beneficial (for privacy against surveillance) or detrimental (e.g., for law enforcement), we will one day face the "end-game" of this contest. This chapter is part of an effort to design ever-more-powerful anonymizers and de-anonymizers. This chapter focuses on the physical layer (PHY).

De-anonymization attacks at the physical layer generally fall into two categories: radiometric signature techniques [101, 102, 103], which examine signals for imperfections characteristic of an individual transmitter chip, and channel-state fingerprinting [104] and localization [105, 106] techniques, which examine the phase and amplitude relationships between the sending and listening radios. We focus on the latter category, positing that a careful sender could calibrate out any unique imperfections in their hardware.

Location privacy is one key to PHY anonymity. It is well known that each transmission of radio energy creates an opportunity for a listener to localize the sender in physical space using angle-of-arrival techniques[1]. Once a packet can be traced to its origin in physical space, then (even supposing its link-layer contents are totally obfuscated) the packet becomes part of a trail of "breadcrumbs" revealing the movements of the sender. If the sender uses a point-of-sale, walks past a camera, or drives past a license plate reader, the entire packet trace is de-anonymized. But if each transmission contributes to de-anonymization, can we place any meaningful limits on the listener's capabilities?

Our first contribution is a geometric framework, **REST**, for modeling powerful physical-layer de-anonymizing listeners. Because wireless channel state information is affected by noise and precoding, it is necessary for the de-anonymizer to separate out

[1] "a person with the right equipment [can] always trace the bark back to the dog" [107].

degrees-of-freedom (DoFs) that are controlled by the sender from those that are not, and to do so in a way that is robust to noise. The DoFs that are retained form the channel-state fingerprint: a concise (though implicit) representation of the location DoFs measurable by a given listening setup.

Unlike cryptographic hashes, which map related inputs to seemingly unrelated outputs, channel-state fingerprints support computing the dissimilarity between fingerprints (they form a metric space). Compared to MUSIC angle-of-arrival pseudospectra [108] or SecureArray signatures [106], REST fingerprints are non-parametric with respect to the geometry of the antenna array, simplifying deployment. They do not require evaluation at a large number of candidate angles, polynomial root-finding, or randomization techniques to perfectly encode the location degrees-of-freedom present in the received signal at the listener. The REST fingerprinting map and dissimilarity metric are deterministic, computationally efficient, and numerically stable. REST's separation of DoFs between those controlled by the sender and those characteristic of the channel is exact at high SNR. By contrast, other fingerprinting algorithms discard either too little information [105], allowing the sender to influence the results, or too much information [108, 104], diminishing the listener's de-anonymizing power. Using experimental data from a live Wi-Fi network, we demonstrate that REST fingerprints are stable over time and show strong clustering.

Our second contribution is an argument, based on geometric considerations, for a countermeasure that the sender can use to increase the adversary's difficulty in obtaining a consistent fingerprint. We call this scheme "exponential ventriloquism": the sender uses a special switchable antenna to "throw its voice", inflating the number of possible fingerprints by a large factor at the cost of lost opportunities for beamforming.

Finally, we derive from electromagnetic theory a pair of upper bounds on the effectiveness of a broad class of antenna- or MIMO-based countermeasures to physical-layer de-anonymization.

4.2 PHY De-anonymizer

4.2.1 Setting

We assume an adversarial setting where a sender does not wish to be de-anonymized, and has carefully calibrated their radio to lack unique or fingerprintable start-up transients, frequency drift effects, and so on. We further assume a passive listening adversary that does not rely on radiometric signatures characteristic of imperfections in the transmitter. That is, in our contest between the anonymizing sender and the de-anonymizing listener, we suppose that the game has played out to a point somewhat ahead of the present state-of-the-art, where radiometric signatures can be reliably defeated through careful transmitter calibration, and transmitters produce signals indistinguishable from the ideal signal prescribed by the wireless standard.

In this setting, we wish to understand how the most powerful passive physical-layer de-anonymization system could work.

4.2.2 Prior Fingerprinting Schemes

The de-anonymizing listener would like to exploit all available information. Some previous channel fingerprinting work [104] used only magnitude and not phase information. Other work [105] used phase information in the form of covariances between received antenna array channels, but permitted the sender to influence the results of fingerprinting through the magnitudes of various MIMO channels. Our de-anonymizing listener will not be satisfied so long as the anonymizing sender can take countermeasures to obscure their fingerprints: for instance, by allocating power among MIMO spatial streams in a random and time-varying way at the cost of some reduction in throughput to the intended recipient.

4.2.3 MIMO Notation

The MIMO channel with linear precoding[2] can be expressed as a map from m-dimensional transmitted vectors $\mathbf{x}$ to n-dimensional received vectors $\mathbf{y}$:

$$\mathbf{y} = \mathbf{HWx} + \mathbf{n},$$

where $\mathbf{n}$ is additive white Gaussian noise, $\mathbf{H}$ is the channel matrix, and $\mathbf{W}$ is an arbitrary precoding matrix chosen by the sender. (Note that $\mathbf{H}$ will be different for the intended recipient and the de-anonymizing listener, but this does not matter in what follows.) At the beginning of each wireless transmission is a sequence of m pre-arranged vectors $\mathbf{x}$ that allow listeners to measure $\mathbf{HW}$ from a sequence of m or more separate observations $\mathbf{y}$. Combining these sequences of m column vectors into matrices, we can write

$$\mathbf{Y} = \mathbf{HWX} + \mathbf{N},$$

where now $\mathbf{X}$ is an $m \times m$ matrix representing the pre-arranged sequence of vectors, and $\mathbf{Y}$ is a $n \times m$ matrix of observations. Without loss of generality, we can take $\mathbf{X}$ to be the identity matrix, so that the listener's observation equation is

$$\widehat{\mathbf{HW}} = \mathbf{Y} = \mathbf{HW} + \mathbf{N},$$

where the $N_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, \sigma^2)$ are complex Gaussians. That is, the first m observed vectors form a noisy estimate of $\mathbf{HW}$. $\mathbf{HW}$ will, with probability 1, have full column rank, provided the sender does not sabotage itself with a singular choice for $\mathbf{W}$. Now any listener, intended or otherwise, can use the pseudo-inverse of $\widehat{\mathbf{HW}}$ to estimate subsequent data signals $\mathbf{x}$ from observations $\mathbf{y}$.

[2]We consider linear precoding because it is transparent to the intended receiver in Wi-Fi or LTE, whereas nonlinear precoding would not be.

4.2.4 Counting Degrees of Freedom

Before we attempt to identify a subset of the degrees of freedom present in $\widehat{\mathbf{HW}}$ that we will use for our fingerprinting scheme, a helpful check is to determine how many of them we should expect to find.

Because $\mathbf{H}$ is a complex $n \times m$ matrix, it can be thought of as a point in a space with $2nm$ real DoFs. Similarly, $\mathbf{W}$ contains $2m^2$ DoFs. Because $\mathbf{H}$ has full column rank m, the map

$$\begin{aligned} f_H : \mathbb{C}^{m\times m} &\to \mathbb{C}^{n\times m} \\ \mathbf{W} &\mapsto \mathbf{HW} \end{aligned}$$

has full column rank m^2. So by controlling $\mathbf{W}$, the sender can cause the observation $\widehat{\mathbf{HW}}$ to vary within a $2m^2$ real dimensional subspace of the full $2nm$ real dimensional space of $n \times m$ complex matrices. We conclude, then, that there are exactly $2nm - 2m^2 = 2m(n-m)$ real DoFs in $\widehat{\mathbf{HW}}$ which are not controlled by the sender, and hence are suitable for fingerprinting in our adversarial setting.

The PHY de-anonymizer must therefore have more antennas than the receiver. Otherwise, the subspace of received signals is of full rank, the number of DoFs suitable for fingerprinting is zero, and there is nothing to fingerprint. The sender is in full control of the received signal across all receive antennas, and can spoof any measurement it likes.

4.2.5 Selecting Degrees of Freedom

We want to identify which of the $2nm$ DoFs in $\widehat{\mathbf{HW}}$ are not controlled by the sender via $\mathbf{W}$. For now, we will ignore measurement noise, so that $\widehat{\mathbf{HW}} = \mathbf{HW}$.

We will find that the only good DoFs in $\mathbf{HW}$ correspond to the space spanned by the columns of $\mathbf{H}$, a m-dimensional subspace in n dimensions. The number of such DoFs turns out to be exactly $2m(n-m)$, as expected from the previous section. We will proceed with some algebraic manipulations to show this.

From the singular value decomposition, we can express the action of any matrix, say $\mathbf{H}$, by a rotation $\mathbf{V}^*$ into the *basis of right singular vectors*, followed by a scaling $\mathbf{\Sigma}$ of each axis by a nonnegative *singular value*, and finally a rotation $\mathbf{U}$ out from the *basis of left singular vectors*. That is, we can write

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*, \text{ where } \mathbf{U}^*\mathbf{U} = \mathbf{I},\ \mathbf{V}^*\mathbf{V} = \mathbf{V}\mathbf{V}^* = \mathbf{I},$$

where $\mathbf{U} \in \mathbb{C}^{n\times m}$, $\mathbf{\Sigma} \in \mathbb{C}^{m\times m}$, and $\mathbf{V}^* \in \mathbb{C}^{m\times m}$ Without loss of generality, we can re-identify the DoFs controlled by the sender as $\mathbf{W}' = \mathbf{u}\mathbf{s}\mathbf{v}^*$, with

$$\mathbf{W} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{W}' = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{u}\mathbf{s}\mathbf{v}^*,$$

with $\mathbf{u}, \mathbf{s}, \mathbf{v}^* \in \mathbb{C}^{m\times m}$, so that

$$\mathbf{H}\mathbf{W} = (\mathbf{U}\mathbf{u})\mathbf{s}\mathbf{v}^*$$

is a singular value decomposition of $\mathbf{HW}$. Now $\mathbf{u}$, $\mathbf{s}$, and $\mathbf{v}^*$ are the pieces of the singular value decomposition of $\mathbf{W}'$, and are controlled by the sender[3].

Two of these three factors are unwanted. So the de-anonymizing listener computes the singular value decomposition of $\mathbf{HW}$, keeps $(\mathbf{U}\mathbf{u})$, the basis of left singular vectors, and discards $\mathbf{s}$, the singular values, and $\mathbf{v}^*$, the basis of right singular vectors. This eliminates half of the $2m^2$ degrees of freedom controlled by the anonymizing sender—where the count takes some care for the non-uniqueness of the singular value decomposition.

The $n \times m$ matrix $\mathbf{U}$ has m orthogonal columns. It belongs to a homogeneous space, the quotient of unitary groups $\mathrm{U}(n)/\mathrm{U}(n-m)$. The $m \times m$ matrix $\mathbf{u}$ belongs to $\mathrm{U}(m)$. Quotienting out the degrees of freedom of $\mathbf{u}$ from $(\mathbf{U}\mathbf{u})$ will take us to a new space: the space of m-dimensional subspaces of $\mathbb{C}^n$, or the complex Grassmannian manifold $\mathrm{Gr}(m, \mathbb{C}^n) \sim \mathrm{U}(n)/(\mathrm{U}(n-m) \times \mathrm{U}(m))$. For our purposes, we will treat this space as the space of orthogonal projection matrices $\mathbf{P} = \mathbf{P}^* = \mathbf{P}^2$. The way we get

[3]Note that $\mathbf{s}$, the scaling piece of both $\mathbf{W}'$ and $\mathbf{HW}$, is controlled by the sender. Intuitively, measuring the power (amplitude) of a received signal does not help us fingerprint the sender when the sender is free to adjust its transmit power.

there is simply $(\mathbf{Uu}) \mapsto \mathbf{P} = (\mathbf{Uu})(\mathbf{Uu})^* = \mathbf{UU}^*$. Note that none of the pieces of $\mathbf{W}'$ survive this mapping; the sender no longer controls any DoFs.

Did the listener lose any good DoFs by this mapping? As expected, the real dimension of the resulting manifold is $2m(n-m)$. The mapping is surjective and smooth. If we choose an arbitrary pre-image $\mathbf{U}'$ for some output $\mathbf{P}$—for instance, the basis of left singular vectors of $\mathbf{P}$—we can get to any other pre-image, including $(\mathbf{Uu})$, via right-multiplication by some $m \times m$ unitary matrix: precisely the degrees of freedom the sender controlled and the listener wished to remove. So the only lost DoFs are bad DoFs.

A suitable and efficient choice of dissimilarity metric after this mapping is the square root of the sum of squared absolute differences between projection matrices: $d(P_1, P_2)^2 = \|P_1 - P_2\|_F^2$. This corresponds to the chordal metric on the Grassmannian manifold.

To summarize in pseudocode, the "good" DoFs are extracted by

function FingerprintNoNoise($\mathbf{HW}$)
  $\mathbf{Uu}, \mathbf{\Sigma}, \mathbf{V}^* \leftarrow$ svd($\mathbf{HW}$)
  $\mathbf{P} \leftarrow (\mathbf{Uu})(\mathbf{Uu}^*)$
  return $\mathbf{P}$
end function
function FingerprintDistance($\mathbf{P}_1$, $\mathbf{P}_2$)
  return $\sqrt{\sum_{ij} |(P_1)_{ij} - (P_2)_{ij}|^2}$
end function

We can equivalently write $\mathbf{P} = (\mathbf{HW})((\mathbf{HW})^*(\mathbf{HW}))^{-1}(\mathbf{HW})^*$, using the matrix inverse, or $\mathbf{P} = (\mathbf{HW})(\mathbf{HW})^+$ using the Moore-Penrose pseudoinverse.

We are not done yet; the algorithm as stated does not consider the effect of noise. Some prior work [108, 106] considers the effects of noise on eigenvalues, but discards relevant information about uncertainty over eigenvectors, and thus overstates its confidence in the localization of the transmitter.

4.2.6 Robust Eigenspace Signal Tracking (REST)

In the absence of noise, we would compute a fingerprint $\mathbf{P} = \mathbf{U}\mathbf{U}^*$ from the basis of left singular vectors of $\widehat{\mathbf{HW}}$, and we would measure distances between fingerprints using squared absolute differences between matrices. In the presence of noise, we can instead ask for the minimum mean-squared error estimator (MMSEE) of $\mathbf{P}$, given the observed noisy estimate $\widehat{\mathbf{HW}}$. In what follows, we will omit $\mathbf{W}$.

The MMSEE $\widehat{\mathbf{P}}$ of the true fingerprint $\mathbf{P}$ given $\mathbf{Y} = \mathbf{H} + \mathbf{N}$ is

$$\begin{aligned}
\widehat{\mathbf{P}} &= \arg\min_{\mathbf{P}} \mathbb{E}_{\mathbf{N}}[\|\mathbf{P} - \mathbf{H}\mathbf{H}^+\|_F^2 | \mathbf{Y}] \\
&= \mathbb{E}_{\mathbf{N}}[\mathbf{H}\mathbf{H}^+ | \mathbf{Y}] \\
&= \mathbb{E}_{\mathbf{N}}[(\mathbf{Y} - \mathbf{N})((\mathbf{Y} - \mathbf{N})^*(\mathbf{Y} - \mathbf{N}))^{-1}(\mathbf{Y} - \mathbf{N})^*],
\end{aligned}$$

where the expectation is with respect to the joint distribution over the elements of the Gaussian noise matrix $\mathbf{N}$. This expectation over a matrix inversion is problematic, so we convert it into a power series by a heuristic argument. It is known that if a matrix $\mathbf{A}$ satisfies $\lim_{n\to\infty}(\mathbf{I} - \mathbf{X}^{-1}\mathbf{A})^n = 0$, then $\mathbf{A}$ is nonsingular and its inverse is

$$\mathbf{A}^{-1} = \sum_{n=0}^{\infty} (\mathbf{X}^{-1}(\mathbf{X} - \mathbf{A}))^n \mathbf{X}^{-1}$$

Letting $\mathbf{X} = \mathbf{Y}^*\mathbf{Y}$ and $\mathbf{A} = (\mathbf{Y} - \mathbf{N})^*(\mathbf{Y} - \mathbf{N})$, we have

$$\begin{aligned}
\lim_{n\to\infty}(\mathbf{I} - \mathbf{X}^{-1}\mathbf{A})^n &= \lim_{n\to\infty}((\mathbf{Y}^*\mathbf{Y})^{-1}(\mathbf{Y}^*\mathbf{N} + \mathbf{N}^*\mathbf{Y} - \mathbf{N}^*\mathbf{N}))^n \\
&\approx \lim_{n\to\infty}(\text{noise/signal})^n, \text{ in expectation} \\
&= 0 \text{ for SNR} > 0 \text{ dB}
\end{aligned}$$

If we permit this slight abuse, we can define $\mathbf{C} = (\mathbf{Y}^*\mathbf{Y})^{-1}$, and

$$\widehat{\mathbf{P}} = \sum_{n=0}^{\infty} \mathbb{E}_{\mathbf{N}}[(\mathbf{Y} - \mathbf{N})(\mathbf{C}(\mathbf{Y}^*\mathbf{N} + \mathbf{N}^*\mathbf{Y} - \mathbf{N}^*\mathbf{N}))^n \mathbf{C}(\mathbf{Y} - \mathbf{N})^*]$$

The terms of this summation are products of constant and complex Gaussian matrices,

and each can be evaluated in closed form by a computer algebra system using Isserlis' theorem (§4.3). (With no further difficulty, a covariance matrix can be added for the elements of $\mathbf{N}$). This leads to

$$
\begin{aligned}
\widehat{\mathbf{P}} &= \mathbf{Y}(\mathbf{M}_1 + \sigma^6(n-m)\mathbf{C}\mathbf{M}_2)\mathbf{Y}^* + \\
&\qquad \mathbf{P}_1 \operatorname{Tr}\{\sigma^2\mathbf{M}_1 - \sigma^6\mathbf{M}_2\} + \mathcal{O}(\sigma^{10}), \text{ where} \\
\mathbf{M}_1 &\triangleq (\mathbf{Y}^*\mathbf{Y} + \sigma^2(n-m)\mathbf{I})^{-1} \\
\mathbf{M}_2 &\triangleq (n-m)\mathbf{C}^2 \operatorname{Tr}\{\mathbf{M}_1\} + \sigma^2(n-m)\mathbf{C}^2(\operatorname{Tr}\{\mathbf{C}\}^2 - \\
&\qquad 2(n-m)\operatorname{Tr}\{\mathbf{C}\}\mathbf{C} + \mathbf{C}^2), \text{ and} \\
\mathbf{P}_1 &\triangleq \mathbf{I} - \mathbf{Y}\mathbf{C}\mathbf{Y}^*. \text{ Alternatively,} \\
\widehat{\mathbf{P}} &= \mathbf{Y}\frac{1}{\mathbf{Y}^*\mathbf{Y} + \sigma^2 \operatorname{Tr}\{\mathbf{P}_1\}}\mathbf{Y}^* + \ \mathbf{P}_1 \operatorname{Tr}\left\{\frac{\sigma^2}{\mathbf{Y}^*\mathbf{Y} + \sigma^2 \operatorname{Tr}\{\mathbf{P}_1\}}\right\} + \mathcal{O}(\sigma^6)
\end{aligned}
$$

In the final step, we have taken the liberty of identifying an expression whose Taylor series expansion matches the obtained expression to high (but finite) order in the signal-to-noise ratio[4]. We draw attention to three features of this (somewhat arbitrary) choice. First, it sidesteps the convergence of the series, which seems to be poor in practice. Second, it is tractable and numerically stable. And third, it suggests a connection between this MMSE estimation problem and the frequency response of the Wiener deconvolution filter. This is the Robust Eigenspace Signal Tracking (REST) fingerprint.

function FingerprintREST($\mathbf{Y}$)

  $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^* \leftarrow$ SVD($\mathbf{Y}$)

  $\widehat{\mathbf{P}} \leftarrow \mathbf{U}$ DIAG($\Sigma_{ii}^2/(\Sigma_{ii}^2 + (n-m)\sigma^2)$) $\mathbf{U}^*$

    $+(\mathbf{I} - \mathbf{U}\mathbf{U}^*)$ Tr DIAG($\sigma^2/(\Sigma_{ii}^2 + (n-m)\sigma^2)$)

  return $\widehat{\mathbf{P}}$

end function

The cost of evaluating a REST fingerprint is $\mathcal{O}(n^2m)$ using the singular value

[4]We've also abused notation by placing a matrix in the denominator of a fraction to indicate an inverse.

decomposition. Wi-Fi has 64-128 individual subcarriers, each of which is fingerprinted separately, and each of which contributes equally to the squared distance metric.

Commentary A REST fingerprint is the mean of a distribution of exact projection matrices corresponding to each possible channel matrix $\mathbf{H}$ that could have given rise to the observation $\mathbf{Y} = \mathbf{H} + \mathbf{N}$, weighted according to the likelihood of the Gaussian matrix $\mathbf{N}$ needed to explain the difference. REST fingerprints may not be projection matrices; they have eigenvalues in $[0, 1]$ rather than $\{0, 1\}$. This is analogous to taking the mean of a set of points on a sphere; the result may be interior to the sphere. The noiser the fingerprint, the more nearly uniform is the distribution of exact fingerprints over the Grassmannian manifold, and the further from a projection matrix $\widehat{\mathbf{P}}$ will be. Note that $\operatorname{Tr} \widehat{\mathbf{P}} = m$, since $\widehat{\mathbf{P}}$ is an expectation over projection matrices each having trace m.

4.2.7 Steps towards Quantitative Evaluation

We built a system to evaluate REST on 802.11n Wi-Fi signals "in the wild", consisting of a 6× 40 MHz MIMO transceiver which streams 7.68 Gb/s of samples onto a host server for storage and offline analysis. We implemented a software decoder for the most common 802.11n modes which provides packet detection and synchronization information to the implementation of our fingerprinting function. Figure 4-1 shows fingerprints extracted from one such packet trace. One can immediately see the repeated patterns representing pairs of transmissions with equivalent channel coefficients and similar fingerprints. While fingerprints with the same pattern may be seen to vary in color (one particular degree of freedom), this variation is gradual over time-scales of multiple seconds, and we expect that tracking will still be possible.

