Distributed Function Approximation over Noisy Channels

With Reliability and Security Guarantees for Applications in Wireless Networks

vorgelegt von
M. Sc.
Matthias Frey
https://orcid.org/0000-0003-3016-2644

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Giuseppe Caire, Ph.D.
Gutachter: Prof. Dr.-Ing. Sławomir Stańczak
Gutachter: Prof. Dr. Michael C. Gastpar
Gutachter: Dr. Jingge Zhu

Tag der wissenschaftlichen Aussprache: 12. Dezember 2022

Berlin 2023

Abstract

Over-the-Air (OTA) computation is the problem of computing functions of distributed data over a wireless channel without transmitting the entirety of the data to a central point. By avoiding such costly transmissions, OTA computation schemes can achieve a better-than-linear (depending on the function, often logarithmic or even constant) scaling of the communication cost as the number of transmitters grows. Among the most common functions computed OTA are linear functions such as weighted sums. In this work, we propose and analyze a method for the approximation of functions of distributed arguments over noisy channels. This method can be used as an analog OTA computation scheme for a class of functions that contains linear functions as well as some nonlinear functions such as p-norms of vectors. We prove error bound guarantees that are valid for fast-fading channels and all distributions of fading and noise contained in the class of sub-Gaussian distributions. This class includes Gaussian distributions, but also many other practically relevant cases such as Class A Middleton noise and fading with dominant line-of-sight components. In addition, there can be correlations in the fading and noise so that the presented results also apply to, for example, block fading channels and channels with bursty interference. We do not rely on any stochastic characterization of the distributed arguments of the OTA computed function; in particular, there is no assumption that these arguments are drawn from identical or independent probability distributions. Our analysis relies on tools from high-dimensional statistics which we adapt so that they are applicable to the scenario at hand. The resulting error guarantees are nonasymptotic and therefore provide error bounds that are valid for a finite number of channel uses.

OTA computation has a huge potential for reducing communication cost in applications such as Machine Learning (ML)-based distributed anomaly detection in large wireless sensor networks. We show this potential with two examples of how our OTA computation scheme can be used to vastly increase the efficiency of Vertical Federated Learning (VFL) over a wireless channel. We also illustrate the efficiency gain with numerical simulations for a few example cases.

Then, we move on to propose a new method to protect OTA computation schemes against passive eavesdropping. Our method uses a friendly jammer whose signal is – contrary to common intuition – stronger at the legitimate receiver than it is at the eaves-

dropper. It works for a large class of analog OTA computation schemes and we give details on the special case of computing an arithmetic average over an Additive White Gaussian Noise (AWGN) channel. The derived secrecy guarantee translates to a lower bound on the eavesdropper's mean square error while the question of how to provide operationally more significant guarantees such as semantic security remains open for future work. As key ingredients in proving the security guarantees, we propose and prove generalizations of known results on channel resolvability and coding for compound channels for the case of continuous channel input and output alphabets.

Zusammenfassung

Funktionsberechnung im Funkkanal ist eine Methode, Funktionen verteilter Daten ohne eine vollständige Übertragung der Daten zu einem zentralen Punkt zu berechnen. Indem solche Übertragungen vermieden werden, kann erreicht werden, dass der Ressourcenverbrauch weniger schnell als linear mit der Anzahl der Sender wächst. Abhängig von der zu berechnenden Funktion kann dieses Wachstum dann in vielen Fällen logarithmisch oder sogar konstant sein (d.h. die zur Funktionsberechnung nötigen Kanalressourcen wachsen überhaupt nicht, wenn die Anzahl der Sender wächst). Zu den am häufigsten im Funkkanal berechneten Funktionen gehören lineare Funktionen wie z.B. gewichtete Summen. In der vorliegenden Arbeit führen wir eine Methode zur verteilten Approximation von Funktionen in verrauschten Kanälen ein. Diese Methode kann als Verfahren zur analogen Funktionsberechnung im Funkkanal genutzt werden. Sie ist auf eine Klasse von Funktionen anwendbar, die zum einen alle linearen Funktionen, zum anderen aber auch einige nichtlineare Funktionen wie die p-Norm von Vektoren enthält. Wir zeigen Fehlerschranken für unser Verfahren, die in Kanälen mit Fast Fading gelten, wobei sowohl die Verteilung des Fadings als auch die des Rauschens zur Klasse der sub-Gauß'schen Wahrscheinlichkeitsverteilungen gehören müssen. Diese Klasse enthält nicht nur alle Normalverteilungen, sondern auch viele andere in der Praxis relevanten Verteilungen wie z.B. Class-A-Rauschen nach Middleton und Fadingverteilungen mit einer dominanten Sichtkomponente. Zudem sind unsere Fehlerschranken auch dann noch gültig, wenn es Korrelationen in der Verteilung des Fadings und/oder Rauschens gibt, sodass unsere Ergebnisse beispielsweise auch auf Kanäle mit Blockfading oder gebündelt auftretender Interferenz anwendbar sind. Dabei verwenden wir keinerlei stochastische Charakterisierung der Argumente der über den Funkkanal zu berechnenden Funktion. Dies bedeutet insbesondere, dass keine Annahme über identische Verteilung oder Unabhängigkeit dieser Argumente nötig ist. Unsere Analyse baut auf Werkzeugen aus der hochdimensionalen Statistik auf, die wir derart anpassen, dass sie im vorliegenden Szenario anwendbar sind. Die sich daraus ergebenden Fehlerschranken sind nichtasymptotisch, d.h. sie sind für jede beliebige (endliche) Anzahl von Kanalnutzungen gültig.

Funktionsberechnung im Funkkanal bietet ein riesiges Potenzial, in Anwendungen wie z.B. der auf maschinellem Lernen basierenden Anomaliedetektion in großen Sensornetzen

die zur Kommunikation nötigen Ressourcen drastisch zu reduzieren. Wir veranschaulichen dieses Potenzial, indem wir zwei Beispiele ausführen, die zeigen, wie unsere Methoden zur Funktionsberechnung im Funkkanal die Effizienz von vertikalem föderierten Lernen enorm steigern können. Wir illustrieren diesen Effizienzgewinn außerdem für einige ausgesuchte Spezialfälle anhand numerischer Simulationen.

Anschließend führen wir ein neues Verfahren ein, um Funktionsberechnungen im Funkkanal gegen passive Lauscher zu schützen. Unsere Methode basiert auf der Kooperation der legitimen Kommunikationspartner mit einem Jammer, dessen Signal – entgegen den normalerweise in diesem Zusammenhang gemachten Annahmen – beim legitimen Empfänger in einer größeren Signalstärke ankommt als beim Lauscher. Sie funktioniert für eine große Klasse von Verfahren zur Funktionsberechnung über den Funkkanal, und wir führen den Spezialfall der Berechnung eines arithmetischen Mittelwerts über einen Kanal mit additivem weißen normalverteilten Rauschen detailliert aus. Dabei erhalten wir eine Sicherheitsgarantie, die den durchschnittlichen quadratischen Fehler des Lauschers nach unten begrenzt. Die Frage, wie stärkere Sicherheitsgarantien (z.B. semantische Sicherheit) gegeben werden können, bleibt dabei für zukünftige Forschungsarbeiten offen. Die wichtigsten informationstheoretischen Zutaten für den Beweis der Sicherheitsgarantie sind Verallgemeinerungen bekannter Ergebnisse zur Resolvability von Kanälen und zur Kommunikation über Compound-Kanäle für den Fall kontinulierlicher Ein- und Ausgabealphabete, die wir im Rahmen dieser Arbeit ebenfalls beweisen.

Acknowledgments

I would like to thank Prof. Dr.-Ing. Sławomir Stańczak for the opportunity to pursue a Ph.D. degree as a member of his research group at Technische Universität Berlin, and for the support and scientific freedom I had during that time. I am also particularly indebted to Dr. Igor Bjelaković who has taken a lot of time and effort to guide and support me in my research work for this thesis, and beyond.

Moreover, I would like to thank Prof. Dr. Michael C. Gastpar and Lecturer Dr. Jingge Zhu for serving as referees for this dissertation.

Finally, I gratefully recognize my co-authors Navneet Agrawal, Dr.-Ing. Zoran Utkovski, Patrick Agostini, Miguel A. Gutierrez-Estevez and Daniel Schäufele for their collaboration, as well as all my current and former colleagues at Technische Universität Berlin and Fraunhofer Heinrich Hertz Institute. It was a pleasure to share many stimulating discussions with them and I am appreciative of this joyful work and research environment.

Copyright Information

Contents

1 Introduction

In this thesis, we study the approximation of a function f of distributed arguments $s_1, \ldots, s_K$. Namely, we consider terminals $\mathfrak{A}_1, \ldots, \mathfrak{A}_K, \mathfrak{B}$ which are connected by a multiple-access channel (MAC). The MAC is described as a stochastic kernel W which randomly maps every possible tuple of inputs $(T_1, \ldots, T_K)$ at $\mathfrak{A}_1, \ldots, \mathfrak{A}_K$ to an output Y at $\mathfrak{B}$. Each terminal $\mathfrak{A}_k$ holds some value s_k from which it generates a sequence of channel inputs $T_k^M = (T_{k,1}, \ldots, T_{k,M})$. The channel is then invoked M times and generates, for each input tuple $(T_{1,m}, \ldots, T_{k,m})$, an output symbol Y_m. At terminal $\mathfrak{B}$, the sequence $Y^M = (Y_1, \ldots, Y_M)$ is processed to yield a value $\tilde{f}$. A distributed function approximation scheme should have the property $\tilde{f} \approx f(s_1, \ldots, s_K)$. The meaning of "$\approx$" can vary, but it usually means that a suitably defined distance[1] between $\tilde{f}$ and $f(s_1, \ldots, s_K)$ is small and approaches 0 as M tends towards infinity. The channels we consider in this work mostly have complex or real inputs and outputs, and they are of the form

$$Y = \sum_{k=1}^{K} H_k T_k + N. \tag{1.1}$$

In channels of this form, $H_1, \ldots, H_K$ are called the *fading coefficients*, and N is called the *(additive) noise*. The fading coefficients can in some cases be deterministic, but in general, both the fading and the noise follow a random distribution. It is this randomness which makes the channel output Y "noisy" in the sense that in general, it is not possible to recover exact information about $T_1, \ldots, T_K$ or their sum from the observed value of Y.

1.1 Nomographic Functions

Given the channel (1.1), it is natural to ask which functions can be approximated from distributed arguments over such channels. In case there is no fading or noise (i.e., $N = 0$ and $H_1 = \cdots = H_K = 1$ almost surely), this turns out to be the class of nomographic functions which we discuss in this section. The connection between nomographic functions and distributed function approximation was first observed in [GS13] and has been discussed

[1] This distance does not need to be a metric.

and analyzed in [GBS13] in more detail than we can in the following summary. Here and in the rest of the thesis, $\mathbb{R}$ denotes the set of real numbers.

Definition 1. *A* nomographic representation *of a function* $f : \mathbb{R}^K \to \mathbb{R}$ *consists of functions* $f_1, \ldots, f_K, F : \mathbb{R} \to \mathbb{R}$ *such that*

$$\forall a_1, \ldots, a_K \in \mathbb{R} : \; f(a_1, \ldots, a_K) = F\left(\sum_{k=1}^{K} f_k(a_k)\right). \tag{1.2}$$

A function $f : \mathbb{R}^K \to \mathbb{R}$ *which has a nomographic representation is called a* nomographic function.

It has been noted in [Buc76, Theorem 8] that every function is nomographic according to this definition. We state a version of this result that fits with Definition 1. Since it illustrates the arguments below very well, we also give a full proof, based on the same idea as in [Buc76].

Theorem 1. (adapted from [Buc76, Theorem 8]). *Every function* $f : \mathbb{R}^K \to \mathbb{R}$ *is nomographic.*

Proof. We first fix an arbitrary bijection $\phi : \mathbb{R} \to (0, 1)$. An example of a possible choice is

$$\phi : a \mapsto \begin{cases} \frac{1}{2} \cdot \frac{1}{a+1}, & a \in (0, \infty) \\ \frac{1}{2} \cdot \left(1 + \frac{1}{|a|+1}\right), & a \in (-\infty, 0) \\ \frac{1}{2}, & a = 0. \end{cases} \tag{1.3}$$

Next, we define for every $a \in (0, 1)$ a sequence of digits $c_{a,1}, c_{a,2}, \cdots \in \{0, \ldots, 9\}$ such that[2]

$$a = 0.\overbrace{c_{a,1}c_{a,2}\ldots}^{\text{dec}} := \sum_{i=1}^{\infty} c_{a,i} \cdot 10^{-i}. \tag{1.4}$$

We make the choice for the sequence $c_{a,1}, c_{a,2}, \ldots$ unique by requiring that it has to contain infinitely many non-zero elements. Let, for all $k \in \{1, \ldots, K\}$,

$$f_k(a) := 0.\overbrace{\underbrace{0\ldots0}_{k-1} c_{\phi(a),1} \underbrace{0\ldots0}_{K-1} c_{\phi(a),2} \underbrace{0\ldots0}_{K-1} c_{\phi(a),3} \underbrace{0\ldots0}_{K-1} \ldots}^{\text{dec}}. \tag{1.5}$$

[2] Of course, there is nothing special about base 10 here, and in fact, [Buc76] uses dyadic representations. We have chosen the base 10 here so that our representation coincides with the usual decimal notation of numbers.

Define $\psi_1, \ldots, \psi_K, F : (0,1) \to \mathbb{R}$ by

$$\psi_k : \quad 0.\overbrace{b_1 b_2 \ldots}^{\text{dec}} \mapsto \phi^{-1}\left(0.\overbrace{b_k b_{k+K} b_{k+2K} \ldots}^{\text{dec}}\right) \tag{1.6}$$

$$F : \quad a \mapsto f\left(\psi_1(a), \ldots, \psi_K(a)\right). \tag{1.7}$$

It is clear from our construction that the maps

$$(a_1, \ldots, a_K) \mapsto \sum_{k=1}^{K} f_k(a_k) \tag{1.8}$$

$$a \mapsto (\psi_1(a), \ldots, \psi_K(a)) \tag{1.9}$$

are inverses of each other and therefore, (1.2) is satisfied, concluding the proof that f is nomographic. □

In order to use the nomographic representation of a function in a wireless communication system, the inner functions $f_1, \ldots, f_K$ should be computed at the transmitter before the actual transmission, while the outer function F should be implemented and evaluated at the receiver. Therefore, $f_1, \ldots, f_K$ are sometimes referred to as the pre-processing functions while F is called a post-processing function. The summation is performed by the wireless channel due to its superposition property. If the receiver has access to $f_1(a_1) + \cdots + f_K(a_K)$, then from (1.8) and (1.9), it is clear that a full reconstruction of $a_1, \ldots, a_K$ is possible and in fact, this full reconstruction is used as an intermediate step in post-processing. However, such an approach does not generalize to noisy channels in an obvious way. Indeed, in (1.5) we can see that arbitrarily significant[3] digits of the transmitted values can be hidden in digits of arbitrarily low significance in the real number that is transmitted over the channel and therefore, even a channel noise of extremely low power can cause arbitrarily strong disruptions. Of course, this is due to the specific construction of F used in the proof of Theorem 1, but we can expect that in general any strong discontinuity of F can cause problems of this kind when the argument of F is noisy.

It appears therefore that in order to apply a nomographic representation to a distributed function approximation problem in the presence of noise, Definition 1 is not strong enough and we need to impose additional constraints on the functions $f_1, \ldots, f_K, F$. A famous result [Arn57, Kol57] states that every continuous function $f : \mathbb{R}^K \to \mathbb{R}$ can be written

[3] The significance of a digit in the decimal representation is an indication of how strongly a change of the digit influences the value of the represented number. For instance, the significance of the digit $c_{a,i}$ in (1.4) is 10^{-i}, since the digit is multiplied with this number to determine the value of a.

as a sum of $2K + 1$ functions with continuous nomographic representations,[4] giving a positive answer in part to the question posed by Hilbert as the thirteenth problem in his list of unresolved mathematical problems of the 20th century [Hil00]. If there was a result implying that for every algebraic function, there is a nomographic representation consisting only of algebraic functions, this would give a positive answer to the as-of-yet unresolved part of Hilbert's thirteenth problem. We can therefore expect that proving such a result would be very hard.[5] Another result worth noting in this context is that the set of functions with a continuous nomographic representation is nowhere dense in the space of continuous functions [Buc82]. This provides another piece of evidence that generic nomographic representations suitable for distributed function approximation may not exist.

1.2 Distributed Function Approximation in Noisy Channels: An Example

From an information-theoretic viewpoint, channels of the form (1.1) with no fading or noise have infinite Shannon capacity. Therefore, they are not realistic in the sense that they cannot be expected to sufficiently accurately model a real-world communication channel. If we focus on finite-capacity channels, however, we can make some basic information-theoretic arguments to further motivate the development of schemes in which $f(s_1, \ldots, s_K)$ is directly approximated at $\mathfrak{B}$ instead of transmitting the values $s_1, \ldots, s_K$ to $\mathfrak{B}$ separately (from which $f(s_1, \ldots, s_K)$ could easily be computed at $\mathfrak{B}$).

If all values $s_1, \ldots, s_K$ are made available at $\mathfrak{B}$, this means that the number of uses of the channel is lower bounded by the quotient of the Shannon entropy of $s_1, \ldots, s_K$ and the sum-rate capacity of the channel. On the other hand, if the entropy of $f(s_1, \ldots, s_K)$ is significantly smaller than that of $s_1, \ldots, s_K$, then a scheme which only makes $f(s_1, \ldots, s_K)$ available at $\mathfrak{B}$ is not subject to the same fundamental bound. The following example illustrates that this difference of entropy can be significant.

Example 1. *(from [7]). Suppose that K nodes send their data $s_1, \ldots, s_n$ to a single receiver through a MAC. For simplicity, we assume that each s_k is an independent random variable uniformly distributed over $\mathcal{S} = \{0, 1\}$. Now if the receiver reconstructs each of these variables, then the entropy or the amount of information available at the receiver is $\sum_{k=1}^{K} H(s_k) = K \log 2$ nats where $\log(\cdot)$ is the natural logarithm with Euler's number*

[4] A continuous nomographic representation is a nomographic representation that consists only of continuous functions $f_1, \ldots, f_K, F$.

[5] Hilbert even hypothesized that the correct answer to the question would be negative [Hil00, Hil27], which was, however, partly disproven in [Arn57, Kol57].

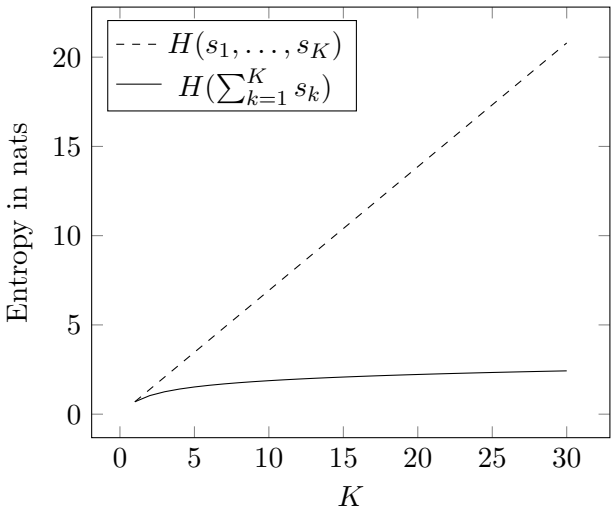

Figure 1.1: Plot of sum vs. tuple entropy in Example 1.

$e \approx 2.71828$ *as the basis,* $H(s_k) := -\sum_{s' \in \mathcal{S}} p_{s_k}(s') \log p_{s_k}(s')$ *is the Shannon entropy*[6] *and* $p_{s_k} : \mathcal{S} \mapsto [0, 1]$ *is the probability mass function of* s_k*. This means that the nodes have to transmit* $K \log 2$ *nats (or, equivalently,* K *bits) to the receiver. Therefore, if the capacity of the communication channel is* 1 *bit per channel use, then* K *channel uses are necessary to convey the full information to the receiver.*[7] *Now we assume that the receiver is only interested in* $f(s_1, \ldots, s_K) = \sum_{k=1}^{K} s_k$ *which can be easily computed from* $s_1, \ldots, s_K$*. By the data processing inequality [EGK11, Section 2.3], this operation cannot increase the amount of information. In fact, the entropy of the function is* $H(\sum_{k=1}^{K} s_k) = K \log 2 - \sum_{k=1}^{K} \binom{K}{k} 2^{-K} \log \binom{K}{k}$ *which is strictly smaller than* $K \log 2$ *for all* $K \geq 2$*. This means that instead of transmitting* K *bits that are necessary to reconstruct each* s_k*, the agents can send significantly less information to the receiver if its objective is to compute the sum function* $f(s_1, \ldots, s_K)$*. In Fig. 1.1, it can be seen that this difference is quite pronounced even for moderately large values of* K*.*

1.3 Over-the-Air Computation

The approximation of functions of distributed arguments over channels of the form (1.1) is of particular interest in the context of wireless communications. Here, use of the channel corresponds to a concurrent transmission of waveforms from all transmitters, and the

[6] We use the convention $0 \cdot \log(1/0) := 0$ in the definition.

[7] In the case of orthogonal channel access, it is necessary to establish K independent (interference-free) communication channels, where each of these has the capacity of 1 bit per channel use.

receiver observes a noisy, superimposed version of these waveforms. In the context of this application, distributed function approximation schemes can be seen as instances of a class of schemes commonly called Computation over MAC (CoMAC), AirComp or OTA computation. The goal of these schemes is to obtain a scaling behavior of the communication cost in the number of transmitters that is better than the linear growth[8] that would ensue from a separation of source and channel coding. Therefore, such schemes exhibit the inherent property that the receiver is unable to fully reconstruct all of the transmitted information.

The idea of a scheme that allows a receiver to reconstruct directly a combined form of two messages, but not the original messages themselves, can be traced back to [KM79] where a source coding problem is formulated in which it is the receiver's task to reconstruct a sequence of modulo-2 sums of encoded bits. An uncoded analog scheme for obtaining a noisy estimate of a function of transmitted values with an application to wireless sensor networks has appeared in [GV03] and is, to the best of our knowledge, the first work that proposes a joint source-channel approach to OTA computation.

The authors in [GV03] take an analog approach in which a certain amount of noise is tolerated in the received value and the function is computed only once.[9] This is in contrast with a class of digital schemes that are closer to [KM79] in the sense that they also consider functions with finite domains and typically give error guarantees for a large number of repeated function computations.

1.3.1 Digital Over-the-Air Computation

In digital OTA computation, the function that is to be computed maps between discrete sets. The computation is carried out repeatedly, and the objective of the corresponding coding scheme is that the probability of a decoding error approaches zero as the number of repetitions tends to infinity.

More formally, [NG07] introduces the problem of digital computation coding in the following way:

Definition 2. *A* digital computation coding problem *consists of the following:*

- *A MAC W which maps channel inputs $T_1, \ldots, T_K$ ranging over the input alphabets $\mathcal{T}_1, \ldots, \mathcal{T}_K$ to a channel output Y which ranges over the channel output alphabet $\mathcal{Y}$.*

[8] If the expense necessary for coordination and scheduling is also considered, this growth can even be superlinear.

[9] The function can be computed multiple times since the scheme can simply be repeated, however, the individual instances do not take advantage of the repeated computation.

- *An* objective function
$$f : \mathcal{S}_1 \times \cdots \times \mathcal{S}_K \to \mathcal{S}, \tag{1.10}$$
where $\mathcal{S}_1, \ldots, \mathcal{S}_K, \mathcal{S}$ *are finite sets.*

- *A probability distribution on* $\mathcal{S}_1 \times \cdots \times \mathcal{S}_K$.

The idea is that, given this problem, the transmitters encode their messages $s_1, \ldots, s_K$ as sequences of channel inputs in such a way that the receiver can, with high probability of success, reconstruct $f(s_1, \ldots, s_K)$ without necessarily being able to draw any further information about $s_1, \ldots, s_K$.

Definition 3. *An* (n, M, ε)-code *for a given digital computation coding problem consists of:*

- *for each* $k \in \{1, \ldots, K\}$, *an* encoder
$$E_k^M : \mathcal{S}_k^n \to \mathcal{T}_k^M \tag{1.11}$$

- *a* decoder
$$D^M : \mathcal{Y}^M \to \mathcal{S}^n \tag{1.12}$$

such that if the sequence of channel inputs is determined by $T_k^M := E_k^M(s_k^M)$, *the error probability at the receiver satisfies*

$$\mathbb{P}\left(D^M(Y^M) \neq (f(s_1^{(1)}, \ldots, s_K^{(1)}), \ldots, f(s_1^{(n)}, \ldots, s_K^{(n)}))\right) \leq \varepsilon. \tag{1.13}$$

These notions can then be used to define the analog of rate and capacity in classical source or channel coding problems.

Definition 4. *The* computation rate *of an* (n, M, ε)*-code is defined as the ratio* n/M. *A computation rate* $\mathcal{R}$ *is called* achievable *if there is a sequence of* (n, M, ε)*-codes of computation rate* $\mathcal{R}$ *where* $M \to \infty$ *and* $\varepsilon \to 0$. *The* computation capacity *is the supremum of all achievable computation rates.*

This framework is extended by allowing the alphabets $\mathcal{S}_1, \ldots, \mathcal{S}_K, \mathcal{S}$ to be infinite and then characterizing the rate-distortion trade-off. In any case, the computation coding problem combines source and channel coding because the encoders simultaneously remove redundancy from the sources and protect the transmission against channel noise. The authors of [NG07] note examples where the rate that separate source and channel coding can achieve is strictly less than the computation capacity.

In the setting with finite alphabets, the typical objective function considered is addition in a finite field, and the main application noted by the authors is physical layer network coding. This idea was seminal to a lot of follow-up research (e.g., [ZNGE09, NG11, OZE+11, NCNC16, GJRM+16]) which has expanded upon and refined the idea of using OTA computation as a means to increase the efficiency of network coding. Notably, there is also a work [GBS14] which proposes schemes that use digital computation codes in conjunction with a quantizer to compute functions that are of interest in other applications, such as the arithmetic mean, the geometric mean and the Euclidean norm.

1.3.2 Analog Over-the-Air Computation

The framework of digital computation codes is promising and its applications to network coding are highly relevant as they can realize impressive performance gains in wireless networks. However, is also has downsides in the context of other applications:

- The notion of computation capacity is an asymptotic one valid only for block lengths tending to infinity. While finite-blocklength results are certainly conceivable, it is nonetheless an inherent property of any approach involving digital coding that a certain number of repeated function computations is necessary in order to guarantee a reasonably low probability of decoding error. This can be problematic in applications where only a few computations are necessary or where protocols are used in which the roles of transmitters and receivers change frequently with only very few computations being done between these changes.

- To the best of our knowledge, the only known digital coding schemes which can deal with channel fading compute sums over finite fields for the application of network coding. Examples of functions that existing digital schemes cannot compute over fading channels include weighted sums which have a high relevance in the context of OTA ML, as well as maxima and various kinds of averages which are important in the context of consensus algorithms and control systems.

- The digital coding schemes can only deal with discrete messages. If real (or floating point) numbers are processed in a certain application, a quantizer needs to be added to the system. Since quantization is a form of source coding, this is somewhat in contrast with the observation that joint source-channel approaches are necessary to achieve optimum system performance.

A way to make OTA computation applicable where these disadvantages hinder the use of digital schemes is to process analog input values directly into an electromagnetic signal without first going through a sequence of bits (or other discrete values) as an intermediary

step. A striking observation in this context is that a standard wireless channel actually performs a summation of the transmitted signals (which, through their inphase-quadrature (IQ) representations, can be seen as points in Euclidean space). This opens the door to the computation both of weighted sums and (as a special case) arithmetic averages, which we have noted above are very relevant functions both for OTA ML and consensus algorithms. There are two important research questions that these observations directly raise:

- If we were able to compute real function values in an analog system without error, this would in the point-to-point case degrade to a possibility to losslessly transmit a real number through the wireless channel which would imply infinite Shannon capacity of the channel. Since this is known to be unrealistic for any real-world channel, we can immediately conclude that a certain amount of noise in the computed function values is unavoidable in any kind of analog OTA computation scheme. But is it possible to control the strength of the noise, for instance by providing tail bounds for its magnitude?
- We can expect from the structure of the wireless channel that it can compute sums in Euclidean space, but can we, with the use of suitable pre- and post-processing schemes, compute a larger class of functions OTA?

A pragmatic way to proceed in light of these difficulties is to attempt to find a subclass of functions that is small enough to permit nomographic representations which are suitable for use with noisy communication systems and at the same time large enough to contain most functions of interest in practical OTA computation problems.