The next step, for future work, will be to complete the quantitative evaluation of REST. Our favored approach is to extract MAC addresses from each packet by fully decoding the raw Wi-Fi signals, and then to compute the fingerprint dissimilarities between all pairs of packets in the trace. Each pair can then be labeled as either a true positive (match) or a true negative by comparing MAC addresses. From this sample

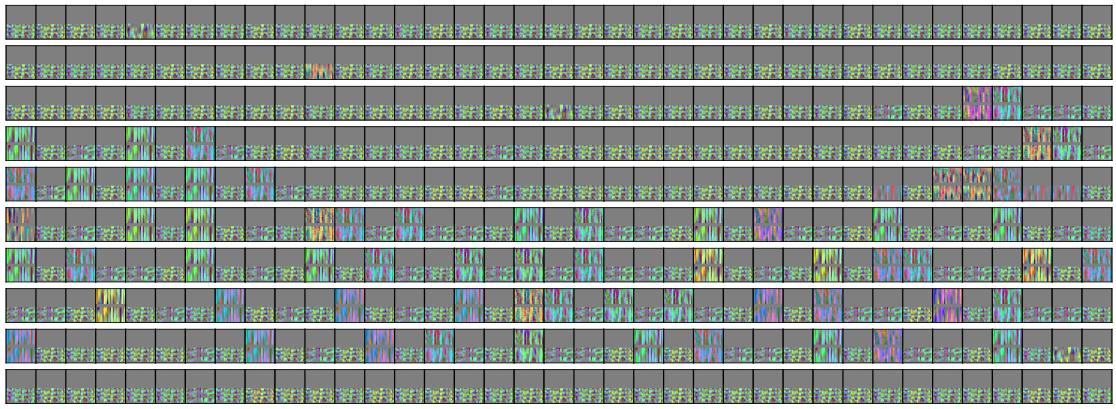


Figure 4-1: REST fingerprints of 370 packets (left-to-right, top-to-bottom) from a 5-second trace in a busy office building. The receiver bandwidth is 40 MHz, and the channel is divided into 128 subcarriers. Each complex 128×6×6 fingerprint is shown as a 128-row color bitmap, with hue and saturation representing phase and amplitude. Of the 377 packets in the full trace, 345 are decodable, and 286 include the transmitter's MAC address. A number of visually distinct sender fingerprints can be identified by their recurring patterns and colors. Quantitative clustering results on these fingerprints are future work.

of labeled pairs and their REST dissimilarity metrics, it would be straightforward to compute a receiver operating characteristic (ROC) curve.

4.3 MMSEE: Rational Gaussian Matrix Moments

We have seen how the computation of an MMSEE MIMO fingerprint requires the evaluation of an expectation of the form

$$\hat{\mathbf{P}} = \mathbb{E}[f(\mathbf{X})], \text{ where} \tag{4.1}$$

$$\begin{aligned} f(\mathbf{X}) &= \mathbf{X}\mathbf{X}^{+} \\ &= \mathbf{X}(\mathbf{X}^{\dagger}\mathbf{X})^{-1}\mathbf{X}^{\dagger} \text{ (when } \mathbf{X} \text{ has full column rank)} \end{aligned} \tag{4.2}$$

and $\mathbf{X}$ is complex Gaussian with

$$\begin{aligned}
\mathbb{E}[X_{ij}] &= \overline{X}_{ij}, \\
\mathbb{E}[(X_{ij} - \overline{X}_{ij})(X_{kl} - \overline{X}_{kl})^*] &= \Gamma_{ik}\delta_{jl}, \\
\mathbb{E}[(X_{ij} - \overline{X}_{ij})(X_{kl} - \overline{X}_{kl})\ ] &= 0.
\end{aligned}$$

Provided that $\mathbf{\Gamma}$ is positive definite, $\mathbf{X}$ has full column rank with probability 1. Furthermore, $\|f(\mathbf{X})\|_{op} \in \{0, 1\}$ is bounded for all $\mathbf{X}$, so the expectation (4.1) exists, and the contribution to this expectation due to the probability-0 set of rank-deficient inputs is zero. So we may safely adopt definition (4.2) and compute the expectation solely over full-rank inputs.

In order to compute the mean squared error of this estimate, we need to evaluate additional expectations of the form

$$\begin{aligned}
\mathbb{E}[\mathrm{MSE}] &= \mathbb{E}[\|f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]\|_F^2] \\
&= \mathbb{E}[\mathrm{Tr}\{(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})])(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})])^\dagger\}] \\
&= \mathrm{Tr}\{\mathbb{E}[f(\mathbf{X})](\mathbf{1} - \mathbb{E}[f(\mathbf{X})])\} \\
&= \mathrm{Tr}\{\hat{\mathbf{P}}(\mathbf{1} - \hat{\mathbf{P}})\}
\end{aligned}$$

In this case, the idempotence of $f(\mathbf{X})$ came to our rescue, and we did not need to do any further work beyond (4.1). But what if $f(\mathbf{X})$ did not have this property? Is there a universal approach to evaluating expectations of "nice" rational functions in Gaussian matrix variables?

Mai and Speicher [109] and Yin [110] have explored the limiting behavior of the eigenvalue distribution of non-commutative rational functions in certain random matrix ensembles; Edelman [111] and Rao [112] have developed algebraic computational tools for constructiong limiting eigenvalue distributions of a broad class of functions as root loci of polynomials. Neither approach directly addresses our problem for three reasons. First, this prior work operates on symmetric matrix spaces, while our inputs are rectangular. Second, these authors studied asymptotic behavior for large

matrices, while our matrices are of a small, fixed size. Third, these authors were interested in the distribution of eigenvalues; we care also about the identification of particular eigenvalues and eigenvectors. That is, even if we knew the correct eigenvalue distribution for $f(\mathbf{X})$ under the Gaussian input distribution, we would require some further reasoning to reconstruct the desired matrix-valued function outputs—for instance, symmetry might constrain the mapping of input singular values to output eigenvalues to be monotone, and we might find a heuristic that the actual output eigenvalues are close to appropriate quantiles of the eigenvalue distribution.

4.3.1 Neumann-Isserlis Estimation

Given a matrix-algebraic expression involving addition, subtraction, multiplication, and inversion, built from a set of constant matrices and matrices whose elements are jointly Gaussian, how can we compute its expectation?

The following heuristic procedure may be correct as a statement about formal power series. It certainly has some potentially-alarming divergences. In any event, however, we may test our results numerically to verify that we are on the right track, and if we demand a more exacting standard, we can fall back to a hypothesis-testing-based same-sender test. The present goal is to compute an MMSEE fingerprint so that the same-sender test becomes simply a computation of root-mean-squared distance between matrices.

Suppose that we have a matrix-algebraic expression $f(\mathbf{X})$ containing the Gaussian random matrix $\mathbf{X}$, and we wish to compute $\mathbb{E}[f(\mathbf{X})]$. We require that the elements X_{ij} have a joint covariance tensor $\Gamma_{ijk\ell} = \mathbb{E}[X_{ij} {X_{k\ell}}^*]$ that factorizes as

$$\Gamma_{ijk\ell} = (\mathbf{\Gamma}_{\mathbf{X},\text{rows}})_{ik} (\mathbf{\Gamma}_{\mathbf{X},\text{cols}})_{j\ell}$$

1. Substitute $\mathbf{X} \to \overline{\mathbf{X}} + \mathbf{Z}$, so that in what follows, the mean $\overline{\mathbf{X}}$ is appended to the set of constant matrices found in f, and $\mathbf{Z}$ is Gaussian and zero-mean.

2. Starting from the innermost matrix-inverse subexpression, factor out the deterministic part (that is, the inverse of the constant part of the subexpression).

Now the inverse is of the form $(\mathbf{I} + \mathcal{O}(\mathbf{Z}))^{-1}$.

3. Rewrite the inverse as a Neumann series in $\mathbf{Z}$.

4. Repeat steps 2 and 3 for each matrix-inverse subexpression. Viewing the expression as a syntax tree, proceed from leaves to roots.

5. Write the final result as a nested sum of matrix monomials, placing the expectation sign inside the summations.

6. Truncate the series to finite order.

7. Compute the expectation of each matrix monomial. (details follow)

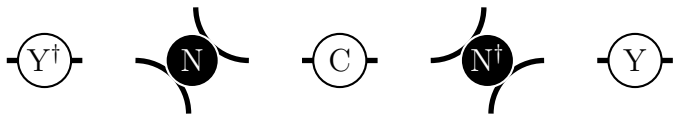
8. By inspection, collect any terms that form (truncated) Neumann series. Find a matrix-algebraic expression with the same Taylor expansion, subject to the constraint that the new expression must evaluate to $f(\overline{\mathbf{X}})$ in the limit where the variance of $\mathbf{Z}$ goes to zero.

This procedure could, presumably, be extended to matrix-algebraic expressions involving exponentiation by rational powers using the generalized binomial coefficients. While there exist closed-form expressions for the matrix inverse in any fixed dimension, the purpose of this procedure is to compute algebraic expectations in variables whose dimension is un-specified. The matrices need not even be square, provided that each sub-expression is well-formed, and the full expression evaluated at the mean $f(\overline{\mathbf{X}})$ is meaningful.

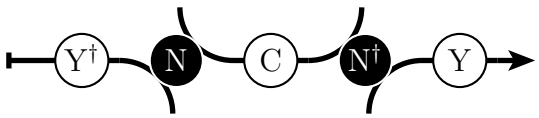
The following procedure evaluates the expectation of a matrix monomial. It is equivalent to the well-known theorem of Isserlis (or Wick), but uses graphical techniques as a short-cut. The alternative is to write the matrix monomial as a nested sum of scalar monomials using the definition of matrix multiplication, and to evaluate the expectation of the scalar monomial using Wick's theorem. The resulting expressions are identical.

1. We start by drawing one node for each factor in the matrix monomial. For instance, take this monomial in the Gaussian random matrix $\mathbf{N}$:

2. Next, we add edges behind each node. Constant matrices get horizontal edges passing through them from left to right. Random matrices get a pair of arcs, either west-to-south, north-to-east if the matrix is not transposed, or west-to-north, south-to-east if the matrix is transposed. For contrast, we color in the nodes that represent random matrices.

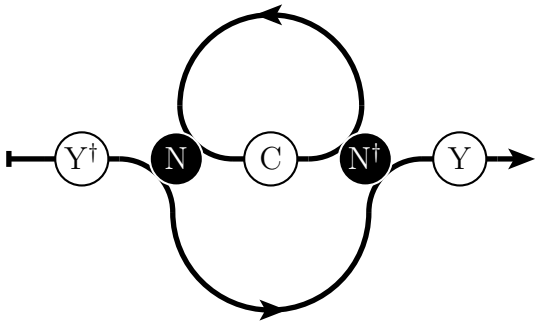


3. We now connect the nodes together side-by-side.



4. We consider all the unconnected arcs on the top (respectively, bottom) of the nodes as vertices. If the number of vertices on the top (resp. bottom) is odd, the expectation of the matrix monomial is zero.

   Otherwise, for each perfect matching over these vertices, we draw a copy of our diagram, connecting the vertices according to the matching. In the example monomial, there is only one matching for the top vertices, and only one matching for the bottom vertices, so there is only one completed diagram.



5. We assign an algebraic expression to each disjoint path through the diagram. The final result will be the product over paths of the expression assigned to each path. The unique open path that reaches all the way from left to right is matrix-valued; the remaining closed cycles are trace- (scalar) valued. The expression for a path is given by fixing a starting point (the left side, for the unique open path; any point, for a closed cycle) and walking the entire path once in the direction shown by the arrows:

   (a) Each time we encounter a white node, we read off the label of the node, and append it as a factor on the right side of the expression we are forming.

   (b) Each time we follow an up-then-down arc connecting two black nodes, we append a factor of $\mathbf{\Gamma}_{\mathbf{X},\text{cols}}$.

   (c) Each time we follow a down-then-up arc connecting two black nodes, we append a factor of $\mathbf{\Gamma}_{\mathbf{X},\text{rows}}$.

   (d) If the path is a closed cycle, we take the trace of the entire expression after walking a complete circuit.

Thus, for the monomial $\mathbf{Y}^\dagger\mathbf{N}\mathbf{C}\mathbf{N}^\dagger\mathbf{Y}$ in the zero-mean random matrix $\mathbf{N}$, we get

$$\mathbb{E}[\mathbf{Y}^\dagger\mathbf{N}\mathbf{C}\mathbf{N}^\dagger\mathbf{Y}] = \mathbf{Y}^\dagger\mathbf{\Gamma}_{\mathbf{N},\text{rows}}\mathbf{Y}\,\mathrm{Tr}(\mathbf{C}\mathbf{\Gamma}_{\mathbf{N},\text{cols}}).$$

For clarity, let us take $\mathbf{\Gamma}_{\mathbf{N},\text{cols}} = \mathbf{I}$ the identity matrix, as in REST, and let us drop the subscript on $\mathbf{\Gamma}_{\mathbf{N},\text{rows}}$. Then

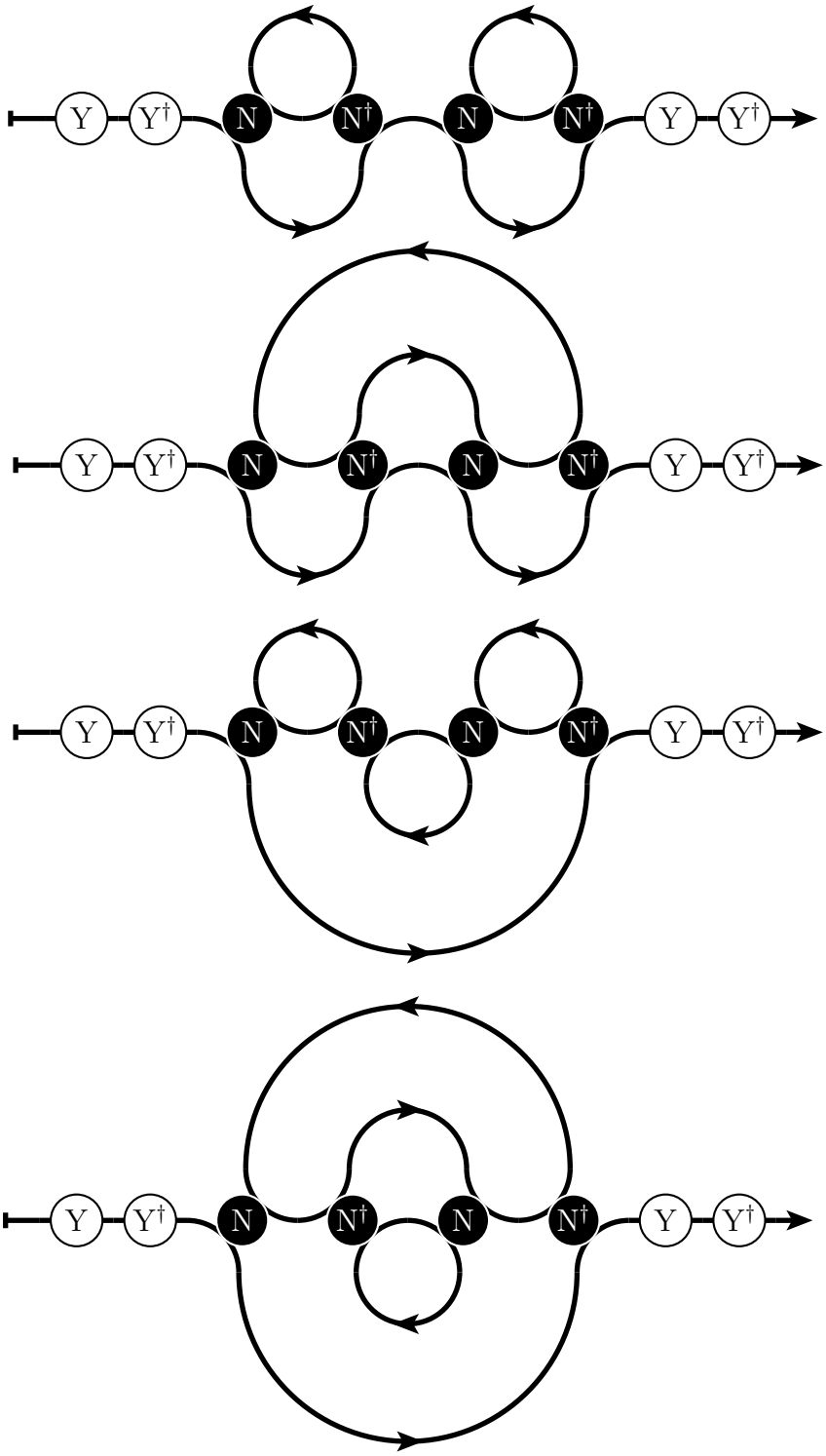
$$\mathbb{E}[\mathbf{Y}^\dagger\mathbf{N}\mathbf{C}\mathbf{N}^\dagger\mathbf{Y}] = \mathbf{Y}^\dagger\mathbf{\Gamma}\mathbf{Y}\ \mathrm{Tr}\,\mathbf{C}.$$

For a monomial with n copies of the un-transposed random variable and n copies of the transposed random variable, there are $(n!!)^2$ diagrams. If there are multiple diagrams, their final expressions add.

Here is a more complicated example. What is the expectation

$$\mathbb{E}[\mathbf{Y}\mathbf{Y}^\dagger\mathbf{N}\mathbf{N}^\dagger\mathbf{N}\mathbf{N}^\dagger\mathbf{Y}\mathbf{Y}^\dagger]?$$

We first draw the diagrams (there are four this time):



Now we read off the result:

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}\mathbf{Y}^\dagger\mathbf{N}\mathbf{N}^\dagger\mathbf{N}\mathbf{N}^\dagger\mathbf{Y}\mathbf{Y}^\dagger] &= \mathbf{Y}\mathbf{Y}^\dagger\mathbf{\Gamma}^2\mathbf{Y}\mathbf{Y}^\dagger(\operatorname{Tr}\mathbf{I})^2 \qquad + \\
&\quad \mathbf{Y}\mathbf{Y}^\dagger\mathbf{\Gamma}^2\mathbf{Y}\mathbf{Y}^\dagger \operatorname{Tr}\mathbf{I}^2 \qquad + \\
&\quad \mathbf{Y}\mathbf{Y}^\dagger\mathbf{\Gamma}\ \mathbf{Y}\mathbf{Y}^\dagger(\operatorname{Tr}\mathbf{I})^2\operatorname{Tr}\mathbf{\Gamma} + \\
&\quad \mathbf{Y}\mathbf{Y}^\dagger\mathbf{\Gamma}\ \mathbf{Y}\mathbf{Y}^\dagger \operatorname{Tr}\mathbf{I}^2 \operatorname{Tr}\mathbf{\Gamma} \\
&= \mathbf{Y}\mathbf{Y}^\dagger(\mathbf{\Gamma}^2 + \mathbf{\Gamma}\operatorname{Tr}\mathbf{\Gamma})\mathbf{Y}\mathbf{Y}^\dagger\left((\operatorname{Tr}\mathbf{I})^2 + \operatorname{Tr}\mathbf{I}\right)
\end{aligned}$$

We recall that $\mathbf{I}$ was introduced as the covariance matrix for the columns of $\mathbf{N}$; so if $\mathbf{N} \in \mathbb{C}^{n\times m}$, then $\operatorname{Tr}\mathbf{I} = m$.

Mathematica code for this procedure is given in Appendix A.

4.4 Maximum-likelihood Estimator

If, rather than the MMSE estimate, we ask for the most likely fingerprint, the problem is more straightforward due to the functional invariance property of the maximum likelihood estimator. In this case,

$$\begin{aligned}
\hat{\mathbf{P}}_{\mathrm{ML}} &= f(\hat{\mathbf{H}}_{\mathrm{ML}}) \\
&= f(\mathbf{Y}),
\end{aligned}$$

since the likelihood of $\mathbf{H}$ given $\mathbf{Y}$ is given by

$$\begin{aligned}
L(\mathbf{H}\,|\,\mathbf{Y}) &= \mathbb{P}(\mathbf{Y}\,|\,\mathbf{H}) \\
&= \mathbb{P}(\mathbf{N} = \mathbf{Y} - \mathbf{H}),
\end{aligned}$$

which is maximized when $\mathbf{N} = 0$. This fingerprint was previously reported by Kuperstein et al [113, 114].

4.5 Anonymizing Countermeasures

4.5.1 Hiding in a Crowd

Connecting the dots from a location trace will not work well in a crowd, or if the user waits a long time between packet transmissions. By a "long time", we mean a time scale long compared to the time required to walk up to and bump into another user. If the area density of users is ρ_A, in persons/m^2, then the distance to another user is typically on the order of $\rho_A^{-1/2}$. If v is a typical walking speed, then mistakes by the de-anonymizer are likely iff the time between packet transmissions from one user's device is $\tau \gg \frac{1}{v\sqrt{\rho_A}}$. So for any given density of "crowd", it is quantitatively possible to hide in that crowd by pausing long enough between transmissions. See Figure 4-2 for an example. In this way, mix zones [115] and silent periods [116] can be viewed as dual techniques.

4.5.2 Antenna-Switching

For the listener to track fingerprints, the sender, listener, and environment geometry must be stable over the inter-arrival time of the wireless link. To defeat localization without physical movement, a sender needs to access a large number of linearly-independent spatial degrees of freedom via independent radiation patterns—at least as many patterns as the listener has antennas.

For a compact, inexpensive, solid-state antenna to access to a large number of linearly-independent radiation patterns, it must be dynamically reconfigurable. One approach is to use an antenna built from small metallic elements connected by electrical switches [117, 118, 119], giving an exponential number of configurations. A manufacturer characterizes each antenna configuration once for a given line of devices and determines the subset with adequate performance. In use, the device cycles through configurations, producing unpredictable and shifting patterns of Wi-Fi energy. The performance vs. anonymity trade-off is then placed in the user's hands.

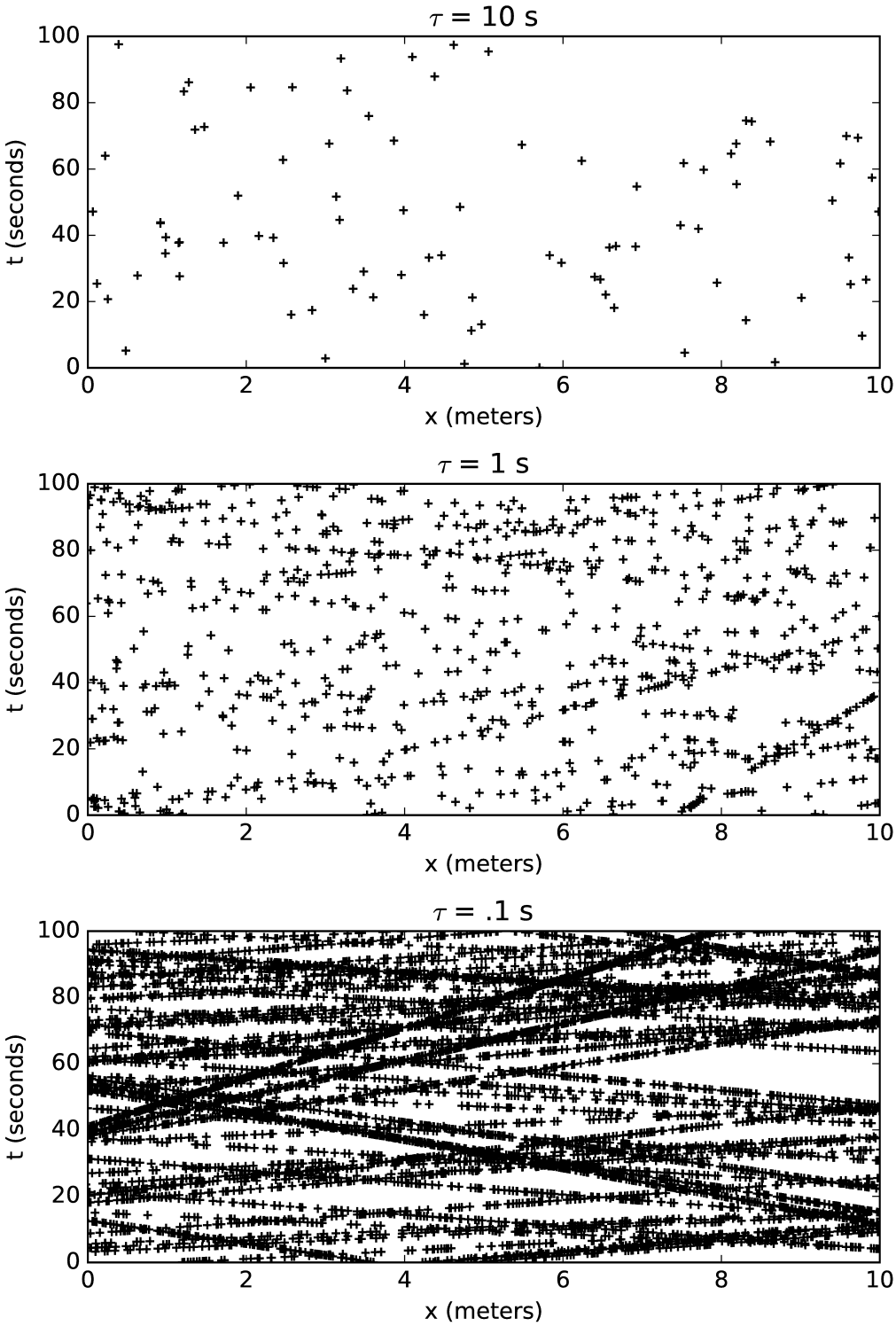


Figure 4-2: With $\rho_A = 1$, $v \in \{-1, 1\}$, shows how trails of packets from a user's device may be concealed (top), marginally clear (middle), or abundantly clear (bottom) depending on the time τ between transmissions.

4.6 Limits on Countermeasures

Is antenna-switching a viable countermeasure? Could it really access an exponential number of degrees of freedom? We offer two lines of evidence to establish that the number of degrees of freedom accessed will be bounded for any device of finite dimensions. The first, heuristic argument will be presented here; the second, more formal argument will be presented at greater length in §4.7.

We abstract a wireless device as a region of space in which currents and charges oscillate at the transmission frequency (say, 2.4 GHz). We first obtain an expression for the radiated fields far (> 1 m) from the wireless device, and then we show how this expression reveals the finiteness of the number of degrees of freedom.

The far-field radiation due to currents $\Re\{e^{j\omega t}\mathbf{J}(\mathbf{x})\}$ is

$$\begin{aligned}
\tilde{\mathbf{J}}(\mathbf{k}) &= \mathcal{F}\{\mathbf{J}\} = \int d^3\mathbf{x}' e^{-j\mathbf{k}\cdot\mathbf{x}'}\mathbf{J}(\mathbf{x}') \\
\mathbf{E}(\mathbf{r}) &= \frac{\mu_0\omega}{4\pi r}\Im\left\{e^{j\omega(t-\frac{r}{c})}(\mathbf{I}-\hat{\mathbf{r}}\hat{\mathbf{r}}^\intercal)\tilde{\mathbf{J}}(\tfrac{-\omega\hat{\mathbf{r}}}{c})\right\} + \mathcal{O}(1/r^2) \\
\mathbf{B}(\mathbf{r}) &= \frac{\mathbf{E}(\mathbf{r})}{c}\times\hat{\mathbf{r}} + \mathcal{O}(1/r^2)
\end{aligned} \tag{4.3}$$

Equation (4.3) is a linear map from the space of current distributions to the space of far-field distributions. We conjecture that, no matter the design of the antenna array, if it is physically confined to the volume of a small device, then this linear map to far-field distributions will be low-rank, and the number of linearly-independent spatial streams (or virtual antennas for location spoofing) will be small. If so, then a de-anonymizing listener with more than this number of receive antennas will be able to characterize the subspace in which this transmitter is to be found, no matter how currents are redirected within the device (e.g. via MIMO or antenna switching).

External scatterers could couple near-field DoFs into additional far-field DoFs. To some approximation, scatterers beyond 10 wavelengths will only sense far-field modes, so they cannot cause a problem. We sidestep the issue of near-field scatterers, supposing that the only objects within a meter or so of the transmitter are e.g. human limbs, which are good absorbers of microwave energy.

4.6.1 Nyquist Bound

$\tilde{\mathbf{J}}(\mathbf{k})$ is the Fourier transform of a function with limited support, since the current distribution is zero outside of the transmitting device. $\tilde{\mathbf{J}}(\mathbf{k})$ is therefore a smooth, "band-limited" function. This smooth function is evaluated only on the sphere of radius ω/c around the origin to determine the far fields. So the dimension of the space of possible radiated far fields is bounded by the dimension of the space of band-limited functions on the sphere.

If we were reconstructing a band-limited function in one dimension, we would need to sample above the Nyquist frequency. In three dimensions, the Petersen-Middleton theorem [120] directs us to sample on a sufficiently fine lattice. If the transmitter has dimensions $a_1 \times a_2 \times a_3$, then an acceptable sampling lattice for $\tilde{\mathbf{J}}(\mathbf{k})$ is a rectangular grid with spacings $1/a_1$, $1/a_2$, and $1/a_3$, and cell volume $1/V = 1/\prod a_i$. The number of sample points in any region is then roughly the volume of the region divided by the grid cell volume; so the number of sample points "near" the sphere of radius ω/c and useful for reconstructing the function on that sphere is proportional to the sample spacing, $V^{-1/3}$, and the area of the sphere:

$$\begin{aligned}\text{points} &\approx 4\pi(\omega/c)^2 V^{-1/3} V \\ &= 16\pi^3 V^{2/3}/\lambda^2\end{aligned}$$

We also need to multiply by the dimension of the current distribution. The projection $(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}^\intercal)$ in Equation (4.3) casts out the irrotational component of the vector field, leaving two DoFs per point.

$$\text{dim} \approx 32\pi^3 V^{2/3}/\lambda^2$$

The number of effective DoFs of a device therefore scales like its surface area over the square of the wavelength. For an iPhone 6 Plus operating at 2.4 GHz, this is about 125.

A serious flaw in this line of reasoning is in the use of Nyquist sampling to

reconstruct a function over a finite region.

4.7 Random-Field Bound on Effective Antennas

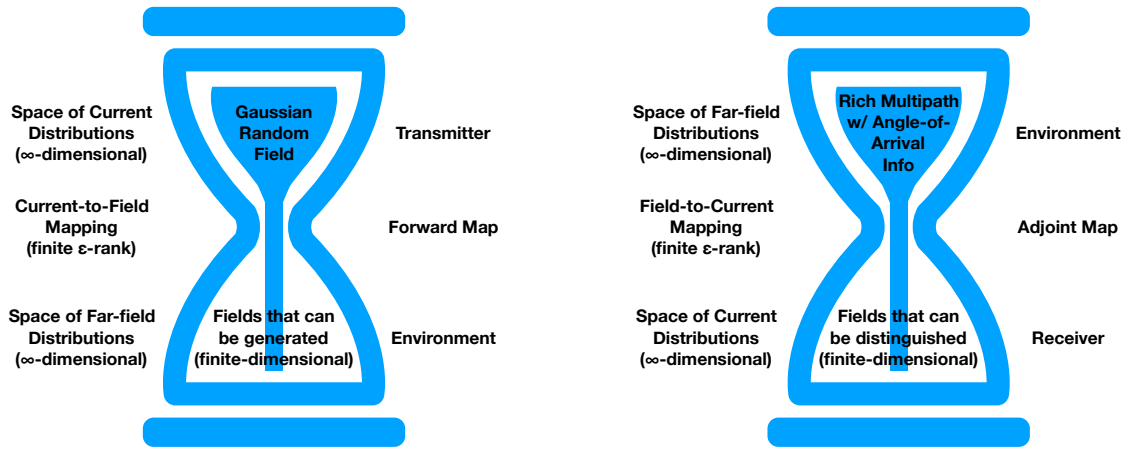
In this section, we seek a bound on the number of degrees of freedom—i.e. radiative modes, or effective antennas—accessible to a radio device of finite dimensions, using finite bandwidth, in free space. This bound applies equally to transmitters and receivers. Maxwell's equations determine the relationship between electrical currents, which are distributed in some (possibly complicated) way across space and time, and the radio waves they generate. This "currents-to-fields" map creates a bottleneck (the so-called *rank* of the map) on the number of degrees of freedom.

Because the currents-to-fields map is *linear*, we may view a radio with a certain antenna design as accessing a pre-determined *linear sub-space* of the space of all possible distributions of electrical current. For each antenna, when it is driven at some frequency, its peculiar distribution of fluctuating currents represents a single point in a space. If two antennas are driven, then we describe the total current as the sum over antennas of the currents in each antenna: a vector sum. With only two antennas, the space of accessible current distributions is two-dimensional.

In the same way, the state of the electric fields at all points far from the radio device may be viewed, in aggregate, as a single point in a large vector space. As the configuration of the fields changes, the trajectory of this point may explore a region in this configuration space. If the explored region is a finite-dimensional linear sub-space of the full configuration space, then we say that the field distribution is finite-dimensional.

Figure 4-3 shows how the currents-to-fields mapping acts as a bottleneck on the dimension of the fields. No matter how large is the dimension of our space of accessible current distributions—that is, no matter how many separate antennas we have within our radio—the dimension of the space of generated electric field configurations is bounded above by the rank of the currents-to-fields mapping.

The rank is more correctly termed an ε-rank, with $\varepsilon \ll 1$. It may be that the



(a) Transmitter (b) Receiver

Figure 4-3: For a physically-small transmitter or receiver, the map between electrical currents and radio waves (electric fields) is *low rank*, meaning that it acts as a bottleneck on the number of degrees-of-freedom we can access.

point in the configuration space of electric fields explores a region that is not quite one-dimensional. The region may, for instance, be a cylinder many times longer than its width. For the purposes of the ε-rank, we treat the cylinder as a one-dimensional space if its width is ε times its length.

The hourglass on the left side of Figure 4-3 represents the relationship between currents in a transmitter (top) and electric far fields (bottom). The currents-to-fields mapping is the narrow neck. To determine the width of this neck, and hence the number of degrees-of-freedom accessible to the transmitter, we supply a very large space of model current distributions, drawn as the "sand" in the top of the hourglass. We allow this space to pass through the mapping to form the corresponding electric field configurations, and we measure the dimension of this output region. Provided that our input space is "large enough", the dimension of the output region is the rank of the bottleneck mapping.

The hourglass on the right side of the figure represents the situation for a receiver, the Helmholtz reciprocal of the transmitter on the left. The knowledge we gain from studying the rank of the currents-to-fields mapping applies directly to its adjoint, the fields-to-currents mapping. In this picture, the (unknown, unmodeled) high-dimensional space of interesting electric field distributions is funneled through the

adjoint map to give the space of fields that are distinguishable within the volume of the receiver. This will, once again, be finite-dimensional; and as a consequence, a listener engaging in signal fingerprinting using measurements collected with a small device will have bounded power to distinguish between senders.

The following metaphor may offer further intuition for the approach.

Suppose that we wish to determine how many notes a piano can play. One way to accomplish this task—though perhaps not the most hygienic—would be to send 1,000,000 monkeys with hammers to strike the instrument at random. By studying the resulting clangor, we might hope to detect each distinct note, since with very high probability, each of the piano's 88 keys (or their corresponding strings) will be struck by at least one monkey.

The hammer-blows that fall on the frame of the piano will fail to radiate sound; and, indeed, the majority of the monkeys will be wasting their efforts. But a positive fraction of the total monkey-power will drive radiative modes of the piano. Once we have determined how the piano may be sounded, a suitable distraction may be employed to remove the monkeys, and a professional musician may prepare a lovely composition that only involves striking the piano on its keys or strings.

The final composition drives only the radiative modes of the piano. We may view such a composition as a particular distribution of mechanical inputs in time and space. Since the composition does not involve striking the piano's frame, we might say that the composition lies in the orthogonal complement of the non-radiating sub-space of mechanical input distributions. This non-radiating sub-space consists of all conceivable hammer-blows that do not strike the keys or strings of the piano.

The number of radiative modes—notes—of the piano is the same in either case: under wasteful, random driving; or under deliberate driving of the radiative sub-space.

In this section, our "piano" is the electromagnetic field itself.

We will draw our conclusions by observing the radiation that results when a large number of Gaussian monkeys perturb the field. Rather than hammers, our monkeys' hands will hold charged particles; and rather than striking hammer-blows, they will generate electric and magnetic fields by shaking their fists.

4.7.1 Approach

We restrict our attention to far-field degrees of freedom—that is, those which decay as $1/r$ in amplitude and $1/r^2$ in power, corresponding to "propagating" electromagnetic waves, as opposed to near-field or evanescent waves which exhibit a more rapid drop in amplitude. The general "rule of thumb" is that near-field effects are negligible at a distance exceeding 10 wavelengths from the source; but we will not attempt to bound these terms directly.

The direct approach is as follows: first, we generate a large number of examples of how electrical currents might be generated inside the transmitter. Next, we solve Maxwell's equations to compute the far fields due to these currents. Now armed with a sample of many fields, we perform principal component analysis to estimate how many independent random variables (degrees of freedom) are needed to account for the co-variance of our sample. Variance in the fields represents radiated power. If we find that all but a negligible quantity of radiated power is accounted for by, say, the 100 most significant degrees of freedom, then the effective number of antennas of the transmitter is 100. A quantity of power is negligible if it would be undetectable at a typical signal-to-noise ratio of e.g. 20 dB.

Directly solving Maxwell's equations for a large number of current distributions makes this approach tedious. In addition, the requirement to make up-front decisions about resolution, bandwidth, center frequency, and transmitter shape and size make generalizing the results challenging.

Fortunately, we can combine several of these steps and carry them out in closed form, leaving only the final step of examining the significant degrees-of-freedom to be performed numerically.

Our strategy will proceed in four phases. First, we will construct a linear map from currents in the transmitter to radiated far fields. Next, we will build a random field model that explores a large space of hypothetical current distributions. Each current distribution corresponds to one way of configuring antennas, reflectors, amplifiers, and other circuitry within the body of the transmitter. Third, supposing our current

distribution is chosen according to this model, we will determine the correlation between the far fields radiated with various polarizations in various directions.

Finally, from this correlation, we will obtain an approximate spectrum. This corresponds to the PCA step, and the spectrum corresponds to the most significant degrees of freedom within the population of far fields. The number of non-negligible eigenvalues in the correlation spectrum will be an estimate of the number of far-field degrees of freedom accessible to the radio transmitter.

4.7.2 Fields from Currents

We follow the approach of Nevels and Jeong [121, (6)]. We wish to know the electric fields due to a given distribution of electric currents. Our problem domain is all of space and all of time.

In general, a solution to this type of problem might include electromagnetic waves that are generated outside our domain and pass through it unaffected. These are of no interest to us, and may be excluded by applying a radiation condition.

In general, the solution may also include pre-existing electromagnetic fields that originate in the distant past; these are of no interest to us, and may be excluded by requiring all of our fields and currents to vanish in the distant past beyond some finite horizon, and finally taking this horizon to infinity at the end of the derivation. We ignore this subtlety in what follows.

The remaining fields are those generated by currents in the interior of the transmitter. These fields can be obtained by a convolution integral with an appropriate Green's function. That is, we suppose that the fields measured at a location $\mathbf{r}$ and a time t arise, in part, from a source current at location $\mathbf{r}'$ and time t'. The Green's function gives the contribution of this source element to the measurement, proportional to the intensity of the source current element.

The convolution integral with the so-called dyadic Maxwell Green's function is

$$\mathbf{F}(\mathbf{r}, t) = \int \mathrm{d}^3\mathbf{r}' \int_{-\infty}^{t+} \mathrm{d}t' \, \overline{G}(\mathbf{r}, t|\mathbf{r}', t') \cdot \mathbf{J}(\mathbf{r}', t') \tag{4.4}$$

with the definitions

$$\mathbf{F} = \begin{bmatrix} \mathbf{E} \\ \frac{1}{\mu_0}\mathbf{B} \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} \frac{1}{\varepsilon_0}\mathbf{J}_e \\ \frac{1}{\mu_0}\mathbf{J}_m \end{bmatrix}, \quad \overline{G} = \begin{bmatrix} \overline{G}_{11} & \overline{G}_{12} \\ \overline{G}_{21} & \overline{G}_{22} \end{bmatrix} U(t-t'), \quad U(\tau) = \begin{cases} 0, & \tau < 0 \\ 1, & \tau > 0 \end{cases}$$

The magnetic (monopole) current $\mathbf{J}_m$ is non-physical, and we may set it to zero.

The magnetic field can be recovered from the electric field by

$$\mathbf{B}(\mathbf{r}, t) = -\int_{-\infty}^{t} \mathrm{d}t' \nabla \times \mathbf{E}(\mathbf{r}, t')$$

and hence the magnetic field does not carry additional degrees of freedom beyond those represented in the electric field. We will therefore focus exclusively on the electric field

$$\mathbf{E}(\mathbf{r}, t) = \int \mathrm{d}^3\mathbf{r}' \int_{-\infty}^{t+} \mathrm{d}t' \, \overline{G}_{11}(\mathbf{r}, t|\mathbf{r}', t') \cdot \mathbf{J}_e(\mathbf{r}', t')/\varepsilon_0 \tag{4.5}$$

with electric Green's function

$$\begin{aligned} \overline{G}_{11}(\mathbf{r}, t \,|\, \mathbf{r}', t') &= U(\tau) \left[\tfrac{1}{c^2}\partial_t^2 - \nabla\nabla^T\right] \frac{U(c\tau - R)}{4\pi R}\mathbf{1} \\ &= U(\tau) \left[+\frac{\delta'(c\tau - R)}{4\pi R}\mathbf{1} - \nabla\nabla^T \frac{U(c\tau - R)}{4\pi R}\right], \text{ where} \\ R &= |\mathbf{r} - \mathbf{r}'| \\ \tau &= t - t' \end{aligned} \tag{4.6}$$

Note 1. (4.6) differs from [121, (51)], which appears to have a sign error on the first term.

This Green's function has components at three levels of dis-continuity: step-function, delta-function, and derivative-of-delta function. Since our sources have finite bandwidth, integrating against this discontinuous Green's function will nevertheless give smooth fields: we can use integration by parts to transfer differentiations from the step function to the source field (in the form of time derivatives), and time derivatives of band-limited signals are band-limited signals.

Expanding the derivatives in (4.6) gives

$$\begin{aligned}
\overline{G}_{11}(\mathbf{r},t\,|\,\mathbf{r}',t') &= \frac{U(\tau)}{4\pi}\Big(+\delta'(c\tau-R)\Big[\mathbf{1}-\tfrac{(\mathbf{r}-\mathbf{r}')(\mathbf{r}-\mathbf{r}')^T}{R^2}\Big]\tfrac{1}{R} \\
&\qquad +\delta\,(c\tau-R)\Big[\mathbf{1}-3\tfrac{(\mathbf{r}-\mathbf{r}')(\mathbf{r}-\mathbf{r}')^T}{R^2}\Big]\tfrac{1}{R^2} \\
&\qquad +U(c\tau-R)\Big[\mathbf{1}-3\tfrac{(\mathbf{r}-\mathbf{r}')(\mathbf{r}-\mathbf{r}')^T}{R^2}\Big]\tfrac{1}{R^3}\Big) \\
&= \frac{U(\tau)\delta'(c\tau-R)}{4\pi R}\left[\mathbf{1}-\frac{(\mathbf{r}-\mathbf{r}')(\mathbf{r}-\mathbf{r}')^T}{R^2}\right]+\mathcal{O}(R^{-2}) \\
&= \frac{U(\tau)\delta'(c\tau-R)(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)}{4\pi|\mathbf{r}|}+\mathcal{O}(|\mathbf{r}|^{-2}) \\
&= \frac{\delta'(c\tau-R)(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)}{4\pi|\mathbf{r}|}+\mathcal{O}(|\mathbf{r}|^{-2})
\end{aligned}$$

where the final step follows provided that we only ever evaluate $\overline{G}_{11}$ with $\mathbf{r}\neq\mathbf{r}'$: in this case, $R>0$ and the derivative-of-delta is nonzero only when $\tau>0$. Now we may substitute into (4.5) to obtain

$$\mathbf{E}(\mathbf{r},t)=\frac{1}{4\pi\varepsilon_0 r}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int \mathrm{d}^3\mathbf{r}'\int_{-\infty}^{t+}\mathrm{d}t'\delta'(c\tau-R)\mathbf{J}_e(\mathbf{r}',t')+\mathcal{O}(1/r^2)$$

and carry out u-substitution and integration by parts to obtain

$$\begin{aligned}
&= \frac{1}{4\pi\varepsilon_0 rc}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int \mathrm{d}^3\mathbf{r}'\int_{-R-}^{+\infty}\mathrm{d}x\,\delta'(x)\mathbf{J}_e(\mathbf{r}',t-\tfrac{x+R}{c})+\mathcal{O}(1/r^2) \\
&= \frac{1}{4\pi\varepsilon_0 rc^2}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int \mathrm{d}^3\mathbf{r}'\int_{-R-}^{+\infty}\mathrm{d}x\,\delta(x)(\partial_t\mathbf{J}_e)(\mathbf{r}',t-\tfrac{x+R}{c})+\mathcal{O}(1/r^2) \\
&= \frac{\mu_0}{4\pi r}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int \mathrm{d}^3\mathbf{r}'(\partial_t\mathbf{J}_e)(\mathbf{r}',t-R/c)+\mathcal{O}(1/r^2)
\end{aligned}$$

We now transform this expression by writing $\mathbf{J}_e$ as an inverse Fourier transform in time, allowing us to carry out the ∂_t. We then make the Fraunhofer diffraction approximation $R\approx(r-\hat{\mathbf{r}}\cdot\mathbf{r}')$.

$$\begin{aligned}
&= \frac{\mu_0}{4\pi r}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int \mathrm{d}^3\mathbf{r}'\int\frac{\mathrm{d}\omega}{2\pi}(j\omega)e^{j\omega(t-R/c)}\mathbf{J}_e(\mathbf{r}',\omega)+\mathcal{O}(1/r^2) \\
&= \frac{\mu_0}{4\pi r}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int\frac{\mathrm{d}\omega}{2\pi}(j\omega)e^{j\omega t}\int \mathrm{d}^3\mathbf{r}'e^{-j\omega R/c}\mathbf{J}_e(\mathbf{r}',\omega)+\mathcal{O}(1/r^2) \\
&\approx \frac{\mu_0}{4\pi r}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int\frac{\mathrm{d}\omega}{2\pi}(j\omega)e^{j\omega t}\int \mathrm{d}^3\mathbf{r}'e^{-j\omega(r-\hat{\mathbf{r}}\cdot\mathbf{r}')/c}\mathbf{J}_e(\mathbf{r}',\omega)+\mathcal{O}(1/r^2)
\end{aligned}$$

Now we may rearrange to identify the $\int \mathrm{d}^3\mathbf{r}$ as a Fourier transform in space:

$$
\begin{aligned}
&= \frac{\mu_0}{4\pi r}(\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T) \int \frac{\mathrm{d}\omega}{2\pi}(j\omega)e^{j\omega(t-r/c)} \int \mathrm{d}^3\mathbf{r}' e^{j(\omega\hat{\mathbf{r}}/c)\cdot\mathbf{r}'}\mathbf{J}_e(\mathbf{r}',\omega) + \mathcal{O}(1/r^2) \\
&= \frac{\mu_0}{4\pi r}(\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T) \int \frac{\mathrm{d}\omega}{2\pi}(j\omega)e^{j\omega(t-r/c)}\tilde{\mathbf{J}}_e(-\omega\hat{\mathbf{r}}/c,\omega) + \mathcal{O}(1/r^2)
\end{aligned}
$$

We could work directly with this quantity, which depends on $\mathbf{r}$ and t. However, its dependence on the radius r does not carry any information that is not already captured by its dependence on t. That is, each shell of radiation $[r, r+\Delta r]$ flies away from the origin at the speed of light, forever diminishing in amplitude but never changing in polarization or phase. The field $\mathbf{E}(\mathbf{r},t)$ is redundant in the sense that it contains infinitely many copies of the information in each shell of radiation, one for each time coordinate after the shell is first emitted at $r = 0$.