There are several prior works that propose approaches to the OTA computation problem for functions particularly relevant to applications for consensus problems in wireless networks and ML over wireless channels. [GBS13] systematically explores the question of what types of functions can efficiently be OTA computed with analog schemes, also taking into account many system-level aspects such as the usage of analog OTA computation schemes in large wireless networks with changing topologies. [GS13] presents a scheme that is able to deal with imperfect synchronization and the presence of fading in OTA computation; extensive theoretical analyses for the asymptotic case is provided for the arithmetic and geometric mean functions. [GS14] presents pre- and post-processing schemes and an asymptotic analysis for the approximation of functions over a fast fading channel with noise in the case of multiple receive antennas. The work covers the case of no instantaneous channel state information (CSI) at the transmitter or receiver as well as two different types of instantaneous CSI at the transmitter. In [RGS16], under the assumption of known fading coefficients at the transmitter, a similar scheme is used for computing the sign of a weighted sum which is the decision function of a linear Support Vector Machine

(SVM) used for classification. As a result, the authors obtain a distributed binary classification scheme that is highly efficient in massively-sized wireless networks. In the more recent work [LZLV20], under the assumption that the sources are independent and the channel state is known at both the receiver and the transmitter, the authors derive analog OTA computation schemes for sums that are optimal in terms of mean square error. In the case of independent and identically distributed (i.i.d.) Gaussian sources the authors of [DSD20] show how to OTA compute sums over slow fading channels where the channel state information is available neither at the transmitter nor the receiver. The work also considers intersymbol interference and provides an asymptotic theoretical analysis as well as numerical results.

1.4 Applications of Over-the-Air Computation

OTA computation has potential applications in every setting in which such a large number of wireless devices share constrained wireless resources that it becomes inefficient or even infeasible to exclusively use traditional scheduling and separate decoding of all transmitted information before it is post-processed at the receiver. Furthermore, even if the available resources are tremendous, but the number of participating devices is so large that traditional scheduling becomes prohibitively expensive, OTA computation can be a useful tool to solve the problem. On the other hand, it inherently fuses concepts that have traditionally been separate in communication systems. We have already discussed the point that from an information theoretic perspective, it is a joint source-channel approach that breaks with the traditional separation paradigm. But also from the perspective of network architecture, it means using schemes on the physical layer that are at least in part tailored to specific applications, and traditional methods of scheduling and routing have to be adapted to be compatible. Therefore, OTA computation can be seen as a cross-layer approach that encompasses the entire network stack from the application layer all the way down to the physical layer. While the pre- and post-processing schemes can be proposed in such a generic manner that they can in principle be used for a large variety of potential applications, they still need to be carefully adapted to each one. There are two main fields of application that have recently motivated the development of OTA computation schemes, namely distributed OTA ML and consensus algorithms. In this section, we give a brief overview of these two applications.

1.4.1 Distributed Machine Learning

In this subsection, we take a look at distributed ML, in particular Federated Learning (FL), describe how this field branches into VFL and Horizontal Federated Learning (HFL) and

cite a few examples from the literature that approach FL problems with OTA computation methods. First, we need to define what ML is for the sake of this thesis, and we follow the formalism in [SC08].

Definition 5. *A* statistical inference problem *is a tuple* $(\mathfrak{X}, \mathfrak{Y}, \mathcal{P}, \mathfrak{L})$*, where*

- *the* feature alphabet $\mathfrak{X}$ *is a Polish space (usually a high-dimensional Euclidean space),*
- $\mathfrak{Y} \subseteq \mathbb{R}$ *is called the* label alphabet*,*
- $\mathcal{P}$ *is a probability measure on* $\mathfrak{X} \times \mathfrak{Y}$*,*
- $\mathfrak{L} : \mathfrak{X} \times \mathfrak{Y} \times \mathbb{R} \to [0, \infty)$ *is called the* loss function*.*

In the usual application setting, only the feature and label alphabets and the loss function are known about the statistical inference problem, while information about $\mathcal{P}$ is only known indirectly through a training sample.

Definition 6. *Given a statistical inference problem* $(\mathfrak{X}, \mathfrak{Y}, \mathcal{P}, \mathfrak{L})$*, a training sample of length* n *is a sequence* $(x_j, y_j)_{j=1}^n \in \mathfrak{X}^n \times \mathfrak{Y}^n$ *where each* (x_j, y_j) *is drawn i.i.d. according to* $\mathcal{P}$*.*

The objective in solving an ML problem is to find an ML model which can make predictions about the labels of newly drawn samples of $\mathcal{P}$, given only the features. An ML model is a mathematical object which provides, given a set of parameters, a labeling function. Examples of ML models are neural networks, SVMs and decision trees.

Definition 7. *Given a statistical inference problem* $(\mathfrak{X}, \mathfrak{Y}, \mathcal{P}, \mathfrak{L})$*, a* labeling function *is a function* $f : \mathfrak{X} \to \mathbb{R}$*. A labeling function induces a* risk *(sometimes also called* loss*)* $\mathfrak{R}_{\mathfrak{L},\mathcal{P}} := \mathbb{E}_{\mathcal{P}}\mathfrak{L}(X, Y, f(X))$*, where* (X, Y) *is the pair of random variables ranging over* $\mathfrak{X} \times \mathfrak{Y}$ *and distributed according to* $\mathcal{P}$*.*

Typically, the objective is to exploit the indirect knowledge that we have about $\mathcal{P}$ through the training sample to obtain a labeling function with low risk, which is usually the measure for how well we have solved the ML problem. To this end, a training procedure for a given ML model takes a training sample as its input and outputs parameters for the ML model. Therefore, in conjunction with the model, it maps training samples to labeling functions.

Distributed ML studies cases of ML problems where some of the information about the statistical inference problem or the training sample are only known at certain locations in a network. Although there are possibilities for communication between the agents in the

network, there are application-specific reasons for not transmitting the entire information to a central point. One particular instance of Distributed ML is called FL [KMRR16, MMR+17]. In FL, the initial main reason for not transmitting all the available information to a central point and then solving the ML problem in the traditional way is to preserve the privacy of the users from whom the training data is collected,[10] but communication efficiency also plays an increasingly important role. FL can be further categorized into HFL and VFL [YLCT19].

In HFL, each agent k out of a total of K agents in the system sees only a subsequence of the training sample $(x_{j_{k,i}}, y_{j_{k,i}})_{i=1}^{n_k}$. In principle, it is possible for each agent to train its own local ML model based on the locally available training sample. Depending on the application at hand, however, this can incur several difficulties:

- The locally available training subsamples may simply be too small to train an ML model and obtain an acceptable risk.
- The way in which the locally available training subsamples are drawn from the overall training sample may be such that the subsamples are not i.i.d. or do not follow $\mathcal{P}$ [ZLL+18]. For instance, it is common for the subsamples to be biased towards certain labels in a way the overall training sample is not.

Distributed optimization algorithms can be used to carry out the training in a decentralized manner. They make, either at one central point or everywhere in the network, a trained ML model available that benefits from the whole training sample without transmitting it through the network in its entirety. There is a huge body of recent research (cf., e.g., [AG20b, ZWH20, ASK19, YJSD20, AG19, GdKS+19, ZDHL20, OUG19, AG20a, ZLD+20, ZCL+19, AOGE20, SC20, SZG20, STL20, ADG19, CT20] and references therein) into ways to perform distributed optimization algorithms such as stochastic gradient descent exploiting OTA computation. This approach can achieve fundamentally more favorable scaling laws than would be possible otherwise.

In VFL, the data is distributed in a different way: In a system with K agents, the statistical inference problem has a feature alphabet $\mathfrak{X} = \mathfrak{X}_1 \times \cdots \times \mathfrak{X}_K$ that is a Cartesian product of K feature spaces. A feature $x \in \mathfrak{X}$ can therefore be written as a tuple $x = (x_1, \ldots, x_K)$ and the training sample is of the form $((x_{1,j}, \ldots, x_{K,j}), y_j)_{j=1}^{n}$ where each agent k has only the local training sample $(x_{k,j}, y_j)_{j=1}^{n}$. Correspondingly, when training is complete and a label needs to be estimated, each agent k sees only the projection to $\mathfrak{X}_k$ of the observed feature. Since the labeling function has the whole feature space $\mathfrak{X} = \mathfrak{X}_1 \times \cdots \times \mathfrak{X}_K$ as its domain, the arguments to compute it are not available at any

[10] A major motivation for introducing the FL framework was Gboard, a software made by Google which is used as the default keyboard on many Android devices [MR17].

single point in the network and it is therefore natural to attempt to compute the labeling function OTA. So there are two important research questions in OTA-VFL:

- Given a training sample that is distributed as described above, how can we carry out a distributed training procedure exploiting OTA computation that scales better than linear in the number of agents involved?
- Given the trained model (which also is available only in a distributed manner), how can we compute the labeling function using the OTA approach?

The first question is quite similar to the main research question in OTA-HFL and there is some hope that tools from this field could be suitably adapted. The second question is more specific to the VFL scenario, and we note that many standard ML labeling functions naturally take the form of (weighted) sums. Examples are layers of neural networks (the activation function can be evaluated afterwards in post-processing if necessary) and the linear SVMs that have been used for OTA-VFL in [RGS16]. Contrary to OTA-HFL, there does not appear to be a large body of research on OTA-VFL. Besides [RGS16] and the work presented in Chapter 3, we are not aware of any works that propose to leverage OTA computation in a VFL scenario. However, there is a related research area called *Type-Based Multiple-Access* [MT06, MNT07] which is concerned with solving problems very similar to those approached by OTA VFL such as anomaly detection in extremely large networks. An important difference is that instead of transmitting analog values directly, this approach relies on a prior quantization step and then exploits the fact that the number of quantization levels usually does not grow with the number of transmitters in the system. Furthermore, this approach uses statistical methods and knowledge about the involved probability distributions, while the VFL approach that we use in Chapter 3 is based on the ML paradigm and hence does not require a priori knowledge of the underlying distributions.

1.4.2 Consensus over Wireless Channels

Consensus problems deal with combining opinions of participating agents to achieve an agreement that encompasses their information about or subjective assessments of an object. They have originally appeared as statistical problems in which the opinions are probability distributions which have to be combined to form a consensus distribution. In [EG59] this is illustrated as a horse race betting problem where the agents' opinions are probability distributions on which horse will win the race. They place their bets according to these opinions and the overall track's odds that result from these bets are considered the consensus which in a certain way combines all the participating agents' opinions. The

problem has subsequently been stated as one of combining various experts' opinions and researched extensively to aid with decision making in the context of management sciences (see, e.g., [Win68, Fre85] and references therein).

The research on this theory has later been applied to problems of multisensor fusion and pattern recognition [BS92] and since found a multitude of other applications in engineering sciences [OSFM07]. In some of the engineering applications the nature of the difficulty of the problem has shifted significantly: Often, an opinion is simply a real number or vector and the way the opinions have to be combined to form the consensus is fully prescribed by the application at hand and is fairly simple compared to the original consensus problem. For instance, the consensus can be the arithmetic average (with applications, e.g., in formation control and flocking of autonomous vehicles [OS06]) or the maximum of the opinions (examples for applications include task assignment [BCH08] and traffic automation [MDR19]). In these applications, the challenge is that it is infeasible to aggregate the opinions in a central point because the communication cost or the time delay incurred would be prohibitive. In these cases, distributed consensus algorithms are used that seek to make the consensus value available to agents in a large network with a minimum of communication required between the agents [OSFM07].

In many applications, the communication links between the agents are wireless channels, and indeed, several agents can be linked to another agent via a wireless broadcast or multiple-access channel. Some works that exploit these properties to reach average or maximum consensus in a way that is more communication-efficient than would be possible with point-to-point communication are [ICJ12,MSR18,MDR19,MASR21]. We expect that theoretical analysis of OTA computation techniques could serve as a building block to enhance the efficiency and in particular the scaling behavior of the communication cost in the number of participating agents. Moreover, this way it would be possible to provide additional theoretical error guarantees for consensus schemes that exploit the superposition of signals in the wireless channel. In [9], we have proposed a maximum consensus scheme which leverages analog OTA computation of sums to make the maximum of the agents' opinions available at the receiver in a wireless MAC with no fading but with additive noise. The OTA computation schemes proposed in this thesis can be used to extend these results to channels exhibiting fast fading [2]. It is in particular worth noting that the scheme proposed in [9] can OTA compute the maximum of the agents' opinions in a wireless channel although we do not expect the maximum function to satisfy Definition 10. This is achieved not through a single OTA computation but through a multi-step protocol that alternates between analog OTA computation of sums and digitally coded broadcast communication. We believe therefore that such multi-step protocols are a potentially promising approach to computing functions that are not amenable to the one-shot OTA

computation methods we propose in this thesis. This is at the cost of higher system and communication complexity, but a favorable scaling of communication cost in extremely large networks would be retained.

1.5 Contribution and Outline

This thesis proposes schemes for the approximation of functions of distributed arguments and deals with the question how OTA computation can be made robust under harsh channel conditions. We also give examples of applications to distributed ML and undertake first steps towards hardening such schemes against eavesdroppers. Although we provide results of numerical simulations for OTA computation under harsh channel conditions and its applications to distributed ML, the main focus of this thesis is on mathematically proving theoretical guarantees. In this sense, our work is complementary to the many predominantly empirical studies in the literature which investigate OTA computation and its applications in specific real-world scenarios through numerical simulations.

The material contained in this introduction is based on the book chapter [7] with the exception of this subsection which uses excerpts from the publications the respective chapters are based on.

In **Chapter 2**, we propose a distributed function approximation scheme which can be applied to wireless channels to perform analog OTA computation without instantaneous CSI. It is not necessary to assume any probability distribution on the data. Therefore, the scheme works equally well for independently distributed data as it does for arbitrarily correlated values. Our error analysis relies on mathematical tools from high-dimensional statistics which we adapt so that they can be applied to the scenario at hand. The resulting error guarantees are nonasymptotic and are valid for any fading and noise in the class of sub-Gaussian distributions. This class contains Gaussian as well as many non-Gaussian distributions of practical interest. Moreover, our scheme can deal with correlated fading and noise and compute a larger class of functions than previous works with theoretically proven bounds. In particular, we do not require linearity of the function to be approximated, which is, e.g., demonstrated by the fact that we can compute p-norms OTA. We conclude the chapter with numerical evaluations for a few selected cases.

The material in Chapter 2 is based on the conference publications [2, 3] and the journal publication [4]. The introductory remarks are in part based on the book chapter [7].

In **Chapter 3**, we propose applications of our scheme to the OTA computation of both regressors and classifiers in VFL and validate the proposed OTA computation schemes and the envisioned applications in ML with extensive numerical simulations for the case of a binary classification problem.

The material in Chapter 3 is based on the conference publications [2, 3] and the journal publication [4].

In **Chapter 4**, we propose a novel framework and result for incorporating security considerations by including a friendly jammer in the system which deteriorates the eavesdropper's SNR while not impacting the legitimate receiver's ability to obtain an approximation of the function value which is to be OTA computed. As an example, we show how this jamming strategy can be applied to an AWGN channel where the OTA computed function is an arithmetic average. We prove that the security guarantee translates to a lower bound on the mean square error of the eavesdropper's function estimate. Our proofs heavily rely on two information-theoretic results which we also prove in this chapter. The first information-theoretic ingredient is a theorem on compound channel coding for continuous alphabets. It is a generalization of the result of the part of [RV68] which considers finite Gaussian channels and we can consequently recover this result as a special case. The second information-theoretic ingredient is an achievability theorem on resolvability of channels with continuous alphabets. Note that the publication [1] on which this part of the chapter is based contains also a converse result and a second-order direct result, both of which are not part of this thesis.

The material in Chapter 4 is based on the conference publications [1, 5], and on the submitted journal paper [6]. A revised version of the latter paper is currently under consideration for publication. The introductory remarks are in part based on the book chapter [7].

Further Work. During my time at Technische Universität Berlin as a PhD student, we were able to obtain further results which are, however, not part of this thesis. They are listed in chronological order.

- In [8], we revisit a secrecy proof for the MAC from the perspective of channel resolvability and refine the approach to obtain novel results on the second-order achievable rates.

- In a work with Navneet Agrawal [9], we propose a multi-step communication protocol based on the idea of OTA computation. The resulting scheme is able to compute the maximum of a set of distributed values in a wireless network. While the communication cost and complexity are greater than in the one-shot method studied in this thesis, they exhibit a similarly favorable scaling behavior as the number of transmitters in the system grows. Computation of the maximum function is relevant in many distributed control systems and moreover, this example shows that OTA computation can potentially be applied even to functions that are not contained in the class of functions studied in this thesis.

- In a series of works with Zoran Utkovski, Patrick Agostini, Miguel Gutierrez-Estevez and Daniel Schäufele [10–12], we investigate how physical layer security methods could be integrated in future cellular networks. The security of this class of communication schemes is always based on the assumption that the channel of the legitimate receiver is stronger than that of the eavesdropper. As a way of ensuring that this assumption holds true in a real-world wireless network, we propose the concept of secrecy maps. They are based on radio maps and help the infrastructure and legitimate users of wireless networks to predict physical locations at which secure communication is possible as well as give guidance on how conditions in other locations can be improved, for instance with the use of intelligent reflective surfaces or friendly jamming.

- In [13], we propose a new proof method for direct coding theorems for wiretap channels where the eavesdropper has access to a quantum version of the transmitted signal on an infinite dimensional Hilbert space. The main part of this proof is a direct coding result on channel resolvability which states that there is only a doubly exponentially small probability that a standard random codebook does not solve the channel resolvability problem for the classical-quantum channel.

2 Distributed Function Approximation in Fading and Correlated Channels

In this section, we discuss our Distributed Function Approximation (DFA) scheme which we proposed in [2] and extended in [3, 4]. The goal in introducing it was to provide a flexible framework that can deal with such a large class of wireless channels that the scheme would be robust to departures from common assumptions on the system model such as Gaussianity of the fading and noise. At the same time, the class of functions for OTA computation should contain the most relevant ones in current applications (which are mainly weighted sums). It should also be large enough to provide flexibility and make the DFA scheme applicable in scenarios where functions that have not yet received much attention are computed OTA. Another important consideration in the design of the scheme was the distribution of the sources. Many existing works on OTA computation assume a particular source distribution for their theoretical analysis, and usually require that the transmitted values are independently distributed between the transmitters. Since this requirement is extremely difficult to check in practice, we have decided to not model the sources stochastically. Instead, we show that the bound on the approximation error is satisfied uniformly over all possible values of the sources. This yields a worst-case analysis with theoretically proven error guarantees that are valid for every distribution of the sources, even if there is arbitrary correlation between them. In addition, the error bounds are nonasymptotic in the sense that they are valid for any number of channel uses, not just for a sufficiently large one.

In theoretical works, it is of particular importance that the considered channel model is sufficiently general so that the assumptions made are met in relevant practical scenarios. The commonly considered Gaussian fading channel is an approximation that is often adopted because it is relatively easy to treat and is quite close to reality in many scenarios of interest.

However, it is well known and has been confirmed in extensive measurement campaigns [MS93, BRB93, Mid99] that there are many natural and artificial sources of noise that do not conform to the assumption of being i.i.d. Gaussian such as automobile ignition, power line emissions, atmospheric disturbances or interfering wireless communica-

tions [MS93].

Moreover, common arguments that the fast fading in wireless channels is Gaussian employ the Central Limit Theorem [YC16, Section 2.4.1] and therefore assume a large number of multipath components, which is not always the case. In scenarios with limited mobility, it is also possible that the fast fading realizations are not independent between channel uses. In the case of the OTA computation scheme proposed in [2], it can deteriorate performance and therefore, it is desirable to be able to quantify this deterioration.

In this work, we therefore consider a channel model that encompasses a large class of possible probability distributions of fading and noise, the class of sub-Gaussian distributions. The analysis provided is valid also for fading and noise exhibiting an arbitrary correlation structure, with practically useful bounds in many relevant cases. In Section 2.1.1, we define precisely what sub-Gaussian means and in Section 2.1.3, we give examples of practically relevant cases that are covered by our system model assumptions.

2.1 System Model

2.1.1 Sub-Gaussian Random Variables

We begin with a short overview of the relevant definitions and properties of sub-Gaussian random variables. More on this topic can be found in Section 2.6 and in [BK00, Wai19, Ver18].

For a random variable X, we define[11]

$$\tau(X) := \inf\left\{t > 0 : \forall \lambda \in \mathbb{R} \;\; \mathbb{E}\exp\left(\lambda(X - \mathbb{E}X)\right) \leq \exp\left(\lambda^2 t^2/2\right)\right\}. \tag{2.1}$$

$\exp(\cdot)$ denotes the exponential function with Euler's number $e \approx 2.71828$ as the basis.

X is called a sub-Gaussian random variable if $\tau(X) < \infty$. The function $\tau(\cdot)$ defines a semi-norm on the set of sub-Gaussian random variables [BK00, Theorem 1.1.2], i.e., it is absolutely homogeneous, satisfies the triangle inequality, and is non-negative. $\tau(X) = 0$ does not necessarily imply $X = 0$ unless we identify random variables that are equal almost everywhere. Examples of sub-Gaussian random variables include Gaussian and bounded random variables.

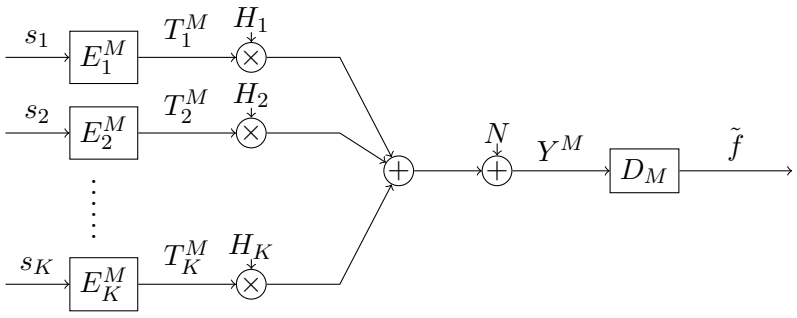

Figure 2.1: System model for DFA.

2.1.2 Channel and Correlation Model

We consider the following channel model with K transmitters and one receiver: For $m = 1, \ldots, M$, the channel output at the m-th channel use is given by

$$Y(m) = \sum_{k=1}^{K} H_k(m) T_k(m) + N(m). \tag{2.2}$$

Here and hereafter, the notation is defined as follows:

- $T_k(m) \in \mathbb{C}$ is a transmit symbol, where $\mathbb{C}$ denotes the set of complex numbers. We assume a peak power constraint $|T_k(m)|^2 \leq \mathfrak{P}$ for $k = 1, \ldots, K$ and $m = 1, \ldots, M$.
- $H_k(m)$, $k = 1 \ldots, K$, $m = 1, \ldots, M$, are complex-valued random variables such that for every $m = 1, \ldots, M$ and $k = 1, \ldots, K$, the real part $H_k^{\mathrm{Re}}(m)$ and the imaginary part $H_k^{\mathrm{Im}}(m)$ of $H_k(m)$ are sub-Gaussian random variables with mean zero and variance 1.
- $N(m)$, $m = 1, \ldots, M$, are complex-valued random variables. We assume that the real and imaginary parts $N^{\mathrm{Re}}(m), N^{\mathrm{Im}}(m)$ of $N(m)$ are sub-Gaussian random variables with mean zero for $m = 1, \ldots, M$.

In order to be able to apply a variation of the Hanson-Wright inequality as a tool, we give the formal description of our dependence model in terms of matrices and vectors with real entries.

We define

$$\mathcal{H} := (\mathcal{H}(1), \ldots, \mathcal{H}(2M))^T \tag{2.3}$$

[11] Note that other norms on the space of sub-Gaussian random variables that appear in the literature are equivalent to $\tau(\cdot)$ (see, e.g., [BK00]). The particular definition we choose here matters, however, because we derive results in which no unspecified constants appear.

where $\cdot^T$ denotes the transpose and for $m = 1, \ldots, M$,

$$\mathcal{H}(2m-1) := (H_1^{\mathrm{Re}}(m), \ldots, H_K^{\mathrm{Re}}(m))$$
$$\mathcal{H}(2m) := (H_1^{\mathrm{Im}}(m), \ldots, H_K^{\mathrm{Im}}(m)).$$

So $\mathcal{H}$ is the vector of all fading coefficients. Similarly, let

$$N := (N^{\mathrm{Re}}(1), N^{\mathrm{Im}}(1), \ldots, N^{\mathrm{Re}}(M), N^{\mathrm{Im}}(M))^T \tag{2.4}$$

be the vector of all the instances of additive noise. The dependence model we consider is such that there is a vector $\mathcal{G}$ of $(2KM + 2M)$ independent random variables with sub-Gaussian norm at most 1 and matrices $\mathcal{A} \in \mathbb{R}^{2KM \times (2KM+2M)}$ and $\mathcal{B} \in \mathbb{R}^{2M \times (2KM+2M)}$ such that

$$\mathcal{H} = \mathcal{A}\mathcal{G}, \; N = \mathcal{B}\mathcal{G}.$$

Our correlation model captures both correlations between users and in the time domain, but the impact of these two types of correlation on the performance of the proposed scheme is different. For this reason, we need the following definition, which describes a scenario where there can be arbitrary correlation in the time domain but no harmful correlation between different users. In order to be as unrestrictive as possible, the following definition prohibits only correlations between different users at the same complex dimension of the same channel use.

Definition 8. *We say that the fading is* user-independent *if for every $k_1 \neq k_2$, $j \in \{\mathrm{Re}, \mathrm{Im}\}$ and m, the random variables $H_{k_1}^j(m)$ and $H_{k_2}^j(m)$ are independent.*

In Section 2.3.1, we discuss the impact of deviations from the user-independence assumption on the performance of the proposed method.

Remark 1. *User-independent fading can be characterized based on the form of $\mathcal{A}$. That is, the fading is user-independent iff it can be written as $\mathcal{H} = \mathcal{A}\mathcal{G}$ with*

$$\mathcal{A} = \begin{pmatrix} \mathcal{A}^{(1)} \\ \vdots \\ \mathcal{A}^{(2M)} \end{pmatrix},$$

where for all m, $\mathcal{A}^{(m)} \in \mathbb{R}^{K \times (2MK+2M)}$ and each $\mathcal{A}^{(m)}$ has at most one nonzero entry per column. This is because $\mathcal{H}(m) = \mathcal{A}^{(m)}\mathcal{G}$ and therefore at most one nonzero entry in a column of $\mathcal{A}^{(m)}$ means that at most one entry in $\mathcal{H}(m)$ depends on the entry in $\mathcal{G}$ to which

the column refers.

2.1.3 Discussion of the System Model Assumptions

The most distinguishing feature of our channel model compared to assumptions made in prior work on OTA computation is that we generalize the distribution of the fading and noise to be sub-Gaussian and to feature arbitrary correlations. We argue that this guarantees robustness against departure from standard assumptions such as independent or Gaussian fading or noise. Regarding the chosen dependence model, we remark that in the case of $\mathcal{G}$ distributed i.i.d. standard Gaussian, this amounts to a standard representation of arbitrarily correlated (and thus arbitrarily interdependent) multivariate Gaussian vectors. Therefore, by replacing $\mathcal{G}$ with a vector of independent sub-Gaussian entries, we obtain a straightforward generalization of the Gaussian case, which specializes to arbitrary correlations (although in the non-Gaussian case, not to arbitrary stochastic dependence). In particular, the following important cases are specializations of our channel model:

- Perfect Gaussian fast fading with i.i.d. bivariate Gaussian fading and additive white Gaussian noise.

- Scenarios where the number of multipath components is not sufficiently large for an appeal to the Central Limit Theorem to argue that the fading is complex Gaussian. For instance, if there is only one multipath component with a length that has small random variations, the resulting complex fading will have a distribution supported on a narrow annulus in the complex plane. Such a distribution is not Gaussian, but sub-Gaussian and therefore covered by our channel model.

- Scenarios where due to limited or slow movement of transmitters, receiver and environment, the independence assumption between subsequent realizations of the channel fading is not satisfied or where the diversity in the radio environment is so small that some of the users have correlated channels. One example of a widely used channel model that is a special case of the one considered in this paper is the block fading channel. In this case, we have a correlation of 1 between fading realizations of the same block and 0 between fading realizations in different blocks.

- Any of the above scenarios where in addition to thermal additive noise, we have interference from users outside the system. E.g., in the case of digital modulated signals, such interference consists of a sequence of transmitted constellation points, which is not Gaussian. Also, such interfering signals are inherently bursty in nature and therefore cannot be argued to be independent between different points in time.

However, signals from realistic transmitters are always bounded in amplitude and therefore sub-Gaussian, which means that they are covered by our system model.