To count degrees of freedom, it is sufficient to consider each shell exactly once, when it passes through some fixed surface r_0. We may further eliminate the amplitude dependence on r_0 by a suitable rescaling. Our new representation of the far fields will therefore be a function of the propagation direction $\hat{\mathbf{r}}$ and the time only:

$$
\begin{aligned}
\mathbf{E}_0(\hat{\mathbf{r}},t) &= \lim_{r\to\infty} r\mathbf{E}(r\hat{\mathbf{r}}, t + r/c) \\
&= \frac{\mu_0}{4\pi}(\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T) \int \frac{\mathrm{d}\omega}{2\pi}(j\omega)e^{j\omega t}\tilde{\mathbf{J}}_e(-\omega\hat{\mathbf{r}}/c,\omega)
\end{aligned} \tag{4.7}
$$

Radiated Power

It will be useful later to have an expression for the flux of radiated power away from the transmitter. The expected Poynting flux density due to the far fields is

$$
\begin{aligned}
\mathbb{E}[\mathbf{S}(\mathbf{r},t)] &= \frac{1}{\mu_0}\mathbb{E}[\mathbf{E}(\mathbf{r},t) \times \mathbf{B}(\mathbf{r},t)] \\
&= -\frac{1}{\mu_0 r^2}\int_{-\infty}^{t} \mathrm{d}t' \mathbb{E}\left[\mathbf{E}_0(\hat{\mathbf{r}}, t - r/c) \times ((\nabla - \hat{\mathbf{r}}/r) \times \mathbf{E}_0(\hat{\mathbf{r}}, t' - r/c))\right]
\end{aligned}
$$

Taking the radial component of this vector field gives

$$\begin{aligned}\hat{\mathbf{r}} \cdot \mathbb{E}[\mathbf{S}(\mathbf{r}, t)] &= -\frac{1}{\mu_0 r^2} \int_{-\infty}^{t} \mathrm{d}t' \mathbb{E}\left[\hat{\mathbf{r}} \cdot (\mathbf{E}_0(\hat{\mathbf{r}}, t - r/c) \times ((\nabla - \hat{\mathbf{r}}/r) \times \mathbf{E}_0(\hat{\mathbf{r}}, t' - r/c)))\right] \\ &= -\frac{1}{\mu_0 r^2} \int_{-\infty}^{t} \mathrm{d}t' \mathbb{E}\left[((\nabla - \hat{\mathbf{r}}/r) \times \mathbf{E}_0(\hat{\mathbf{r}}, t' - r/c)) \cdot (\hat{\mathbf{r}} \times \mathbf{E}_0(\hat{\mathbf{r}}, t - r/c))\right]\end{aligned}$$

One can show by manipulation in spherical coordinates that for two differentiable vector fields $\mathbf{A}$, $\mathbf{B}$ s.t. $\hat{\mathbf{r}} \cdot \mathbf{A} = \hat{\mathbf{r}} \cdot \mathbf{B} = 0$,

$$((\nabla - \hat{\mathbf{r}}/r) \times \mathbf{A}) \cdot (\hat{\mathbf{r}} \times \mathbf{B}) = (\partial_r \mathbf{A}) \cdot \mathbf{B}$$

We take $\mathbf{A} = \mathbf{E}_0(\hat{\mathbf{r}}, t' - r/c)$ and $\mathbf{B} = \mathbf{E}_0(\hat{\mathbf{r}}, t - r/c)$ to get

$$\begin{aligned}\hat{\mathbf{r}} \cdot \mathbb{E}[\mathbf{S}(\mathbf{r}, t)] &= -\frac{1}{\mu_0 r^2} \int_{-\infty}^{t} \mathrm{d}t' \mathbb{E}\left[\mathbf{E}_0(\hat{\mathbf{r}}, t - r/c) \cdot (\partial_r \mathbf{E}_0(\hat{\mathbf{r}}, t' - r/c))\right] \\ &= \frac{1}{\mu_0 r^2 c} \int_{-\infty}^{t} \mathrm{d}t' \mathbb{E}\left[\mathbf{E}_0(\hat{\mathbf{r}}, t - r/c) \cdot \partial_{t'} \mathbf{E}_0(\hat{\mathbf{r}}, t' - r/c)\right] \\ &= \frac{1}{\mu_0 r^2 c} \mathbb{E}\left[\mathbf{E}_0(\hat{\mathbf{r}}, t - r/c) \cdot \int_{-\infty}^{t} \mathrm{d}t' \partial_{t'} \mathbf{E}_0(\hat{\mathbf{r}}, t' - r/c)\right] \\ &= \frac{1}{\mu_0 r^2 c} \mathbb{E}\left[|\mathbf{E}_0(\hat{\mathbf{r}}, t - r/c)|^2\right]\end{aligned}$$

To get the radiated power due to currents at time t, we evaluate the outgoing Poynting flux at $t + r/c$ and integrate over the sphere of radius r:

$$\begin{aligned}\mathbb{E}[P_{\text{radiated}}(t)] &= \int r^2 \mathrm{d}\Omega \left(\hat{\mathbf{r}} \cdot \mathbb{E}[\mathbf{S}(\mathbf{r}, t + r/c)]\right) \\ &= \int r^2 \mathrm{d}\Omega \left(\frac{1}{\mu_0 r^2 c} \mathbb{E}\left[|\mathbf{E}_0(\hat{\mathbf{r}}, t)|^2\right]\right) \\ &= \frac{1}{\mu_0 c} \int \mathrm{d}\Omega \, \mathbb{E}\left[|\mathbf{E}_0(\hat{\mathbf{r}}, t)|^2\right] \qquad (4.8)\end{aligned}$$

4.7.3 Random Field Current Model

We seek an upper bound on the number of effective antennas that the transmitter may have. We therefore wish to show that, for some upper bound of N antennas, there does not exist any way of packing antennas, reflectors, amplifiers, and circuitry

into the transmitter such that we achieve an effective antenna count $M > N$.

It is important, then, to determine what current distributions are possible. Otherwise, the space of current distributions we seek to bound would include, for instance, a distribution consisting of one thousand tiny dipole radiators carrying 100 amperes of current each. Such a distribution would be impossible to create in a real mobile device for power reasons.

In this section, we model the current distribution $\mathbf{J}_e(\mathbf{r}, \omega)$ as a Gaussian random field. That is, the value of the current density at any given point at any given time is a vector whose components are random variables; and all the components of the vectors at all points at all times are, jointly, zero-mean complex Gaussian. We will impose some further restrictions in order to motivate our choice of a functional form for the correlation statistics of the Gaussian random field.

A question that we must pose and leave for future work is whether this model captures all the current distributions that it needs to capture. That is, do there exist high-dimensional sub-spaces of current distributions that have measure zero in this random field model, yet achieve a large number of effective antennas?

Bounded Power

We'd like to impose the constraint that the transmitter is using a bounded amount of electrical power. By the conservation of energy, the time-averaged power consumed in the transmitter will be the sum of ohmic losses and radiative losses—that is, resistive dissipation in the conductive materials of the transmitter, and far-field radiated power. We will begin by upper-bounding the ohmic losses only: a necessary but not sufficient condition. Later, we will show that radiative losses are also bounded.

If our current density were constant over some volume of material $\ell \times A$ with resistivity ρ, we might compute the total ohmic power dissipation via Ohm's law and

Pouillet's law as

$$\begin{aligned} P_{\text{ohmic}} &= I^2 R \\ &= |\mathbf{J}_e A|^2 (\rho \ell / A) \\ &= \rho |\mathbf{J}|^2 \ell A \end{aligned}$$

By dividing up the volume of the transmitter into small regions of constant current density and constant material properties, we can compute the ohmic power dissipation for a non-constant current distribution as

$$P_{\text{ohmic}}(t) = \int \mathrm{d}^3\mathbf{r} \rho(\mathbf{r}) |\mathbf{J}_e(\mathbf{r}, t)|^2 \tag{4.9}$$

The dependence of the ohmic dissipation on the resistivity complicates matters. Since we want an upper bound on the number of effective antennas, we may be generous in our assumptions about the transmitter. To maximize the transmitter's freedom to design its current distribution to generate fields, we may take ρ to be a constant equal to the lowest resistivity component of the device, which we denote by ρ_{min}. That way, under any given bound on ohmic losses, the penalty per unit $|\mathbf{J}_e|^2$ is as low as possible, and the currents may be as large as possible.

This suggests the bound

$$\rho_{\text{min}} \int \mathrm{d}^3\mathbf{r} |\mathbf{J}_e(\mathbf{r}, t)|^2 \leq P_{\text{ohmic,max}}$$

This is not quite what we need. If our current distribution is a Gaussian random field, then we cannot impose any finite bound on its squared magnitude at any point. To make our bound sensible, we may average over realizations of the random field:

$$\begin{aligned} \mathbb{E}[P_{\text{ohmic}}(t)] &= \rho_{\text{min}} \int \mathrm{d}^3\mathbf{r}\, \mathbb{E}\left[|\mathbf{J}_e(\mathbf{r}, t)|^2\right] \\ &\leq P_{\text{ohmic,max}} \end{aligned}$$

We have established a functional form for the bounded-ohmic-dissipation constraint:

an upper bound on the expected squared current density. We will show later that the average radiated power is also bounded under these conditions.

Band-limitedness

Since real transmitters contain filters to prevent the amplification and radiation of unwanted frequencies, we may require the current distribution to be a band-limited function of time. So, in a sense, the current distribution varies slowly over time-scales comparable to the inverse bandwidth. We denote the center frequency by $f_0 = \omega_0/2\pi$ and the bandwidth by $1/\Delta t$. (The slowly-varying quantity is actually $e^{-j\omega_0 t}\mathbf{J}_e$, not $\mathbf{J}_e$).

It might seem plausible to further assume that the current distribution is smooth over spatial scales comparable to a wavelength. After all, we are used to treating the electromagnetic fields in a particular band of frequencies as if they are smooth over spatial scales short compared to a wavelength. This is, however, not a correct assumption.

One way to see this is to note that skin effects localize high-frequency alternating currents into very thin sheets. In a copper wire at 2.4 GHz, the current density decays in amplitude by a factor $1/e$ for every 1.33 μm of depth beneath the surface of the metal. Within the metal, then, the current distribution is, indeed, smooth—but over a length scale five orders of magnitude shorter than the wavelength of 12.5 cm! The decay of the current distribution to zero above the surface of the metal may be even more abrupt, depending on the properties of the adjacent insulating dielectric.

Bounded Support

We require the current distribution to be zero outside the volume of the transmitter. (If we were not working in free space, there might be currents on nearby scatterers or reflectors). If the transmitter is rectangular and aligned with the coordinate axes, then let its dimensions be $a_1 \times a_2 \times a_3$.

Isotropy

We choose for the vectorial components of the current density to be independent and identically distributed.

Time-whiteness

We choose for the time correlations to behave as if each vector component at each point in space, considered in isolation as a function of time, is a pulse-amplitude modulated signal with i.i.d. pulse amplitudes and a pulse shape of $p(t) = e^{j\omega_0 t}\,\mathrm{sinc}(\pi t/\Delta t)$:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{J}_e(\mathbf{r},t))_i\,(\mathbf{J}_e(\mathbf{r},t'))_i^*\right] &= \mathbb{E}\left[\left(\sum_{n=-\infty}^{\infty} x_n p(t-n\Delta t)\right)\left(\sum_{m=-\infty}^{\infty} x_m p(t'-m\Delta t)\right)^*\right] \\
&= \sum_{n,m} \mathbb{E}\left[x_n {x_m}^*\right] p(t-n\Delta t)p(t'-m\Delta t)^* \\
&\propto \sum_n p(t-n\Delta t)p(t'-n\Delta t)^* \\
&= e^{j\omega_0\tau} \sum_n \mathrm{sinc}\left(\frac{\pi(t-n\Delta t)}{\Delta t}\right)\mathrm{sinc}\left(\frac{\pi(t'-n\Delta t)}{\Delta t}\right) \\
&= e^{j\omega_0\tau}\,\mathrm{sinc}\left(\frac{\pi\tau}{\Delta t}\right) \qquad (4.10) \\
&= \int \frac{\mathrm{d}\omega}{2\pi} e^{j\omega\tau}\left(\Delta t\, U(\pi/\Delta t - |\omega-\omega_0|)\right)
\end{aligned}$$

We will need this correlation in the Fourier domain:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{J}_e(\mathbf{r},\omega))_i\,(\mathbf{J}_e(\mathbf{r},\omega'))_i^*\right] &= \mathbb{E}\left[\left(\int \mathrm{d}t e^{-j\omega t}\mathbf{J}_e(\mathbf{r},t)\right)_i\left(\int \mathrm{d}t' e^{-j\omega' t'}\mathbf{J}_e(\mathbf{r},t')\right)_i^*\right] \\
&= \int \mathrm{d}t \int \mathrm{d}t' e^{j(\omega' t'-\omega t)}\mathbb{E}\left[(\mathbf{J}_e(\mathbf{r},t))_i\,(\mathbf{J}_e(\mathbf{r},t'))_i^*\right] \\
&\propto \int \mathrm{d}t \int \mathrm{d}t' e^{j(\omega' t'-\omega t)} \int \frac{\mathrm{d}\omega''}{2\pi} e^{j\omega''\tau}\Delta t\, U(\pi/\Delta t - |\omega''-\omega_0|) \\
&= \int \frac{\mathrm{d}\omega''}{2\pi}(2\pi)\delta(\omega''-\omega)(2\pi)\delta(\omega''-\omega')\Delta t\,\cdot \\
&\qquad U(\pi/\Delta t - |\omega''-\omega_0|) \\
&= (2\pi)\delta(\omega-\omega')\Delta t\, U\left(\frac{\pi}{\Delta t} - |\omega-\omega_0|\right)
\end{aligned}$$

The last factor can be identified as the band limit.

Current Correlation Function

To complete our Gaussian random field model, we need to provide an auto-correlation function for the current distribution. There are four ways we might specify this object: in the spatial or the spatial-frequency domain, and in the time or temporal-frequency domain. That is, the arguments may be $(\mathbf{r}, t)$, $(\mathbf{k}, t)$, $(\mathbf{r}, \omega)$, or $(\mathbf{k}, \omega)$. Our assumptions guarantee that all four forms exist, and that we can convert between them.

Because our band-limited assumption applies in the temporal-frequency domain, and our bounded-support assumption applies in the spatial domain, the most convenient form will be the $(\mathbf{r}, \omega)$ form. We will assume that any covariances which are not otherwise constrained are zero. We therefore have

$$\mu_{\mathbf{J}}(\mathbf{r}, \omega) = \mathbb{E}[\mathbf{J}_e(\mathbf{r}, \omega)] = 0 \tag{4.11}$$

$$\begin{aligned} \mathbf{S}_{\mathbf{JJ}}(\mathbf{r}, \omega, \mathbf{r}', \omega') &= \mathbb{E}[\mathbf{J}_e(\mathbf{r}, \omega)\mathbf{J}_e(\mathbf{r}', \omega')^\dagger] \\ &= \underline{\alpha\ \mathbf{1}} \ \underline{(2\pi)\delta(\omega - \omega')\Delta t\, U\left(\frac{\pi}{\Delta t} - |\omega - \omega_0|\right)} \cdot \\ &\quad \underline{(2\pi)^3 \delta^{(3)}(\mathbf{r} - \mathbf{r}')} \ \underline{\prod_{i=1}^{3} U(a_i/2 - |r_i|)} \end{aligned} \tag{4.12}$$

where the underlined factors, in order, correspond to the power bound (normalization), isotropy (i.i.d. of vector components), time-whiteness and band-limitedness (i.i.d. of distinct frequencies), i.i.d. of distinct points in space, and bounded support in space.

We need to perform one more Fourier transformation, this time into the $(\mathbf{k}, \omega)$ domain. We get

$$\mu_{\mathbf{J}}(\mathbf{k}, \omega) = \int \mathrm{d}^3\mathbf{r} e^{-j\mathbf{k}\cdot\mathbf{r}} \mu_{\mathbf{J}}(\mathbf{r}, \omega) = 0 \tag{4.13}$$

$$\tilde{\mathbf{S}}_{\mathbf{JJ}}(\mathbf{k}, \omega, \mathbf{k}', \omega') = \int \mathrm{d}^3\mathbf{r} e^{-j\mathbf{k}\cdot\mathbf{r}} \int \mathrm{d}^3\mathbf{r}' e^{+j\mathbf{k}'\cdot\mathbf{r}'} \mathbf{S}_{\mathbf{JJ}}(\mathbf{r}, \omega, \mathbf{r}', \omega') \tag{4.14}$$

$$\begin{aligned} &= \alpha \mathbf{1} (2\pi)\delta(\omega - \omega')\Delta t\, U\left(\frac{\pi}{\Delta t} - |\omega - \omega_0|\right) \cdot \\ &\quad \prod_{i=1}^{3} (2\pi) a_i \operatorname{sinc}\left(\frac{a_i}{2}(\mathbf{k} - \mathbf{k}')_i\right) \end{aligned} \tag{4.15}$$

4.7.4 Correlation of Electric Fields

We write the two-point auto-correlation of the electric far fields as

$$\begin{aligned}
\mathbf{C}_{\hat{\mathbf{r}}\,t\,\hat{\mathbf{r}}'t'} &= \mathbb{E}\left[\mathbf{E}_0(\hat{\mathbf{r}},t)\mathbf{E}_0(\hat{\mathbf{r}}',t')^\dagger\right] \\
&= \frac{{\mu_0}^2}{(4\pi)^2}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)\int\frac{\mathrm{d}\omega}{2\pi}\int\frac{\mathrm{d}\omega'}{2\pi}\omega\omega' e^{j(\omega t-\omega' t')}\tilde{\mathbf{S}}_{\mathbf{JJ}}(-\omega\hat{\mathbf{r}}/c,\omega,-\omega'\hat{\mathbf{r}}'/c,\omega')\cdot \\
&\qquad (\mathbf{1}-\hat{\mathbf{r}}'\hat{\mathbf{r}}'^T) \\
&= \frac{\alpha{\mu_0}^2}{4}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)(\mathbf{1}-\hat{\mathbf{r}}'\hat{\mathbf{r}}'^T)\int_{\omega_0-\frac{\pi}{\Delta t}}^{\omega_0+\frac{\pi}{\Delta t}}\mathrm{d}\omega\omega^2 e^{j\omega\tau}\Delta t\prod_{i=1}^{3}a_i\,\mathrm{sinc}\left(\frac{a_i\omega}{2c}(\hat{\mathbf{r}}-\hat{\mathbf{r}}')_i\right)
\end{aligned}$$

If $\omega_0 \gg \frac{\pi}{\Delta t}$, then

$$\begin{aligned}
&\approx \frac{\alpha{\mu_0}^2{\omega_0}^2}{4}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)(\mathbf{1}-\hat{\mathbf{r}}'\hat{\mathbf{r}}'^T)\left(\int_{\omega_0-\frac{\pi}{\Delta t}}^{\omega_0+\frac{\pi}{\Delta t}}\mathrm{d}\omega e^{j\omega\tau}\Delta t\right)\prod_{i=1}^{3}a_i\,\mathrm{sinc}\left(\frac{a_i\omega_0}{2c}(\hat{\mathbf{r}}-\hat{\mathbf{r}}')_i\right) \\
&= \left(e^{j\omega_0\tau}\,\mathrm{sinc}(\pi\tau/\Delta t)\right)\frac{\alpha{\mu_0}^2{\omega_0}^2\pi}{2}(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)(\mathbf{1}-\hat{\mathbf{r}}'\hat{\mathbf{r}}'^T)\prod_{i=1}^{3}a_i\,\mathrm{sinc}\left(\frac{a_i\omega_0}{2c}(\hat{\mathbf{r}}-\hat{\mathbf{r}}')_i\right) \\
&= R_{yy}(\tau)\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'},
\end{aligned}$$

with

$$\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} = \left(\frac{\alpha{\mu_0}^2{\omega_0}^2\pi}{2}\right)(\mathbf{1}-\hat{\mathbf{r}}\hat{\mathbf{r}}^T)(\mathbf{1}-\hat{\mathbf{r}}'\hat{\mathbf{r}}'^T)\prod_{i=1}^{3}a_i\,\mathrm{sinc}\left(\frac{a_i\omega_0}{2c}(\hat{\mathbf{r}}-\hat{\mathbf{r}}')_i\right)\ \text{and} \tag{4.16}$$

$$R_{yy}(\tau) = e^{j\omega_0\tau}\,\mathrm{sinc}\left(\frac{\pi\tau}{\Delta t}\right). \tag{4.17}$$

We may recognize the temporal correlation factor $R_{yy}(\tau)$ from (4.10) as the auto-correlation of band-limited white noise.

We have now observed that the two-point function approximately factorizes (as a Kronecker product) into independent spatial and temporal parts. Why is this significant? It indicates that the vector of all the degrees of freedom in our problem can be organized into independent sub-vectors, or "slices", one for each time interval Δt; that the full problem in continuous time can be perfectly reconstructed using the information from these discrete slices, in accordance with the Nyquist-Shannon sampling theorem; and that the joint distribution over the variables in a slice is the

same for every slice.

As a consequence, the total number of degrees of freedom in our population of electric far fields is given by the number of degrees of freedom per slice, times the number of slices. We may now eliminate time from the problem, and consider a single slice.

Total Power

We can now re-visit (4.8) for radiated power.

$$\begin{aligned}
\mathbb{E}[P_{\text{radiated}}(t)] &= \frac{1}{\mu_0 c}\int \mathrm{d}\Omega \,\mathrm{Tr}\, \mathbf{C}_{\hat{\mathbf{r}}\,t\,\hat{\mathbf{r}}\,t} \\
&= \frac{1}{\mu_0 c}\int \mathrm{d}\Omega R_{yy}(0)\,\mathrm{Tr}\,\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}} \\
&= \frac{1}{\mu_0 c}\left(\frac{\alpha {\mu_0}^2 {\omega_0}^2 \pi}{2}\right) V \int \mathrm{d}\Omega\,\mathrm{Tr}(\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T) \\
&= \frac{4\pi^2 \alpha \mu_0 {\omega_0}^2 V}{c},
\end{aligned}$$

where V is the volume $a_1 \cdot a_2 \cdot a_3$ of the device. We can also evaluate the ohmic dissipation of (4.9):

$$\begin{aligned}
\mathbb{E}[P_{\text{ohmic}}(t)] &= \rho_{\min}\int \mathrm{d}^3\mathbf{r}\,\mathbb{E}\left[|\mathbf{J}_e(\mathbf{r},t)|^2\right] \\
&= \rho_{\min}\int \mathrm{d}^3\mathbf{r}\,\mathbb{E}\left[|\mathbf{J}_e(\mathbf{r},t)|^2\right] \\
&= \rho_{\min}\int \mathrm{d}^3\mathbf{r}\int\frac{\mathrm{d}\omega}{2\pi}\int\frac{\mathrm{d}\omega'}{2\pi}e^{+j(\omega'-\omega)t}\mathbb{E}\left[\mathbf{J}_e(\mathbf{r},\omega)^\dagger\mathbf{J}_e(\mathbf{r},\omega')\right] \\
&= \rho_{\min}\int \mathrm{d}^3\mathbf{r}\int\frac{\mathrm{d}\omega}{2\pi}\int\frac{\mathrm{d}\omega'}{2\pi}e^{+j(\omega'-\omega)t}\,\mathrm{Tr}\,\mathbf{S}_{\mathbf{JJ}}(\mathbf{r},\omega',\mathbf{r},\omega) \\
&= \rho_{\min} 3\alpha(2\pi)^3\delta^{(3)}(\mathbf{r}-\mathbf{r})V
\end{aligned}$$

How are we to interpret the infinite quantity $(2\pi)^3\delta^{(3)}(\mathbf{r}-\mathbf{r})$? It seems that the delta-function spatial auto-correlation, which we introduced without offering any justification, has come back to haunt us.

Such a factor may arise naturally via an integral of the form $\int \mathrm{d}^3\mathbf{k}e^{j\mathbf{k}\cdot\mathbf{r}}$ with $\mathbf{r} = 0$. If, rather than integrating from $-\infty$ to ∞, we had instead imposed a cut-off wavenumber

$k_{\max}$ on each dimension, then we would have obtained $(2\pi)^3\delta^{(3)}(\mathbf{r}-\mathbf{r}) \to (2k_{\max})^3$. So we may write

$$\mathbb{E}[P_{\text{ohmic}}(t)] = 24\alpha {k_{\max}}^3 V \rho_{\min}$$

Combining the results of this subsection for ohmic and radiative dissipation, we have

$$\mathbb{E}[P_{\text{total}}(t)] = \alpha V \left(\underbrace{24{k_{\max}}^3 \rho_{\min}}_{\text{ohmic}} + \underbrace{\frac{4\pi^2 \mu_0 {\omega_0}^2}{c}}_{\text{radiative}} \right) \tag{4.18}$$

For 2.4 GHz signals, copper conductors, and $k_{\max} \approx \frac{1}{1\,\mu\text{m}}$, the ohmic term in the parentheses is $4 \times 10^{11}\,\Omega/\text{m}^2$, while the radiative term is only $4 \times 10^7\,\Omega/\text{m}^2$. These quantities are in proportion to the power consumed by the two mechanisms. Evidently, the spatially-white Gaussian random field current distribution is a terrible way to design a mobile device: only 0.1% of the total power radiates into the far field!

Fortunately for us, efficiency is not a goal in this context. If 99.9% of the source power of the Gaussian random field is driving non-radiative modes, the radio designer can simply choose a sub-space of current distributions that lies in the orthogonal complement of the non-radiating sub-space—just as the musician may prepare a composition that only involves striking the piano on the keys or strings.

4.7.5 Correlation Spectrum

If the two-point function $\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'}$ were a simple covariance matrix, then the number of spatial degrees of freedom would be determined by its number of large eigenvalues. Unlike a square matrix, however, whose rows and columns are indexed by natural numbers in some range $1 \ldots n$, the "rows" and "columns" of the two-point function are indexed by vectors on the unit sphere. It would seem that our "matrix" has a continuum of rows and columns.

Viewed as a linear operator, the two-point function acts on the space of far-field

distributions, and returns a far-field distribution:

$$\begin{aligned} \mathbf{C}\colon \mathcal{H} &\to \mathcal{H} \\ (\mathbf{C}[\mathbf{x}])_{\hat{\mathbf{r}}} &= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \mathbf{x}_{\hat{\mathbf{r}}'} \end{aligned} \tag{4.19}$$

Dealing with a continuum of rows and columns is less than ideal. We can improve matters considerably by restricting to square-integrable far-field distributions—that is, those which contain finite radiated power:

$$\mathcal{H} = \left(L^2(S^2, \mathrm{d}\Omega/4\pi)\right)^3$$

$L^2(\mathcal{X}, \mu)$ denotes the Hilbert space of functions from the measure space S to $\mathcal{C}$ which are square-integrable under integration measure μ. Properly speaking, the objects in this space are not functions, but equivalence classes of functions which are equal almost everywhere. In our case, the measure space $\mathcal{X}$ is the two-sphere S^2, and the integration measure is normalized solid angle on the sphere. $\mathcal{H}$ consists of three copies of $L^2(\mathcal{X}, \mu)$, since our far-field distributions assign a three-dimensional vector to each point on the sphere.

Note 2. Valid far-field distributions do not permit a longitudinal polarization $\mathbf{E} \parallel \hat{\mathbf{r}}$. These transverse-polarized distributions form a linear sub-space of $\mathcal{H}$.

Note 3. The inner product on $\mathcal{H}$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}} = \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathbf{y}^{\dagger}_{\hat{\mathbf{r}}} \mathbf{x}_{\hat{\mathbf{r}}}$$

Separability

The advantage of this choice of space is that $\mathcal{H}$ admits a countable basis: the (vectorial) spherical harmonic functions. That is, we can represent any square-integrable far-field

distribution $\mathbf{x}_{\hat{\mathbf{r}}}$ via a semi-infinite sequence of coefficients:

$$\mathbf{x}_{\hat{\mathbf{r}}} = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{+\ell} \mathbf{x}_\ell^m Y_\ell^m(\hat{\mathbf{r}})$$

We may place these coefficients into a linear sequence by sorting first in order of increasing ℓ, then in order of increasing m, then by vector component in $\mathbb{C}^3$.

Properties of the Two-Point Function

By construction, the two-point function $\mathbf{C}\colon \mathcal{H} \to \mathcal{H}$ given by (4.16) is a self-adjoint, non-negative linear operator.

To simplify notation, we let

$$A = V\alpha{\mu_0}^2{\omega_0}^2\pi = \frac{\mu_0 c}{4\pi}\mathbb{E}[P_{\text{radiated}}]$$

Then we have the following results. Proofs are given in Appendix B.

Theorem 4.7.1 $\mathbf{C}$ *is trace class (and hence compact) with* $\operatorname{Tr}_{\mathcal{H}} \mathbf{C} = A$.

This immediately implies that the Schatten norms (p-norms of the sequence of eigenvalues of $\mathbf{C}$) with $p \geq 1$ are finite and $\leq A$, and hence the operator is bounded ($p = \infty$). However, we obtain tighter bounds by considering each norm separately.

Theorem 4.7.2 *The sum of squared eigenvalues of* $\mathbf{C}$ *is*

$$\operatorname{Tr}_{\mathcal{H}} \mathbf{C}^2 \leq \frac{A^2}{3}$$

Theorem 4.7.3 *The largest eigenvalue of* $\mathbf{C}$ *satisfies*

$$\|\mathbf{C}\|_{\mathcal{H},op} \leq \frac{A}{3}$$

It follows that the eigenvalues λ of $\mathbf{C}$ lie in the interval

$$0 \leq \lambda \leq \|\mathbf{C}\|_{\mathcal{H},\mathrm{op}} \leq \frac{A}{3},$$

with units of ohm·watt.

Consider the basis of elements $\mathbf{f}_{\ell mn} = Y_\ell^m \otimes e_n$ whose angular dependence is a spherical harmonic function Y_ℓ^m and whose polarization is a constant vector e_n. The elements of order up to ℓ span a sub-space $\mathcal{G}_\ell \subset \mathcal{H}$ of dimension $3\ell^2$. If we evaluate the trace of $\mathbf{C}$ restricted to this subspace, we can compute

$$\left\| \mathbf{C} - \mathbf{C}|_{\mathcal{G}_\ell} \right\|_{\mathcal{H},op} \leq A - \mathrm{Tr}_{\mathcal{G}_\ell} \mathbf{C}$$

If this quantity is small, then we may conclude that the amount of radiated power found in the complementary subspace $\mathcal{G}_\ell^\perp$ is small.

A laborious calculation involving the absolutely-convergent Taylor series expansion of the expression in (4.16) to order 24 in a_1, a_2, and a_3 allows us to compute $\mathrm{Tr}_{\mathcal{G}_\ell}$. This is what we plotted in Figure 1-2, which we reproduce as Figure 4-4 here for comparison with the results for other, non-cubical transmitter shapes (Figures 4-5, 4-6).

Because the two-point function is a bounded self-adjoint linear operator on a separable Hilbert space, it is quasi-diagonal [122, 123], and its spectrum may be obtained by the finite section method [124]:

There is guaranteed to exist a special nested sequence of finite-dimensional linear sub-spaces $\{\mathcal{H}_n\}$ converging to $\mathcal{H}$ (henceforth, the "filtration"). The restriction of the two-point function onto these sub-spaces gives a sequence $\{\mathbf{C}_n\}$ of finite matrices. The eigenvalue spectra of these finite matrices form a sequence of sets $\{\sigma(\mathbf{C}_n)\}$. The eigenvalues of $\mathbf{C}$ correspond to the limit points of Cauchy sequences constructed from this sequence of sets.

Put another way, as we examine larger and larger finite-dimensional approximations to the infinite-dimensional operator $\mathbf{C}$, it will eventually be the case that (some of) the eigenvalues of each approximation are arbitrarily close to eigenvalues of $\mathbf{C}$, and

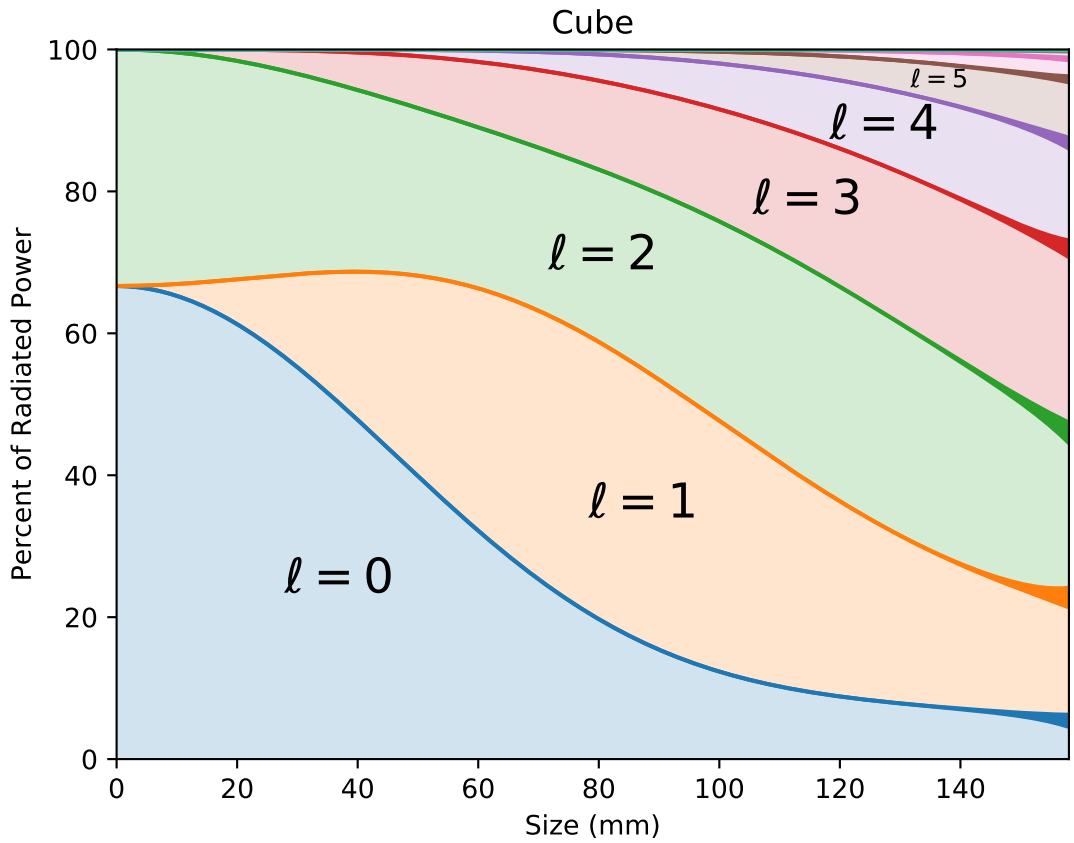


Figure 4-4: Proportion of 2.4 GHz radiated energy in each subspace $\mathcal{G}_\ell$ for a small cube-shaped ($a_1 = a_2 = a_3 > 0$) transmitter or receiver, as a function of size a_3. The thickening of the lines at the right edge of the plot represents the difference between the Taylor series at orders 22 and 24.

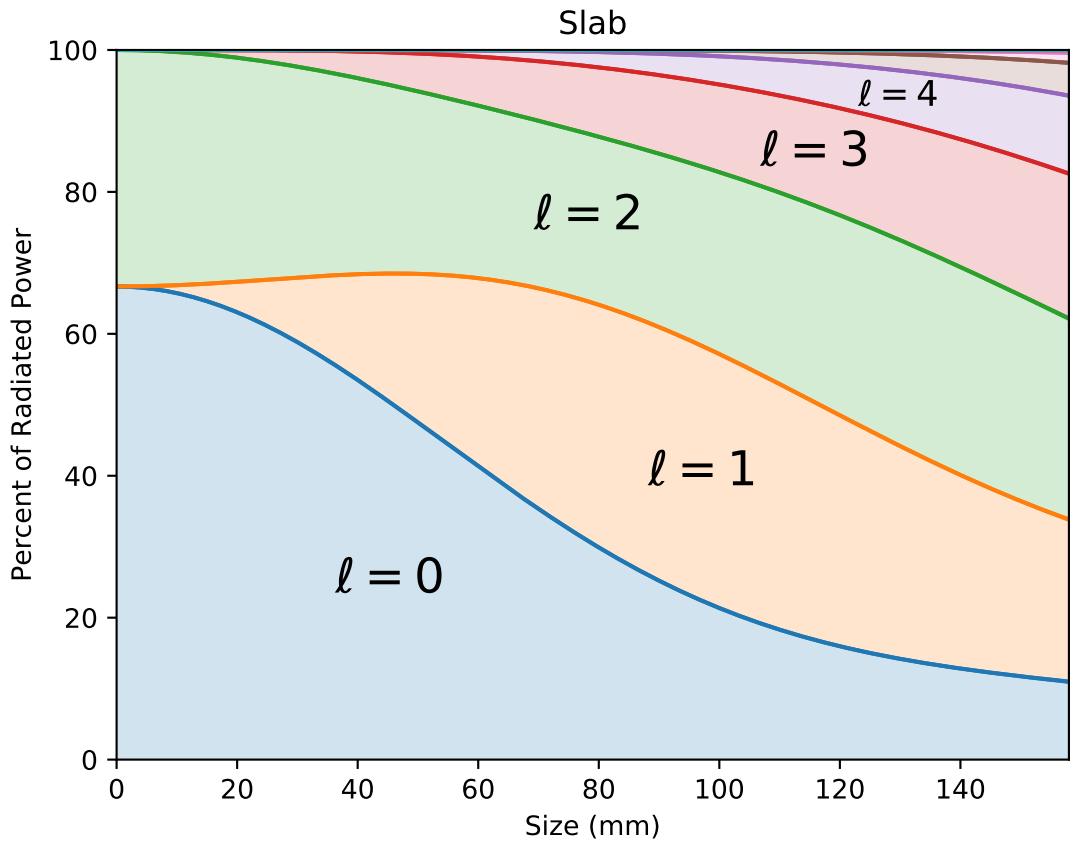


Figure 4-5: Proportion of 2.4 GHz radiated energy in each subspace $\mathcal{G}_\ell$ for a small slab-shaped ($a_1 \to 0$, $a_2 = a_3 > 0$) transmitter or receiver, as a function of size a_3.

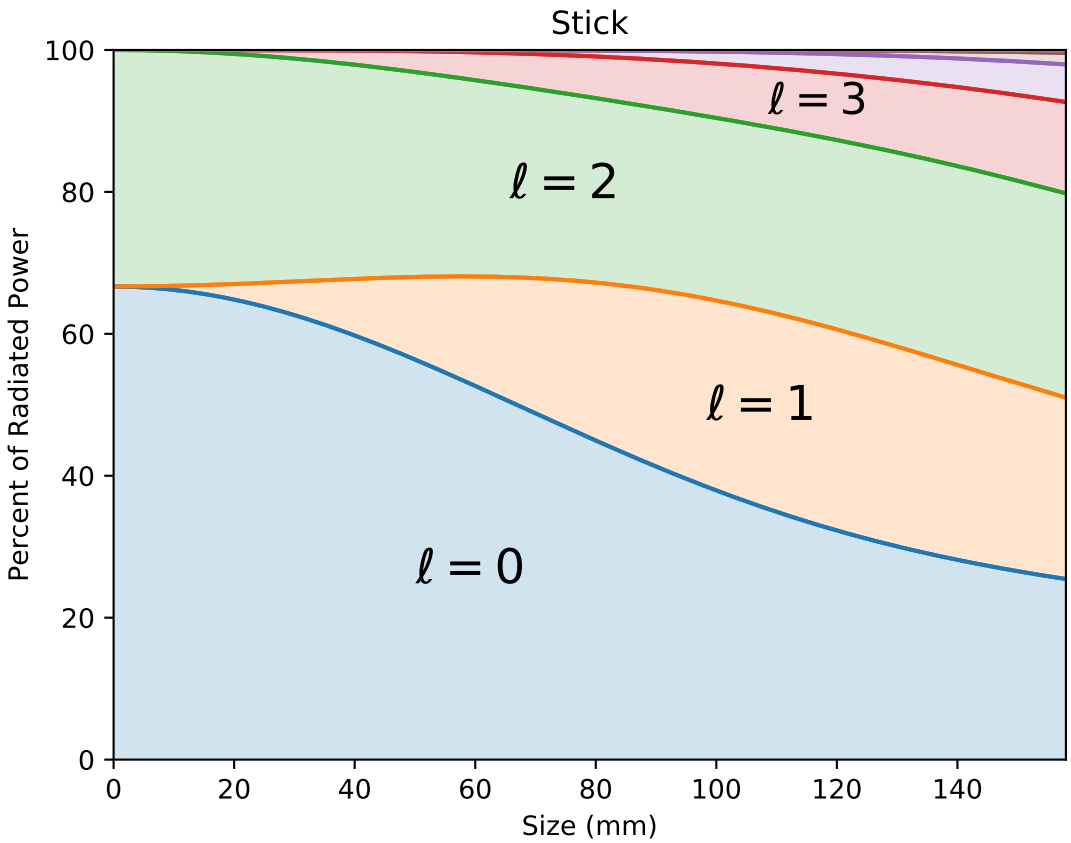


Figure 4-6: Proportion of 2.4 GHz radiated energy in each subspace $\mathcal{G}_\ell$ for a small stick-shaped ($a_1 = a_2 \to 0$, $a_3 > 0$) transmitter or receiver, as a function of size a_3.

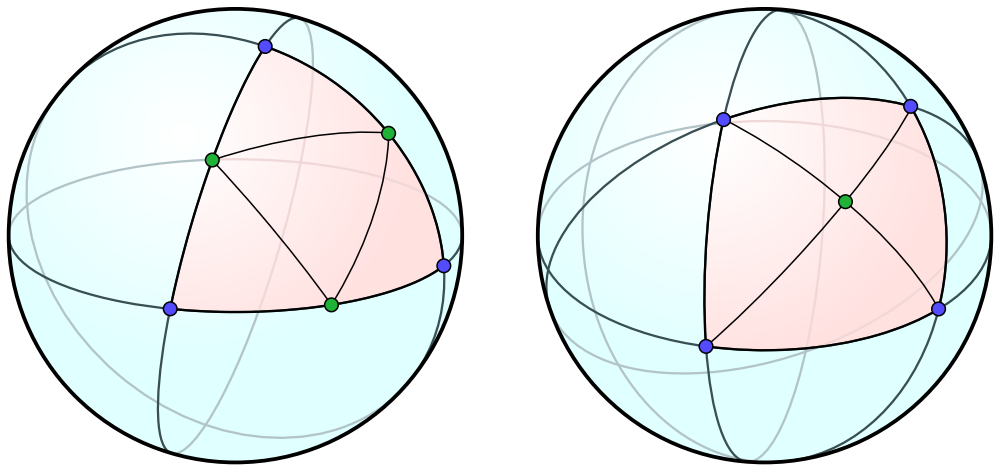


Figure 4-7: Sub-division of spherical polygons. The vertices (blue) at one level of refinement are used to construct new vertices (green), which are connected to form the triangles of the next level of refinement. The non-triangular case (right) appears only at the first level of refinement.

each eigenvalue of $\mathbf{C}$ will eventually appear in this way.