- Any of various types of artificial and natural interference that is not necessarily Gaussian, but limited in power. [MS93, BRB93, Mid99] investigate various sources of non-Gaussian interference of this type, both correlated and uncorrelated over time, through theoretical modeling as well as extensive experiments and measurements confirming the theoretical models and the non-Gaussianity of the sources. [MS93, Table 2.1] enumerates several examples of this type of interference, including, e.g., solar radiation, automobile ignition and power line EM emissions.

The other main difference between our system model and the models used in existing works is that we do not make any assumption on a distribution of the input data at the transmitters. This means in particular that we cover arbitrary dependencies in the input data, which can be very important for applications, because the input data can depend, e.g., on local sensor readings recorded by devices or on training data collected for the same ML problem. Therefore, in many relevant application scenarios, it is important that the transmission schemes employed are robust even to high, but unknown levels of correlations in the transmitted data.

2.2 Problem Statement

2.2.1 Distributed Approximation of Functions

Our goal is to approximate functions $f : \mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ in a distributed setting. The sets $\mathcal{S}_1, \ldots \mathcal{S}_K \subseteq \mathbb{R}$ are assumed to be closed and endowed with their natural Borel σ-algebras, and we consider the product σ-algebra on the set $\mathcal{S}_1 \times \ldots \times \mathcal{S}_K$. Furthermore, the functions $f : \mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ under consideration are assumed to be measurable in what follows.

Definition 9. *An admissible DFA scheme for $f : \mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ with M channel uses, depicted in Fig. 2.1, is a pair (E^M, D^M), consisting of:*

1. *A pre-processing function $E^M = (E_1^M, \ldots, E_K^M)$, where each E_k^M is of the form*

$$E_k^M(s_k) = (E_k(m, s_k, U_k(m)))_{m=1}^M \in \mathbb{C}^M$$

where $U_k(1), \ldots, U_k(M)$ are random variables and the map

$$(s_k, u_1, \ldots, u_M) \mapsto (E_k(m, s_k, u_m))_{m=1}^M \in \mathbb{C}^M$$

is measurable for s_k ranging over $\mathcal{S}_k$ and $u_1, \ldots, u_M$ ranging over the same sets as the i.i.d. random variables $U_k(1), \ldots, U_k(M)$. The encoder E_k^M is subject to the peak power constraint $|E_k(m, s_k, U_k(m))|^2 \leq \mathfrak{P}$ for all $k = 1, \ldots, K$ and $m = 1, \ldots, M$.

2. *A post-processing function D^M: The receiver is allowed to apply a measurable recovery function $D^M : \mathbb{C}^M \to \mathbb{R}$ upon observing the output of the channel.*

So in order to approximate f, the transmitters apply their pre-processing maps to $(s_1, \ldots, s_K) \in \mathcal{S}_1 \times \ldots \times \mathcal{S}_K$ resulting in $E_1^M(s_1), \ldots, E_K^M(s_K)$, which are sent to the receiver using the channel M times. The receiver observes the output of the channel and applies the recovery map D^M. The whole process defines an estimate $\tilde{f}$ of f.

Let $\varepsilon > 0, \delta \in (0,1)$ and $f : \mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ be given. We say that f is ε-approximated after M channel uses with confidence level δ if there is a DFA scheme (E^M, D^M) such that the resulting estimate $\tilde{f}$ of f satisfies

$$\mathbb{P}(|\tilde{f}(s^K) - f(s^K)| \geq \varepsilon) \leq \delta \tag{2.5}$$

for all $s^K := (s_1, \ldots, s_K) \in \mathcal{S}_1 \times \ldots \times \mathcal{S}_K$. Let $M(f, \varepsilon, \delta)$ denote the smallest number of channel uses such that there is an approximation scheme (E^M, D^M) for f satisfying (2.5). We call $M(f, \varepsilon, \delta)$ the communication cost for approximating a function f with accuracy ε and confidence δ.

2.2.2 The Class of Functions to be Approximated

A measurable function $f : \mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ is called a *generalized linear function* if there are bounded measurable functions $(f_k)_{k \in \{1,\ldots,K\}}$, with $f(s_1, \ldots, s_K) = \sum_{k=1}^K f_k(s_k)$, for all $(s_1, \ldots, s_K) \in \mathcal{S}_1 \times \ldots \times \mathcal{S}_K$. The set of generalized linear functions from $\mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ is denoted by $\mathcal{F}_{K,\text{lin}}$. Our main object of interest will be the following class of functions.

Definition 10. *A measurable function $f : \mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ is said to belong to $\mathcal{F}_{mon}$ if there exist bounded and measurable inner functions functions $(f_k)_{k \in \{1,\ldots,K\}}$, a measurable set $A \subseteq \mathbb{R}$ with the property $f_1(\mathcal{S}_1) + \ldots + f_K(\mathcal{S}_K) \subseteq A$, a measurable outer function $F : A \to \mathbb{R}$ such that for all $(s_1, \ldots, s_K) \in \mathcal{S}_1 \times \ldots \times \mathcal{S}_K$ we have*

$$f(s_1, \ldots, s_K) = F\left(\sum_{k=1}^K f_k(s_k)\right), \tag{2.6}$$

and there is a strictly increasing measurable function $\Phi : [0, \infty) \to [0, \infty)$ with $\Phi(0) = 0$ and

$$|F(a_1) - F(a_2)| \leq \Phi(|a_1 - a_2|) \tag{2.7}$$

for all $a_1, a_2 \in A$. *We call the function* Φ *an* increment majorant *of* f.

Some examples of functions in $\mathcal{F}_{\text{mon}}$ are:

1. Obviously, all $f \in \mathcal{F}_{K,\text{lin}}$ belong to $\mathcal{F}_{\text{mon}}$.
2. For any $f \in \mathcal{F}_{K,\text{lin}}$ and B-Lipschitz function $F : \mathbb{R} \to \mathbb{R}$ we have $F \circ f \in \mathcal{F}_{\text{mon}}$ with $\Phi : [0, \infty) \to [0, \infty)$, $a \mapsto Ba$.
3. If $f \in \mathcal{F}_{K,\text{lin}}$ and F is (C, α)-Hölder continuous, i.e., for all $a_1, a_2 \in A$, $|F(a_1) - F(a_2)| \leq C\,|a_1 - a_2|^{\alpha}$, then $F \circ f \in \mathcal{F}_{\text{mon}}$ with $\Phi : a \mapsto Ca^{\alpha}$.
4. For any $p \geq 1$ and $\mathcal{S}_1, \ldots, \mathcal{S}_K$ compact, $||\cdot||_p \in \mathcal{F}_{\text{mon}}$. In this example we have $f_k(s_k) = |s_k|^p$, $k = 1, \ldots, K$, $F : [0, \infty) \to [0, \infty)$, $a \mapsto a^{1/p}$, and $F = \Phi$.
   This can be seen as follows. We have to show that for all nonnegative $a_1, a_2 \in \mathbb{R}$ and $p \geq 1$ we have
   $$|a_1^{1/p} - a_2^{1/p}| \leq |a_1 - a_2|^{1/p}. \tag{2.8}$$
   We can assume w.l.o.g. that $a_1 < a_2$ holds. Then since
   $$|a_1^{1/p} - a_2^{1/p}| = |a_2|^{1/p} \left(1 - \left(\frac{a_1}{a_2}\right)^{1/p}\right)$$
   it suffices to prove that for all $a \in [0, 1]$ and $p \geq 1$ we have $1 - a^{1/p} \leq (1 - a)^{1/p}$. Now since $a^{1/p} + (1 - a)^{1/p} \geq a + (1 - a) = 1$ for $a \in [0, 1]$ and $p \geq 1$, we can conclude that (2.8) holds.

Given a function $f \in \mathcal{F}_{\text{mon}}$, we implicitly fix a representation (2.6) and define the total spread of the inner part of $f \in \mathcal{F}_{\text{mon}}$ as

$$\bar{\Delta}(f) := \sum_{k=1}^{K} (\phi_{\max,k} - \phi_{\min,k}), \tag{2.9}$$

along with the max-spread

$$\Delta(f) := \max_{1 \leq k \leq K} (\phi_{\max,k} - \phi_{\min,k}), \tag{2.10}$$

where

$$\phi_{\min,k} := \inf_{s \in \mathcal{S}_k} f_k(s), \quad \phi_{\max,k} := \sup_{s \in \mathcal{S}_k} f_k(s). \tag{2.11}$$

We define the relative spread of f with power constraint $\mathfrak{P}$ as

$$\Delta(f \| \mathfrak{P}) := \mathfrak{P} \cdot \frac{\bar{\Delta}(f)}{\Delta(f)}. \tag{2.12}$$

2.3 Main Result

We are now in a position to state our main theorem on approximation of functions in $\mathcal{F}_{\text{mon}}$. We use $\|\cdot\|_{\text{op}}$ and $\|\cdot\|_{\text{F}}$ to denote the operator and Frobenius norm of matrices, respectively.

Theorem 2. *Let $f \in \mathcal{F}_{mon}$, $M \in \mathbb{N}$, and the power constraint $\mathfrak{P} \in (0, \infty)$ be given. Let Φ be an increment majorant of f. Assume the fading and noise are correlated as determined by matrices $\mathcal{A}$ and $\mathcal{B}$. Let $\mathcal{A}_i \in \mathbb{R}^{2MK \times (2MK+2M)}$ be a matrix in the form of Remark 1 which generates user-independent fading that approximates $\mathcal{A}$ in the sense that*

$$\|(\mathcal{A} + \mathcal{A}_i)(\mathcal{A} - \mathcal{A}_i)^T\|_{op} \leq \eta.$$

Then there exist pre- and post-processing operations such that

$$\begin{aligned}\mathbb{P}\left(\left|\bar{f} - f(s_1, \ldots, s_K)\right| \geq \varepsilon\right) \\ \leq 2\exp\left(-\frac{M\Phi^{-1}(\varepsilon)^2}{16\mathbf{F} + \mathbf{D} + 4\Phi^{-1}(\varepsilon)\mathbf{L}}\right) + 2\exp\left(-\frac{M\Phi^{-1}(\varepsilon)^2}{256\mathbf{F} + 32\Phi^{-1}(\varepsilon)\mathbf{L}}\right), \quad (2.13)\end{aligned}$$

where

$$\begin{aligned}\mathbf{L} &= \left(\sqrt{\bar{\Delta}(f)}\|\mathcal{A}\|_{op} + \sqrt{\frac{\Delta(f)}{\mathfrak{P}}}\|\mathcal{B}\|_{op}\right)^2 \\ \mathbf{F} &= \mathbf{L}\left(\sqrt{\frac{\bar{\Delta}(f)}{M}}\|\mathcal{A}\|_F + \sqrt{\frac{\Delta(f)}{\mathfrak{P}M}}\|\mathcal{B}\|_F\right)^2 \\ \mathbf{D} &= \left(4\sqrt{2M}\bar{\Delta}(f)\eta + 4\frac{\Delta(f)}{\sqrt{\mathfrak{P}M}}\|\mathcal{A}\mathcal{B}^T\|_F\right)^2.\end{aligned}$$

In the following, we sketch how the pre- and post-processing schemes of Theorem 2 work. For a full formal definition, we refer the reader to the proof in Section 2.5. The pre- and post-processing schemes are similar (although not identical) to the ones that have appeared in [GS14, RGS16]. They are based on the same idea of combining random phase shifts at the transmitter and averaging at the receiver to mitigate the impact of the unknown fading and noise.

We consider fast fading (which can have any sub-Gaussian distribution) and assume that CSI is available neither at the transmitter nor at the receiver. One example is i.i.d. complex standard normal fading, which has a uniformly distributed phase. In this case, since there is no CSI, the phase of the received signal cannot carry any useful information. On the other hand, the phase difference between the signals from different transmitters

plays a crucial role, since it determines whether the signals constructively or destructively overlap at the receiver. We mitigate the impact of a destructive superposition by applying a random phase shift at the transmitters and averaging the transmission over multiple channel uses. This has the added benefit of averaging out additive noise.

In summary, the pre-processing is performed by applying the following steps at each transmitter k (see Section 2.5.1):

- Apply the inner function f_k from the nomographic representation (2.6).
- Shift and rescale to satisfy the power constraint.
- Apply a random phase shift $U_k(m)$ that is independent for each channel use m.

One option for the random phase shift $U_k(m)$ is to draw it uniformly from the complex unit circle, but as we argue in the proofs, it is actually sufficient to draw it uniformly from $\{-1, 1\}$.

As described above, the phase of the received signal carries no useful information due to the absence of CSI. Moreover, to compensate for the phase differences between the transmitters and reduce the influence of additive noise, some form of averaging is required. We therefore perform the following steps (see Section 2.5.2):

- Compute the total energy of the received signal.
- Subtract the energy of the additive noise over the receive time slot (this is statistical information and does not require knowledge of the instantaneous noise realizations).
- Invert the rescaling and shift that have been applied during pre-processing.
- Apply the outer function F from the nomographic representation (2.6).

We remark that these pre- and post-processing steps, while specific to the fast fading scenario, are unmodified compared to the steps used in [2], where we considered only the case of uncorrelated fading and noise. On the other hand, the error bound of Theorem 2 and the statistical tools used to prove it are different. Moreover, the error bound depends on the correlation structure of the fading and noise, while the pre- and post-processors do not need this information.

Remark 2. *If no suitable user-independent approximation for $\mathcal{A}$ is available, we can always choose $\mathcal{A}_i := 0$, which results in $\eta = \|\mathcal{A}\|_{op}^2$.*

Remark 3. *In order to gain more freedom for optimizing the bound for a given correlation structure, it is possible to replace $\mathcal{A}_i$ in Theorem 2 with $\mathcal{A}_i\mathcal{A}_U$, where $\mathcal{A}_U \in \mathbb{R}^{(2MK+2M)\times(2MK+2M)}$ is a unitary matrix. This requires only minor adaptations in the proof of the theorem.*

Corollary 1. *For the approximation communication cost, we have*

$$M(f,\varepsilon,\delta) \leq \frac{\log 4 - \log \delta}{\Phi^{-1}(\varepsilon)^2}\Gamma, \tag{2.14}$$

where

$$\Gamma := \max\left(16\mathbf{F} + \mathbf{D} + 4\Phi^{-1}(\varepsilon)\mathbf{L}, 256\mathbf{F} + 32\Phi^{-1}(\varepsilon)\mathbf{L}\right).$$

Proof. We upper bound (2.13) as

$$\mathbb{P}(|\bar{f} - f(s^K)| \geq \varepsilon) \leq 4\exp\left(-\frac{M\Phi^{-1}(\varepsilon)^2}{\Gamma}\right),$$

and solve the expression for M concluding the proof. □

Remark 4. *If F is B-Lipschitz continuous, we can replace $\Phi^{-1}(\varepsilon)$ in (2.14) and the expression for Γ with ε/B.*

2.3.1 Special Cases of Theorem 2

In this subsection we discuss the bound of Theorem 2 and illustrate it with two examples. In order to be a useful bound, (2.13) should approach 0 as $M \to \infty$. Clearly, it does so exponentially whenever $\mathbf{L}$, $\mathbf{F}$ and $\mathbf{D}$ are bounded for $M \to \infty$.

For $\mathbf{D}$, we observe that $\mathcal{A}\mathcal{B}^T$ is 0 if the fading is independent of the additive noise (which we usually expect to be the case in practically relevant scenarios) and that η is 0 in the case of user-independent fading in the sense of Definition 8, which means that the fading of any one user is independent of the fading of the other users (arbitrary correlations in the time domain are still allowed). Since $\mathbf{D}$ grows proportionally with $M\eta^2$, we can see that our bound is not useful for the case of strong correlations between users. Therefore, in the presence of user-correlated fading, the usefulness of the bound depends on the scaling behavior of $M\eta^2$. When this term exhibits sublinear growth, the error bound of Theorem 2 approaches 0 as $M \to 0$, and if it is additionally upper bounded, the error bound does so exponentially. In this sense, the bound is robust to small deviations from the assumption of user-independence. However, it is important to note that even the user-independent case covers relevant cases of correlation in the time domain such as the block fading channel. In the expression of $\mathbf{F}$, we can see that the Frobenius norm of both $\mathcal{A}$ and $\mathcal{B}$ should not grow faster than $\sqrt{M}$ and finally, in the expression of $\mathbf{L}$, we see that the operator norms of $\mathcal{A}$ and $\mathcal{B}$ should not grow with M. We illustrate that this is the case in scenarios of interest with the following two examples.

Corollary 2. *In the setting of Theorem 2 with uncorrelated fading and noise, i.e.,*

$$\mathcal{A} := \begin{pmatrix} \sigma_F \mathbf{id}_{2MK} & 0 \end{pmatrix}, \ \mathcal{B} := \begin{pmatrix} 0 & \sigma_N \mathbf{id}_{2M} \end{pmatrix}, \tag{2.15}$$

where $\mathbf{id}_n$ *denotes the* $n \times n$ *identity matrix, we have*

$$\begin{aligned} &\mathbb{P}\left(\left|\bar{f} - f(s_1, \ldots, s_K)\right| \geq \varepsilon\right) \\ &\qquad \leq 2\exp\left(-\frac{M\Phi^{-1}(\varepsilon)^2}{16\mathbf{F}' + 4\Phi^{-1}(\varepsilon)\mathbf{L}'}\right) + 2\exp\left(-\frac{M\Phi^{-1}(\varepsilon)^2}{256\mathbf{F}' + 32\Phi^{-1}(\varepsilon)\mathbf{L}'}\right), \end{aligned} \tag{2.16}$$

where

$$\begin{aligned} \mathbf{L}' &= \left(\sqrt{\bar{\Delta}(f)}\sigma_F + \sqrt{\frac{\Delta(f)}{\mathfrak{P}}}\sigma_N\right)^2 \\ \mathbf{F}' &= \mathbf{L}'\left(\sqrt{2K\bar{\Delta}(f)}\sigma_F + \sqrt{\frac{2\Delta(f)}{\mathfrak{P}}}\sigma_N\right)^2. \end{aligned}$$

Proof. Note that $\mathcal{A}\mathcal{B}^T = 0$, $\|\mathcal{A}\|_{\mathrm{op}} = \sigma_F$, $\|\mathcal{B}\|_{\mathrm{op}} = \sigma_N$, $\|\mathcal{A}\|_{\mathrm{F}} = \sqrt{2MK}\sigma_F$ and $\|\mathcal{B}\|_{\mathrm{F}} = \sqrt{2M}\sigma_N$; pick $\mathcal{A}_i := \mathcal{A}$ and substitute this into (2.13). □

Corollary 3. *In the setting of Theorem 2 where each user has a block fading channel with block length* β*, i.e.,*

$$\mathcal{A} := \sigma_F \left(\begin{array}{cccc:c} \mathbf{id}_{2K} & & & & \\ \vdots & & & & \\ \mathbf{id}_{2K} & & & 0 & \\ & \mathbf{id}_{2K} & & & \\ & \vdots & & & \\ & \mathbf{id}_{2K} & & & 0 \\ & & \ddots & & \\ 0 & & & \mathbf{id}_{2K} & \\ & & & \vdots & \\ & & & \mathbf{id}_{2K} & \end{array}\right) \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\}\beta \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\}\beta \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\}\beta \end{matrix}$$

$$\mathcal{B} := \begin{pmatrix} 0 & \sigma_N \mathbf{id}_{2M}, \end{pmatrix}$$

we have a bound of the form (2.16), where

$$\mathbf{L}' = \left(\sqrt{\bar{\Delta}(f)\beta}\sigma_F + \sqrt{\frac{\Delta(f)}{\mathfrak{P}}}\sigma_N \right)^2$$
$$\mathbf{F}' = \mathbf{L}' \left(\sqrt{2K\bar{\Delta}(f)}\sigma_F + \sqrt{\frac{2\Delta(f)}{\mathfrak{P}}}\sigma_N \right)^2 .$$

Proof. Note that $\mathcal{A}\mathcal{B}^T = 0$, $\|\mathcal{A}\|_{\mathrm{op}} = \sigma_F\sqrt{\beta}$, $\|\mathcal{B}\|_{\mathrm{op}} = \sigma_N$, $\|\mathcal{A}\|_{\mathrm{F}} = \sqrt{2MK}\sigma_F$ and $\|\mathcal{B}\|_{\mathrm{F}} = \sqrt{2M}\sigma_N$; pick $\mathcal{A}_i := \mathcal{A}$ and substitute this into (2.13). □

2.3.2 Sharpness of the Bound in Theorem 2

We do not expect the bound (2.13) to be sharp in the sense that there are non-trivial examples in which it holds with equality. This, we believe, is in part a price that we pay for using a very general system model, but it is also due to the underlying tools from high-dimensional statistics that we employ. A further sharpening of this bound could be an interesting question for future research, but it would hinge on optimizing the bounds of some of the basic results that we use (such as Lemma 2). In some special cases (such as uncorrelated Gaussian noise and fading) it is not hard, however, to compute exact bounds, as can for instance be seen in (2.18) below. In the sequel, we argue that in a sense that will be made precise, the bound (2.13) is sharp "up to absolute constants". The example case for which we show that the bound holds with equality up to constants is uncorrelated Gaussian fading and noise so that the bound specializes to (2.16). For the purpose of this section, we focus on the behavior with varying M and ε, while we consider everything else constant system parameters.

Theorem 3. *In the case of uncorrelated Gaussian fading and noise; i.e., (2.15) is satisfied and the entries of $\mathcal{G}$ are i.i.d. standard Gaussian, there are constants c and C such that the estimate $\bar{f}$ obtained by the pre- and post-processing schemes described in Sections 2.5.1 and 2.5.2 satisfies*

$$\mathbb{P}\left(\left|\bar{f} - f(s_1, \ldots, s_K)\right| \geq \varepsilon\right) \geq c\exp\left(-CM\min(\Phi^{-1}(\varepsilon), \Phi^{-1}(\varepsilon)^2)\right) \tag{2.17}$$

for suitable choices of F and Φ such as $F = \Phi = \mathbf{id}$ (the identity function).

Note that the upper tail bound (2.16) also has the same form for suitably chosen c and C, so that we can conclude that it is sharp up to the values of these constants.

Proof. The proof is relatively straightforward, so we only sketch it. Under the assumptions

made in Theorem 3, we readily compute from the equations in Sections 2.5.1 and 2.5.2

$$\frac{\|Y\|_2^2}{\sigma_F^2 \sum_{k=1}^K \mathfrak{a}_k + \sigma_N^2} = \|\mathcal{G}\|_2^2.$$

Since the entries of $\mathcal{G}$ are i.i.d. standard Gaussian and the vector has $2M$ entries, $\|\mathcal{G}\|_2^2$ clearly follows a chi-square distribution with $2M$ degrees of freedom. We therefore have in parallel to (2.24)

$$\mathbb{P}\left(\left|\bar{f} - f(s_1, \ldots, s_K)\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\|\mathcal{G}\|_2^2 - \mathbb{E}\|\mathcal{G}\|_2^2\right| \geq \frac{2\mathfrak{P} M \Phi^{-1}(\varepsilon)}{\Delta(f)(\sigma_F^2 \sum_{k=1}^K \mathfrak{a}_k + \sigma_N^2)}\right). \quad (2.18)$$

The bound here is sharp in the sense that it holds with equality in case $F = \Phi = \mathbf{id}$. We can now use [ZZ20, Corollary 3] to conclude that in this case, (2.17) holds for suitable c and C. □

2.4 Numerical Results

We have simulated the DFA scheme for Rayleigh fading channels with varying noise power, number of users and amount of channel resources. The simulations were done for two different functions, with the function arguments in both cases confined to the unit interval $[0, 1]$, to highlight different aspects and properties of the scheme: The arithmetic mean function is linear and maps only to the interval $[0, 1]$ (which means that no scheme can have an error larger than 1), while the Euclidean norm function maps to $[0, \sqrt{K}]$ and can show how the DFA scheme deals with nonlinearities.

We compare with a simple Time Division Multiple Access (TDMA) scheme, in which each user transmits separately in its designated slot, protecting the analog transmission against channel noise in the same fashion as the DFA scheme, but not sharing the channel use with other transmitters. In the case where the number of channel uses available is much larger than the number of users sharing the resources, this form of a TDMA scheme is of course highly suboptimal, as the transmitters could use source and channel coding to achieve a higher reliability. However, such an approach is infeasible if the number of users is so high in comparison to the number of channel uses that only a few or possibly even less than one channel use is available to each user, and in this work we are mainly interested in the scaling behavior of our schemes in the number of users K. Therefore, this comparison provides an insight into the gain achieved by exploiting the superposition properties of the wireless channel while keeping in mind that for the regime of low K, there are better coded schemes available. We also remark that the DFA scheme only needs coordination

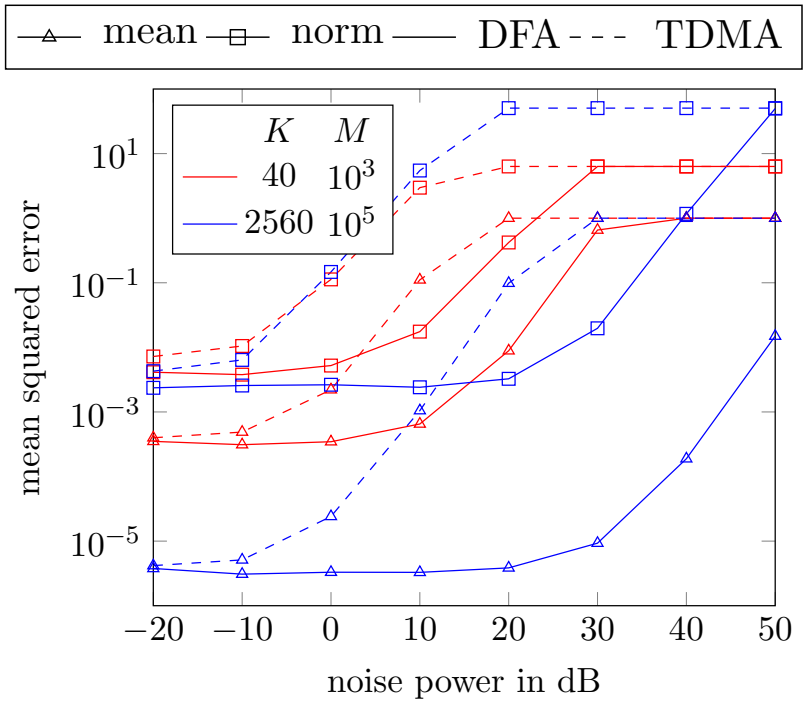

Figure 2.2: Mean squared error of the approximation schemes dependent on the channel noise power.

between the transmitters insofar as all users need to transmit roughly at the same time, while a TDMA scheme necessitates an allocation of the channel uses to the individual transmitters, which can be costly in the case of high K. The simulations carried out in this section do not consider this scheduling problem and assume for the TDMA scheme that the time slots have already been allocated, and this knowledge is available at both the transmitters and the receiver. If $M < K$, there is not at least one channel use available to each user and the TDMA scheme can therefore not be carried out. We set the error in such cases to the maximum of 1 or $\sqrt{K}$, respectively.

For the simulations, we assume a normalized peak transmitter power constraint of 1 and channels with fading normalized to a variance of 1 per complex dimension. The power of the additive noise is given in dB per complex dimension and its negative can therefore be considered as the signal-to-noise ratio (SNR). Each plotted data point is based on an average of 1000 simulation runs.

The messages transmitted by the users are generated in the following way: First, we draw a value μ, which is common to all transmitters, uniformly at random from $[0, 1]$. We then draw the messages of all the users from a convex combination of the uniform distributions on $[0, \mu]$ and $[\mu, 1]$ where we choose the weights in such a way that each

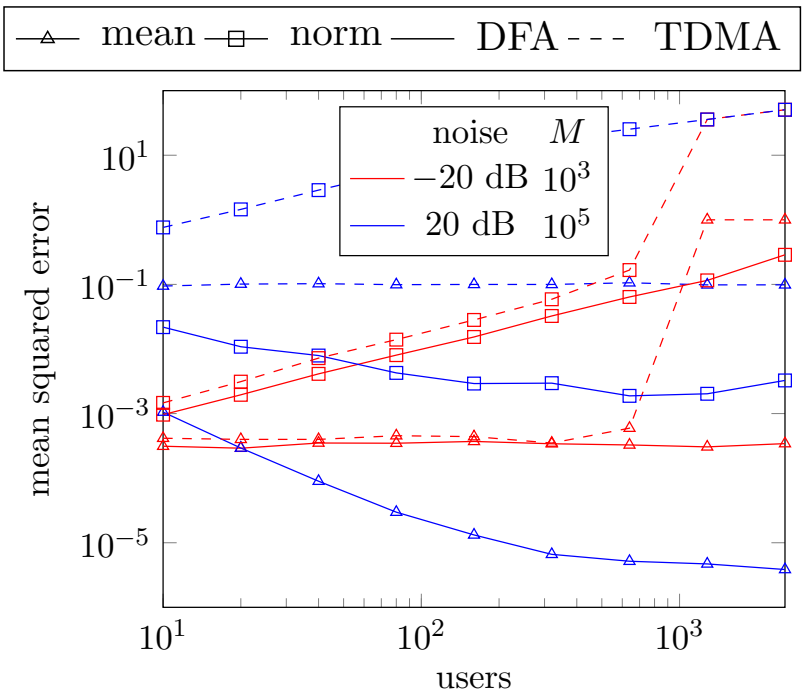

Figure 2.3: Mean squared error of the approximation schemes dependent on the number of participating transmitters.