The trouble is finding this filtration (special nested sequence of sub-spaces). Hansen [123] offers a tridiagonalization algorithm for this purpose.

Filtration Heuristic

For our purposes, we will employ a heuristic choice of nested sub-spaces, motivated by the idea of approximating spherical integrals by Riemann sums over the faces of a geodesic polyhedron. We will provide numerical evidence of convergence.

Let p_0 be the truncated icosahedron inscribed in the unit sphere, and let the polyhedron p_{n+1} be constructed from p_n by the following sub-division procedure, as depicted in Figure 4-7: For each triangular face of p_n, construct the midpoints of its edges, and connect them to form four similar triangles. For each $n > 3$-gon face, construct the arithmetic mean of the vertices, and connect it with the n original vertices to form n triangles. In either case, normalize each new vertex so that it lies on the unit sphere.

We now convert the inscribed polygons of p_n into their corresponding spherical polygons. That is, for each edge of p_n, we draw the great-circle arc connecting its endpoints. The vertices of p_n, together with the arcs connecting them, form a spherical

polyhedron s_n that sub-divides the sphere.

Let

$$\mathcal{H}_n \subset \mathcal{H} \tag{4.20}$$

be the vector sub-space of square-integrable functions $S^2 \to \mathcal{C}^3$ which are piecewise constant on the faces of s_n. We will denote the value of such a function $\mathbf{x} \in \mathcal{H}_n$ on face m of s_n by $\mathbf{x}_m$.

Note 4. By the definition of L^2 spaces, two functions which are equal almost everywhere are considered equal in the function space. The edges and vertices of s_n form a set of measure 0 in the sphere, so w.l.o.g. we may require that the values of the functions in $\mathcal{H}_n$ on edges and vertices of s_n should be equal to the means of the values on the adjacent faces.

The polyhedra p_n, s_n and the sub-spaces $\mathcal{H}_n$ are characterized by the statistics in Table 4.1.

	$n = 0$	$n > 0$
Vertices	60	$45/2 \cdot 4^n + 2$
Edges	90	$135/2 \cdot 4^n$
Faces	32	$45 \cdot 4^n$
Dimension	96	$135 \cdot 4^n$

Table 4.1: Results of sub-division procedure.

Every function which is piecewise-constant over a coarser sub-division s_n is also piecewise constant over each finer sub-division s_m with $m > n$. Thus, our sub-spaces $\mathcal{H}_n$ are nested.

We do not offer a proof that the sub-spaces $\mathcal{H}_n$ converge to $\mathcal{H}$; but one might proceed as follows. Every continuous bounded function on a compact domain may be approximated arbitrarily closely (in L^2 norm) by a piecewise-constant function, provided that the mesh of faces nowhere fails to become arbitrarily fine. By this reasoning, the projections onto $\mathcal{H}_n$ of all of the spherical harmonic basis functions up to a given order converge uniformly in L^2 norm to their respective smooth limits as

$n \to \infty$. Since the spherical harmonic functions are a complete basis for $\mathcal{H}$, and our piecewise-constant functions are a complete basis for the spherical harmonic functions, $\mathcal{H}_n \to \mathcal{H}$.

Now let t_{nm} be the mth face of s_n, and let $\hat{\mathbf{r}}_{nm}$ be the (spherical) centroid of t_{nm}. Let $M_n(\hat{\mathbf{r}})$ be the set of faces in s_n which contain $\hat{\mathbf{r}}$, and let $m_n(\hat{\mathbf{r}})$ be an arbitrary pick from $M_n(\hat{\mathbf{r}})$. The inner product on the sub-space $\mathcal{H}_n$ is then given by

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}_n} &= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathbf{y}^{\dagger}_{m_n(\hat{\mathbf{r}})} \mathbf{x}_{m_n(\hat{\mathbf{r}})} \\
&= \sum_m \left(\int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \right) \mathbf{y}^{\dagger}_m \mathbf{x}_m \\
&\triangleq \sum_m w_{nm} \mathbf{y}^{\dagger}_m \mathbf{x}_m
\end{aligned}$$

and the natural injective mapping from $\mathcal{H}_n \to \mathcal{H}$ is given by

$$\begin{aligned}
B_n &\colon \mathcal{H}_n \to \mathcal{H} \\
(B_n[\mathbf{x}])_{\hat{\mathbf{r}}} &= \operatorname*{mean}_{m \in M_n(\hat{\mathbf{r}})} \mathbf{x}_m = \mathbf{x}_{m_n(\hat{\mathbf{r}})},
\end{aligned}$$

where the second equality holds because the two functions are equal almost everywhere in $\hat{\mathbf{r}}$.

Let

$$\mathbf{e}_{nm\ell} \in \mathcal{H}_n \tag{4.21}$$

denote orthonormal basis elements of $\mathcal{H}_n$, and let e_i denote the standard basis for $\mathbb{C}^3$. In particular, $\mathbf{e}_{nm\ell} \colon S^2 \to \mathbb{C}^3$ is a function that is zero unless $\hat{\mathbf{r}} \in t_{nm}$, where it is equal to the constant vector $(1/\sqrt{w_{nm}})e_\ell$. This ensures that $\langle \mathbf{e}_{nm\ell}, \mathbf{e}_{nij} \rangle_{\mathcal{H}} = \delta_{mi}\delta_{\ell j}$, in terms of the Kronecker delta. With this basis, we can write

$$(B_n[\mathbf{x}])_{\hat{\mathbf{r}}} = \operatorname*{mean}_{m \in M_n(\hat{\mathbf{r}})} \mathbf{x}_m = \mathbf{x}_{m_n(\hat{\mathbf{r}})},$$

The adjoint operator of $B_n^\dagger \colon \mathcal{H} \to \mathcal{H}_n$ given by

$$(B_n^\dagger \mathbf{y})_m = \frac{1}{w_{nm}} \int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathbf{y}_{\hat{\mathbf{r}}} = \frac{\int_{t_{nm}} \mathrm{d}\Omega(\hat{\mathbf{r}}) \mathbf{y}_{\hat{\mathbf{r}}}}{\int_{t_{nm}} \mathrm{d}\Omega(\hat{\mathbf{r}})}$$

$B_n^\dagger$ is the "averaging" map that assigns to each constant region of its output the average value of the input function over the same region. It is straightforward to show that $B_n^\dagger B_n = I_n$, and $P_n \colon \mathcal{H} \to \mathcal{H} = B_n B_n^\dagger$ is an orthogonal projection onto $\mathcal{H}_n$.

Now let $\mathbf{C}_n \colon \mathcal{H}_n \to \mathcal{H}_n$ be the finite-dimensional matrix formed by restricting $\mathbf{C}$ to $\mathcal{H}_n$ as $\mathbf{C} = B_n^\dagger \mathbf{C} B_n$. We first compose $\mathbf{C}$ with B_n to get

$$(\mathbf{C} B_n \mathbf{x})_{\hat{\mathbf{r}}} = \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} (B_n \mathbf{x})_{\hat{\mathbf{r}}'} = \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \mathbf{x}_{m_n(\hat{\mathbf{r}}')}$$

Next, we compose $B_n^\dagger$ with the result to get

$$\begin{aligned}
(B_n^\dagger \mathbf{C} B_n \mathbf{x})_m &= \frac{1}{w_{nm}} \int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} (\mathbf{C} B_n \mathbf{x})_{\hat{\mathbf{r}}} \\
&= \frac{1}{w_{nm}} \int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \mathbf{x}_{m_n(\hat{\mathbf{r}}')} \\
&= \frac{1}{w_{nm}} \sum_\ell \left(\int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int_{t_{n\ell}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \right) \mathbf{x}_\ell \\
&= \sum_\ell w_{n\ell} \left(\operatorname*{mean}_{\hat{\mathbf{r}} \in t_{nm}} \operatorname*{mean}_{\hat{\mathbf{r}}' \in t_{n\ell}} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \right) \mathbf{x}_\ell
\end{aligned}$$

Taking the inner product with a test vector, we get

$$\langle B_n^\dagger \mathbf{C} B_n \mathbf{x}, \mathbf{y} \rangle_{\mathcal{H}_n} = \sum_m w_{nm} \mathbf{y}_m^\dagger \sum_\ell w_{n\ell} \left(\operatorname*{mean}_{\hat{\mathbf{r}} \in t_{nm}} \operatorname*{mean}_{\hat{\mathbf{r}}' \in t_{n\ell}} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \right) \mathbf{x}_\ell$$

Finally, we can evaluate the matrix elements of the desired restricted two-point matrix

by choosing $\mathbf{x} = \mathbf{e}_{nm\ell}$ and $\mathbf{y} = \mathbf{e}_{nij}$:

$$
\begin{aligned}
(\mathbf{C}_n)_{(3m+\ell),(3i+j)} &= \langle B_n^\dagger \mathbf{C} B_n \mathbf{e}_{nm\ell}, \mathbf{e}_{nij} \rangle_{\mathcal{H}_n} \\
&= \sum_p w_{np} (\mathbf{e}_{nij})_p^\dagger \sum_q w_{nq} \left(\operatorname*{mean}_{\hat{\mathbf{r}} \in t_{np}} \operatorname*{mean}_{\hat{\mathbf{r}}' \in t_{nq}} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \right) (\mathbf{e}_{nm\ell})_q \\
&= \sqrt{w_{ni} w_{nm}} e_j^\dagger \left(\operatorname*{mean}_{\hat{\mathbf{r}} \in t_{ni}} \operatorname*{mean}_{\hat{\mathbf{r}}' \in t_{nm}} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \right) e_\ell \\
&= \sqrt{w_{ni} w_{nm}} \left(\operatorname*{mean}_{\hat{\mathbf{r}} \in t_{ni}} \operatorname*{mean}_{\hat{\mathbf{r}}' \in t_{nm}} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \right)_{j\ell} \\
&\approx \sqrt{w_{ni} w_{nm}} \left(\mathbf{C}_{\hat{\mathbf{r}}_{ni}\,\hat{\mathbf{r}}_{nm}} \right)_{j\ell}
\end{aligned}
$$

where the approximation is justified by the fact that the element $\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'}$ is a smooth function of its two indices, and our sub-division will eventually become sufficiently fine that the approximation error becomes negligible.

Using $\omega_0 = 2.4\,\mathrm{GHz}$ and transmitter dimensions of $7.1\,\mathrm{mm} \times 77.8\,\mathrm{mm} \times 158.1\,\mathrm{mm}$, we may finally evaluate this finite matrix, compute its spectrum, and produce the following results demonstrating the convergence of the largest eigenvalues (power distribution over modes) and eigenvectors (far-field mode shapes) (Figures 4-8 through 4-12).

The eigenvectors displayed in Figures 4-9 through 4-12 take the form of far-field radiation patterns. On the left of each subplot, a heat-map of field intensity over spherical coordinates (elevation, el; azimuth, az) is overlayed with polarization arrows. Because $\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'}$ is real and the eigenvalues are real, the eigenvectors can be chosen to be real, and hence the polarization arrows do not need to be split into real and imaginary parts. On the right of each subplot, the distribution of radiated energy over the sphere is rendered as a wireframe, with radial distance from the origin representing (relative) power. A bounding box is drawn in black to indicate the orientation and aspect ratio of the transmitter; the overall scale of this box is meaningless.

The number of far-field modes with power no less than $-50\,\mathrm{dB}$ relative to the strongest mode is 77. This, finally, is our estimate of the maximum effective number of antennas of the radio transmitter.

For comparison, the surface area of the device in units of $(\lambda/2\pi)^2$ is 70.7. (We will

call these units "square radians"). Is this a coincidence, or is this evidence in favor of the sampling-based heuristic argument in the previous section?

Exploring the parameter space of device dimensions in the range of 4 mm to 200 mm in each axis, we find that ratio of antennas-per-square-radian for rectangular devices is a decreasing function of each dimension. It peaks at 45 for the case of a 4 mm cube, which has just 11 effective antennas, and reaches a minimum value on the search space of 0.45 for a 200 mm cube, which has 274 effective antennas. For cubes, the antennas-per-square-radian ratio first drops below 2 at a side-length of 36 mm, or 1.8 radians. For 4 mm-thick tablets of at least 70 mm square, the ratio is less than 2. For tablets of at least 135 mm square, the ratio is less than 1.

The take-away, then, is that a radio transmitter requires $\mathcal{O}(1)$ square radians of bounding-box surface area per effective antenna. For mobile devices of typical dimensions on the 2.4 GHz band, the constant factor is between 1/2 and 3/2.

4.8 Conclusion

In this chapter, we have provided details of **REST**: a geometric framework for modeling powerful de-anonymizing listeners at the physical layer. REST is constructed to use all, and only, the degrees of freedom that the sender does not control. REST does not require randomization, is efficient, and provides a simple dissimilarity metric.

We have also outlined a countermeasure against PHY de-anonymizers, and we have given bounds on the effectiveness of a class of similar countermeasures involving reconfigurable antennas. As wireless networks move to higher frequencies, e.g. 60 GHz, our bounds loosen, giving grounds for optimism that wireless networks which preserve user privacy are growing more feasible with time.

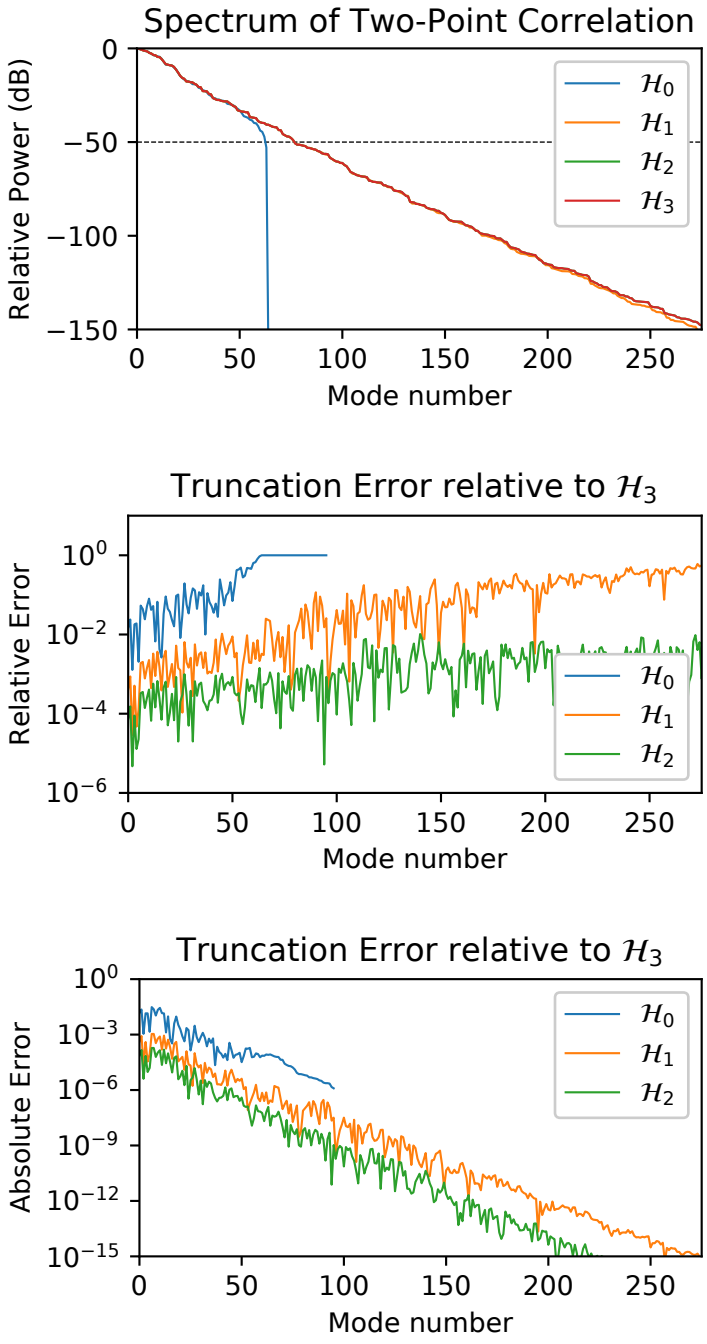


Figure 4-8: Convergence of two-point correlation spectrum estimates at four levels of refinement, corresponding to square matrices of dimension 96, 540, 2160, and 8640. Non-negligible modes have a relative power no less than $-50\,\mathrm{dB}$. Using the results for $\mathcal{H}_3$ as ground truth, the relative and absolute errors in the estimated eigenvalues are shown in the lower plots.

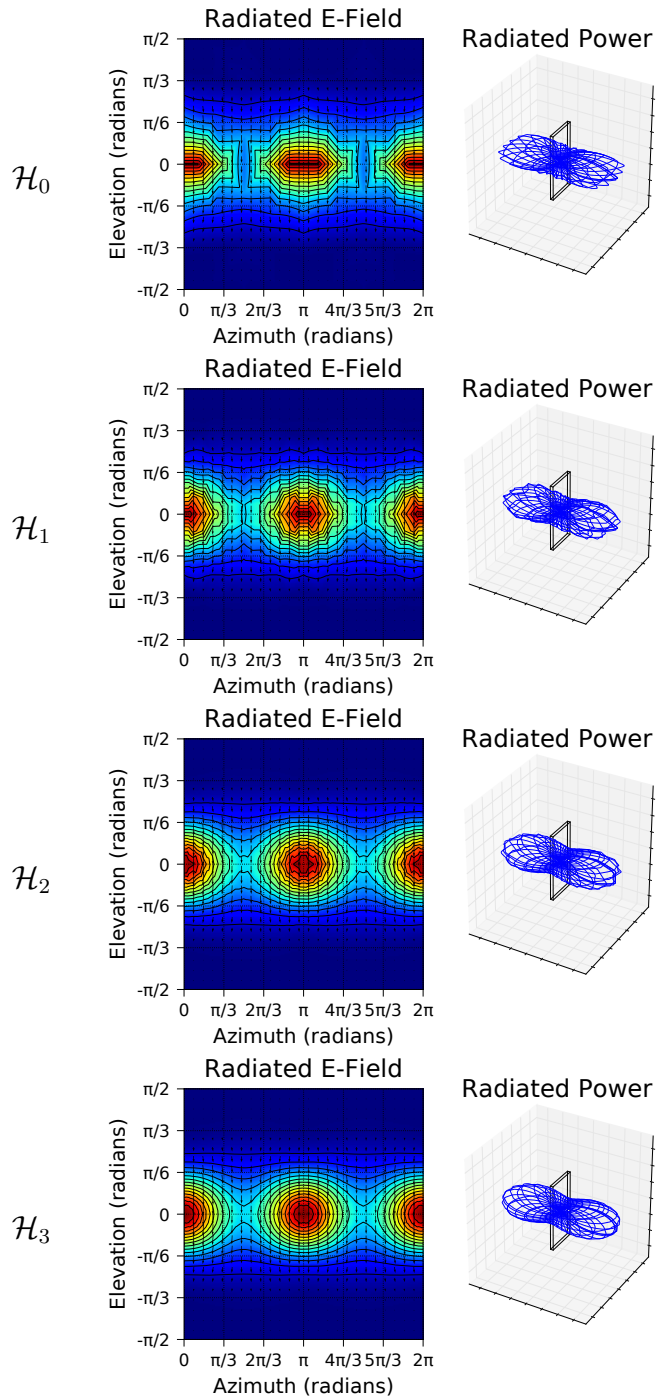


Figure 4-9: Convergence of two-point correlation eigenvector estimates for the first eigenvector.

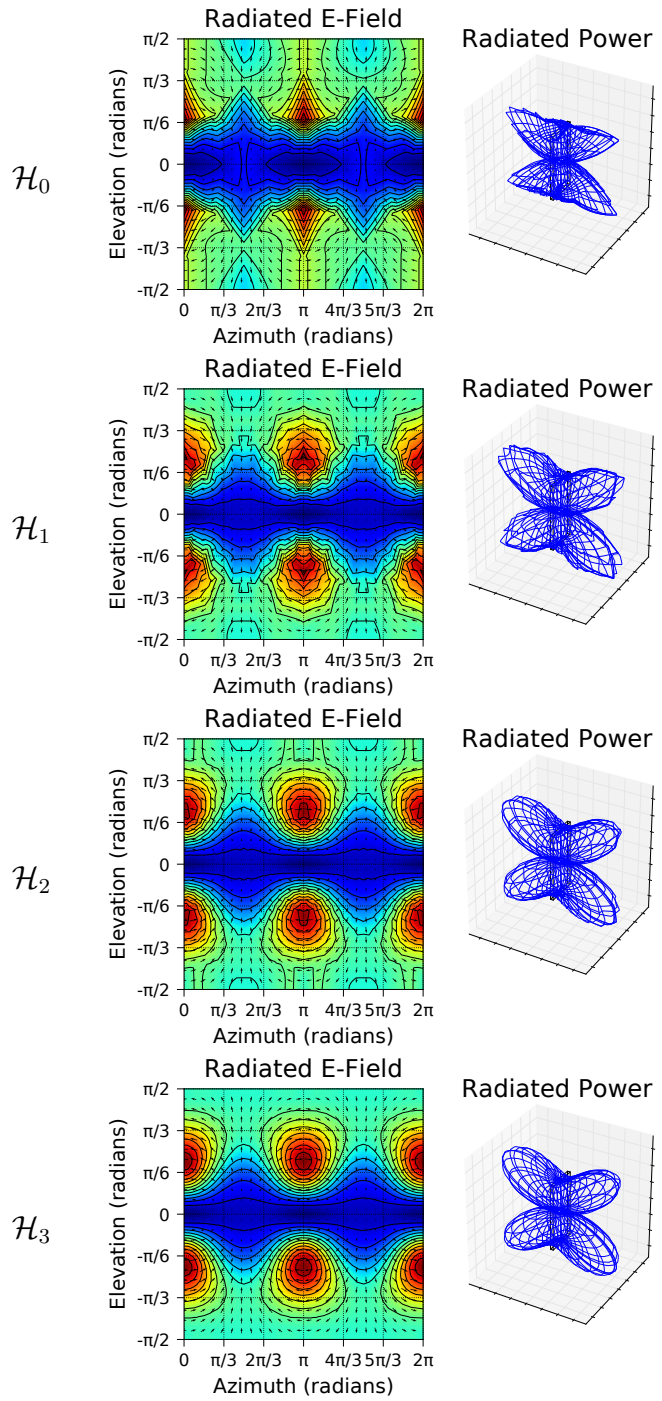


Figure 4-10: Convergence of two-point correlation eigenvector estimates for the second eigenvector.

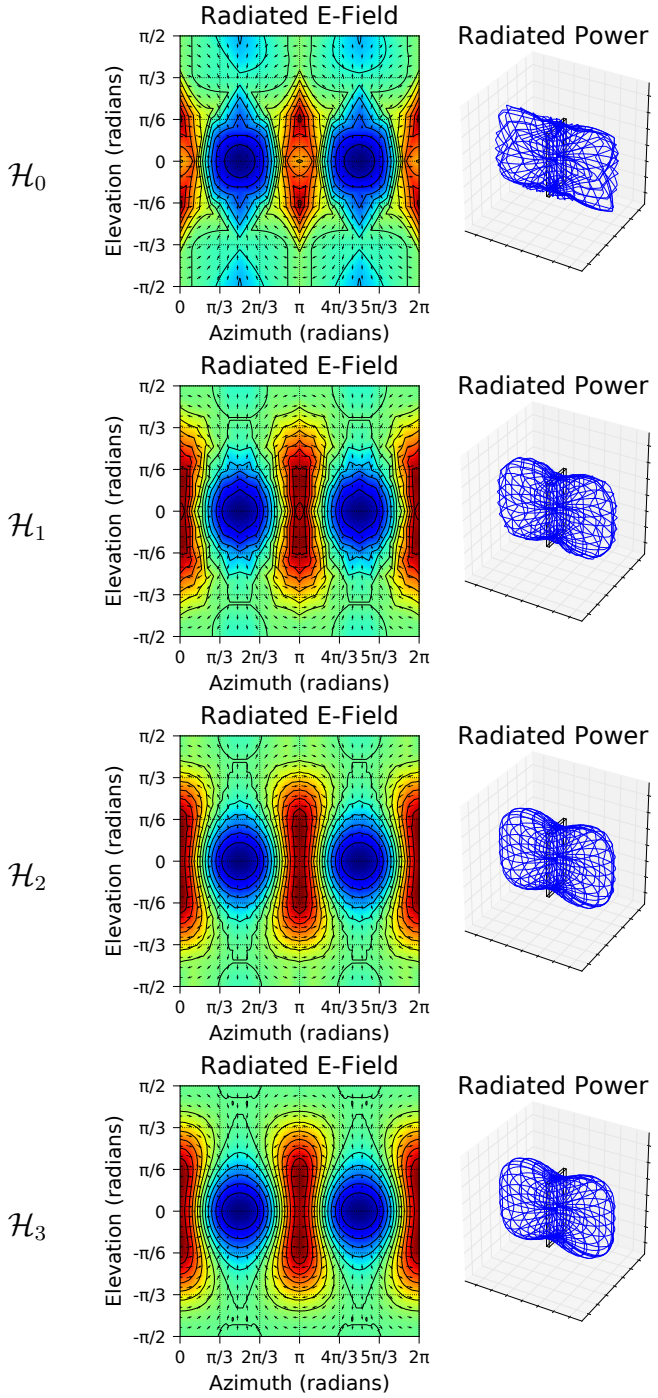


Figure 4-11: Convergence of two-point correlation eigenvector estimates for the third eigenvector.

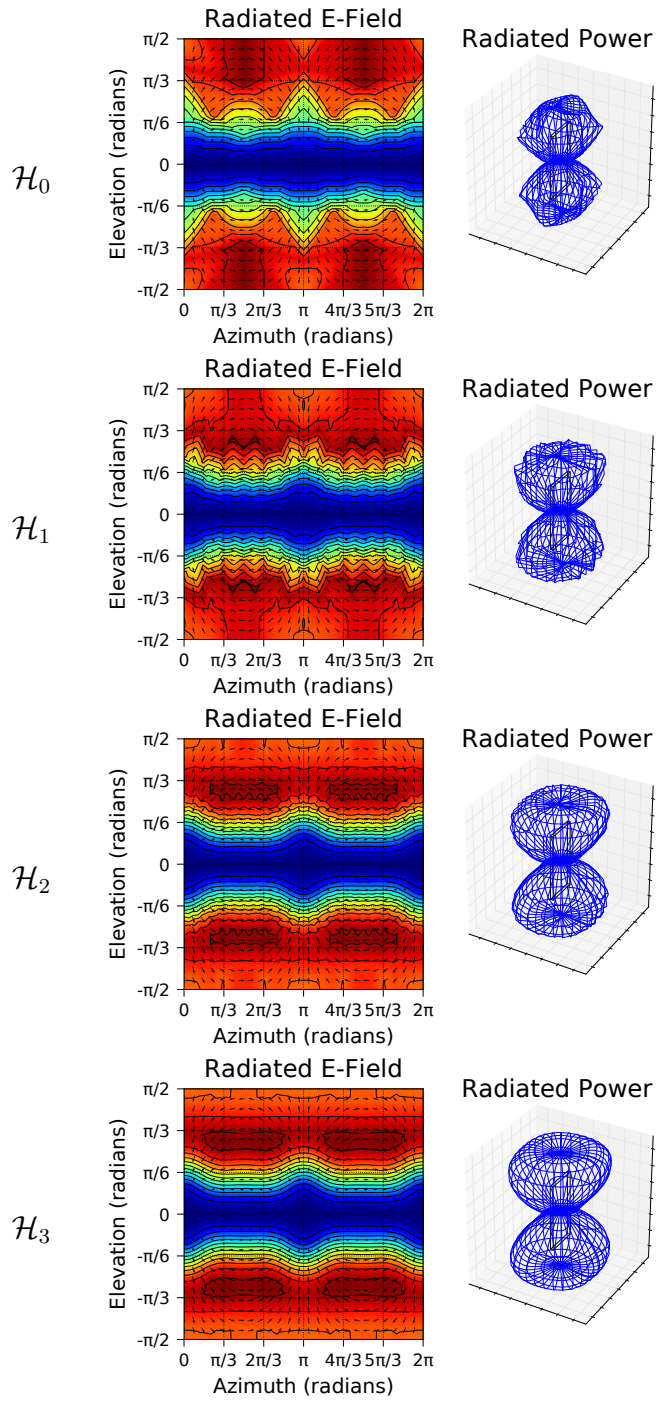


Figure 4-12: Convergence of two-point correlation eigenvector estimates for the fourth eigenvector.

Bibliography

[1] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, Mar. 2006.

[2] Science & Technology Facilities Council, "HSL. A collection of Fortran codes for large scale scientific computation." 2014. [Online]. Available: http://www.hsl.rl.ac.uk/

[3] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Transactions on Information Theory*, 2011.

[4] J. Tellado and J. M. Cioffi, "Peak Power Reduction for Multicarrier Transmission," Stanford University, Tech. Rep., 1998.

[5] A. Aggarwal and T. H. Meng, "Minimizing peak-to-average power ratio of OFDM signals via convex optimization," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 3099–3110, Aug. 2006.

[6] P. Iannucci, J. Perry, H. Balakrishnan, and D. Shah, "No symbol left behind: A link-layer protocol for rateless codes," in *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM*, 2012.

[7] C. Matte, M. Cunche, F. Rousseau, and M. Vanhoef, "Defeating MAC Address Randomization Through Timing Attacks," in *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks - WiSec '16*, 2016, pp. 15–20.

[8] P. Iannucci and H. Balakrishnan, "Challenges to PHY anonymity for Wi-Fi," in *Proceedings of the 10th ACM Conference on Security and Privacy in Wireless and Mobile Networks, WiSec 2017*, 2017.

[9] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[10] C. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[11] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[12] D. Tse and V. Pramod, *Fundamentals of wireless communication.* Cambridge University Press, 2005.

[13] V. Sethuraman, L. Wang, B. Hajek, and A. Lapidoth, "Low-SNR Capacity of Noncoherent Fading Channels," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1555–1574, Apr. 2009.

[14] L. Zheng and D. N. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Transactions on Information Theory*, 2002.

[15] A. Lapidoth and S. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2426–2467, Oct. 2003.

[16] M. Dörpinghaus, G. Koliander, G. Durisi, E. Riegler, and H. Meyr, "Oversampling Increases the Pre-Log of Noncoherent Rayleigh Fading Channels," May 2014. [Online]. Available: http://arxiv.org/abs/1405.0370

[17] V. I. Morgenshtern, G. Durisi, and H. Bölcskei, "The SIMO pre-log can be larger than the SISO pre-log," in *IEEE International Symposium on Information Theory - Proceedings.* IEEE, Jun. 2010, pp. 320–324.

[18] Y. Liang and V. Veeravalli, "Capacity of Noncoherent Time-Selective Rayleigh-Fading Channels," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3095–3110, Dec. 2004.

[19] U. Salim and D. Slock, "Asymptotic capacity of underspread and overspread doubly selective MIMO channels," in *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications.* IEEE, Sep. 2008, pp. 1–5.

[20] G. Durisi, U. G. Schuster, H. Bölcskei, and S. Shamai, "Noncoherent capacity of underspread fading channels," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 367–395, Jan. 2010.

[21] I. Bergel and S. Benedetto, "Bounds on the capacity of OFDM underspread frequency selective fading channels," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6446–6470, Jul. 2012.

[22] G. Durisi, V. I. Morgenshtern, and H. Bolcskei, "On the sensitivity of continuous-time noncoherent fading channel capacity," pp. 6372–6391, Oct. 2012.

[23] M. V. Burnashev, "Data Transmission over a Discrete Channel with Feedback. Random Transmission Time," *Problems of Information Transmission*, vol. 12, no. 4, pp. 250–265, 1976.

[24] S. Verdú and S. Shamai, "Variable-Rate Channel Capacity," *IEEE TRANSACTIONS ON INFORMATION THEORY*, vol. 56, no. 6, 2010.

[25] H. Holma and A. Toskala, *LTE for UMTS - OFDMA and SC-FDMA based radio access*. Chichester, West Sussex, England: John Wiley & Sons, 2009.

[26] S. Dawkins, G. Montenegro, M. Kojo, V. Magret, and N. Vaidya, "RFC3155, End-to-end Performance Implications of Links with Errors," Network Working Group, Internet Engineering Task Force, Tech. Rep., 2001. [Online]. Available: https://tools.ietf.org/html/rfc3155

[27] G. U. o. A. Fairhurst and L. C. S. L. Wood, "RFC3366, Advice to link designers on link Automatic Repeat reQuest (ARQ)," Network Working Group, Internet Engineering Task Force, Tech. Rep., 2002.

[28] J. Perry, P. Iannucci, K. Fleming, H. Balakrishnan, and D. Shah, "Spinal codes," in *SIGCOMM'12 - Proceedings of the ACM SIGCOMM 2012 Conference Applications, Technologies, Architectures, and Protocols for Computer Communication*, vol. 42, no. 4, 2012.

[29] A. Gudipati and S. Katti, "Strider: Automatic Rate Adaptation and Collision Handling," *ACM SIGCOMM Computer Communication Review*, 2011.

[30] U. Erez, M. D. Trott, and G. W. Wornell, "Rateless coding for gaussian channels," *IEEE Transactions on Information Theory*, 2012.

[31] D. E. Lucani, M. Médard, and M. Stojanovic, "On coding for delay - Network coding for time-division duplexing," *IEEE Transactions on Information Theory*, vol. 58, no. 4, pp. 2330–2348, 2012.

[32] D. E. Lucani, M. Stojanovic, and M. Médard, "Random Linear Network Coding For Time Division Duplexing: When To Stop Talking And Start Listening," p. 9, Sep. 2008. [Online]. Available: http://arxiv.org/abs/0809.2350

[33] K. Lea, "Wham City Lights," http://whamcitylights.com/, Mar. 2014.

[34] C. Roeding and A. Emigh, "Method and system for presence detection," U.S. Patent 8 489 112, Jul. 16, 2013.

[35] Google, "Guest mode FAQs – Chromecast help," https://support.google.com/chromecast/answer/6109297?hl=en, 2015.

[36] A. Huibers, "Bump validation," U.S. Patent 8 577 292, Nov. 05, 2013.

[37] Building Robotics, "Comfy," https://gocomfy.com, Apr. 2015.

[38] S. Dawson-Haggerty, A. Krioukov, J. Taneja, S. Karandikar, G. Fierro, N. Kitaev, and D. Culler, "Boss: building operating system services," in *NSDI*, 2013.

[39] Apple, "GKSession class reference," http://developer.apple.com/library/ios/documentation/GameKit/Reference/GKSession_Class/, Jun. 2009.

[40] A. Kauffmann and B. Smus, "Tone: An experimental Chrome extension for instant sharing over audio," http://googleresearch.blogspot.com/2015/05/tone-experimental-chrome-extension-for.html, May 2015.

[41] P. Bergel and A. Steed, "Data communication system," U.S. Patent 10 025 870, Jul. 17, 2018.

[42] H. Lashkari, H. Le, and D. Borges, "Peer-to-peer file transfer between computer systems and storage devices," U.S. Patent 8 838 697, Sep. 16, 2014.

[43] Android, "NFC basics," http://developer.android.com/guide/topics/connectivity/nfc/nfc.html#p2p.

[44] Apple, "Multipeer connectivity framework reference," http://developer.apple.com/library/ios/documentation/MultipeerConnectivity/Reference/MultipeerConnectivityFramework, Sep. 2013.

[45] B. Smus, "Ultrasonic networking on the web," http://smus.com/ultrasonic-networking/, Aug. 2013.

[46] S. Kumar, S. Gil, D. Katabi, and D. Rus, "Accurate Indoor Localization with Zero Start-up Cost," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '14. New York, NY, USA: ACM, 2014, pp. 483–494.

[47] J. Xiong, K. Jamieson, and K. Sundaresan, "Synchronicity: Pushing the Envelope of Fine-grained Localization with Distributed Mimo," in *Proceedings of the 1st ACM Workshop on Hot Topics in Wireless*, ser. HotWireless '14. New York, NY, USA: ACM, 2014, pp. 43–48.

[48] J. Xiong, K. Sundaresan, and K. Jamieson, "ToneTrack: Leveraging Frequency-Agile Radios for Time-Based Indoor Wireless Localization," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. New York, NY, USA: ACM, 2015, pp. 537–549.

[49] V. Bychkovsky, B. Hull, A. Miu, H. Balakrishnan, and S. Madden, "A measurement study of vehicular internet access using in situ wi-fi networks," in *Proceedings of the 12th annual international conference on Mobile computing and networking*. ACM, 2006, pp. 50–61.

[50] D. Hadaller, S. Keshav, T. Brecht, and S. Agarwal, "Vehicular opportunistic communication under the microscope," in *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, 2007, pp. 206–219.

[51] E. Uzun, N. Saxena, and A. Kumar, "Pairing devices for social interactions: a comparative usability evaluation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 2315–2324.

[52] T. S. Rappaport and S. Sandhu, "Radio wave propagation for emerging wireless personal communication systems," in *Wireless Personal Communications*. Springer, 1995, pp. 1–27.

[53] J. Grubor, V. Jungnickel, and K.-D. Langer, "Adaptive-modulation technique in wireless infrared indoor communications," in *Photonic Networks, 2006 ITG Symposium on*, Apr. 2006, pp. 1–8.

[54] M. Crawford, "Transmission of IPv6 Packets over Ethernet Networks," RFC 2464 (Proposed Standard), Internet Engineering Task Force, Dec. 1998, updated by RFC 6085. [Online]. Available: http://tools.ietf.org/html/rfc2464

[55] R. Gilligan, S. Thomson, J. Bound, J. McCann, and W. Stevens, "Basic Socket Interface Extensions for IPv6," RFC 3493 (Informational), Internet Engineering Task Force, Feb. 2003. [Online]. Available: http://tools.ietf.org/html/rfc3493

[56] B. Carpenter, S. Cheshire, and R. Hinden, "Representing IPv6 Zone Identifiers in Address Literals and Uniform Resource Identifiers," RFC 6874 (Proposed Standard), Internet Engineering Task Force, Feb. 2013. [Online]. Available: http://tools.ietf.org/html/rfc6874

[57] J. Hui and P. Thubert, "Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks," RFC 6282 (Proposed Standard), Internet Engineering Task Force, Sep. 2011. [Online]. Available: http://tools.ietf.org/html/rfc6282

[58] M. Dong, Q. Li, D. Zarchy, P. B. Godfrey, and M. Schapira, "PCC: Re-architecting congestion control for consistent high performance," in *NSDI*, 2015.

[59] G. Montenegro, N. Kushalnagar, J. Hui, and D. Culler, "Transmission of IPv6 Packets over IEEE 802.15.4 Networks," RFC 4944 (Proposed Standard), Internet Engineering Task Force, Sep. 2007, updated by RFCs 6282, 6775, 8025, 8066. [Online]. Available: http://tools.ietf.org/html/rfc4944

[60] S. Deering and R. Hinden, "Transmission of IPv6 Packets over Ethernet Networks," RFC 2460 (Proposed Standard), Internet Engineering Task Force, Dec. 1998, updated by RFCs 5095, 5722, 5871, 6437, 6564, 6935, 6946, 7045, 7112. [Online]. Available: http://tools.ietf.org/html/rfc2460

[61] P. Iannucci, "Blurt codebase," http://www.github.com/piannucci/blurt.

[62] R. Nandakumar, K. K. Chintalapudi, V. Padmanabhan, and R. Venkatesan, "Dhwani: Secure peer-to-peer acoustic nfc," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 63–74, Aug. 2013.

[63] K. Murphy, "Quietnet," http://github.com/Katee/quietnet, Jan. 2014.

[64] "Ultrasound networking," http://www.anfractuosity.com/projects/ultrasound-networking, Feb. 2014.

[65] "Ultrasound networking," http://github.com/anfractuosity/ultrasonicnetworking, Feb. 2014.

[66] J. L. Northern, M. P. Downs, W. Rudmose, A. Glorig, and J. L. Fletcher, "Recommended high-frequency audiometric threshold levels (8000–18 000 hz)," *The Journal of the Acoustical Society of America*, vol. 52, no. 2B, pp. 585–595, 1972.

[67] S. Ehrensvřd, "Yubikey standard," http://www.yubico.com/products/yubikey-hardware/yubikey/, Mar. 2014.

[68] P. Lesh, *Searching for the Sound: My Life with the Grateful Dead*. New York: Little, Brown and Company, 2005.

[69] D. McNally, *A Long Strange Trip: the Inside History of the Grateful Dead*. New York: Broadway Books, 2002.

[70] The Grateful Dead, "The Grateful Dead's Sound System," DeadHead Newsletter, Ice Nine Publishing Co., San Rafael, CA, p. 7, Dec. 1974.

[71] H. S. Hee and L. J. Hong, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Communications*, vol. 12, no. 2, pp. 56–65, Apr. 2005.

[72] S. Thompson, "Constant Envelope OFDM Phase Modulation," Ph.D. dissertation, UC San Diego, 2005.

[73] C. Köse, K. M. Chugg, and T. R. Halford, "Constant modulus orthogonal frequency division multiplexing," in *Proceedings - IEEE Military Communications Conference MILCOM*. IEEE, Oct. 2010, pp. 1624–1631.

[74] D. Falconer, S. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Communications Magazine*, vol. 40, no. 4, pp. 58–66, Apr. 2002.

[75] H. G. Myung, J. Lim, and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 30–38, Sep. 2006.

[76] M. Sharif and B. Hassibi, "Towards reducing the gap between PMEPR of multicarrier and single carrier signals," *IEEE 6th Workshop on Signal Processing Advances in Wireless Communications, 2005.*, pp. 380–384, 2005.

[77] S. Litsyn and A. Shpunt, "A balancing method for PMEPR reduction in OFDM signals," *IEEE Transactions on Communications*, vol. 55, no. 4, pp. 683–691, Apr. 2007.

[78] M. Sharif and B. Hassibi, "Existence of Codes With Constant PMEPR and Related Design," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2836–2846, Oct. 2004.

[79] C.-L. Wang, Y. Ouyang, and F.-H. Huang, "A Low-Complexity Peak-to-Average Power Ratio Reduction Technique for OFDM Systems Using Guided Scrambling Coding," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, vol. 6, no. C. IEEE, Apr. 2007, pp. 2837–2840.

[80] M. Laabidi, R. Zayani, and R. Bouallegue, "A quick convergence active constellation extension projection onto convex sets algorithm for reducing the PAPR of OFDM system," in *Proceedings - IEEE 30th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2016*. IEEE, Mar. 2016, pp. 439–443.

[81] J. Tellado, L. M. Hoo, and J. M. Cioffi, "Maximum-likelihood detection of nonlinearly distorted multicarrier symbols by iterative decoding," *IEEE Transactions on Communications*, vol. 51, no. 2, pp. 218–228, 2003.

[82] H. Karner, M. Auer, and C. W. Ueberhuber, "Top Speed FFTs for FMA Architectures," Institute for Applied and Numerical Mathematics, Technical University of Vienna, Tech. Rep., 2007.

[83] D. S. Atkinson and P. M. Vaidya, "A cutting plane algorithm for convex programming that uses analytic centers," *Mathematical Programming*, vol. 69, no. 1-3, pp. 1–43, Jul. 1995.

[84] J.-L. Goffin, Z.-Q. Luo, and Y. Ye, "Complexity Analysis of an Interior Cutting Plane Method for Convex Feasibility Problems," *SIAM Journal on Optimization*, vol. 6, no. 3, pp. 638–652, Aug. 1996.

[85] Z.-Q. Luo, "Analysis of a Cutting Plane Method That Uses Weighted Analytic Center and Multiple Cuts," *SIAM Journal on Optimization*, vol. 7, no. 3, pp. 697–716, Aug. 1997.

[86] J.-L. Goffin and J.-P. Vial, "Multiple Cuts in the Analytic Center Cutting Plane Method," *SIAM Journal on Optimization*, vol. 11, no. 1, pp. 266–288, Jan. 2000.

[87] A. Iouditski, "Convex optimization, theory and algorithms: Lecture 3, Convex Functions," Nov. 2015.

[88] A. Nedić, "Lecture Notes: Optimization I," in *Lecture Notes: Optimization I*. Urbana, IL: University of Illinois at Urbana-Champaign, 2008, ch. 2, pp. 42–52. [Online]. Available: http://www.ifp.illinois.edu/~angelia/L5_exist_optimality.pdf

[89] E. L. Allgower and K. Georg, *Introduction to numerical continuation methods*. SIAM, 2003, vol. 45.

[90] M. A. Goberna and M. A. López, "Linear semi-infinite programming theory: An updated survey," *European Journal of Operational Research*, vol. 143, no. 2, pp. 390–405, Dec. 2002.