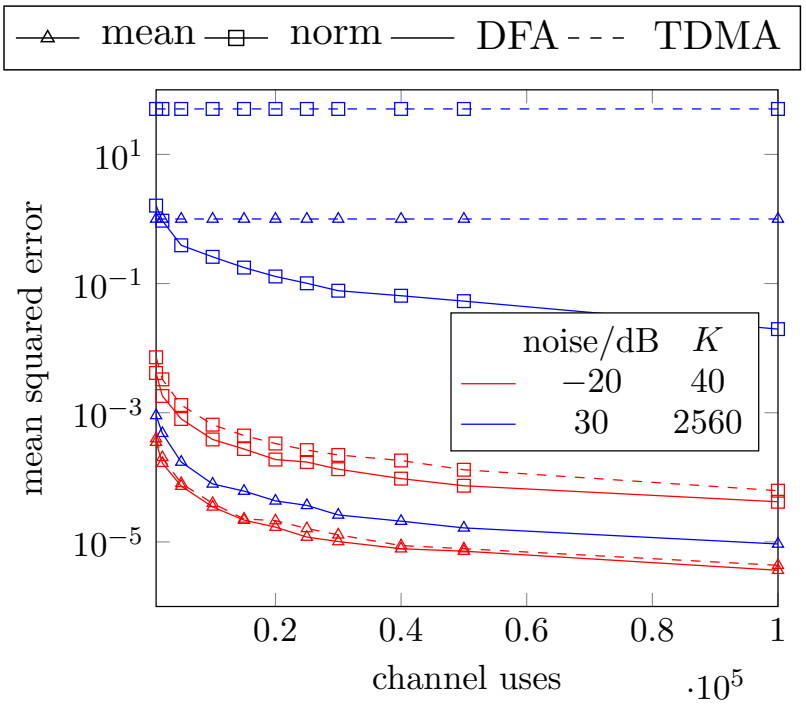

Figure 2.4: Mean squared error of the approximation schemes dependent on the number of channel uses.

message has expectation μ. The reason for choosing this procedure although the DFA scheme also performs well for more natural distributions such as i.i.d. uniform in $[0, 1]$ for all users is that in case of messages distributed according to a known i.i.d. distribution, the problem is too easy in the sense that both the mean and the Euclidean norm concentrate around values that depend only on the distribution and K, and therefore even without any communication at all, the function value can be quite accurately guessed if K is large. On the other hand, we intend the DFA scheme for applications in which the messages can be correlated and distributed according to unknown distributions, so we opt for this form of correlation between the messages for the sake of the numerical evaluation.

In Fig. 2.2, we can see that the DFA scheme is at least as good as the TDMA schemes for all the plotted data points and outperforms it in most cases, achieving a gain of up to 30 dB for $K = 2560$. For low powers of the additive noise, the effect of the multiplicative fading dominates, and therefore, the error saturates as the additive noise grows weaker. Fig. 2.3 illustrates that the DFA scheme performs significantly better if the number of users is not too low, which is due to the superposition of the signals in the wireless channel resulting in a combined signal strength that grows with the number of users. We can also see the TDMA scheme performing similarly to the DFA scheme for low numbers of users, while quickly deteriorating in performance or even becoming infeasible as their number grows. In Fig. 2.4, we can observe the exponential decay of the error as the amount of channel resources used increases. Once again, we can observe that the TDMA scheme performs similarly to DFA for a low number of users, but becomes infeasible for larger K.

2.5 Proof of Theorem 2

2.5.1 Pre-Processing

In the pre-processing step we encode the function values $f_k(s_k)$, $k = 1, \ldots, K$ as transmit power:

$$T_k(m) := \sqrt{\mathfrak{a}_k} U_k(m), 1 \leq m \leq M$$

with $\mathfrak{a}_k = \mathfrak{g}_k(f_k(s_k))$, where $\mathfrak{g}_k : [\phi_{\min,k}, \phi_{\max,k}] \to [0, \mathfrak{P}]$ such that

$$\mathfrak{g}_k(a) := \frac{\mathfrak{P}}{\Delta(f)}(a - \phi_{\min,k}), \tag{2.19}$$

where $\Delta(f)$ is given in (2.10) and $\phi_{\min,k}$ is defined in (2.11).
$U_k(m)$, $k = 1, \ldots, K$, $m = 1, \ldots, M$ are i.i.d. with the uniform distribution on $\{-1, +1\}$. We assume the random variables $U_k(m)$, $k = 1, \ldots, K$, $m = 1 \ldots, M$, are independent of $H_k(m)$, $k = 1, \ldots, K$, $m = 1, \ldots, M$, and $N(m)$, $m = 1, \ldots, M$.

We write the vector of transmitted signals at channel use m as

$$T(m) := (T_1(m), \ldots, T_K(m))$$

and combine them in a matrix

$$\mathcal{Q} := \begin{pmatrix} T(1) & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & T(1) & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & T(2) & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & T(2) & 0 & \ldots & 0 \\ & & & & \ddots & & \\ 0 & 0 & 0 & \ldots & 0 & T(M) & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & T(M) \end{pmatrix}.$$

2.5.2 Post-Processing

The vector Y of received signals across the M channel uses can be written as $Y = \mathcal{Q} \cdot \mathcal{H} + N$, where $\mathcal{H}$ and N are given in (2.3) and (2.4). The post-processing is based on receive energy which has the form

$$\|Y\|_2^2 = Y^T Y = (\mathcal{Q}\mathcal{A}\mathcal{G} + \mathcal{B}\mathcal{G})^T (\mathcal{Q}\mathcal{A}\mathcal{G} + \mathcal{B}\mathcal{G}) = \mathcal{G}^T \mathcal{W} \mathcal{G},$$

where we use

$$\mathcal{W} := (\mathcal{Q}\mathcal{A} + \mathcal{B})^T (\mathcal{Q}\mathcal{A} + \mathcal{B}) = \mathcal{A}^T \mathcal{Q}^T \mathcal{Q} \mathcal{A} + \mathcal{A}^T \mathcal{Q}^T \mathcal{B} + \mathcal{B}^T \mathcal{Q} \mathcal{A} + \mathcal{B}^T \mathcal{B}. \tag{2.20}$$

Equivalently, we can phrase this as

$$\|Y\|_2^2 = \sum_{k=1}^{K} \mathfrak{a}_k \|H_k\|_2^2 + \bar{N}_{s^K}, \tag{2.21}$$

where $H_k = (H_k(1), \ldots, H_k(M))$ is a vector consisting of complex fading coefficients, and $\bar{N}_{s^K} = \sum_{m=1}^{M} \bar{N}_{s^K}(m)$ where the overbar notation is used to denote the complex

conjugate. The random variables $\bar{N}_{s^K}(m)$, $m = 1, \ldots, M$, are given by

$$\bar{N}_{s^K}(m) := \sum_{\substack{k,k'=1 \\ k \neq l}}^{K} \sqrt{\mathfrak{a}_k \mathfrak{a}_{k'}} H_k(m) \overline{H_{k'}(m)} U_k(m) U_{k'}(m) + 2 \left(\overline{N(m)} \sum_{k=1}^{K} \sqrt{\mathfrak{a}_k} H_k(m) U_k(m) \right)^{\mathrm{Re}} + |N(m)|^2. \quad (2.22)$$

The receiver computes its estimate $\bar{f}$ of $f(s_1, \ldots, s_K)$ as

$$\bar{f} := F(\bar{\mathfrak{g}}(\|Y\|_2^2 - \mathbb{E}\|N\|_2^2)),$$

where

$$\bar{\mathfrak{g}}(a) := \frac{\Delta(f)}{2 \cdot M \cdot \mathfrak{P}} a + \sum_{k=1}^{K} \phi_{\min,k}.$$

2.5.3 The Error Event

Clearly, $\mathbb{E}\bar{N}_{s^K}(m) = \mathbb{E}|N(m)|^2$ (since all the other summands in (2.22) are centered). We can therefore conclude

$$\mathbb{E}\left(\bar{\mathfrak{g}}\left(\|Y\|_2^2 - \mathbb{E}\|N\|_2^2\right)\right) = \bar{\mathfrak{g}}\left(\mathbb{E}\|Y\|_2^2 - \mathbb{E}\|N\|_2^2\right) = \sum_{k=1}^{K} f_k(s_k).$$

We use this to argue

$$\begin{aligned} \left|\bar{f} - f(s_1, \ldots, s_K)\right| &= \left| F\left(\bar{\mathfrak{g}}\left(\|Y\|_2^2 - \mathbb{E}\|N\|_2^2\right)\right) - F\left(\sum_{k=1}^{K} f_k(s_k)\right) \right| \\ &\leq \Phi\left(\left| \bar{\mathfrak{g}}\left(\|Y\|_2^2 - \mathbb{E}\|N\|_2^2\right) - \sum_{k=1}^{K} f_k(s_k) \right| \right) \\ &= \Phi\left(\left| \bar{\mathfrak{g}}\left(\|Y\|_2^2 - \mathbb{E}\|N\|_2^2\right) - \bar{\mathfrak{g}}\left(\mathbb{E}\|Y\|_2^2 - \mathbb{E}\|N\|_2^2\right) \right| \right) \\ &= \Phi\left(\frac{\Delta(f)}{2M\mathfrak{P}} \left| \|Y\|_2^2 - \mathbb{E}\|Y\|_2^2 \right| \right) \end{aligned} \quad (2.23)$$

and therefore

$$\mathbb{P}\left(\left|\bar{f} - f(s_1, \ldots, s_K)\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left| \|Y\|_2^2 - \mathbb{E}\|Y\|_2^2 \right| \geq \frac{2M\mathfrak{P}}{\Delta(f)} \Phi^{-1}(\varepsilon) \right). \quad (2.24)$$

2.5.4 Performance Bounds

Our objective is now to establish the concentration of $\|Y\|_2^2$ around its expectation and thus obtain an upper bound for the right hand side of (2.24). To this end, we first need to establish a series of lemmas that we will use as tools.

We will split the deviation from the mean into a diagonal and an off-diagonal part. The first lemma will later help us bound the diagonal part of the error.

Lemma 1. *Let $\mathcal{V}_1, \dots, \mathcal{V}_n$ be independent random variables and centered with sub-Gaussian norm at most 1. Let $\mathcal{A}_1, \dots, \mathcal{A}_n$ be random variables independent of $\mathcal{V}_1, \dots, \mathcal{V}_n$ but not necessarily of each other, and assume that for all k, $|\mathcal{A}_k| \leq \tilde{L}$ and $\sum_{k=1}^n \mathcal{A}_k^2 \leq \tilde{F}$ almost surely. Then we have for any $c \in (0,1)$ and any $\lambda \in (-c/(2\tilde{L}), c/(2\tilde{L}))$,*

$$\mathbb{E}\exp\left(\lambda \sum_{k=1}^n \left(\mathcal{A}_k\mathcal{V}_k^2 - \mathbb{E}(\mathcal{A}_k\mathcal{V}_k^2)\right)\right) \leq \exp\left(\frac{\lambda^2}{2}\cdot\frac{8\tilde{F}}{1-c}\right)\mathbb{E}\exp\left(\lambda\sum_{k=1}^n\left(\mathcal{A}_k - \mathbb{E}(\mathcal{A}_k)\right)\right).$$

The proof of this lemma relies on some other technical results and is therefore relegated to Section 2.7.

The next lemma is a slight variation of the Hanson-Wright inequality as phrased in [Ver18, Theorem 6.2.1] and will help us bound the off-diagonal part of the error.

Lemma 2. *Let $\mathcal{V}$ be an $\mathbb{R}^n$-valued random variable with independent, centered entries and assume that for all $k \in \{1, \dots, n\}$, the k-th entry of $\mathcal{V}$ satisfies $\tau(\mathcal{V}_k) \leq \tau_{\max}$. Let $\mathcal{A} \in \mathbb{R}^{n\times n}$ with zeros on the diagonal and $\varepsilon > 0$. Suppose further that $\|\mathcal{A}\|_{op} \leq A_{\mathrm{op}}$ and $\|\mathcal{A}\|_F \leq A_{\mathrm{F}}$. Then $\mathbb{E}\left(\mathcal{V}^T\mathcal{A}\mathcal{V}\right) = 0$ and*

$$\mathbb{P}\left(\left|\mathcal{V}^T\mathcal{A}\mathcal{V}\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{16\tau_{\max}^2\varepsilon A_{\mathrm{op}} + 256\tau_{\max}^4 A_{\mathrm{F}}^2}\right). \tag{2.25}$$

This lemma differs from [Ver18, Theorem 6.2.1] mainly in that we require the diagonal entries of $\mathcal{A}$ to be 0 and that all the constants are explicit. The proof follows [Ver18] closely and is given in Section 2.7. We remark that it is not hard to follow the proof in [Ver18] further and expand the result to matrices with non-zero diagonal elements, however, this is not relevant for the present work.

Mainly because the matrix $\mathcal{W}$ contains randomness, we need a slight modification of this lemma as well as two more lemmas exploring some specific properties of $\mathcal{W}$.

Corollary 4. *Assume a setting as in Lemma 2, but let $\mathcal{A}$ be an $\mathbb{R}^{n\times n}$-valued random variable independent of $\mathcal{V}$ such that almost surely, the diagonal entries of $\mathcal{A}$ are 0, $\|\mathcal{A}\|_{op} \leq A_{\mathrm{op}}$ and $\|\mathcal{A}\|_F \leq A_{\mathrm{F}}$. Then $\mathbb{E}\left(\mathcal{V}^T\mathcal{A}\mathcal{V}\right) = 0$ and (2.25), considering joint expectation, respectively probability of $\mathcal{V}$ and $\mathcal{A}$, still hold.*

Proof. $\mathbb{E}\left(\mathcal{V}^T\mathcal{A}\mathcal{V}\right) = 0$ as well as (2.25) hold conditional on any realization of $\mathcal{A}$ (except possibly in a null set) and therefore, the Corollary follows by the laws of total expectation and total probability. □

Lemma 3. *We have almost surely*

$$\|\mathcal{W}\|_F \leq \left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_{op} + \|\mathcal{B}\|_{op}\right)\left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_F + \|\mathcal{B}\|_F\right),$$
$$\|\mathcal{W}\|_{op} \leq \left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_{op} + \|\mathcal{B}\|_{op}\right)^2.$$

Proof. In order to bound the norm of $\mathcal{W}$, we first note that

$$\mathcal{Q}\mathcal{Q}^T = \sum_{k=1}^{K} a_k \mathbf{id}_{2M}. \tag{2.26}$$

Therefore, we can conclude that all singular values of $\mathcal{Q}$ are bounded by $\sqrt{\Delta(f||\mathfrak{P})}$ and thus $\|\mathcal{Q}\|_{\mathrm{op}} \leq \sqrt{\Delta(f||\mathfrak{P})}$.

Noting that $\|AB\|_{\mathrm{F}} \leq \|A\|_{\mathrm{op}}\|B\|_{\mathrm{F}}$ for all matrices A, B of compatible dimensions and further noting the submultiplicativity of the operator norm and the triangle inequality of both norms, we get

$$\begin{aligned}
\|\mathcal{W}\|_{\mathrm{F}} &\leq \|\mathcal{Q}\mathcal{A}+\mathcal{B}\|_{\mathrm{op}}\|\mathcal{Q}\mathcal{A}+\mathcal{B}\|_{\mathrm{F}} \leq \left(\|\mathcal{Q}\|_{\mathrm{op}}\|\mathcal{A}\|_{\mathrm{op}} + \|\mathcal{B}\|_{\mathrm{op}}\right)\left(\|\mathcal{Q}\|_{\mathrm{op}}\|\mathcal{A}\|_{\mathrm{F}} + \|\mathcal{B}\|_{\mathrm{F}}\right) \\
&\leq \left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_{\mathrm{op}} + \|\mathcal{B}\|_{\mathrm{op}}\right)\left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_{\mathrm{F}} + \|\mathcal{B}\|_{\mathrm{F}}\right), \\
\|\mathcal{W}\|_{\mathrm{op}} &= \|\mathcal{Q}\mathcal{A}+\mathcal{B}\|_{\mathrm{op}}^2 \leq \left(\|\mathcal{Q}\|_{\mathrm{op}}\|\mathcal{A}\|_{\mathrm{op}} + \|\mathcal{B}\|_{\mathrm{op}}\right)^2 \leq \left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_{\mathrm{op}} + \|\mathcal{B}\|_{\mathrm{op}}\right)^2
\end{aligned}$$ □

Lemma 4. *We have*

$$\tau\left(\mathrm{tr}\mathcal{W}\right) \leq 4M\Delta(f||\mathfrak{P})\|(\mathcal{A}+\mathcal{A}_i)(\mathcal{A}-\mathcal{A}_i)^T\|_{op} + 2\sqrt{2\mathfrak{P}}\|\mathcal{A}\mathcal{B}^T\|_F, \tag{2.27}$$

where the trace $\mathrm{tr}(\cdot)$ *is the sum of elements on the diagonal of a square matrix.*

Proof. With an addition of zero, we can rewrite

$$\begin{aligned}
\mathrm{tr}\left(\mathcal{A}^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}\right) &= \mathrm{tr}\left(\mathcal{A}^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}\right) + \mathrm{tr}\left(\mathcal{A}^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}_i\right) - \mathrm{tr}\left((\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}\right) \\
&\quad - \mathrm{tr}\left((\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}_i\right) + \mathrm{tr}\left((\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}_i\right) \\
&= \mathrm{tr}\left((\mathcal{A}-\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}(\mathcal{A}+\mathcal{A}_i)\right) + \mathrm{tr}\left((\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}_i\right)
\end{aligned}$$

and use this together with (2.20) to conclude

$$\mathrm{tr}\mathcal{W} = \mathrm{tr}\left((\mathcal{A}-\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}(\mathcal{A}+\mathcal{A}_i)\right) + 2\mathrm{tr}\left(\mathcal{B}^T\mathcal{Q}\mathcal{A}\right) + \mathrm{tr}\left((\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}_i\right) + \mathrm{tr}\left(\mathcal{B}^T\mathcal{B}\right). \tag{2.28}$$

Next, we argue that the last two summands are almost surely constant. For $\mathrm{tr}(\mathcal{B}^T\mathcal{B})$ this is immediately clear. Moreover, we have $\mathrm{tr}\Big((\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}\mathcal{A}_i\Big) = \|\mathcal{Q}\mathcal{A}_i\|_{\mathrm{F}}^2$. We note that as per Remark 1 and using corresponding notation, we have

$$\mathcal{Q}\mathcal{A}_i = \begin{pmatrix} T(1)\mathcal{A}_i^{(1)} & 0 & 0 & \dots & 0 \\ 0 & T(1)\mathcal{A}_i^{(2)} & 0 & \dots & 0 \\ & & \ddots & & \\ 0 & \dots & 0 & T(M)\mathcal{A}_i^{(2M-1)} & 0 \\ 0 & \dots & 0 & 0 & T(M)\mathcal{A}_i^{(2M)} \end{pmatrix}$$

and because each $\mathcal{A}_i^{(m)}$ has only one nonzero entry per column, each entry of $\mathcal{Q}\mathcal{A}_i$ is the product of $U_k(m)$ with a deterministic term for some m, k and therefore, its square can take only one value almost surely, and consequently, $\|\mathcal{Q}\mathcal{A}_i\|_{\mathrm{F}}^2$ also takes only one value almost surely.

We can use this in (2.28) and incorporate the triangle inequality to obtain

$$\tau\left(\mathrm{tr}\mathcal{W}\right) \leq \tau\left(\xi_1\right) + 2\tau\left(\xi_2\right),$$

where

$$\begin{aligned} \xi_1 &:= \mathrm{tr}\left((\mathcal{A}-\mathcal{A}_i)^T\mathcal{Q}^T\mathcal{Q}(\mathcal{A}+\mathcal{A}_i)\right) \\ \xi_2 &:= \mathrm{tr}\left(\mathcal{B}^T\mathcal{Q}\mathcal{A}\right). \end{aligned}$$

To the end of bounding $\tau\left(\xi_1\right)$, we argue

$$\begin{aligned} \xi_1 &= \mathrm{tr}\left(\mathcal{Q}(\mathcal{A}+\mathcal{A}_i)(\mathcal{A}-\mathcal{A}_i)^T\mathcal{Q}^T\right) \leq \|(\mathcal{A}+\mathcal{A}_i)(\mathcal{A}-\mathcal{A}_i)^T\|_{\mathrm{op}}\mathrm{tr}\left(\mathcal{Q}\mathcal{Q}^T\right) \\ &\leq 2M\Delta(f||\mathfrak{P})\|(\mathcal{A}+\mathcal{A}_i)(\mathcal{A}-\mathcal{A}_i)^T\|_{\mathrm{op}}. \end{aligned}$$

The first inequality holds because for any square matrix A and compatible column vector v, we have

$$v^T(\|A\|_{\mathrm{op}}\mathbf{id} - A)v = \|v\|_2^2\left(\|A\|_{\mathrm{op}} - \left(\frac{v}{\|v\|_2}\right)^T A\frac{v}{\|v\|_2}\right) \geq 0$$

(see, e.g., [Bha97, Exercise I.2.10]) and therefore $\|A\|_{\mathrm{op}}\mathbf{id} - A$ is positive semidefinite. The second inequality directly follows from (2.26). It follows, e.g., from [BK00, Example 1.1.2], that $\tau(\xi_1)$ is upper bounded by the first summand on the right hand side in (2.27). Note that in an analogous way, we can derive the same bound for $-\xi_1$.

In order to bound the sub-Gaussian norm of ξ_2, we view it as a function of $(U_k(m))_{k,m=1}^{K,M}$ and use part of the proof of the Bounded Differences Inequality [BLM13, Theorem 6.2] to bound the moment generating function. To this end, we define

$$(\mathbf{E}_{i,j})_{i',j'} = \begin{cases} 1, & i' = i \text{ and } j' = j \\ 0, & \text{otherwise.} \end{cases}$$

and note that a change in the value of $U_k(m)$ changes the value of ξ_2 by

$$\begin{aligned} & 2\sqrt{a_k}\mathrm{tr}\left(\mathcal{B}^T(\mathbf{E}_{2m-1,K(2m-2)+k} + \mathbf{E}_{2m,K(2m-1)+k})\mathcal{A}\right) \\ &= 2\sqrt{a_k}\mathrm{tr}\left(\mathcal{A}\mathcal{B}^T(\mathbf{E}_{2m-1,K(2m-2)+k} + \mathbf{E}_{2m,K(2m-1)+k})\right) \\ &= 2\sqrt{a_k}\left((\mathcal{A}\mathcal{B}^T)_{K(2m-2)+k,2m-1} + (\mathcal{A}\mathcal{B}^T)_{K(2m-1)+k,2m}\right) \\ &\leq 2\sqrt{\mathfrak{P}}\left((\mathcal{A}\mathcal{B}^T)_{K(2m-2)+k,2m-1} + (\mathcal{A}\mathcal{B}^T)_{K(2m-1)+k,2m}\right) \end{aligned}$$

Following the proof of the Bounded Differences Inequality [BLM13, Theorem 6.2], we can now conclude

$$\begin{aligned} \tau(\xi_2)^2 &\leq \frac{1}{4}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(2\sqrt{\mathfrak{P}}(\mathcal{A}\mathcal{B}^T)_{(2m-2)K+k,2m-1} + 2\sqrt{\mathfrak{P}}(\mathcal{A}\mathcal{B}^T)_{(2m-1)K+k,2m}\right)^2 \\ &\leq \frac{1}{4}\cdot 2\cdot 4\cdot\mathfrak{P}\|\mathcal{A}\mathcal{B}^T\|_{\mathrm{F}}^2, \end{aligned}$$

concluding the proof of the lemma. □

Proof of Theorem 2. What remains to be established is the concentration of $\|Y\|_2^2$ around its expectation. To this end, we observe

$$\mathbb{P}\left(\left|\|Y\|_2^2 - \mathbb{E}\|Y\|_2^2\right| \geq \varepsilon\right) = \mathbb{P}\left(\left|\mathcal{G}^T\mathcal{W}\mathcal{G} - \mathbb{E}(\mathcal{G}^T\mathcal{W}\mathcal{G})\right| \geq \varepsilon\right) \leq \mathbb{P}\left(|\Sigma_1| \geq \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|\Sigma_2| \geq \frac{\varepsilon}{2}\right) \tag{2.29}$$

where

$$\Sigma_1 := \sum_{i=1}^{2KM+2M}\left(\mathcal{G}_i^2\mathcal{W}_{i,i} - \mathbb{E}\left(\mathcal{G}_i^2\mathcal{W}_{i,i}\right)\right)$$

$$\Sigma_2 := \sum_{\substack{i,j=1 \\ i \neq j}}^{2KM+2M} \mathcal{G}_i \mathcal{G}_j \mathcal{W}_{i,j}.$$

We use Lemma 1, Lemma 3 and Lemma 4 to bound the moment generating function of Σ_1 as

$$\mathbb{E} \exp(\lambda \Sigma_1) \leq \exp\left(\frac{\lambda^2}{2}\left(\frac{8\tilde{F}_1}{1-c} + \tilde{F}_2\right)\right) \leq \exp\left(\frac{\lambda^2}{2} \cdot \frac{8\tilde{F}_1 + \tilde{F}_2}{1-c}\right)$$

for any $c \in (0,1)$ and $\lambda \in (-c/(2\tilde{L})), c/(2\tilde{L}))$, where

$$\begin{aligned}
\tilde{L} &:= \left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_{\mathrm{op}} + \|\mathcal{B}\|_{\mathrm{op}}\right)^2 \\
\tilde{F}_1 &:= \tilde{L}\left(\sqrt{\Delta(f||\mathfrak{P})}\|\mathcal{A}\|_{\mathrm{F}} + \|\mathcal{B}\|_{\mathrm{F}}\right)^2 \\
\tilde{F}_2 &:= \left(4M\Delta(f||\mathfrak{P})\|(\mathcal{A}+\mathcal{A}_i)(\mathcal{A}-\mathcal{A}_i)^T\|_{\mathrm{op}} + 2\sqrt{2\mathfrak{P}}\|\mathcal{A}\mathcal{B}^T\|_{\mathrm{F}}\right)^2
\end{aligned}$$

By Lemma 9, this yields

$$\mathbb{P}\left(|\Sigma_1| \geq \frac{\varepsilon}{2}\right) \leq 2\exp\left(-(1-c)\frac{\varepsilon^2}{64\tilde{F}_1 + 8\tilde{F}_2}\right)$$

in case $0 < \varepsilon \leq \frac{c}{1-c} \cdot \frac{8\tilde{F}_1 + \tilde{F}_2}{\tilde{L}}$ and

$$\mathbb{P}\left(|\Sigma_1| \geq \frac{\varepsilon}{2}\right) \leq 2\exp\left(-\frac{c\varepsilon}{8\tilde{L}}\right)$$

otherwise. Since the first case term is increasing with c and the second case term is decreasing, the optimal value for c is where the two cases meet, which is at

$$c = \frac{\tilde{L}\varepsilon}{\tilde{L}\varepsilon + 8\tilde{F}_1 + \tilde{F}_2}.$$

Substituting this, we get

$$\mathbb{P}\left(|\Sigma_1| \geq \frac{\varepsilon}{2}\right) \leq 2\exp\left(-\frac{\varepsilon^2}{64\tilde{F}_1 + 8\tilde{F}_2 + 8\tilde{L}\varepsilon}\right).$$

Turning our attention to Σ_2, we note that by [BCD89, Theorem 2.1] the operator norm of the off-diagonal part of $\mathcal{W}$ can be upper bounded by $2\|\mathcal{W}\|_{\mathrm{op}}$ and thus by $2\tilde{L}$. Therefore,

we can directly apply Lemma 2 and get

$$\mathbb{P}\left(|\Sigma_2| \geq \frac{\varepsilon}{2}\right) \leq 2\exp\left(-\frac{\varepsilon^2}{1024\tilde{F}_1 + 64\tilde{L}\varepsilon}\right).$$

Substituting these into (2.29) and using (2.24) concludes the proof. □

2.6 Preliminaries on Sub-Gaussian and Sub-Exponential Random Variables

We begin with a definition that is adapted from [Ver18, Definition 3.4.1]. For $\mathbb{R}^n$-valued random variables $\mathcal{V}$, we define the sub-Gaussian norm as

$$\tau(\mathcal{V}) := \inf\left\{a : \forall v \in S^{n-1}\ \forall \lambda \in \mathbb{R}\ \mathbb{E}\exp(\lambda\langle\mathcal{V}, v\rangle) \leq \exp\left(\frac{a^2\lambda^2}{2}\right)\right\} \tag{2.30}$$

and we observe that if all entries of $\mathcal{V}$ have a sub-Gaussian norm bounded by $\tau_{\max}$ and are independent, we have for any $v \in S^{n-1}$

$$\begin{aligned}\mathbb{E}\exp(\lambda\langle\mathcal{V}, v\rangle) &= \mathbb{E}\exp\left(\lambda\sum_{k=1}^{n}\mathcal{V}_k v_k\right) = \prod_{k=1}^{n}\mathbb{E}\exp(\lambda\mathcal{V}_k v_k)\\ &\leq \prod_{k=1}^{n}\exp\left(\frac{\tau_{\max}^2 v_k^2\lambda^2}{2}\right) = \exp\left(\frac{\tau_{\max}^2\lambda^2}{2}\right)\end{aligned}$$

and therefore $\tau(\mathcal{V}) \leq \tau_{\max}$.

In the following, we recall some basic definitions and results from [BK00, Chapter 1]. For a random variable X we define[12]

$$\theta(X) := \sup_{k\geq 1}\left(\frac{\mathbb{E}(|X|^k)}{k!}\right)^{\frac{1}{k}} \tag{2.31}$$

If $\theta(X) < \infty$ then X is called a sub-exponential random variable. $\theta(\cdot)$ defines a semi-norm on the vector space of sub-exponential random variables [BK00, Remark 1.3.2]. Typical examples of sub-exponential random variables are bounded random variables and random variables with exponential distribution. We collect some useful properties of and interrelations between the sub-exponential and sub-Gaussian norms in the following lemma.

[12] Note that as with our definition of the sub-Gaussian norm, other norms on the space of sub-exponential random variables that appear in the literature are equivalent to $\theta(\cdot)$ (see, e.g., [BK00]). The particular definition we choose here matters, however, because we derive results in which no unspecified constants appear.

Lemma 5. *Let X, Y be random variables. Then:*

1. *If X follows the normal distribution with expectation μ and variance σ^2 then we have*

$$\tau(X) = \sigma. \tag{2.32}$$

2. *(Rotation Invariance) If $X_1, \ldots, X_M$ are independent, sub-Gaussian and centered, we have*

$$\tau\left(\sum_{m=1}^{M} X_m\right)^2 \leq \sum_{m=1}^{M} \tau(X_m)^2 \tag{2.33}$$

3. *If X is a random variable with $|X| \leq 1$ with probability 1 and if Y is independent of X and sub-Gaussian then we have*

$$\tau(X \cdot Y) \leq \tau(Y). \tag{2.34}$$

4. *If X and Y are sub-Gaussian and centered, then $X \cdot Y$ is sub-exponential and*

$$\theta(X \cdot Y) \leq 2 \cdot \tau(X) \cdot \tau(Y). \tag{2.35}$$

5. *(Centering) If X is sub-exponential and $X \geq 0$ almost surely, then*

$$\theta(X - \mathbb{E}(X)) \leq \theta(X). \tag{2.36}$$

Proof. (2.32) follows in a straightforward fashion by calculating the moment generating function of X. (2.33) is e.g. proven in [BK00, Lemma 1.1.7]. (2.34) follows directly from the definition conditioning on X. We show (2.35) first for $X = Y$. In this case, we have

$$\theta\left(X^2\right) = \sup_{k\geq 1}\left(\frac{\mathbb{E}X^{2k}}{k!}\right)^{\frac{1}{k}} \leq \sup_{k\geq 1}\left(\frac{2^{k+1}k^k\tau(X)^{2k}}{e^k k!}\right)^{\frac{1}{k}} = 2\tau(X)^2 \sup_{k\geq 1}\left(\frac{2^{\frac{1}{k}}k}{e(k!)^{\frac{1}{k}}}\right) \leq 2\tau(X)^2,$$

where the first inequality is by [BK00, Lemma 1.1.4] and the second follows from $2k^k/k! \leq e^k$, which is straightforward to prove for $k \geq 1$ by induction. In the general case, we have

$$\begin{aligned}
\theta(XY) &= \tau(X)\tau(Y)\,\theta\left(\frac{XY}{\tau(X)\tau(Y)}\right) \\
&\leq \tau(X)\tau(Y)\,\theta\left(\frac{1}{2}\left(\frac{X}{\tau(X)}\right)^2 + \frac{1}{2}\left(\frac{Y}{\tau(Y)}\right)^2\right) \\
&\leq 2\tau(X)\tau(Y),
\end{aligned}$$

where the first inequality can be verified in (2.31), considering that $ab \leq a^2/2 + b^2/2$ for all $a, b \in \mathbb{R}$, and the second inequality follows from the triangle inequality and the special case $X = Y$.

For (2.36), we assume without loss of generality $\mathbb{E}X = 1$ (otherwise we can scale X), and note that for all $a \in [0, \infty)$ and $k \geq 1$, $a^k - |a-1|^k > a - 1$ and thus $\mathbb{E}(X^k - |X-1|^k) \geq \mathbb{E}(X - 1) = 0$. □

2.7 Proof of Lemmas 1 and 2

The proofs closely follow the proof of the Hanson-Wright inequality in [Ver18, Theorem 6.2.1]. We carry out the changes that are necessary to arrive at explicit constants. To this end, we begin with some slightly modified versions of lemmas used as ingredients in the proof of Bernstein's inequality in [BK00, Theorem 1.5.2].

Lemma 6. *Let X be a random variable with $\mathbb{E}(X) = 0$ and $\theta(X) < +\infty$. For any $\lambda \in \mathbb{R}$ with $|\lambda\theta(X)| < 1$ we have*

$$\mathbb{E}(\exp(\lambda X)) \leq 1 + |\lambda|^2 \theta(X)^2 \cdot \frac{1}{1 - |\lambda\theta(X)|}.$$

Proof. Let $\lambda \in \mathbb{R}$ satisfy $|\lambda\theta(X)| < 1$. Then

$$\begin{aligned}
\mathbb{E}(\exp(\lambda X)) &= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}(X^k)}{k!} \\
&\leq 1 + \sum_{k=2}^{\infty} \frac{|\lambda|^k \mathbb{E}(|X|^k)}{k!} \leq 1 + \sum_{k=2}^{\infty} |\lambda|^k \theta(X)^k \\
&= 1 + |\lambda|^2 \theta(X)^2 \left(\sum_{k=0}^{\infty} |\lambda\theta(X)|^k \right) \\
&= 1 + |\lambda|^2 \theta(X)^2 \cdot \frac{1}{1 - |\lambda\theta(X)|}, \qquad (2.37)
\end{aligned}$$

where in the last line we have used $|\lambda\theta(X)| < 1$. □

In the next lemma we derive an exponential bound depending on $\theta(X)$ on the moment generating function of the random variable X.

Lemma 7. *Let X be a random variable with $\mathbb{E}(X) = 0$ and $\theta(X) < +\infty$. For any $c \in (0, 1)$ and $\lambda \in \left(-\frac{c}{\theta(X)}, \frac{c}{\theta(X)}\right)$ we have*

$$\mathbb{E}(\exp(\lambda X)) \leq \exp\Big(\frac{\lambda^2}{2} \frac{2 \cdot \theta(X)^2}{1 - c}\Big).$$

Proof. For $\lambda \in \left(-\frac{c}{\theta(X)}, \frac{c}{\theta(X)}\right)$ we have

$$|\lambda\theta(X)| < c < 1, \tag{2.38}$$

therefore by Lemma 6

$$\begin{aligned}\mathbb{E}(\exp(\lambda X)) &\leq 1 + |\lambda|^2\theta(X)^2 \cdot \frac{1}{1-|\lambda\theta(X)|} \\ &\leq 1 + |\lambda|^2\theta(X)^2 \cdot \frac{1}{1-c} \\ &\leq \exp\Big(\frac{\lambda^2}{2}\frac{2\cdot\theta(X)^2}{1-c}\Big),\end{aligned}$$

where in the second line we have used the first inequality in (2.38) and the last line is by the numerical inequality $1 + a \leq \exp(a)$ valid for $a \geq 0$. □

We are now ready to prove Lemma 1.

Proof of Lemma 1. The lemma follows by a straightforward calculation

$$\begin{aligned}&\mathbb{E}\exp\left(\lambda\sum_{k=1}^{n}\left(\mathcal{A}_k\mathcal{V}_k^2 - \mathbb{E}(\mathcal{A}_k\mathcal{V}_k^2)\right)\right) \\ &= \mathbb{E}\left(\exp\left(\lambda\sum_{k=1}^{n}\left(\mathcal{A}_k(\mathcal{V}_k^2 - \mathbb{E}(\mathcal{V}_k^2))\right)\right)\cdot\exp\left(\lambda\sum_{k=1}^{n}\left(\mathbb{E}(\mathcal{V}_k^2)(\mathcal{A}_k - \mathbb{E}(\mathcal{A}_k))\right)\right)\right) &(2.39)\\ &= \mathbb{E}_{\mathcal{A}}\left(\exp\left(\lambda\sum_{k=1}^{n}\left(\mathbb{E}(\mathcal{V}_k^2)(\mathcal{A}_k - \mathbb{E}(\mathcal{A}_k))\right)\right)\cdot\prod_{k=1}^{n}\mathbb{E}_{\mathcal{V}}\exp\left((\lambda\mathcal{A}_k)\left(\mathcal{V}_k^2 - \mathbb{E}(\mathcal{V}_k^2)\right)\right)\right) &(2.40)\\ &\leq \mathbb{E}_{\mathcal{A}}\left(\exp\left(\lambda\sum_{k=1}^{n}\left(\mathbb{E}(\mathcal{V}_k^2)(\mathcal{A}_k - \mathbb{E}(\mathcal{A}_k))\right)\right)\cdot\prod_{k=1}^{n}\exp\left(\frac{\lambda^2}{2}\cdot\frac{8\mathcal{A}_k^2}{1-c}\right)\right) &(2.41)\\ &\leq \exp\left(\frac{\lambda^2}{2}\cdot\frac{8\tilde{F}}{1-c}\right)\mathbb{E}\left(\exp\left(\lambda\sum_{k=1}^{n}(\mathcal{A}_k - \mathbb{E}\mathcal{A}_k)\right)\right), &(2.42)\end{aligned}$$

where (2.40) follows by the independence assumptions, (2.41) is an application of Lemma 7 and (2.42) holds because $\sum_{k=1}^{n}\mathcal{A}_k^2 \leq \tilde{F}$ almost surely. For this last step, also note that the variance of a sub-Gaussian random variable is upper bounded by the squared sub-Gaussian norm [BK00, Lemma 1.1.2]. □

Lemma 8. *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}(X_i) = 0$ and $\theta(X_i) < +\infty$, $i = 1, \ldots, n$. Let $L := \max_{1\leq i\leq n}\theta(X_i)$, $c \in (0,1)$, and $\lambda \in \left(-\frac{c}{L}, \frac{c}{L}\right)$. Then for*

$\Sigma_M := \sum_{i=1}^n X_i$ *we have*

$$\mathbb{E}(\exp(\lambda\Sigma_M)) \leq \exp\Big(\frac{\lambda^2}{2}\frac{2\cdot\sum_{i=1}^n \theta\left(X_i\right)^2}{1-c}\Big). \tag{2.43}$$

Proof. By independence of $X_1, \ldots, X_n$, we have

$$\mathbb{E}(\exp(\lambda\Sigma_n)) = \prod_{i=1}^M \mathbb{E}(\exp(\lambda X_i)).$$

Combining this with Lemma 7 proves the lemma. □

The next lemma establishes the basic tail bound for random variables satisfying inequalities of type (2.43). The proof can be found in [BK00, Lemma 1.4.1].

Lemma 9. *Let X be a random variable with $\mathbb{E}(X) = 0$. If there exist $\tau \geq 0$ and $\Lambda > 0$ such that*

$$\mathbb{E}(\exp(\lambda X)) \leq \exp\Big(\frac{\lambda^2}{2}\tau^2\Big),$$

holds for all $\lambda \in (-\Lambda, \Lambda)$, then for any $t \geq 0$, we have

$$\mathbb{P}(|X| \geq t) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\tau^2}\right), & 0 < t \leq \Lambda\tau^2 \\ 2\exp\left(-\frac{\Lambda t}{2}\right), & \Lambda\tau^2 \leq t. \end{cases}$$

The following lemma is a slightly modified version of [Ver18, Lemma 6.2.3]. $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and covariance matrix Σ.

Lemma 10 (Comparison Lemma)**.** *Let $\mathcal{V}$ and $\mathcal{V}'$ be independent, $\mathbb{R}^n$-valued, centered and sub-Gaussian random variables, and let G, G' be independent and distributed according to $\mathcal{N}(0, \mathbf{id}_n)$. Let $\mathcal{A} \in \mathbb{R}^{n\times n}$ and $\lambda \in \mathbb{R}$. Then*

$$\mathbb{E}\exp(\lambda\mathcal{V}^T\mathcal{A}\mathcal{V}') \leq \mathbb{E}\exp(\lambda\tau\left(\mathcal{V}\right)\tau\left(\mathcal{V}'\right)G^T\mathcal{A}G'.)$$

Proof. We first observe that for any $v \in \mathbb{R}^n$,

$$\begin{aligned} \mathbb{E}(\exp(\lambda\langle\mathcal{V}, v\rangle)) &= \mathbb{E}\left(\exp\left(\lambda\left\|v\right\|_2\left\langle\mathcal{V}, \frac{v}{\left\|v\right\|_2}\right\rangle\right)\right) \\ &\leq \exp\left(\frac{\lambda^2\left\|v\right\|_2^2\tau\left(\mathcal{V}\right)^2}{2}\right) \\ &= \mathbb{E}\Big(\exp\big(\lambda\tau\left(\mathcal{V}\right)\langle G, v\rangle\big)\Big), \end{aligned} \tag{2.44}$$

where the inequality in (2.44) is by the definition of vector-valued sub-Gaussian random variables and the equality is obtained by calculating the moment-generation function of $\langle G, v\rangle$. We can now conclude the proof from the following:

$$\begin{aligned}
\mathbb{E}\Big(\exp\left(\lambda \mathcal{V}^T \mathcal{A}\mathcal{V}'\right)\Big) &= \mathbb{E}_{\mathcal{V}'}\Big(\mathbb{E}_{\mathcal{V}}\Big(\exp\left(\lambda\langle \mathcal{V}, \mathcal{A}\mathcal{V}'\rangle\right)\Big)\Big) && (2.45)\\
&\leq \mathbb{E}_{\mathcal{V}'}\Big(\mathbb{E}_{G}\Big(\exp\left(\lambda\tau\left(\mathcal{V}\right)\langle G, \mathcal{A}\mathcal{V}'\rangle\right)\Big)\Big) && (2.46)\\
&= \mathbb{E}_{G}\Big(\mathbb{E}_{\mathcal{V}'}\Big(\exp\left(\lambda\tau\left(\mathcal{V}\right)\langle \mathcal{V}', \mathcal{A}^T G\rangle\right)\Big)\Big) && (2.47)\\
&\leq \mathbb{E}_{G}\Big(\mathbb{E}_{G'}\Big(\exp\left(\lambda\tau\left(\mathcal{V}\right)\tau\left(\mathcal{V}'\right)\langle G', \mathcal{A}^T G\rangle\right)\Big)\Big) && (2.48)\\
&= \mathbb{E}\Big(\exp\left(\lambda\tau\left(\mathcal{V}\right)\tau\left(\mathcal{V}'\right) G^T \mathcal{A} G'\right)\Big), && (2.49)
\end{aligned}$$

where (2.45), (2.47) and (2.49) are due to Fubini's theorem and elementary transformations and (2.46) and (2.48) are both instances of the observation (2.44). □

Proof of Lemma 2. We can write

$$\mathcal{V}^T \mathcal{A}\mathcal{V} = \sum_{k,k'=1, k\neq k'}^{n} \mathcal{V}_k \mathcal{A}_{k,k'} \mathcal{V}_{k'}, \tag{2.50}$$

and since $\mathcal{V}$ is centered, $\mathbb{E}\left(\mathcal{V}^T \mathcal{A}\mathcal{V}\right) = 0$ immediately follows. Let $\mathcal{V}'$ be an independent copy of $\mathcal{V}$, and let G and G' be independently distributed according to $\mathcal{N}(0, \mathbf{id}_n)$. We denote the singular values of $\mathcal{A}$ with $s_1, \ldots, s_n$. With these definitions, we bound the moment-generating function of $\mathcal{V}^T \mathcal{A}\mathcal{V}$ as

$$\begin{aligned}
\mathbb{E}\exp\left(\lambda \mathcal{V}^T \mathcal{A}\mathcal{V}\right) &= \mathbb{E}\exp\left(\lambda \mathcal{V}^T \mathcal{A}\mathcal{V}\right) && (2.51)\\
&\leq \mathbb{E}\exp\left(4\lambda \mathcal{V}^T \mathcal{A}\mathcal{V}'\right) && (2.52)\\
&\leq \mathbb{E}\exp\left(4\lambda\tau\left(\mathcal{V}\right)^2 G^T \mathcal{A} G'\right) && (2.53)\\
&= \mathbb{E}\exp\left(4\lambda\tau\left(\mathcal{V}\right)^2 \sum_{k=1}^{n} \hat{G}_k \hat{G}'_k s_k\right) && (2.54)\\
&\leq \exp\left(\frac{\lambda^2}{2} \cdot \frac{128\tau\left(\mathcal{V}\right)^4 \sum_{k=1}^{n} s_k^2}{1-c}\right), && (2.55)
\end{aligned}$$

where (2.52) is due to the Decoupling Theorem [Ver18, Theorem 6.1.1], (2.53) is an application of Lemma 10, (2.54) holds for suitably transformed versions $\hat{G}, \hat{G}'$ of G, G' (note that they are still independent and follow the same distribution) and (2.55) is true if

$c \in (0,1)$ and $|\lambda| < c/(8\tau(\mathcal{V})^2 \max_{1\le k\le n} s_k)$ according to Lemma 8. So we can apply Lemma 9 to obtain

$$\mathbb{P}\left(\left|\mathcal{V}^T\mathcal{A}\mathcal{V}\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{\varepsilon^2(1-c)}{256\tau(\mathcal{V})^4 \sum_{k=1}^n s_k^2}\right) \tag{2.56}$$

in case $\varepsilon \le \frac{c}{1-c} \cdot \frac{16\tau(\mathcal{V})^2 \sum_{k=1}^n s_k^2}{\max_{1\le k\le n} s_k}$ and

$$\mathbb{P}\left(\left|\mathcal{V}^T\mathcal{A}\mathcal{V}\right| \ge \varepsilon\right) \le 2\exp\left(-c \cdot \frac{\varepsilon}{16\tau(\mathcal{V})^2 \max_{1\le k\le n} s_k}\right)$$

otherwise. We next choose c so as to minimize the upper bound on the tail probability. Because the bound in the first case is increasing with c while it is decreasing in the second case, the optimal choice for c is where the two cases meet. We can therefore calculate the optimal c as

$$c = \frac{\varepsilon \max_{1\le k\le n} s_k}{\varepsilon \max_{1\le k\le n} s_k + 16\tau(\mathcal{V})^2 \sum_{k=1}^n s_k^2}$$

and substituting this in (2.56), we obtain

$$\mathbb{P}\left(\left|\mathcal{V}^T\mathcal{A}\mathcal{V}\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{\varepsilon^2}{16\varepsilon\tau(\mathcal{V})^2 \max\limits_{1\le k\le n} s_k + 256\tau(\mathcal{V})^4 \sum_{k=1}^n s_k^2}\right).$$

The bounds $\tau(\mathcal{V}) \le \tau_{\max}$, $|s_k| \le \|\mathcal{A}\|_{\mathrm{op}}$, and identity $\|\mathcal{A}\|_{\mathrm{F}}^2 = \sum_{k=1}^n s^2$ allow us to conclude the proof of the lemma. □

3 Applications of Distributed Function Approximation to Vertical Federated Learning

In this chapter, we give examples of how the results of Chapter 2 can be applied to ML problems. We focus on a case that is particularly simple on the communication side (since only one weighted sum is OTA computed) but which we expect to gain significant relevance in practical systems. The application examples show how our DFA scheme can be leveraged to vastly increase the efficiency of the prediction phase of VFL both in terms of time and bandwidth resources (i.e., in our model, channel uses) and in terms of energy resources expended. For the training phase, we will either assume centralized offline training or use more communication-efficient decentralized methods that do not, however, leverage any form of OTA computation. Developing distributed training algorithms for VFL which can leverage the full power of OTA computation remains open as an interesting future research direction. However, we argue that in many cases of interest the communication cost incurred in the prediction phase can dominate that incurred in training and it is therefore worthwhile to focus on the prediction. This can, e.g., be the case when the training can be conducted offline and the models do not or only infrequently have to be re-trained; or when the number of training samples is small, but the number of features observed in the system is large and prediction tasks have to be carried out very frequently.

Contrary to [YLCT19], where the main focus in VFL is on providing privacy and security guarantees, in this work we focus on the communication efficiency of such schemes under the use of OTA computation. Since the OTA-computed predictors as well as the distributed training procedures we describe do not aggregate the observed features in a central point, it is reasonable to expect that these methods have good inherent privacy properties, and for some of the envisioned applications, such as, e.g., e-health, it is an important open question for future research how these privacy guarantees can be formalized and perhaps strengthened in the context of OTA-computed ML predictors. In Chapter 4, we address some of these concerns by providing formal security guarantees in the presence of eavesdroppers.

In the following, we expand upon the idea of OTA VFL by showing in Section 3.1 how the SVM approach can be generalized to the use of additive kernel SVMs and applied to regression and classification problems, and in Section 3.2 how classifiers that do not necessarily have to rely on SVMs at all can be constructed to solve binary classification problems. In Section 3.3, we present simulation results of two classification schemes constructed as described in Section 3.2 and compare them to two baseline approaches.

Both in Section 3.1 and in Section 3.2, we construct an ML regressor or predictor that has the form of a weighted sum, because such a function can be straightforwardly computed OTA using the DFA scheme described in Section 2.5. If the loss is Lipschitz-continuous, it can play the role of the function F in Definition 10 so that Theorem 2 provides a tail bound on the additional ML loss that the OTA-computed classifiers can incur in addition to the loss they would have in case of noiseless communication. The detailed technical statements and proofs of these facts can be found in Corollaries 5 and 6. We also give some examples of applicable Lipschitz-continuous losses for regression and classification that are commonly used in practice in Subsection 3.1.

3.1 Support Vector Machines with Additive Kernels for Regression and Classification

In this section, we give an example of additive, and therefore OTA computable, SVM regressors, focusing on Lipschitz-continuous losses. We say that the loss $\mathfrak{L}$ is B-Lipschitz-continuous if $\mathfrak{L}(x, y, \cdot)$ is Lipschitz-continuous for all $x \in \mathfrak{X}$ and $y \in \mathfrak{Y}$ with a Lipschitz constant uniformly bounded by B. Lipschitz-continuity of a loss function is a property that is also often needed in other contexts. Fortunately, many loss functions of practical interest possess this property. For instance, the absolute distance loss, the logistic loss, the Huber loss and the ε-insensitive loss, all of which are commonly used in regression problems [SC08, Section 2.4], are Lipschitz-continuous. Even in scenarios in which the naturally arising loss is not Lipschitz-continuous, for the purpose of designing the ML model, it is often replaced with a Lipschitz-continuous alternative. For instance, in binary classification, we have $\mathfrak{Y} = \{-1, 1\}$ and the loss function is given by

$$(x, y, t) \mapsto \begin{cases} 0, & \operatorname{sign}(y) = \operatorname{sign}(t) \\ 1, & \text{otherwise.} \end{cases}$$

This loss is not even continuous, which makes it hard to deal with. So for the purpose of designing the ML model, it is commonly replaced with the Lipschitz-continuous hinge loss or logistic loss [SC08, Section 2.3].

Here, we consider the case in which the features are K-tuples and the SVM can be trained in a centralized fashion. The actual predictions, however, are performed in a distributed setting; i.e., there are K users each of which observes only one component of the features. The objective is to make an estimate of the label available at the receiver while using as little communication resources as possible.

To this end, we consider the case of additive models which is described in [CH12, Section 3.1]. We have $\mathfrak{X} = \mathfrak{X}_1 \times \cdots \times \mathfrak{X}_K$ and a kernel $\kappa_k : \mathfrak{X}_k \times \mathfrak{X}_k \to \mathbb{R}$ with an associated reproducing kernel Hilbert space $\mathfrak{H}_k$ of functions mapping from $\mathfrak{X}_k$ to $\mathbb{R}$ for each $k \in \{1, \ldots, K\}$. Then by [CH12, Theorem 2]

$$\kappa : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R},\ ((x_1, \ldots, x_K), (x'_1, \ldots, x'_K)) \mapsto \kappa_1(x_1, x'_1) + \cdots + \kappa_K(x_K, x'_K) \tag{3.1}$$

is a kernel and the associated reproducing kernel Hilbert space is

$$\mathfrak{H} := \{f_1 + \cdots + f_K : f_1 \in \mathfrak{H}_1, \ldots, f_K \in \mathfrak{H}_K\}. \tag{3.2}$$

So this model is appropriate whenever the function to be approximated is expected to have an additive structure. We know [SC08, Theorem 5.5] that an SVM estimator has the form

$$f(x) = \sum_{j=1}^{n} \alpha_j \kappa(x, x^j), \tag{3.3}$$

where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $x^1, \ldots x^n \in \mathfrak{X}$. In our additive model, this is

$$f(x_1, \ldots, x_k) = \sum_{k=1}^{K} f_k(x_k), \tag{3.4}$$

where for each k,

$$f_k(x_k) = \sum_{j=1}^{n} \alpha_j \kappa_k(x_k, x_k^j). \tag{3.5}$$

We now state a result for the distributed approximation of the estimator of such an additive model as an immediate consequence of Theorem 2.

Corollary 5. *Consider an additive ML model, i.e., we have an estimator of the form (3.4), and assume that $\mathfrak{L}$ is a B-Lipschitz-continuous loss. Suppose further that all the f_K have bounded range such that the quantities $\bar{\Delta}(f)$ and $\Delta(f)$ as defined in (2.9) and (2.10) exist and are finite. Let $\varepsilon, \delta > 0$ and $M \geq M(f, \varepsilon, \delta)$ as defined in (2.14), where*

$\Phi := \mathbf{id}$ *and thus* $\Phi^{-1}(\varepsilon) = \varepsilon$*. Then, given* $x^K = (x_1, \ldots, x_K)$ *drawn from an arbitrary distribution*[13,14] *at the transmitters and any* $y \in \mathfrak{Y}$*, through* M *uses of the channel (2.2), the receiver can obtain an estimate* $\bar{f}$ *of* $f(x^K)$ *satisfying*

$$\mathbb{P}(\left|\mathfrak{L}(x^K, y, \bar{f}) - \mathfrak{L}(x^K, y, f(x^K))\right| \geq B\varepsilon) \leq \delta. \tag{3.6}$$

Proof. The Lipschitz continuity of $\mathfrak{L}$ yields

$$\mathbb{P}(\left|\mathfrak{L}(x^K, y, \bar{f}) - \mathfrak{L}(x^K, y, f(x^K))\right| \geq B\varepsilon) \leq \mathbb{P}(\left|\bar{f} - f(x^K)\right| \geq \varepsilon),$$

from which (3.6) follows by the definition of $M(f, \varepsilon, \delta)$. □

While [CH12] provides some examples of applications of SVMs with additive kernels to regression problems, the example for anomaly detection described in [RGS16] can be recovered as a special case of the framework described in this subsection, where the employed SVMs have linear kernels.

We conclude this subsection with a brief discussion of the feasibility of the condition that $f_1, \ldots, f_K$ have bounded ranges in the case of the additive SVM model discussed above. The coefficients $\alpha_1, \ldots, \alpha_n$ are a result of the training step and can therefore be considered constant, so all we need is that the ranges of $\kappa_1, \ldots, \kappa_K$ are bounded. This heavily depends on $\mathfrak{X}_1, \ldots, \mathfrak{X}_K$ and the choices of the kernels, but we remark that the boundedness criterion is satisfied in many cases of interest. The range of Gaussian kernels is always a subset of $(0, 1]$, and while other frequent choices such as exponential, polynomial and linear kernels can have arbitrarily large ranges, they are nonetheless continuous which means that as long as the input alphabets are compact topological spaces (e.g., closed hyperrectangles or balls), the ranges are also compact, and therefore bounded.