[91] M. Sharif, M. Gharavi-Alkhansari, and B. Khalaj, "New results on the peak power of OFDM signals based on oversampling," in *2002 IEEE International Conference on Communications. Conference Proceedings. ICC 2002 (Cat. No.02CH37333)*, vol. 2. IEEE, 2002, pp. 866–871.

[92] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[93] A. Beck, *First-Order Methods in Optimization*. Philadelphia: Society for Industrial and Applied Mathematics : Mathematical Optimization Society, 2017.

[94] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 2011.

[95] M. L. N. Gonçalves, J. G. Melo, and R. D. C. Monteiro, "Extending the ergodic convergence rate of the proximal ADMM," 2016. [Online]. Available: http://arxiv.org/abs/1611.02903

[96] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Püschel, "A Proof of Convergence For the Alternating Direction Method of Multipliers Applied to Polyhedral-Constrained Functions," pp. 1–7, 2011. [Online]. Available: https://arxiv.org/abs/1112.2295

[97] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. I. Jordan, "A General Analysis of the Convergence of ADMM," 2015. [Online]. Available: http://arxiv.org/abs/1502.02009

[98] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Nashua, NH: Athena Scientific, 1989.

[99] B. He and X. Yuan, "On the O(1/n) Convergence Rate of the Douglas–Rachford Alternating Direction Method," *SIAM Journal of Numerical Analysis2*, vol. 55, no. 2, pp. 700–709, 2012.

[100] K. E. Batcher, "Sorting networks and their applications," in *Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference*. ACM, 1968, pp. 307—-314.

[101] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless device identification with radiometric signatures," in *Proceedings of the 14th ACM international conference on Mobile computing and networking*. ACM, 2008, pp. 116–127.

[102] Y. Shi and M. Jensen, "Improved radiometric identification of wireless devices using MIMO transmission," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 4, pp. 1346–1354, Dec 2011.

[103] T. D. Vo-Huu, T. D. Vo-Huu, and G. Noubir, "Fingerprinting Wi-Fi devices using software defined radios," in *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, ser. WiSec '16. New York, NY, USA: ACM, 2016, pp. 3–14.

[104] K. Bauer, D. McCoy, B. Greenstein, D. Grunwald, and D. Sicker, "Using wireless physical layer information to construct implicit identifiers," in *Privacy Enhancing Technologies, 8th International Symposium, PETS 2008, Leuven, Belgium, July 23-25, 2008, Proceedings*, Jul 2008.

[105] E. Kupershtein and M. Wax, "Indoor localization based on multipath fingerprinting," Ph.D. dissertation, Technion-Israel Institute of Technology, Faculty of Electrical Engineering, 2013.

[106] J. Xiong and K. Jamieson, "SecureArray: Improving wifi security with fine-grained physical-layer information," in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, ser. MobiCom '13. New York, NY, USA: ACM, 2013, pp. 441–452.

[107] R. Brandom, "The Dragnet: How a man accused of million-dollar fraud uncovered a never before seen, secret surveillance device," *The Verge*, Jan 2016. [Online]. Available: http://www.theverge.com/2016/1/13/10758380/stingray-surveillance-device-daniel-rigmaiden-case

[108] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[109] T. Mai and R. Speicher, "Free Probability, Random Matrices, and Representations of Non-Commutative Rational Functions," in *Abel Symposium on Computation and Combinatorics in Dynamics, Stochastics and Control*, Barony Rosendal, Norway, 2016.

[110] S. Yin, "Non-commutative rational functions in strong convergent random variables," *Advances in Operator Theory*, vol. 3, no. 1, pp. 178–192, Jan. 2018.

[111] A. Edelman and N. R. Rao, "Random matrix theory," *Acta Numerica*, vol. 14, pp. 233–297, May 2005.

[112] N. R. Rao and A. Edelman, "The polynomial method for random matrices," *Foundations of Computational Mathematics*, vol. 8, no. 6, pp. 649–702, 2007.

[113] E. Kupershtein, M. Wax, and I. Cohen, "Single-site emitter localization via multipath fingerprinting," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 10–21, 2013.

[114] E. Kupershtein, "Indoor Localization based on Multipath Fingerprinting," Ph.D. dissertation, Technion–Israel Institute of Technology, 2013.

[115] S. Du, H. Zhu, X. Li, K. Ota, and M. Dong, "Mixzone in motion: Achieving dynamically cooperative location privacy protection in delay-tolerant networks," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 9, pp. 4565–4575, Nov 2013.

[116] L. Huang, K. Matsuura, H. Yamane, and K. Sezaki, "Enhancing wireless location privacy using silent period," in *Wireless Communications and Networking Conference, 2005 IEEE*, vol. 2, March 2005, pp. 1187–1192 Vol. 2.

[117] L. N. Pringle, P. H. Harms, S. P. Blalock, G. N. Kiesel, E. J. Kuster, P. G. Friederich, R. J. Prado, J. M. Morris, and G. S. Smith, "A reconfigurable aperture antenna based on switched links between electrically small metallic patches," *Antennas and Propagation, IEEE Transactions on*, vol. 52, no. 6, pp. 1434–1445, 2004.

[118] P. Mookiah, "Reconfigurable antennas for wireless network security," Ph.D. dissertation, Drexel University, 2011.

[119] C. G. Christodoulou, Y. Tawk, S. Lane, S. R. Erwin *et al.*, "Reconfigurable antennas for wireless and space applications," *Proceedings of the IEEE*, vol. 100, no. 7, pp. 2250–2261, 2012.

[120] D. P. Petersen and D. Middleton, "Sampling and reconstruction of wave-number-limited functions in n-dimensional euclidean spaces," *Information and Control*, vol. 5, no. 4, pp. 279–323, 1962.

[121] R. Nevels and J. Jeong, "The time domain Green's function and propagator for Maxwell's equations," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 11, pp. 3012–3018, 2004.

[122] I. D. Berg, "An extension of the Weyl-von Neumann theorem to normal operators," *Trans. Amer. Math. Soc.*, vol. 160, pp. 365–371, 1971.

[123] A. C. Hansen, "On the approximation of spectra of linear operators on Hilbert spaces," *Journal of Functional Analysis*, vol. 254, no. 8, pp. 2092–2126, Apr. 2008.

[124] N. P. Brown, "Quasi-diagonality and the finite section method," *Mathematics of Computation*, vol. 76, no. 257, pp. 339–361, Jan. 2007.

Appendix A

Computer Algebra Code

This code carries out the diagrammatic matrix-monomial-expectation procedure of §4.3.1.

```
(* Preprocessor alias for prefix form of ** operator *)
$PreRead=ReplaceAll["NCM"→"NonCommutativeMultiply"];

(*
  DeclareNC[x,n,m] tags x with metadata to represent a linear
  operator from an m-dimensional space to an n-dimensional space.

  NonCommutativeQ is a predicate which identifies linear
  expressions involving objects so tagged and scalars.
*)
SetAttributes[DeclareNC,HoldFirst]
SetAttributes[NonCommutativeQ,HoldFirst]
DeclareNC[List[a___]]:=(Map[DeclareNC @@ # &,{a}];)
DeclareNC[x_,n_,m_]:=(
    x/:NonCommutativeQ[x]:=True;
    x/:Ident[x,Left]:=Ident[n];
    x/:Ident[x,Right]:=Ident[m];
    x/:Format[x,TraditionalForm]:=Style[SymbolName[x],Bold];
)
NonCommutativeQ[a_+__]:=NonCommutativeQ[a]
NonCommutativeQ[(_?NumberQ)*a_]:=NonCommutativeQ[a]
```

```
NonCommutativeQ[NCM[__]]:=True
NonCommutativeQ[_]:=False

(*
  Modify pretty-printing for infix ** and Tr[] operators in
  TraditionalForm.
*)
Unprotect[NCM,Tr];
Format[NCM[a__],TraditionalForm]:=
    (If[Length[#]>1,Infix[#,""],#[[1]]] &)
    [If[Length[#]>1,Superscript[#[[1]],Length[#]],#[[1]]] &
     /@ Split[{a}]]
Format[Tr[a_],TraditionalForm]:=Prefix[{{a}},Tr]
Protect[NCM,Tr];

(*
  Given a non-commutative linear expression and a space (Right for
  domain, Left for range), obtain a symbolic representation of the
  identity operator on this space.
*)
Ident[Plus[a_,__],side_]:=Ident[a,side]
Ident[a_**__,Left]:=Ident[a,Left]
Ident[__**a_,Right]:=Ident[a,Right]
Ident[(_?NumberQ)*a_,side_]:=Ident[a,side]
Ident[a_†,Left]:=Ident[a,Right]
Ident[a_†,Right]:=Ident[a,Left]
Ident[Ident[n_],side_]:=Ident[n]
Format[Ident[n_],TraditionalForm]:=Subscript[Style["I",Bold],n]

(*
  Add rules to ExpandAll and Power for non-commutative expressions.
*)
Unprotect[ExpandAll,Power];
ExpandAll[NCM[a___,b_Plus,c___]]:=
    Distribute[NCM[a,b,c],Plus,NCM,Plus,ExpandAll[NCM[##]]&]
ExpandAll[NCM[a___,b_Times,c___]]:=
```

```
    Most[b]ExpandAll[NCM[a,Last[b],c]]
Power[a_?NonCommutativeQ,0]:=Ident[a,Left]
Power[a_?NonCommutativeQ,n_]:=Nest[NCM[a,#]&,a,n-1]
Protect[ExpandAll,Power];

(*
  Rules for simplifying linear non-commutative expressions.
  Given a list of pairs (A, B) of linear operators such that A is
  the inverse of B, generates rules collapsing expressions AB->I,
  BA->I, IX->X, XI->X, Tr{I_n}->n.  In addition, rotates the
  factors inside Tr if the rightmost term is conjugated, and
  unwraps non-commutative products of a single factor.
*)
NCMRules[inversePairs_List]:={
  Sequence @@ Map[
    (#[[1]] ** #[[2]] → Ident[#[[1]],Left])&,
    inversePairs
  ],
  Sequence @@ Map[
    (#[[2]] ** #[[1]] → Ident[#[[1]],Right])&,
    inversePairs
  ],
  (a_ ** b_) /; (a===Ident[b,Left]) → b,
  (a_ ** b_) /; (Ident[a,Right]===b) → a,
  Tr[Ident[n_]] → n,
  Tr[a_ ** b___ ** a_†] → Tr[a† ** a ** b],
  NCM[a_] → a
}

(*
  Computes the matrix-valued expectation of a non-commutative
  linear expression in a given (possibly rectangular) zero-mean
  complex Gaussian random matrix with the provided row and column
  covariance matrices.
*)
MatrixMoment[NCM[a__],
```

```
            RV[rv_, rvct_, leftcov_, rightcov_, orderparam_],
            inversePairs_:{}]
  := Module[
    {i,j,ll,cycles,chains},
    i=Flatten[Position[{a},rv,1]];
    j=Flatten[Position[{a},rvct,1]];
    If[Length[i]!=Length[j],0,
      Plus @@ Function[k,
        ll=ReplacePart[
          Range[1,Length[{a}]],
          MapThread[Rule,{Join[i,k],Join[k,i]}]
        ];
        cycles=Map[List,
          PermutationCycles[Append[ll+1,1],Identity],
          {2}];
        chains=NCM @@@ (
          Extract[
            {a,Null}/.{rv→leftcov,rvct→rightcov}, #] &
          /@ cycles);
        (Most[First[chains]]) *
        (Times @@ Tr /@ Rest[chains]) *
        orderparam^Length[i] //. NCMRules[inversePairs]
      ] /@ Permutations[j]
    ]
  ]
MatrixMoment[a_+b__,args___]:=MatrixMoment[#,args] & /@ (a+b)
MatrixMoment[(a_?NumberQ)*b_,args___]:=a*MatrixMoment[b,args]
MatrixMoment[a_?NonCommutativeQ,args___]:=MatrixMoment[NCM[a],args]
```

Appendix B

Norm Bounds for the Electric Two-Point Function

In §4.7.5, we found it useful to establish certain properties regarding the two-point correlation function of the electric far field generated by the Gaussian random field of current density. Proofs of these properties follow.

To simplify notation, we let

$$
\begin{aligned}
A &= V\alpha{\mu_0}^2{\omega_0}^2\pi \\
b_i &= \frac{a_i\omega_0}{2c},
\end{aligned}
$$

where b_i is the half-width of the transmitter along axis i, in radians of phase.

In what follows, we will need to compute traces of linear operators over $\mathcal{H}$. Using the sub-spaces $\mathcal{H}_n$ of (4.20) and the basis of (4.21), we write for $\mathbf{X}\colon \mathcal{H} \to \mathcal{H}$,

$$
\begin{aligned}
\operatorname{Tr}_{\mathcal{H}} \mathbf{X} &= \lim_{n\to\infty} \operatorname{Tr}_{\mathcal{H}_n} \mathbf{X}_n \\
&= \lim_{n\to\infty} \sum_{m,\ell} \langle \mathbf{X}\mathbf{e}_{nm\ell}, \mathbf{e}_{nm\ell}\rangle_{\mathcal{H}} \\
&= \lim_{n\to\infty} \sum_{m} \frac{1}{w_{nm}} \int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \left(\operatorname{Tr}_{\mathbb{C}^3} \mathbf{X}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \right)
\end{aligned}
$$

In the limit, the ranges of integration over $\hat{\mathbf{r}}$ and $\hat{\mathbf{r}}'$ become arbitrarily small, and so

the difference $|\hat{\mathbf{r}} - \hat{\mathbf{r}}'|$ becomes arbitrarily small. Provided that $\mathrm{Tr}_{\mathbb{C}^3} \mathbf{X}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'}$ is continuous in $\hat{\mathbf{r}}'$, we may therefore evaluate it at $\hat{\mathbf{r}}' = \hat{\mathbf{r}}$ without altering the limit. This gives

$$\begin{aligned}
\mathrm{Tr}_{\mathcal{H}} \mathbf{X} &= \lim_{n\to\infty} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \left(\frac{1}{w_{nm}} \int_{t_{nm}} \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \right) \left(\mathrm{Tr}_{\mathbb{C}^3} \mathbf{X}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}} \right) \\
&= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathrm{Tr}_{\mathbb{C}^3} \mathbf{X}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}},
\end{aligned}$$

Proof of Theorem 4.7.1: We must show that

$$\mathrm{Tr}_{\mathcal{H}} \mathbf{C} = A$$

We have

$$\begin{aligned}
\mathrm{Tr}_{\mathcal{H}} \mathbf{C} &= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathrm{Tr}_{\mathbb{C}^3} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}} \\
&= \frac{A}{2} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathrm{Tr}_{\mathbb{C}^3} (\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T) \prod_{i=1}^{3} \mathrm{sinc}\,(b_i(\hat{\mathbf{r}} - \hat{\mathbf{r}}')_i) \\
&\leq \frac{A}{2} \mathrm{Tr}_{\mathbb{C}^3} \left(\int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} (\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T) \right) = A
\end{aligned}$$

■

Proof of Theorem 4.7.2: We must show that

$$\mathrm{Tr}_{\mathcal{H}} \mathbf{C}^2 \leq \frac{A^2}{3}$$

We have

$$\begin{aligned}
\mathrm{Tr}_{\mathcal{H}} \mathbf{C}^2 &= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathrm{Tr}_{\mathbb{C}^3} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \mathbf{C}_{\hat{\mathbf{r}}'\,\hat{\mathbf{r}}} \\
&= \frac{A^2}{4} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathrm{Tr}_{\mathbb{C}^3} (\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T)(\mathbf{1} - \hat{\mathbf{r}}'\hat{\mathbf{r}}'^T) \prod_{i=1}^{3} \mathrm{sinc}^2\,(b_i(\hat{\mathbf{r}} - \hat{\mathbf{r}}')_i) \\
&\leq \frac{A^2}{4} \mathrm{Tr}_{\mathbb{C}^3} \left(\int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} (\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T) \right)^2 = \frac{A^2}{3}
\end{aligned}$$

■

Proof of Theorem 4.7.3: We must show that for all $\mathbf{x} \in \mathcal{H}$,

$$\|\mathbf{C}[\mathbf{x}]\|_{\mathcal{H}} \leq \frac{A}{3}\|\mathbf{x}\|_{\mathcal{H}} \tag{B.1}$$

Now

$$\begin{aligned}
\|\mathbf{C}[\mathbf{x}]\|_{\mathcal{H}}^2 &= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \left|(\mathbf{C}[\mathbf{x}])_{\hat{\mathbf{r}}}\right|^2 \\
&= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \left| \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \mathbf{x}_{\hat{\mathbf{r}}'} \right|^2 \\
&= \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}'')}{4\pi} \left(\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}'} \mathbf{x}_{\hat{\mathbf{r}}'}\right)^\dagger \left(\mathbf{C}_{\hat{\mathbf{r}}\,\hat{\mathbf{r}}''} \mathbf{x}_{\hat{\mathbf{r}}''}\right) \\
&= \left(\frac{A}{2}\right)^2 \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}'')}{4\pi} \\
&\qquad \mathbf{x}_{\hat{\mathbf{r}}'}^\dagger (\mathbf{1} - \hat{\mathbf{r}}'\hat{\mathbf{r}}'^T)(\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T)(\mathbf{1} - \hat{\mathbf{r}}''\hat{\mathbf{r}}''^T)\mathbf{x}_{\hat{\mathbf{r}}''} \\
&\qquad \prod_{i=1}^{3} \operatorname{sinc}\left(b_i(\hat{\mathbf{r}} - \hat{\mathbf{r}}')_i\right) \operatorname{sinc}\left(b_i(\hat{\mathbf{r}} - \hat{\mathbf{r}}'')_i\right)
\end{aligned}$$

Because of the three projection matrices $(\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T)$ lying between $\mathbf{x}^\dagger$ and $\mathbf{x}$ in this expression, we may conclude that $\|\mathbf{C}[\mathbf{x}]\|_{\mathcal{H}}^2$ is not maximized for a given $\|\mathbf{x}\|_{\mathcal{H}}^2$ unless $\mathbf{x}$ falls in the sub-space with transverse polarization, i.e. $\hat{\mathbf{r}} \cdot \mathbf{x}_{\hat{\mathbf{r}}} = 0$ for all $\hat{\mathbf{r}}$. That is, the inequality (B.1) is tightest for $\mathbf{x}$ in this sub-space. For such a vector,

$$\begin{aligned}
\|\mathbf{C}[\mathbf{x}]\|_{\mathcal{H}}^2 &= \left(\frac{A}{2}\right)^2 \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}'')}{4\pi} \mathbf{x}_{\hat{\mathbf{r}}'}^\dagger (\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T)\mathbf{x}_{\hat{\mathbf{r}}''} \\
&\qquad \prod_{i=1}^{3} \operatorname{sinc}\left(b_i(\hat{\mathbf{r}} - \hat{\mathbf{r}}')_i\right) \operatorname{sinc}\left(b_i(\hat{\mathbf{r}} - \hat{\mathbf{r}}'')_i\right) \\
&\leq \left(\frac{A}{2}\right)^2 \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}'')}{4\pi} \mathbf{x}_{\hat{\mathbf{r}}'}^\dagger (\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T)\mathbf{x}_{\hat{\mathbf{r}}''} \\
&= \left(\frac{A}{2}\right)^2 \left(\int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{x}_{\hat{\mathbf{r}}'}\right)^\dagger \left(\int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} (\mathbf{1} - \hat{\mathbf{r}}\hat{\mathbf{r}}^T)\right) \left(\int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}'')}{4\pi} \mathbf{x}_{\hat{\mathbf{r}}''}\right) \\
&= \frac{2}{3}\left(\frac{A}{2}\right)^2 \left| \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}}')}{4\pi} \mathbf{x}_{\hat{\mathbf{r}}'} \right|^2
\end{aligned}$$

Once again, the inequality (B.1) will be tightest if we maximize this expression over $\mathbf{x}$ subject to a constraint on $\|\mathbf{x}\|_{\mathcal{H}}^2$.

The transverse polarization constraint is linear, so the last expression is maximized for some vector $\mathbf{x}$ that is as nearly constant over $\hat{\mathbf{r}}$ as possible. Let $\hat{z}$ be the unit vector in the direction of $(\int \mathrm{d}\Omega/4\pi \mathbf{x}_{\hat{\mathbf{r}}})$, and establish a spherical coordinate system with polar angle θ and azimuthal angle ϕ.

Now we have for all $\hat{\mathbf{r}}$, $\hat{\mathbf{r}} \cdot \mathbf{x}_{\hat{\mathbf{r}}} = 0$. In addition, if $\hat{\phi} \cdot \mathbf{x}_{\hat{\mathbf{r}}} \neq 0$, then we can increase $|\int \mathrm{d}\Omega/4\pi \mathbf{x}|^2$ without changing $\int \mathrm{d}\Omega/4\pi |\mathbf{x}|^2$ by rotating $\mathbf{x}_{\hat{\mathbf{r}}}$ away from $\hat{\phi}$.

We can further simplify the maximization problem by observing that if $\mathbf{x}$ does not possess radial symmetry, i.e. if $\hat{\theta} \cdot \mathbf{x}_{\hat{\mathbf{r}}}$ is not independent of ϕ, then by re-distributing amplitude to be more uniform over ϕ, we can increase $|\int \mathrm{d}\Omega/4\pi \mathbf{x}|^2$ without changing $\int \mathrm{d}\Omega/4\pi |\mathbf{x}|^2$. So we have for the $\mathbf{x}$ that maximizes the LHS of our desired bound for a given RHS,

$$\mathbf{x}_{\hat{\mathbf{r}}} = \hat{\theta}\, x(\theta)$$

$$\left| \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathbf{x}_{\hat{\mathbf{r}}} \right|^2 = \left(\int (\sin^2\theta \mathrm{d}\theta/2)\; x(\theta) \right)^2 \tag{B.2}$$

$$\int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \left|\mathbf{x}_{\hat{\mathbf{r}}}\right|^2 = \int (\sin\theta \mathrm{d}\theta/2)\; x(\theta)^2 \tag{B.3}$$

Using the calculus of variations, it follows that the optimal polar magnitude profile $x(\theta)$ that maximizes (B.2) for a given (B.3) has the form

$$x(\theta) = \beta \sin\theta$$

$$\left| \int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \mathbf{x}_{\hat{\mathbf{r}}} \right|^2 = 4\beta^2/9$$

$$\int \frac{\mathrm{d}\Omega(\hat{\mathbf{r}})}{4\pi} \left|\mathbf{x}_{\hat{\mathbf{r}}}\right|^2 = 2\beta^2/3$$

and hence that for any $\mathbf{x} \in \mathcal{H}$,

$$\|\mathbf{C}[\mathbf{x}]\|_{\mathcal{H}} \leq \frac{A}{3} \|\mathbf{x}\|_{\mathcal{H}}$$

■