3.2 Model-Agnostic Approach to Over-the-Air-Computed Classifiers

In this subsection, we focus on classification problems. The general approach is model-agnostic in the sense that arbitrary and even different ML models can be used in the distributed agents, but we have decentralized classifiers with a low computational burden in mind, as is exemplified in the numerical simulations discussed in the following subsection.

[13] Arbitrary distribution means in particular that the components can be arbitrarily correlated.

[14] Note that Theorem 2 actually provides for a stronger result, since it allows arbitrary deterministic values, which implies the applicability to arbitrarily distributed random variables through the law of total probability.

We consider a feature alphabet $\mathfrak{X} = \mathfrak{X}_1 \times \cdots \times \mathfrak{X}_K$ and a label alphabet $\mathfrak{Y} = \{-1, 1\}$ as well as an unknown, but fixed probability distribution $\mathcal{P}$ on $\mathfrak{X} \times \mathfrak{Y}$. In the training phase, each user k observes J training samples

$$T_k = \left(\left(x_k^{(1)}, y^{(1)} \right), \ldots, \left(x_k^{(J)}, y^{(J)} \right) \right),$$

where for all k, j, we have $x_k^{(j)} \in \mathfrak{X}_k$, $y^{(j)} \in \mathfrak{Y}$ and $(x_1^{(j)}, \ldots, x_K^{(j)}, y^{(j)})$ is drawn according to $\mathcal{P}$.

Each user k can train its own model based on T_k which is distributed according to the marginal of $\mathcal{P}$ with respect to $\mathfrak{X}_k \times \mathfrak{Y}$. We propose to use a slight variation of the well-known boosting technique and define a classifier

$$f := \sum_{k=1}^{K} \alpha_k f_k, \tag{3.7}$$

where f_k is the base classifier locally trained at user k and α_k is a nonnegative weight. As an immediate corollary to Theorem 2 parallel to Corollary 5, f can be approximated at a central node in a distributed manner.

Corollary 6. *Assume that $\mathfrak{L}$ is a B-Lipschitz-continuous loss. Let $\varepsilon, \delta > 0$ and $M \geq M(f, \varepsilon, \delta)$ as defined in (2.14), where $\Phi^{-1}(\varepsilon) = \varepsilon$, noting that*

$$\bar{\Delta}(f) = 2 \sum_{k=1}^{K} \alpha_k, \;\; \Delta(f) = 2 \max_{k=1}^{K} \alpha_k.$$

Then, given any $x^K = (x_1, \ldots, x_K)$ drawn from an arbitrary[13,14] distribution at the transmitters and any $y \in \mathfrak{Y}$, through M uses of the channel (2.2), the receiver can obtain an estimate $\bar{f}$ of $f(x^K)$ satisfying

$$\mathbb{P}(\left| \mathfrak{L}(x^K, y, \bar{f}) - \mathfrak{L}(x^K, y, f(x^K)) \right| \geq B\varepsilon) \leq \delta. \tag{3.8}$$

The proof is the same as for Corollary 5.

It is important to remark here that the predictor f can only be approximated at the receiver up to a residual error (which can, however, be controlled) and thus, a guarantee in terms of the 0-1-loss is not sufficient to apply Corollary 6 and we instead need it to be in terms of a Lipschitz-continuous loss.

This is a relatively generic framework that can in principle work with any particular boosting technique which determines weights $\alpha_1, \ldots, \alpha_K$ and guarantees a bound on the loss of the predictor f dependent on the errors of the base classifiers $f_1, \ldots, f_K$. In the

following, we describe two variations of this general approach, both based on well-known ideas from ML (cf., e.g., [MRT12, Chapter 6]).

The first one, *equal majority vote*, amounts to setting $\alpha_1 = \cdots = \alpha_K = 1$ and using local classifiers f_k trained only on the locally available features. This method has the advantage that the whole training procedure can be carried out in a fully decentralized way without any form of coordination or exchange of information between the agents (given that the labels for the training phase are already known everywhere; but they could, e.g., be broadcast from a central point at a cost independent of the number of agents or dimensionality of the feature space).

If we have the possibility to exchange some data between the agents, we can use the following adaptation of the AdaBoost scheme [MRT12, Figure 6.1] to the distributed setting. The algorithm runs through $L \leq K$ iterations, choosing a user $\mathbf{h}_\ell$ at iteration ℓ to provide a base classifier $f_{\mathbf{h}_\ell}$ and assigning a corresponding weight $\alpha_{\mathbf{h}_\ell}$. It also computes probability distributions $p_1, \ldots, p_{L+1}$ on the index set of the training data $\{1, \ldots, J\}$, initializing p_1 as the uniform distribution, as well as base classifier errors $\epsilon_1, \ldots, \epsilon_L$ and normalization constants $\mathbf{Z}_1, \ldots, \mathbf{Z}_L$. Each iteration ℓ consists of the following steps:

1. The central node chooses a user $\mathbf{h}_\ell$ and broadcasts the choice.
2. User $\mathbf{h}_\ell$ trains a base classifier $f_{\mathbf{h}_\ell} : \mathfrak{X}_{\mathbf{h}_\ell} \to \{-1, 1\}$ on the training sample with distribution p_ℓ and broadcasts the indices of the training samples incorrectly classified by $f_{\mathbf{h}_\ell}$.
3. From this information, every node in the system computes the following, where $\mathbf{1}_{\cdot}$ denotes the indicator function which is 1 if the condition in the index is true and 0 otherwise:
   - $\epsilon_\ell := \sum_{j=1}^{J} p_\ell(j) \mathbf{1}_{f_{\mathbf{h}_\ell}(x^{(j)}_{\mathbf{h}_\ell}) \neq y^{(j)}}$
   - $\alpha_{\mathbf{h}_\ell} := \frac{1}{2} \log \frac{1-\epsilon_\ell}{\epsilon_\ell}$
   - $\mathbf{Z}_\ell := 2\sqrt{\epsilon_\ell(1-\epsilon_\ell))}$
   - $p_{\ell+1}(j) := p_\ell(j) \exp(-\alpha_{\mathbf{h}_\ell} f_{\mathbf{h}_\ell}(x^{(j)}_{\mathbf{h}_\ell}) y^{(j)}) / \mathbf{Z}_\ell$

The resulting classifier is then as defined in (3.7), where we assign $\alpha_k := 0$ whenever $k \neq \mathbf{h}_\ell$ for all ℓ. [MRT12, Theorem 6.1] guarantees that the empirical 0-1-loss of f is at most

$$\exp\left(-2\sum_{\ell=1}^{L}\left(\frac{1}{2} - \epsilon_\ell\right)^2\right), \tag{3.9}$$

which unfortunately is insufficient to apply Corollary 6, because the 0-1-loss is not Lipschitz-continuous. However, the proof of the theorem relies only on the inequality $\mathbf{1}_{f(x^K)y \leq 0} \leq$

$\exp(-f(x^K)y)$ for the instantaneous 0-1-loss. Since the inequality $\log(1+\exp(-f(x^K)y)) \leq \exp(-f(x^K)y)$ also clearly holds, we can replace the 0-1-loss in the proof with the logistic loss $\mathfrak{L}(x^K, y, \hat{y}) := \log(1 + \exp(-y\hat{y}))$ (or, indeed, any other loss which satisfies this inequality). This yields the same bound (3.9) on the 1-Lipschitz-continuous logistic loss and thus we can apply Corollary 6 with $B := 1$ to derive a guarantee on the logistic loss of the distributed approximation of our AdaBoost classifier.

We conclude with some remarks on the distributed training. The choice in step 1 could, e.g., be predetermined (in which case no communication in this step is necessary) or random, but we could also greedily select the classifier with smallest error using an instance of ScalableMax [9] [2, Section IV]. As for the communication cost of the distributed training, step 1 exhibits a favorable scaling which is linear in L and logarithmic in K, however, step 2 has a cost linear in the number of training samples. There is a conceptually simpler alternative to this distributed scheme in which we communicate the full training set to the central node and perform the training in a centralized manner. The advantage in communication cost of the distributed scheme over this centralized alternative is only a constant factor. On the other hand, since only one bit per training sample and user is transmitted, this constant gain could potentially be quite large, depending on the complexity of the feature spaces. Also, in the distributed training scheme, the computational load of training the base classifiers is distributed across all nodes, which may in practice also be an advantage wherever the computational capabilities of the central node are limited. However, since the distributed training currently leverages no OTA computation and leaves that for the computation of the trained classifier itself, finding a distributed scheme which can exploit OTA computation to achieve a gain in asymptotic behavior as opposed to only a constant factor could be a worthwhile question for future research.

3.3 Numerical Results for Over-the-Air-Computed Decision Tree Classifiers

In order to illustrate how the scheme analyzed in this work can be used to compute classifiers for anomaly detection problems in large sensor networks, we have conducted numerical simulations on a synthetic binary classification problem generated by the `make_classification` function in the *datasets* package of the scikit-learn toolbox [P+11] for Python. It places clusters for the two classes at the edges of a hypercube in a Euclidian space of informative features, adds redundant features that are linear combinations as well as useless features that are pure noise and applies various kinds of noise and nonlinearities. The resulting features are then shuffled randomly, partitioned and assigned to the distributed agents. For the training set, the agents also learn the correct corresponding

labels. We construct two different OTA-computed classifiers as described in the preceding subsection:

1. For the equal majority vote classifier, each agent trains a decision tree model of height at most 2 based on the locally available features only. The OTA-computed classifier is then as put forth in equation (3.7), where $\alpha_1 = \cdots = \alpha_K = 1$.

2. For the AdaBoost classifier, the agents train their models cooperatively as described in the preceding section, using decision tree classifiers as the local base classifiers. The next agent at each iteration is picked uniformly at random from among the agents which have not yet been selected. This procedure yields not only differently trained local models compared to the equal majority vote, but also weights $\alpha_1, \ldots, \alpha_K$ which can be used for the OTA-computed classifier as in equation (3.7).

We assume that the agents are connected to a central receiver through a fast-fading wireless MAC, where no instantaneous CSI is available. The only kind of information we assume is available is the average power of the complex Gaussian channel gain at the transmitters and the average power of the additive noise at the receiver. The distributed classification is simulated for noise and fading drawn from i.i.d. Gaussian distributions and for a scenario exhibiting various degrees of correlation and non-Gaussian components:

- For the fading, we achieve this by passing the fading coefficients through a lowpass filter, which cuts off all but a given percentage of the energy (the cutoff percentage) and then re-normalizes the remaining signal.

- For the noise, we simulate Middleton class A noise (also called impulsive noise), which is a commonly used noise model for power line communications [FLNS10, Section 2.6.3.1] but has also been empirically found to be a relevant phenomenon in wireless communications [MS93]. We simulate it as described in [FLNS10, eq. (2.49) ff.]: In order to create one sample of noise, a random variable $\mathbf{m}$ is drawn from a Poisson distribution with intensity $\mathbf{A}$, a parameter called the *impulsive index*, and then a centered Gaussian random variable with variance

$$2\mathfrak{P}_N \frac{\mathbf{m}/\mathbf{A} + \mathbf{\Gamma}}{1 + \mathbf{\Gamma}}$$

  is drawn, where $\mathfrak{P}_N$ is the overall power of the noise per complex dimension and the parameter $\mathbf{\Gamma}$ is called the *Gaussian-to-impulsive power ratio*. Finally, a phase shift is applied drawn uniformly from the complex unit circle. This process defines non-Gaussian, but i.i.d. noise. We have therefore modified it slightly and draw one

m for every 4 channel realizations so that we create a correlation also in the additive noise.

We simulate the computation of each of the two distributed classifiers described above in three different ways:

1. The DFA scheme as described in Section 2.5.
2. A TDMA scheme with average power normalization in which only one of the agents can transmit at a time. The information is also transmitted in analog form, since each agent conveys only one bit of information and therefore digital coded schemes would not be suitable. Since the agents transmit during a much shorter time than in the DFA scheme, we normalize their transmission power so that the average energy consumed per channel use equals that in the DFA scheme. The only exception to this is the case when some agents have to be allocated zero channel uses, since the number of total channel uses available is smaller than the number of agents in the system. In this case, obviously, the agents allocated zero channel uses also have zero energy consumption. This scheme has a significantly higher peak transmission power than the DFA scheme.
3. A TDMA scheme with peak power normalization. It works as the one under item 2, but the transmission power is normalized so that it has the same peak power as the DFA scheme, which means that it has a significantly lower average consumption.

We also show two baselines to make the error contribution of the compared communication schemes clearer:

1. a noiseless version of the majority vote classifier, and
2. a noiseless version of the AdaBoost-inspired classifier.

The training set consists of $50,000$ samples and the test set of $200,000$. We have generated two different binary classification problems, one for 10 transmitters and one for 25 transmitters.

Comparison of DFA and TDMA schemes for equal majority vote and AdaBoost We have simulated the DFA scheme as well as the TDMA baseline comparisons for a scenario with moderate correlation and non-Gaussianity. The cutoff percentage for the lowpass filter applied to the fading was chosen at 80%, and the parameters for the Middleton Class A noise were $\mathbf{A} = 3$ and $\mathbf{\Gamma} = 3$. In Fig. 3.1, we plot the classification error on the test set as a function of the number of complex channel uses for a fixed SNR of -6

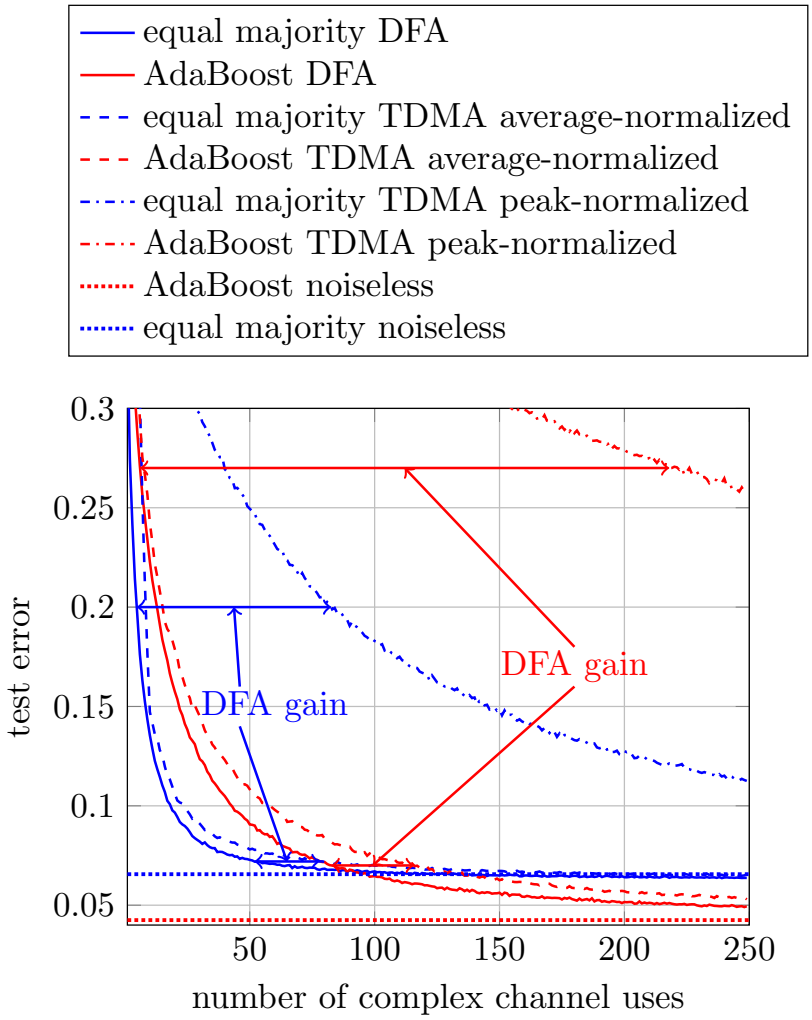

Figure 3.1: Comparison of the classification error on the test set of DFA/TDMA and equal majority/AdaBoost schemes. 10 transmitters, cutoff percentage 80%, $\mathbf{A} = 3$, $\mathbf{\Gamma} = 3$, SNR -6 dB.

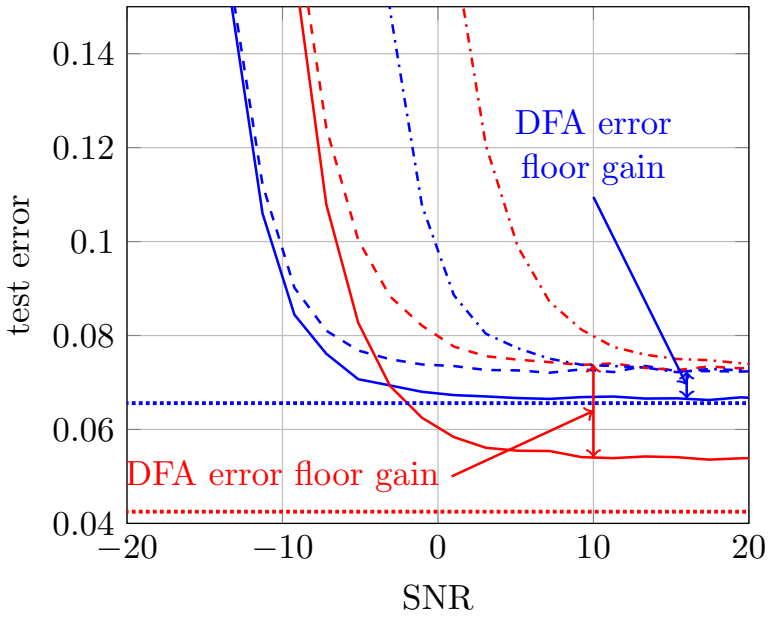

Figure 3.2: Comparison of the classification error on the test set of DFA/TDMA and equal majority/AdaBoost schemes. 10 transmitters, cutoff percentage 80%, $\mathbf{A} = 3$, $\mathbf{\Gamma} = 3$, 50 complex channel uses.

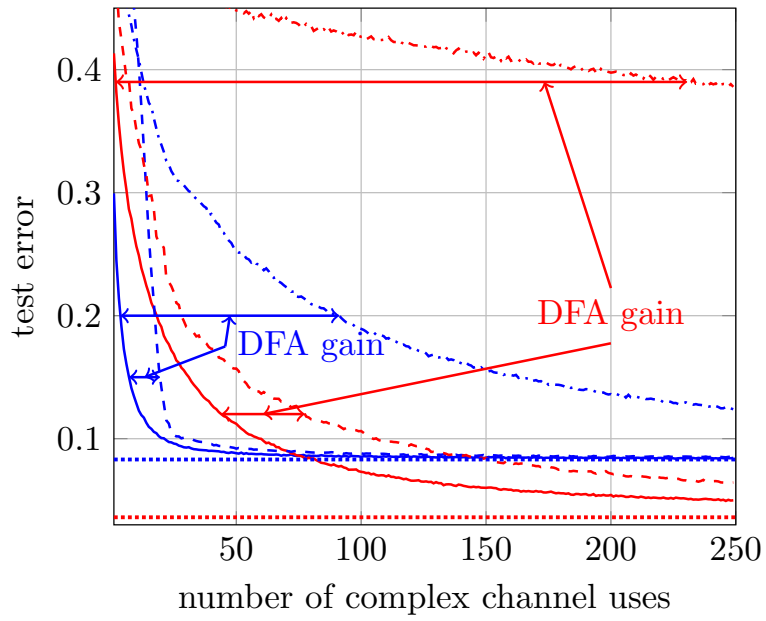

Figure 3.3: Simulation results for 25 transmitters at a fixed SNR of −6 dB, correlation parameters are the same as in Fig. 3.1.

dB and 10 transmitters, and in Fig. 3.2, we plot the error as function of SNR for a fixed number of 50 complex channel uses. We can see that, as the effect of the multiplicative fading dominates that of the additive noise, the schemes reach an error floor that cannot be lowered with an increase of the transmission power. When the number of complex channel uses is increased, on the other hand, the error curves approach the noiseless classification error even if the SNR is kept fixed.

For instance, to obtain a classification error of 0.07 or better, both for the equal majority vote and AdaBoost, the average-power normalized TDMA scheme needs over 30 channel uses more than the DFA scheme. Since we compare with a TDMA scheme that uses the same average energy per channel use as the DFA scheme, this means that the TDMA scheme not only consumes more wireless spectrum and/or time, but also significantly more energy. For the case of the same peak power consumption (which means that TDMA consumes less average power since transmitters are silent most of the time), the difference is huge and can be several hundred channel uses depending on the error level.

The advantage of the DFA scheme over the TDMA alternatives is quite pronounced even at a relatively low number of only 10 transmitters. In Fig. 3.3, we show the same plot as in Fig. 3.1, but for a different ML problem with 25 transmitters. It can be seen in the plot that as the number of transmitters increases, the difference in performance between the DFA and TDMA schemes becomes even stronger. This is due to the different scaling behaviors of the schemes.

We have run these simulations for many instantiations of the randomly generated classification schemes (not depicted for lack of space) and note that while in some cases the equal majority vote scheme performs similarly as the AdaBoost scheme in the noiseless case, in many cases the error behavior of the AdaBoost scheme is much better and it is more robust, e.g. in the case that a large number of agents observes only useless features while only few agents observe the informative and repetitive features that can be used to solve the classification problem. That being said, in the case in which the equal majority vote performs similarly to AdaBoost in the noiseless case, its error behavior in the communication schemes is better since it better utilizes the available peak transmission power.

Synchronization errors Since the pre-processing described in Section 2.5.1 creates a sequence of i.i.d. random variables (conditioned under $s_1, \ldots, s_K$), we can expect that the scheme is quite robust to synchronization errors between the transmitters. This is important since perfect synchronization of the transmitted signals at the receiver would be a very hard task to achieve in practice. In order to substantiate this argument, we have run simulations with relatively large synchronization errors of several symbol durations:

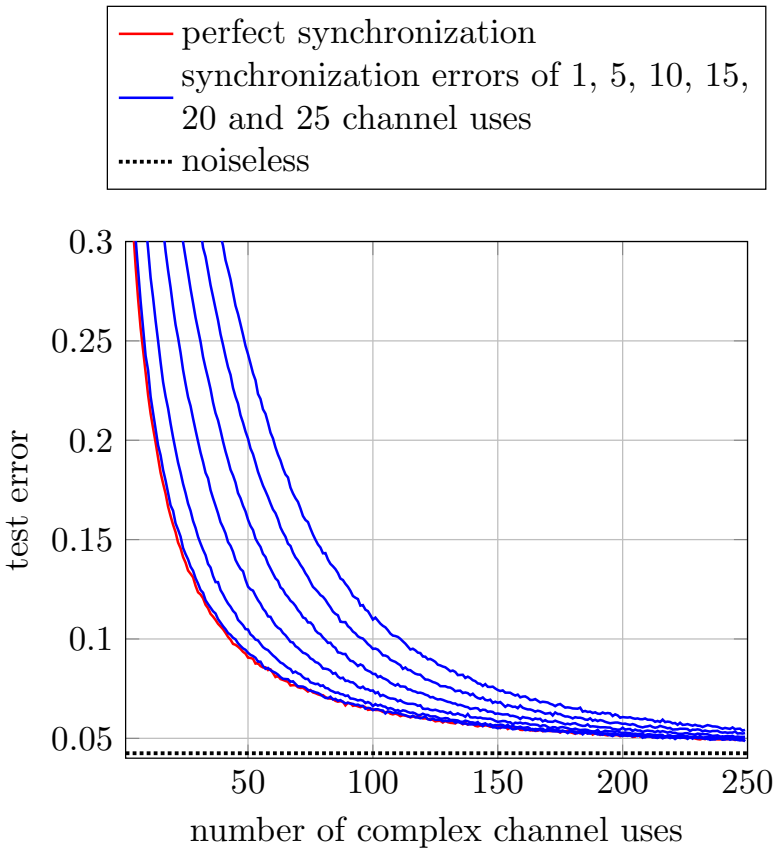

Figure 3.4: Impact of synchronization errors on the DFA AdaBoost scheme in the scenario of Fig. 3.1.

For the starting time of the transmission for each transmitter, we have added a uniformly random number of channel uses in a range of up to 1, up to 5, up to 10, up to 15, up to 20 and up to 25 channel uses. In Fig. 3.4, we show the impact of the synchronization errors for the AdaBoost DFA scheme and the same choice of parameters as for Fig. 3.1. The solid red curves (representing the case of perfect synchronization) are therefore the same in both figures while the blue curves in Fig. 3.4 depict the performance for various values of the maximum synchronization error. The performance degrades gracefully even for extremely large synchronization errors and the number of additional channel uses that needs to be expended to maintain the same classification error is about twice the value of the maximum synchronization error. We remark that we expect the number of additional channel uses that needs to be expended to scale with the synchronization error and not with the number of transmitters. Moreover, for synchronization errors of the same order of magnitude as the symbol duration, the drop in performance is barely noticeable.

Comparison of different correlation scenarios In order to get an idea of how strongly the correlation and non-Gaussianity impact on the performance of the scheme, we have compared the AdaBoost DFA scheme (the solid red curve from Fig. 3.1) for various choices of the correlation and non-Gaussianity parameters. Qualitatively, the higher the values of $\mathbf{A}$ and $\mathbf{\Gamma}$, the more closely the additive noise is to a Gaussian distribution and the lower the values, the stronger pronounced is the non-Gaussianity of the noise. For the fading, a cutoff percentage of 100% corresponds to i.i.d. Gaussian fading, while lower cutoff percentages mean that the fading changes more slowly over time. The results of our comparison in Fig. 3.5 show that the scheme performs best for the Gaussian i.i.d. case but, as is expected from the theoretical analysis, the exponential decay of the error is retained even for strongly pronounced correlation and non-Gaussianity.

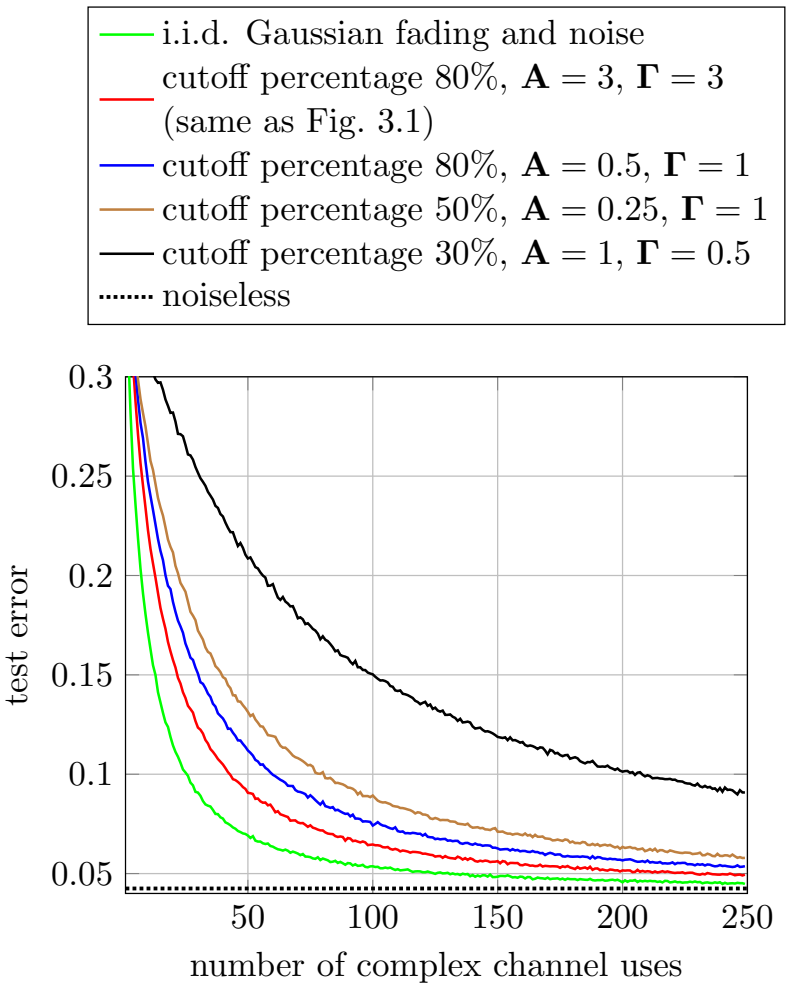

Figure 3.5: Comparison of the performance of AdaBoost DFA for various choices of the non-Gaussianity and correlation parameters.

4 Security in Over-the-Air Computation

OTA computation schemes carry the promise that they can improve communication efficiency so dramatically in many cases of practical interest that they can be seen as an enabler for applications in massive wireless networks for which the communication cost or the time delay incurred would otherwise be prohibitive. However, there is also a flip side that has the potential to hinder widespread adoption: Some tools that enhance the properties of communication and are frequently used as building blocks in communication systems inherently rely on the principle of source-channel separation. Therefore, they cannot be adapted to work in a scenario where a joint source-channel approach is taken such as in OTA computation. One example of such a building block that is particularly important in modern communication systems is cryptography. OTA communication schemes as described in Chapter 2 are vulnerable to a number of attacks such as malicious transmitters participating in the scheme or attackers eavesdropping on the transmission, and it is unclear whether and how state-of-the-art cryptographic security could be adapted to defend against such threats.

At least for the latter kind of threat – attackers eavesdropping on the communication – information-theoretic security, while not adaptable in a straightforward fashion, provides a set of tools with which a defense can be developed. The ultimate goal in this direction should be full semantic security [BTV12]. As a first step, we propose to extend the system model with a jammer as depicted in Fig. 4.1. This shows how information-theoretic security tools can be adapted to the OTA computation setting. The jammer can increase the variance of the eavesdropper's estimate of the quantity of interest, but not fully prevent it from obtaining an estimate.

The key assumption we make is that the received jamming signal must be stronger for the legitimate receiver than it is for the eavesdropper. This way, the legitimate receiver $\mathfrak{B}$ can exploit the dependencies which we carefully introduce into the jamming signal to reconstruct it exactly. To the eavesdropper, the received signal is almost equivalent to an i.i.d. jammed transmission. With the knowledge of the full jamming signal, the legitimate receiver can then cancel it from its received signal or at least mitigate its impact. The approximation of the OTA computed function value at the legitimate receiver can then be carried out independently of the security scheme. It can, e.g., follow the method described

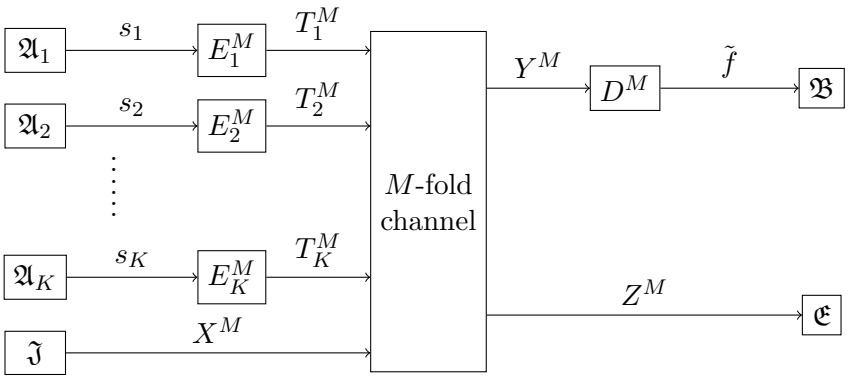

Figure 4.1: System model for distributed function approximation with security constraints.

in Chapter 2. In the present chapter, we give an example for how to combine the security and function approximation schemes in the simpler case of an AWGN channel without fast fading.

While our results on OTA computation rely heavily on the particular structure of wireless channels, the class of channels for which the reconstruction of the jamming signal is possible is much more general. This will become apparent in the following.

We are not aware of a similar system model having been proposed before for OTA computation, but we draw heavily from existing tools in information theory. As the main building blocks for proving security guarantees in a DFA scheme with friendly jamming, we use two information theoretic tools, namely *coding for the compound channel* and *channel resolvability*. The latter ensures that the jamming signal can be reconstructed fully by the legitimate receiver[15] and thus be canceled from the received signal. Therefore, it has no impact on the quality of the objective function estimate. Channel resolvability guarantees that the jamming signal is virtually indistinguishable from white noise for the eavesdropper. This virtual indistinguishability is phrased in terms of variational distance between the actually observed distribution and a superposition of the transmitters' signal with white noise from the jammer. The coding result for the continuous compound channel which we derive may also be of interest as an independent result.

4.1 Prior Work

To the best of our knowledge, the OTA computation problem over a wiretap channel has not yet been considered in the literature. Therefore, in this subsection we briefly summarize the literature on the building blocks other than OTA computation we use for

[15]For a formal statement, see Definition 12.

the approach to the wiretap OTA computation channel that we propose in this work as well as for literature on concepts that are closely related to the ones presented in this paper.

Coding for compound channels. The compound channel problem was introduced independently in [Dob59, BBT59, Wol59], while first independent results for the capacity expression can be found in [BBT59, Wol59]. These works, however, explore mainly the case of finite input and output alphabets. The *semicontinuous* case in which only the input alphabet is assumed to be finite is briefly touched upon in [Wol59] and studied in more detail in [Kes61] which provides an example showing that the capacity expression from the finite case does not carry over to the semicontinuous case in general. The semicontinuous case was further explored in [Yos65, Ahl67]. In many cases of practical interest, the capacity expression from the finite case can be generalized to the *continuous* case in which neither input nor output alphabets are assumed to be finite, as was found in [RV68] for a class of Gaussian compound channels.

Channel Resolvability and Semantic Security. The concept of channel resolvability was introduced in [Wyn75a, HV93]. Further results relevant in the context of this work appeared, e.g., in [Csi96, Dev05, HM16, Cuf16]. We use our generalization proposed in Theorem 6 for continuous channels as a basis for our proposed scheme. Although we cannot provide full semantic security guarantees in this work, we also heavily draw from the idea of obtaining semantic security by means of channel resolvability, which is laid out in [Hay06, CK11, BTV12, BL13].

Friendly Jamming The idea of friendly jamming has been used in [NG05] to aid a transmitter-receiver pair in protecting a point-to-point transmission from a passive eavesdropper. Distributed and centralized beamforming techniques are used so that the jamming signal impacts the signal-to-noise ratio at the eavesdropper but not at the legitimate receiver. Several more recent works (cf., e.g., [VBBM10, VBBM11, SY12]) have expanded upon this idea and refined the friendly jamming techniques, but to the best of our knowledge, they have not yet been used to protect OTA computation against eavesdropping.

Physical Layer Security The concept of information theoretic secrecy was introduced in [Sha49] and the wiretap channel model together with a weaker, but more tractable notion of secrecy was introduced in [Wyn75b]. Based on this, various stronger secrecy notions have been introduced and investigated (e.g., [Mau94, HK14, BTV12]). All of these existing works investigate how digitally coded transmissions can be protected against

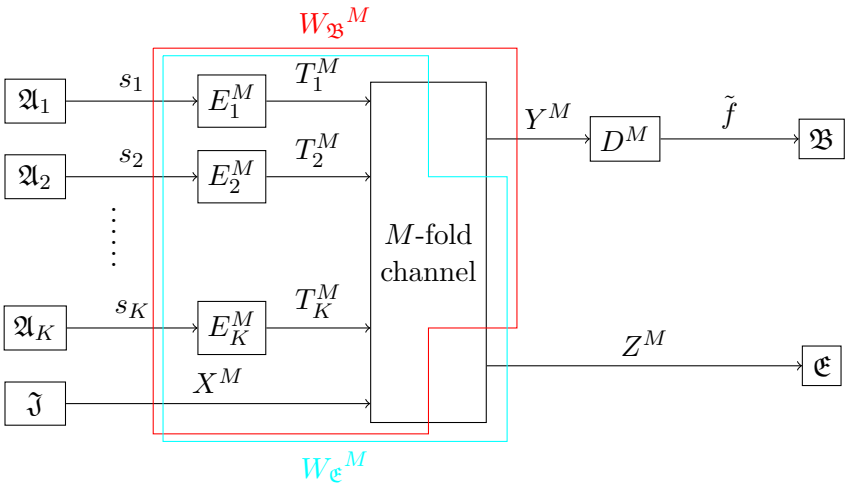

Figure 4.2: System model for distributed function approximation with jamming described in Section 4.2.2.

eavesdropping, while in the present work, we focus on uncoded analog transmissions over multiple-access channels.

4.2 System Model

4.2.1 Distributed Approximation of Functions

The system model and problem statement for DFA are the same as in Section 2.2.1.

Depending on the application at hand, there are multiple ways in which the quality of the estimate $\tilde{f}$ can be quantified. Besides the notion of ε-approximation of a function at confidence level δ defined in (2.5), we also define the approximation by criterion of the mean square error (MSE).

Definition 11. *We say that f is V-MSE-approximated if, under a uniform distribution of $f(s_1, \ldots, s_K)$, we have*

$$\mathbb{E}\left(\left(\tilde{f} - f(s_1, \ldots, s_K)\right)^2\right) \leq V,$$

where the expectation is over the joint distribution of $s_1, \ldots, s_K$ and $\tilde{f}$ which is induced by the distributed function approximation scheme and the channel.

4.2.2 Secrecy Extension to Distributed Approximation of Functions

In order to incorporate security aspects into the framework, we consider the extended system model depicted in Fig. 4.2.

The first addition to the model is an attacker $\mathfrak{E}$ which attempts to eavesdrop on the transmission and would like to gain knowledge about $s_1, \ldots, s_K$. At each channel use, $\mathfrak{E}$ observes an output Z ranging over the eavesdropper's alphabet $\mathcal{Z}$. As a counter-measure, we add a friendly jammer $\mathfrak{J}$ which transmits some jamming sequence X^M with the objective to prevent $\mathfrak{E}$ from obtaining information while still allowing $\mathfrak{B}$ to obtain a good estimate of $f(s_1, \ldots, s_K)$.

Definition 12. *A DFA scheme with jamming consists of:*

- *A distributed function approximation scheme as described in Section 2.2.1; i.e., pre- and post-processing schemes*
- *A jamming strategy given by a probability distribution on $\mathcal{X}^M$.*

We say that a distributed function approximation scheme with jamming allows reconstruction of the jamming signal with probability δ *if there is a decoding function $\vartheta : \mathcal{Y}^M \to \mathcal{X}^M$ such that*

$$\sup_{s_1 \in \mathcal{S}_1, \ldots, s_K \in \mathcal{S}_K} \mathbb{P}_{s_1,\ldots,s_K} \left(\vartheta(Y^M) \neq X^M \right) \leq \delta$$

and δ is the smallest number with this property.

The objective is to find admissible pre- and post-processing strategies as well as a jamming strategy such that $\mathfrak{B}$ can obtain a good approximation $\tilde{f}$ of $f(s_1, \ldots, s_K)$ while bounding the usefulness of any information that $\mathfrak{E}$ can obtain about $s_1, \ldots, s_K$.

Together with the channel, a distributed function approximation scheme with jamming induces a probability distribution $\hat{R}_{s_1,\ldots,s_K}$ on $\mathcal{Z}^M$ for each $(s_1, \ldots, s_K) \in \mathcal{S}$. How secure the scheme is depends on how strongly $\hat{R}_{s_1,\ldots,s_K}$ depends on $s_1, \ldots, s_K$. In the following, we formalize this notion.

Any measurable function $g : \mathcal{S} \to \hat{\mathcal{S}}$, where $\hat{\mathcal{S}}$ is a measurable space, is called an *eavesdropper's objective.*

Definition 13. *1. Given a real number $\eta \geq 0$, we say that a distributed function approximation scheme with jamming is* η-semantically secure *if there is a probability measure ν on $\mathcal{Z}^M$ such that for all $(s_1, \ldots, s_K) \in \mathcal{S}$,*

$$\left\| \hat{R}_{s_1,\ldots,s_K} - \nu \right\|_{\mathrm{TV}} \leq \eta, \tag{4.1}$$

where $\|\mu\|_{\mathrm{TV}} := \sup\{\mu(\mathcal{E}) : \mathcal{E}$ *is an event*$\}$ *denotes the total variation norm on signed measures.*

2. *Let* $g : \mathcal{S} \to \hat{\mathcal{S}}$*, where* $\hat{\mathcal{S}} \subseteq \mathbb{R}$ *is measurable and bounded, be an eavesdropper's objective. Let* $V \geq 0$ *be a real number. We say that a distributed function approximation scheme with jamming is* (g, V)-MSE-secure *if under a uniform distribution of* $g(s_1, \ldots, s_K)$*, for every estimator* $d : \mathcal{Z}^M \to \hat{\mathcal{S}}$*, we have*

$$\mathbb{E}\left(\left(d(Z^M) - g(s_1, \ldots, s_K)\right)^2\right) \geq V,$$

where the expectation is over the joint distribution of $s_1, \ldots, s_K$ *and* Y^M *which results from the application of the distributed function approximation scheme with jamming and the channel.*

In statistical terms, that a scheme is (g, V)-MSE-secure means that all estimators the eavesdropper can apply have MSE at least V under a uniformly distributed objective. A uniform distribution of the objective means that $s_1, \ldots, s_K$ are randomly distributed in such a way that $g(s_1, \ldots, s_K)$ follows a uniform distribution.

In a sense made explicit by the following lemma, semantic security is the stronger of the two security notions from Definition 13.

Lemma 11. *Let* $\hat{\mathcal{S}} := [a, b]$*, let* $g : \mathcal{S} \to \hat{\mathcal{S}}$ *be an eavesdropper's objective and* $\eta \geq 0$ *a real number.*

Then, any distributed function approximation scheme with jamming that is η*-semantically secure is also* $(g, (1/12 - \eta)(b-a)^2)$*-MSE-secure.*

Proof. Let $d : \mathcal{Z}^M \to \hat{\mathcal{S}}$. Then, assuming the distribution of $s_1, \ldots, s_K$ corresponds to a uniform distribution on $[a, b]$ of $g(s_1, \ldots, s_K)$, we have

$$\begin{aligned}
&\mathbb{E}_{s_1,\ldots,s_K}\mathbb{E}_{\hat{R}_{s_1,\ldots,s_K}}\left(\left(d(Z^M) - g(s_1, \ldots, s_K)\right)^2\right) \\
&= \mathbb{E}_{s_1,\ldots,s_K}\int_0^{(b-a)^2} \hat{R}_{s_1,\ldots,s_K}\left(\left(d(Z^M) - g(s_1, \ldots, s_K)\right)^2 > a\right) da \\
&\overset{(4.1)}{\geq} \mathbb{E}_{s_1,\ldots,s_K}\int_0^{(b-a)^2} \left(\nu\left(\left(d(Z^M) - g(s_1, \ldots, s_K)\right)^2 > a\right) - \eta\right) da \\
&= \mathbb{E}_{s_1,\ldots,s_K}\mathbb{E}_{\nu}\left(\left(d(Z^M) - g(s_1, \ldots, s_K)\right)^2\right) - \eta(b-a)^2 \\
&\geq \left(\frac{1}{12} - \eta\right)(b-a)^2,
\end{aligned}$$

where the last step is because under ν, Z^M is independent of $s_1, \ldots, s_K$. Therefore, the posterior distribution of $g(s_1, \ldots, s_K)$ given Z^M is uniform on $[a, b]$, which implies that

the minimum MSE is the variance of the uniform distribution. Denoting with $\mathcal{U}$ a random variable distributed uniformly in $[a, b]$, we can calculate its variance as

$$\begin{aligned}
\mathbb{E}\left(\mathcal{U}^2\right) - \left(\mathbb{E}\mathcal{U}\right)^2 &= \int_a^b \frac{g^2}{b-a} dg - \frac{(a+b)^2}{4} \\
&= \frac{b^3 - a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\
&= \frac{b^2 + a^2 + ab}{3} - \frac{a^2 + 2ab + b^2}{4} \\
&= \frac{4b^2 + 4a^2 + 4ab - 3a^2 - 6ab - 3b^2}{12} \\
&= \frac{1}{12}(b-a)^2. \qquad \square
\end{aligned}$$

4.2.3 Special case $K = 1$

We conclude this section with a brief discussion of the important special case $K = 1$. While one of the main motivations of the methods developed in this paper is their scalability to large values of K, the case of low values of K can also be interesting in many practical applications and be instructive to understand the nature of our results better.

For the special case of only a single transmitter ($K = 1$), the problem reduces to a point-to-point transmission of the real number $f(s_1)$ in the presence of an eavesdropper and a friendly jammer. In our results in this paper, there is no assumption that K has to be large; in particular, they remain applicable also when $K = 1$. However, since in this case no function of *distributed* values has to be computed over the channel, it is possible to separately source and channel encode $f(s_1)$. After the source coding step has been performed, the remaining problem is very similar to jammer-aided secret communication as treated for instance in [NG05, VBBM10, VBBM11].

But although this approach is applicable to the same communication task, it is important to note that the way in which the friendly jammer has to be placed differs significantly. In the approach of this paper, the jamming signal has to be stronger at the legitimate receiver than it is at the eavesdropper. As long as this condition is satisfied, the legitimate receiver has the ability to almost completely cancel the jamming signal. This means that our method remains applicable even if the gap in terms of jammer signal strength between the legitimate receiver and the eavesdropper is relatively small. In [NG05,VBBM10,VBBM11], on the other hand, it is necessary that the jamming signal is stronger at the eavesdropper than it is at the legitimate receiver. Moreover, this gap between signal strengths has to be as large as possible since the jammer's signal strength at the legitimate receiver diminishes the capacity of the main channel. Therefore, our results in this case are more suitable for

scenarios where is is possible to assure a high jamming signal strength at the legitimate receiver while results from [NG05, VBBM10, VBBM11] are more suitable in cases where all possible eavesdropper locations can be covered with strong jamming signals that have very low strength at the location of the legitimate receiver.

4.3 Specialization to the Additive White Gaussian Noise Channel

In general, the approximation scheme even without an eavesdropper or jammer highly depends on the particular structure of the channel and f. It is therefore instructive to consider a specialization of the DFA framework with jamming to the computation of arithmetic means over AWGN channels. Specifically, the objective function is given as

$$f : (s_1, \ldots, s_K) \mapsto \frac{1}{K} \sum_{k=1}^{K} s_k, \tag{4.2}$$

where for all k, $\mathcal{S}_k = [-1, 1]$. The channel is given by

$$Y = h_{\mathfrak{AB}} \sum_{k=1}^{K} T_k + h_{\mathfrak{JB}} X + N_{\mathfrak{B}} \tag{4.3}$$

$$Z = h_{\mathfrak{AE}} \sum_{k=1}^{K} T_k + h_{\mathfrak{JE}} X + N_{\mathfrak{E}}, \tag{4.4}$$

where each of the transmitters $\mathfrak{A}_1, \ldots, \mathfrak{A}_K$ is subject to a total power constraint of $\mathfrak{P}_{\mathfrak{A}}$, $\mathfrak{J}$ is subject to an average power constraint $\mathfrak{P}_{\mathfrak{J}}$, $N_{\mathfrak{B}}$ is centered normal with variance $\sigma^2_{\mathfrak{B}}$ and $N_{\mathfrak{E}}$ is centered normal with variance $\sigma^2_{\mathfrak{E}}$. The real channel coefficients $h_{\mathfrak{AB}}, h_{\mathfrak{JB}}, h_{\mathfrak{AE}}, h_{\mathfrak{JE}}$ are assumed deterministic and known everywhere. The channel is used M times with transmitter input sequences T_k^M for each $k \in \{1, \ldots, K\}$ and X^M for the jammer. The input sequences are subject to the average power constraints

$$\frac{1}{M} \sum_{m=1}^{M} (T_{k,m})^2 \leq \mathfrak{P}_{\mathfrak{A}}, \ \frac{1}{M} \sum_{m=1}^{M} (X_m)^2 \leq \mathfrak{P}_{\mathfrak{J}}.$$

The problem is easily approached if we can assume that $\mathfrak{B}$ has full knowledge of X^M, while $\mathfrak{E}$ knows only how X^M is distributed.

In this case, we have the following result.

Lemma 12. *Consider the wiretap channel given by (4.3) and (4.4) and the objective*

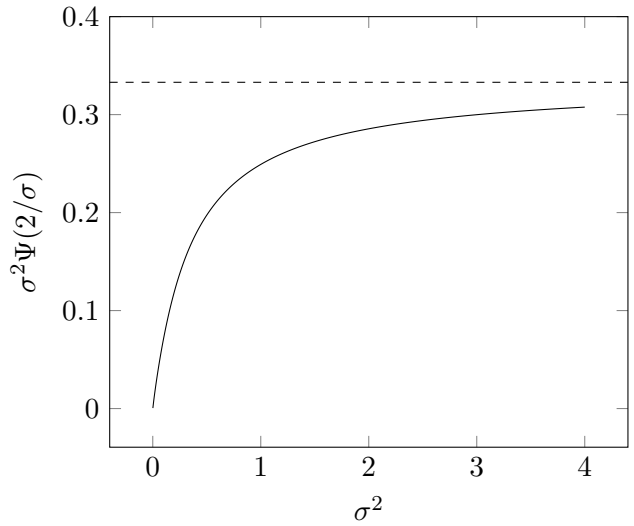

Figure 4.3: Illustration of the MSE guarantees of Lemma 12. The dashed line is the MSE which an eavesdropper would have without any received signal (i.e., guessing the middle of the interval).

function f defined in (4.2). Define

$$\sigma^2_{\mathrm{eff},\mathfrak{B}} := \frac{\sigma^2_{\mathfrak{B}}}{h^2_{\mathfrak{A}\mathfrak{B}} K^2 \mathfrak{P}_{\mathfrak{A}}},\ \sigma^2_{\mathrm{eff},\mathfrak{E}} := \frac{\sigma^2_{\mathfrak{E}} + h^2_{\mathfrak{J}\mathfrak{E}} \mathfrak{P}_{\mathfrak{J}}}{h^2_{\mathfrak{A}\mathfrak{E}} K^2 \mathfrak{P}_{\mathfrak{A}}}$$

Assume that the jamming sequence X^M is perfectly known at the legitimate receiver while the eavesdropper has only statistical information. Define

$$\Psi(a) := \int_0^a \int_{-\infty}^{\infty} \left(c + \frac{\varphi_{\mathcal{N}}(-c) - \varphi_{\mathcal{N}}(a-c)}{\Phi_{\mathcal{N}}(a-c) - \Phi_{\mathcal{N}}(-c)} - b \right)^2 \frac{1}{a} \varphi_{\mathcal{N}}(b-c) dc db, \tag{4.5}$$

where $\varphi_{\mathcal{N}}$ denotes the probability density function and $\Phi_{\mathcal{N}}$ the cumulative distribution function of the standard normal distribution, respectively. Then there is a distributed function approximation scheme with jamming which is $(f, \sigma^2_{\mathrm{eff},\mathfrak{E}} \Psi(2/\sigma_{\mathrm{eff},\mathfrak{E}}))$-MSE-secure and $(\sigma^2_{\mathrm{eff},\mathfrak{B}} \Psi(2/\sigma_{\mathrm{eff},\mathfrak{B}}))$-MSE-approximates f at the receiver.

The proof is based on a few facts from statistics. We only state the relevant lemmas here. For the sake of completeness, we include the proofs of the following two lemmas in Section 4.6.1.

Lemma 13. *If $\mathcal{U}$ is distributed uniformly in $[a, b]$ and, conditioned on $\mathcal{U}$, $\mathcal{V}_1, \ldots, \mathcal{V}_M$ are i.i.d. normally distributed with mean $\mathcal{U}$ and variance σ^2, then the minimum MSE*

estimator for estimating $\mathcal{U}$ from the observations $\mathcal{V}_1, \dots, \mathcal{V}_M$ is

$$\hat{\mathcal{U}} := \bar{\mathcal{V}} + \frac{\sigma}{\sqrt{M}} \cdot \frac{\varphi_{\mathcal{N}}\left(\frac{a-\bar{\mathcal{V}}}{\sigma/\sqrt{M}}\right) - \varphi_{\mathcal{N}}\left(\frac{b-\bar{\mathcal{V}}}{\sigma/\sqrt{M}}\right)}{\Phi_{\mathcal{N}}\left(\frac{b-\bar{\mathcal{V}}}{\sigma/\sqrt{M}}\right) - \Phi_{\mathcal{N}}\left(\frac{a-\bar{\mathcal{V}}}{\sigma/\sqrt{M}}\right)}, \tag{4.6}$$

where $\bar{\mathcal{V}} := \frac{1}{M}\sum_{m=1}^{M} \mathcal{V}_m$.

Lemma 14. *Under the assumptions of Lemma 13, the estimator $\hat{\mathcal{U}}$ satisfies*

$$\mathbb{E}\left(\left(\mathcal{U} - \hat{\mathcal{U}}\right)^2\right) = \frac{\sigma^2}{M}\Psi\left(\frac{b-a}{\sigma/\sqrt{M}}\right),$$

with Ψ as defined in (4.5).

Proof of Lemma 12. We use the following transmission strategy:

$$X_m : \text{Gaussian with mean 0 and variance } \mathfrak{P}_{\mathfrak{J}}, \tag{4.7}$$

$$E_k^M : s_k \mapsto (1, \dots, 1) \cdot s_k \sqrt{\frac{\mathfrak{P}_{\mathfrak{A}}}{M}} \tag{4.8}$$

The receiver can obtain

$$\begin{aligned} Y'_m &:= \frac{Y_m - h_{\mathfrak{J}\mathfrak{B}} X_m}{h_{\mathfrak{A}\mathfrak{B}} K \sqrt{\mathfrak{P}_{\mathfrak{A}}/M}} \\ &= \frac{h_{\mathfrak{A}\mathfrak{B}} \sum_{k=1}^{K} T_k + N_{\mathfrak{B},m}}{h_{\mathfrak{A}\mathfrak{B}} K \sqrt{\mathfrak{P}_{\mathfrak{A}}/M}} \\ &= \frac{h_{\mathfrak{A}\mathfrak{B}} \sum_{k=1}^{K} s_k \sqrt{\mathfrak{P}_{\mathfrak{A}}/M} + N_{\mathfrak{B},m}}{h_{\mathfrak{A}\mathfrak{B}} K \sqrt{\mathfrak{P}_{\mathfrak{A}}/M}} \\ &= f(s_1, \dots, s_K) + \frac{N_{\mathfrak{B},m}}{h_{\mathfrak{A}\mathfrak{B}} K \sqrt{\mathfrak{P}_{\mathfrak{A}}/M}}. \end{aligned}$$

We define the post-processing operation D^M at the receiver as first obtaining $Y'_1, \dots, Y'_M$ and then computing the MSE estimator from Lemma 13. With this choice, Lemma 14 yields the claimed reconstruction error guarantee.

On the other hand, the output at $\mathfrak{E}$ is given by

$$\begin{aligned} Z_m &\overset{(4.4)}{=} h_{\mathfrak{A}\mathfrak{E}} \sum_{k=1}^{K} T_k + h_{\mathfrak{J}\mathfrak{E}} X_m + N_{\mathfrak{E},m} \\ &\overset{(4.8)}{=} h_{\mathfrak{A}\mathfrak{E}} \sum_{k=1}^{K} s_k \sqrt{\mathfrak{P}_{\mathfrak{A}}/M} + h_{\mathfrak{J}\mathfrak{E}} X_m + N_{\mathfrak{E},m} \end{aligned}$$

$$\overset{(4.2)}{=} f(s_1, \ldots, s_K) \cdot K\sqrt{\mathfrak{P}_{\mathfrak{A}}/M} h_{\mathfrak{A}\mathfrak{E}} + h_{\mathfrak{J}\mathfrak{E}} X_m + N_{\mathfrak{E},m}.$$

From this, Lemmas 13 and 14 yield the claimed MSE-security of the scheme. □

The assumption that the legitimate receiver has full knowledge of the jamming signal seems quite strong. So the main part of this chapter is devoted to setting out and analyzing a jamming strategy in Theorem 4 which does not need any form of shared randomness or additional communication between the friendly jammer and the legitimate receiver. This jamming strategy has "almost" the same implications on the overall system performance as the assumption that the legitimate receiver has full knowledge of the jamming signal, while the eavesdropper only has knowledge about its distribution. In Corollary 9, we formalize this notion for the AWGN case. In principle, however, this jamming strategy is not restricted to the AWGN scenario; but in fact, it can be combined with any class of channel models in which it is possible to cancel or at least mitigate the effect of the jamming signal if exact knowledge of it is available.

One example where not a full cancellation but at least a good mitigation is possible is the fast-fading scenario treated in Chapter 2.

4.4 Main Results

In this section, we formally state the main results of this chapter.

A function approximation scheme is specific to a particular channel model, which among other things influences how the pre- and post-processing operations have to be designed as well as which class of functions can be approximated. In Section 4.3, we have described such a scheme for the AWGN channel and only a singleton class of functions, namely for the arithmetic average, which is a particularly simple case. The strategy for the legitimate receiver to counter the signal of the friendly jammer given that it is known is also particularly simple in the AWGN case and can be done perfectly, as we have seen. The missing part of the secrecy extension, which is the method for the legitimate receiver to obtain the necessary knowledge of the jamming signal while the eavesdropper cannot, on the other hand, can be phrased and proven to work in somewhat greater generality. In order to formally state our main results, therefore, we have to introduce a few technical concepts first.

For any channel W, we denote the joint input-output distribution under the input distribution P by $Q_{P,W}$ and the marginal for $\mathcal{Y}$ by $R_{P,W}$. With these conventions, we define the *information density* of tuples of elements of the input and output alphabets

under the channel W and an input distribution P as

$$\mathbf{i}_{P,W}(x^M; y^M) := \log \frac{dW^M(x^M, \cdot)}{dR^M_{P,W}}(y^M).$$

Correspondingly, the *mutual information* is defined as

$$\mathbf{I}_{P,W} := \mathbb{E}_{Q_{P,W}} \mathbf{i}_{P,W}(X; Y).$$

Moreover, given two probability measures μ and ν, we define the *Rényi divergence* of order $\alpha \in (0,1) \cup (1,\infty)$ between them as

$$\mathbf{D}_\alpha(\mu||\nu) := \frac{1}{\alpha - 1} \log \mathbb{E}_\mu \left(\left(\frac{d\mu}{d\nu} \right)^{\alpha - 1} \right).$$

$\mathbf{D}_1(\mu||\nu) := \lim_{\alpha \nearrow 1} \mathbf{D}_\alpha(\mu||\nu)$ is the *Kullback-Leibler divergence*.

A *compound channel* is a family $(W_s)_{s \in \mathcal{S}}$ of memoryless time-discrete point-to-point channels with common input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The transmitter's channel input is passed through a fixed W_s for the entire block length, but the transmitter does not control the choice of s, nor is it governed by a probability distribution. In this work, we assume neither the transmitter nor the receiver knows s. A *compound channel code* with block length M and rate $\mathcal{R}$ consists of an encoder $E^M_k : \{1, \ldots, \exp(M\mathcal{R})\} \to \mathcal{X}^M$ and a decoder $D^M : \mathcal{Y}^M \to \{1, \ldots, \exp(M\mathcal{R})\}$. We say that it has error probability δ if under a uniform distribution of $\mathfrak{M} \in \{1, \ldots, \exp(M\mathcal{R})\}$, the following is true: Let Y^M be constructed by passing the components of $X^M := E^M(\mathfrak{M})$ independently through W_s. Then, we have

$$\sup_{s \in \mathcal{S}} \mathbb{E}_{\mathfrak{M}} \mathbb{P}_s(\mathfrak{M} \neq D^M(Y^M)) \leq \delta,$$

where δ is the smallest number with this property.

Our secrecy scheme will hinge on the capacity of compound channels with possibly continuous alphabets (such as Gaussian compound channels). As mentioned in Section 4.1, it is shown in [Kes61] that even in the case that only the output alphabet is countably infinite, the capacity expressions from the finite case [BBT59, Wol59] do not carry over. It is therefore clear that an additional assumption on the compound channel is needed. In existing literature (e.g., [BBT59, RV68]), the problem is often approached by proving that the compound channel can be approximated by a finite class of channels in which case classical channel coding techniques such as joint typicality decoding can be adapted in a straightforward manner. In this work, we choose to directly pose the approximability of the compound channel by a finite class of channels as an assumption of our coding

theorem. In Section 4.5.1, we justify the usefulness of results involving this assumption by proving that a large class of practically relevant channels can indeed be approximated in the sense of the following definition.

Given measures μ and ν, we say that μ is *absolutely continuous* with respect to ν, or $\mu \ll \nu$, if all ν-null sets are also μ-null sets.

Definition 14. *Given a compound channel $(W_s)_{s\in\mathcal{S}}$ with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, we say that it can be* (η, J)-approximated *under a probability distribution P on $\mathcal{X}$ if there is a sequence $(\hat{W}_{\eta,j})_{j=1}^{J}$ of channels from $\mathcal{X}$ to $\mathcal{Y}$ such that for every $s \in \mathcal{S}$, there is $j \in \{1, \ldots, J\}$ such that*

$$\mathbb{E}_P \mathbf{D}_1 \left(W_s(X,\cdot) || \hat{W}_{\eta,j}(X,\cdot) \right) \leq \eta \tag{4.9}$$

$$\exists \alpha > 1 \ \forall x \in \mathcal{X} : \mathbf{D}_\alpha \left(W_s(x,\cdot) || \hat{W}_{\eta,j}(x,\cdot) \right) < \infty \tag{4.10}$$

$$\forall x \in \mathcal{X} : \hat{W}_{\eta,j}(x,\cdot) \ll W_s(x,\cdot) \tag{4.11}$$

$$\mathbf{I}_{P,\hat{W}_{\eta,j}} - \mathbf{I}_{P,W_s} \leq \eta, \tag{4.12}$$

and for every $j \in \{1, \ldots, J\}$ there is $s \in \mathcal{S}$ such that

$$\mathbf{I}_{P,W_s} - \mathbf{I}_{P,\hat{W}_{\eta,j}} \leq \eta. \tag{4.13}$$

The discussion in Section 4.5.1 provides sufficient topological conditions for (η, J)-approximability and shows that an example of channels with this property are (possibly fading) Gaussian channels.

We use a standard random codebook construction: Given a channel input alphabet $\mathcal{X}$, a distribution P on $\mathcal{X}$, a block length M and a rate $\mathcal{R}$, we define the $(P, M, \mathcal{R})$-ensemble of codebooks as a random experiment in which $\exp(M\mathcal{R})$ codewords of length M are drawn randomly and independently according to P for each component of each codeword.

A codebook $\mathcal{C}$ induces a jamming strategy in the following way: The jammer draws a codeword index $\mathfrak{M}$ uniformly at random and transmits $\mathcal{C}(\mathfrak{M})$, the codeword in $\mathcal{C}$ indexed by $\mathfrak{M}$. Therefore, the number of codewords in the codebook controls the amount of randomness contained in the jamming signal.

In order to be able to impose an average power constraint on the jammer, we define an *additive cost constraint* $(\mathfrak{c}, \mathfrak{C})$ for an input alphabet $\mathcal{X}$ consisting of a function $\mathfrak{c} : \mathcal{X} \to \mathbb{R}_0^+$ and a number $\mathfrak{C} \in \mathbb{R}_0^+$. Given any M, we say that $x^M \in \mathcal{X}^M$ *satisfies* the cost constraint if $\sum_{m=1}^{M} \mathfrak{c}(x_m) \leq M\mathfrak{C}$.

The specialization of this definition to a usual average power constraint would be to pick the square function as $\mathfrak{c}$ and the maximum admissible average power as $\mathfrak{C}$.

As long as there is at least one $x^M \in \mathcal{X}^M$ which satisfies the cost constraint $(\mathfrak{c}, \mathfrak{C})$, given any codebook $\mathcal{C}$ of block length M, we can define an associated *cost-constrained codebook* $\mathcal{C}_{\mathfrak{c},\mathfrak{C}}$ which is generated from $\mathcal{C}$ by replacing all codewords that do not satisfy the cost constraint with x^M. Obviously, all codewords in a cost-constrained codebook satisfy the cost constraint. We say that a cost constraint $(\mathfrak{c}, \mathfrak{C})$ is *compatible* with an input distribution P if for a random variable X distributed according to P, $\mathfrak{c}(X)$ has a finite moment generating function in an interval containing 0 in its interior and $\mathfrak{C} > \mathbb{E}_P \mathfrak{c}(X)$.

We assume a given pre-processing scheme which is admissible in the sense of Definition 9 and consider effective channels incorporating both the pre-processing at the transmitters and the physical channel. We denote the legitimate user's effective channel, which is a stochastic kernel mapping from $\mathcal{S}_1 \times \cdots \times \mathcal{S}_K \times \mathcal{X}$ to $\mathcal{Y}$, by $W_{\mathfrak{B}}$ and the eavesdropper's effective channel, which is a stochastic kernel mapping from $\mathcal{S}_1 \times \cdots \times \mathcal{S}_K \times \mathcal{X}$ to $\mathcal{Z}$, by $W_{\mathfrak{E}}$. The M-fold products of these effective channels are outlined in the system model in Fig. 4.2. With these concepts and notations defined, we are ready to state the main result of this work, which gives sufficient conditions for the existence of a jamming scheme that can simultaneously ensure that the legitimate receiver is able to reconstruct the full jamming signal and limit the usefulness of the eavesdropper's received signal.

Theorem 4. *Let P be a jammer input distribution. Suppose that for every $\eta > 0$, there is some $J(\eta)$ such that the compound channel $(W_s)_{s\in\mathcal{S}}$ defined by $W_{(s_1,\ldots,s_K)} := W_{\mathfrak{B}}(s_1, \ldots, s_K, \cdot, \cdot)$ can be $(\eta, J(\eta))$-approximated under P. Suppose further that for all $s_1 \in \mathcal{S}_1, \ldots, s_K \in \mathcal{S}_K$, the moment-generating function*

$$\mathbb{E} \exp(a \cdot \mathbf{i}_{P, W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)}(X; Z))$$

of the information density exists and is finite at some point $a > 0$. Let $(\mathfrak{c}, \mathfrak{C})$ be an additive cost constraint compatible with P, and let $\mathcal{C}$ be a random codebook from the $(P, M, \mathcal{R})$-ensemble. Let $\mathcal{R} \in (0, \infty)$ such that

$$\sup_{s_1\in\mathcal{S}_1,\ldots,s_K\in\mathcal{S}_K} \mathbf{I}_{P,W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)} < \mathcal{R} < \inf_{s_1\in\mathcal{S}_1,\ldots,s_K\in\mathcal{S}_K} \mathbf{I}_{P,W_{\mathfrak{B}}(s_1,\ldots,s_K,\cdot,\cdot)}. \tag{4.14}$$

Then there are numbers $\gamma_1, \gamma_2, \gamma_3, \gamma_4 > 0$ such that for sufficiently large M,

$$\mathbb{P}_{\mathcal{C}}\left(\left\| \hat{R}_{W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)^M, \mathcal{C}_{\mathfrak{c},\mathfrak{C}}} - R^M_{P, W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)} \right\|_{\mathrm{TV}} \geq \exp(-M\gamma_1) \right) < \exp(-\exp(M\gamma_2)) \tag{4.15}$$

and

$$\mathbb{P}_{\mathcal{C}}(\mathcal{E}) < \exp(-M\gamma_4), \tag{4.16}$$

where $\mathcal{E}$ is the event that the jamming strategy induced by $\mathcal{C}_{\mathfrak{c},\mathfrak{C}}$ does not allow reconstruction of the jamming signal with error at most $\exp(-M\gamma_3)$.

Obviously, the theorem is only useful if there exists some $\mathcal{R}$ satisfying (4.14). The condition that such an $\mathcal{R}$ exists is the formalization of the notion that the jamming signal has to be stronger at the legitimate receiver than it is at the eavesdropper.

In Section 4.5.2, we discuss in more detail how the guarantee in (4.15) can be used to arrive at a MSE security guarantee for the scheme.

Theorem 4 needs a compound channel coding result as an ingredient for its proof, and since this result is slightly more general than results available in the literature, it may be of independent interest. Therefore, we also state it in this section.

Theorem 5. *Let $(W_s)_{s\in\mathcal{S}}$ be a compound channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, and let P be a probability distribution on $\mathcal{X}$ such that for every $\eta > 0$, there is a $J(\eta)$ such that $(W_s)_{s\in\mathcal{S}}$ can be $(\eta, J(\eta))$-approximated under P. Let*

$$0 < \mathcal{R} < \inf_{s\in\mathcal{S}} \mathbf{I}_{P,W_s}, \tag{4.17}$$

and let $\mathcal{C}$ be a random codebook from the $(P, M, \mathcal{R})$-ensemble. Define an encoder $\mathfrak{m} \mapsto \mathcal{C}(\mathfrak{m})$. Then there is a decoder such that the average error probability δ of the resulting compound channel code satisfies

$$\mathbb{E}_{\mathcal{C}}(\delta) < \exp(-M\gamma), \tag{4.18}$$

for some $\gamma > 0$ and sufficiently large M.

With standard techniques, this theorem can be extended to the case of cost-constrained codebooks. We provide the full details of the proof of the following corollary in Section 4.6.5.

Corollary 7. *In the setting of Theorem 5, and given an additive cost constraint $(\mathfrak{c}, \mathfrak{C})$ compatible with P, there are $\gamma_1, \gamma_2 > 0$ such that for sufficiently large M,*

$$\mathbb{P}_{\mathcal{C}_{\mathfrak{c},\mathfrak{C}}}\big(\delta \geq \exp(-M\gamma_1)\big) < \exp(-M\gamma_2). \tag{4.19}$$

As the other main technical ingredient, we need the following result on channel resolvability.

Theorem 6. *Given a channel W from $\mathcal{X}$ to $\mathcal{Y}$, an input distribution P such that the moment-generating function $\mathbb{E}_{Q_{P,W}} \exp(a \cdot \mathbf{i}_{P,W}(X;Y))$ of the information density exists*

and is finite for some $a > 0$, and $\mathcal{R} > \mathbf{I}_{P,W}$, there exist $\gamma_1 > 0$ and $\gamma_2 > 0$ such that for large enough block lengths M, the $(P, M, \mathcal{R})$-ensemble satisfies

$$\mathbb{P}_{\mathcal{C}}\left(\|\hat{R}_{W,\mathcal{C}} - Q_{P,W}^{M}\|_{\mathrm{TV}} > \exp(-\gamma_1 M)\right) \leq \exp\left(-\exp\left(\gamma_2 M\right)\right), \tag{4.20}$$

where $\hat{R}_{W,\mathcal{C}}$ is the output distribution of channel W given that a uniformly random codeword from $\mathcal{C}$ is transmitted.

Similarly as with the compound channel coding theorem, we can use known methods to incorporate an additive cost constraint and argue the following corollary. For full details, we refer the reader to Section 4.6.5.

Corollary 8. *Let P be an input distribution on $\mathcal{X}$ and $(\mathfrak{c}, \mathfrak{C})$ an additive cost constraint compatible with P. Then the statement of Theorem 6 is valid even if the codebook $\mathcal{C}$ is replaced with its associated cost-constrained version $\mathcal{C}_{\mathfrak{c},\mathfrak{C}}$.*

4.5 Implications of the Main Results

In this section, we show that the main result on secure OTA computation, Theorem 4, implies a MSE security guarantee in the case of AWGN channels discussed in Section 4.3. To this end, we show in Section 4.5.1 that AWGN compound channels satisfy (among other channel models) the approximability criterion of Definition 14. In Section 4.5.2, we show how Theorem 4 can be applied to carry the MSE security result of Lemma 12 over to the case in which the legitimate receiver does not share randomness with the jammer.

4.5.1 Feasibility of Channel Approximation

In this subsection, we provide some tools and examples to argue that many compound channels of practical interest can indeed be (η, J)-approximated so that Theorem 5 may be applied to them. We begin with an observation that shows how our result specializes to the known results [BBT59, Wol59] for channels with finite alphabets.

Remark 5. *[BBT59, Lemma 4] implies that for every compound channel $(W_s)_{s\in\mathcal{S}}$ with finite input and output alphabets and every $\eta > 0$, there is an integer $J(\eta)$ such that $(W_s)_{s\in\mathcal{S}}$ can be $(\eta, J(\eta))$-approximated.*

We repeat the construction here and discuss how this fact is proved.

Let M be an integer which satisfies

$$M \geq \max\left(\frac{4|\mathcal{Y}|^3}{\eta^2}, \frac{2|\mathcal{Y}|^2}{\eta}\right).$$

Given $s \in \mathcal{S}$, we construct a channel W'_s. To this end, given any $x \in \mathcal{X}$, we fix an enumeration $(y_i)_{i=1}^{|\mathcal{Y}|}$ such that the finite sequence $(W_s(x,\{y_i\}))_{i=1}^{|\mathcal{Y}|}$ is nondecreasing. For every $i < |\mathcal{Y}|$, we can then uniquely choose a value for $W'_s(x,\{y_i\})$ such that it is an integer multiple of $1/M$ and

$$W_s(x,\{y_i\}) \leq W'_s(x,\{y_i\}) < W_s(x,\{y_i\}) + \frac{1}{M}. \tag{4.21}$$

It is argued in [BBT59] that this leaves a positive probability mass for $W'_s(x,\{y_{|\mathcal{Y}|}\})$ and therefore, this construction fully defines a channel W'_s. We define the approximation sequence $(\hat{W}_{\eta,j})_{j=1}^{J(\eta)}$ as an enumeration of the set $\{W'_s : s \in \mathcal{S}\}$. The cardinality of this set is upper bounded by $(M+1)^{|\mathcal{X}||\mathcal{Y}|}$ since all singleton probabilities are integer multiples of $1/M$.

For finite alphabets, (4.10) is trivially satisfied since Rényi divergence is in this case always finite [vEH14]. Regarding the absolute continuity criterion (4.11), we recall that $W'_s(x,\{y_{|\mathcal{Y}|}\})$ always has a positive probability, and for $i < |\mathcal{Y}|$, the assumption $W_s(x,\{y_i\}) = 0$ immediately implies $W'_s(x,\{y_{|\mathcal{Y}|}\}) = 0$ by (4.21), since 0 is the only integer multiple of $1/M$ which is strictly smaller than $1/M$. The proof in [BBT59] exploits (4.21) to prove that the absolute difference between the information of W_s and W'_s under any input distribution is at most $2|\mathcal{Y}|^{3/2}M^{-1/2}$ (statement (c) of the lemma) which by our choice of M immediately implies (4.12) and (4.13). Moreover, it is shown that (4.21) also implies that for all $x \in \mathcal{X}, y \in \mathcal{Y}$,

$$\log \frac{W_s(x,\{y\})}{W'_s(x,\{y\})} \leq \frac{2|\mathcal{Y}|^2}{M}$$

(statement (b) of the lemma) which by our choice of M implies (4.9).

For many channels of interest, $(\eta, J(\eta))$-approximability can be shown directly by going through properties (4.9) – (4.13). However, it is often easier to make an argument involving topological properties of $\mathcal{S}$. The following lemma provides some machinery to this end.

Lemma 15. *Let $(W_s)_{s\in\mathcal{S}}$ be a compound channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, let P be a probability distribution on $\mathcal{X}$ and assume that there is a topology on $\mathcal{S}$ such that $\mathcal{S}$ is compact and*

$$\forall s_0 \in \mathcal{S} : \ s \mapsto \mathbb{E}_P \mathbf{D}_1\left(W_s(X,\cdot)||W_{s_0}(X,\cdot)\right) \text{ is upper semi-continuous at } s_0 \tag{4.22}$$

$$\forall s_1, s_2 \in \mathcal{S}\ \exists \alpha > 1\ \forall x \in \mathcal{X} : \mathbf{D}_\alpha\left(W_{s_1}(x,\cdot)||W_{s_2}(x,\cdot)\right) < \infty \tag{4.23}$$

$$s \mapsto \mathbf{I}_{P,W_s} \text{ is lower semi-continuous.} \tag{4.24}$$

Then, for every $\eta > 0$, there is $J(\eta)$ such that $(W_s)_{s\in\mathcal{S}}$ can be $(\eta, J(\eta))$-approximated under P.

Proof. Fix some $\eta > 0$. For a given $s \in \mathcal{S}$, consider

$$\{s' : \mathbb{E}_P \mathbf{D}_1 \left(W_{s'}(X,\cdot)||W_s(X,\cdot)\right) < \eta\} \cap \{s' : \mathbf{I}_{P,W_s} - \mathbf{I}_{P,W_{s'}} < \eta\}.$$

Clearly, (4.22) and (4.24) ensure that this intersection is a neighborhood of s, so we can find an open neighborhood $\mathcal{D}_s$ contained in it. Thus, $(\mathcal{D}_s)_{s\in\mathcal{S}}$ is an open cover of $\mathcal{S}$ and therefore, the compactness of $\mathcal{S}$ yields a finite subcover $\mathcal{D}_{s_1}, \ldots, \mathcal{D}_{s_{J(\eta)}}$. We set $\hat{W}_{\eta,j} := W_{s_j}$ and given any $s \in \mathcal{S}$, we choose j such that $s \in \mathcal{D}_{s_j}$ and argue that $\hat{W}_{\eta,j}$ satisfies (4.9), (4.10) and (4.12). To this end, we note that (4.10) and (4.11) follow from (4.23), while (4.9) and (4.12) are ensured by the definition of $\mathcal{D}_{s_j}$. Finally, (4.13) is trivially satisfied, concluding the proof. □

We now make use of Lemma 15 to prove that a large class of Gaussian fading multiple-input and multiple-output channels can actually be $(\eta, J(\eta))$-approximated and thus Theorem 5 can be applied to them. The class of compound channels covered in the following theorem contains the class considered in [RV68, Sections 3 and 4] as a proper subset. We denote the set of symmetric, positive semidefinite $n \times n$-matrices with Sym_+^n and the set of symmetric, positive definite $n \times n$-matrices with Sym_{++}^n.

Theorem 7. *Let $\mathcal{X} = \mathbb{R}^{n_2}$, $\mathcal{Y} = \mathbb{R}^{n_1}$, let $\mathcal{S}$ be a compact subset of $\mathbb{R}^{n_1 n_2} \times \mathrm{Sym}_+^{n_1 n_2} \times \mathbb{R}^{n_1} \times \mathrm{Sym}_{++}^{n_1}$ (under the topology induced by the Frobenius norm). For any $s = (\mu_H, \Sigma_H, \mu_N, \Sigma_N) \in \mathcal{S}$, let W_s be the channel given by*

$$Y = HX + N,$$

where the channel input X has range $\mathbb{R}^{n_2}$, the channel output Y has range $\mathbb{R}^{n_1}$, the entries of the $n_1 \times n_2$ fading matrix H follow the distribution $\mathcal{N}(\mu_H, \Sigma_H)$ and the additive noise N is independent of H and follows the distribution $\mathcal{N}(\mu_N, \Sigma_N)$. Let P be a distribution on $\mathcal{X}$ and assume that either P is a multivariate Gaussian with positive definite covariance matrix or that the support of P is contained in some compact set. Then, given any $\eta > 0$, there is $J(\eta)$ such that $(W_s)_{s\in\mathcal{S}}$ can be $(\eta, J(\eta))$-approximated under P.

Proof. We show that the conditions of Lemma 15 are met. [Gil11] provides closed-form expressions for Rényi and Kullback-Leibler divergences between multivariate normal distributions. The only fact that we are going to use and which is apparent from these expressions, however, is that the Rényi and Kullback-Leibler divergences between two multivariate normal distributions are finite and continuous in the mean vectors and covariance matrices of the distributions wherever the covariance matrices are positive definite or, equivalently, both distributions are absolutely continuous with respect to the Lebesgue measure.

$\Sigma_N \in \mathrm{Sym}_{++}^{n_1}$ and therefore, given any $x \in \mathcal{X}$, $W_s(x,\cdot)$ is absolutely continuous with respect to the Lebesgue measure and thus has a positive definite covariance matrix and a density $r_{W_s(x,\cdot)}$, which implies (4.23).

Next, from the well-known closed-form expression of the multivariate normal density, we know that for any x and y, $r_{W_s(x,\cdot)}(y)$ is continuous in s. The boundedness of $\mathcal{S}$ implies a uniform upper bound on $r_{W_s(x,\cdot)}(y)$, so we can use the theorem of dominated convergence to argue that the marginal density $r_{R_{P,W_s}}(y) = \mathbb{E}_P r_{W_s(X,\cdot)}(y)$ depends continuously on s for any fixed y. We write

$$\mathbf{I}_{P,W_s} = \mathbb{E}_{PR_{P,W_s}} \left(\frac{r_{W_s(X,\cdot)}(Y)}{r_{R_{P,W_s}}(Y)} \log \frac{r_{W_s(X,\cdot)}(Y)}{r_{R_{P,W_s}}(Y)} \right).$$

Since the integrand is lower bounded by $-1/e$, (4.24) follows as an application of Fatou's lemma.

Finally, in order to argue (4.22), we distinguish between the two cases in the statement of the theorem.

First, suppose that there is a compact subset $\hat{\mathcal{X}} \subseteq \mathcal{X}$ with $P(\mathcal{X} \setminus \hat{\mathcal{X}}) = 0$. For any fixed s_0, the map

$$(s,x) \mapsto \mathbf{D}_1 \left(W_s(x,\cdot) || W_{s_0}(x,\cdot) \right)$$

is continuous, therefore the image of $\mathcal{S} \times \hat{\mathcal{X}}$ is compact and hence bounded. We can therefore invoke the theorem of dominated convergence and argue that (4.22) is satisfied.

Now, suppose that P is multivariate Gaussian with positive definite covariance matrix. We write

$$\begin{aligned}
\mathbb{E}_P \mathbf{D}_1 \left(W_s(X,\cdot) || W_{s_0}(X,\cdot) \right) &= \mathbb{E}_P \mathbb{E}_{W_s(X,\cdot)} \log \frac{r_P(X) r_{W_s(X,\cdot)}(Y)}{r_P(X) r_{W_{s_0}(X,\cdot)}(Y)} \\
&= \mathbb{E}_{Q_{P,W_s}} \log \frac{r_{Q_{P,W_s}}(X,Y)}{r_{Q_{P,W_{s_0}}}(X,Y)} \\
&= \mathbf{D}_1 \left(Q_{P,W_s} || Q_{P,W_{s_0}} \right).
\end{aligned}$$

From our arguments above, given any s, the distribution Q_{P,W_s} is multivariate Gaussian with positive definite covariance matrix, which implies that (4.22) is satisfied. □

4.5.2 Back to the Additive White Gaussian Noise case: Calculating Mean Square Error Security Guarantees

Revisiting the AWGN example from Section 4.3, Theorem 4 and Lemma 12 imply MSE security and reconstruction guarantees in case the legitimate receiver does not know the

jamming sequence, as we show in the following corollary.

Corollary 9. *Make the same assumptions and definitions as in Lemma 12, but do not assume that the legitimate receiver has knowledge of the jamming sequence $X_1, \dots, X_M$. Assume in addition that the channel from $\mathfrak{J}$ to $\mathfrak{B}$ is stronger than the channel from $\mathfrak{J}$ to $\mathfrak{E}$, i.e., $h_{\mathfrak{JB}}/\sigma_{\mathfrak{B}} > h_{\mathfrak{JE}}/\sigma_{\mathfrak{E}}$. Then there is a distributed approximation scheme with jamming and there are constants $\gamma_1, \gamma_2 > 0$ such that for sufficiently large M, the following hold:*

- *$\mathfrak{B}$ can approximate the objective function $f(s_1, \dots, s_K)$ with a MSE not exceeding*

$$\sigma^2_{\text{eff},\mathfrak{B}} \Psi\left(\frac{2}{\sigma^2_{\text{eff},\mathfrak{B}}}\right) + \exp(-M\gamma_1) \tag{4.25}$$

- *The scheme is (f, V)-MSE-secure, where*

$$V := \sigma^2_{\text{eff},\mathfrak{E}} \Psi\left(\frac{2}{\sigma^2_{\text{eff},\mathfrak{E}}}\right) - \exp(-M\gamma_2). \tag{4.26}$$

Proof. For the pre-processing at the transmitters, we use the same scheme as in the proof of Lemma 12 and begin by verifying that the resulting effective channels $W_{\mathfrak{B}}$ and $W_{\mathfrak{E}}$ with the input distribution P chosen to be Gaussian with mean 0 and variance $\mathfrak{P}_{\mathfrak{J}}$ satisfy the assumptions of Theorem 4. Since the defined compound channel is a class of Gaussian channels with different means taking values in the compact set $[-1, 1]$, the approximability of the channel is an immediate consequence of Theorem 7. The finiteness of the moment-generating function of the information density can be seen by straightforward applications of the definitions of information density and Rényi divergence:

$$\begin{aligned}
&\mathbb{E}\exp(a \cdot \mathbf{i}_{P,W_{\mathfrak{E}}(s_1,\dots,s_K,\cdot,\cdot)}(X;Z)) \\
&= \mathbb{E}\left(\left(\frac{dW_{\mathfrak{E}}(s_1,\dots,s_K,X,\cdot)}{dR_{P,W_{\mathfrak{E}}(s_1,\dots,s_K,\cdot,\cdot)}}(Z)\right)^a\right) \\
&= \exp\left(a \cdot \frac{1}{a}\log\mathbb{E}\left(\left(\frac{dW_{\mathfrak{E}}(s_1,\dots,s_K,X,\cdot)}{dR_{P,W_{\mathfrak{E}}(s_1,\dots,s_K,\cdot,\cdot)}}(Z)\right)^a\right)\right) \\
&= \exp\left(a\mathbf{D}_{a+1}\left(Q_{P,W_{\mathfrak{E}}(s_1,\dots,s_K,\cdot,\cdot)}||PR_{P,W_{\mathfrak{E}}(s_1,\dots,s_K,\cdot,\cdot)}\right)\right)
\end{aligned}$$

The Rényi divergence appearing at the end is between two multivariate Gaussian distributions and can be seen to be finite from the expressions given in [Gil11]. In order to verify (4.14), we first note that the information expressions appearing are the capacities of the effective channels $W_{\mathfrak{B}}$ and $W_{\mathfrak{E}}$. Since $s_1, \dots, s_K$ change the mean of the channel only, they do not influence the capacity. Therefore, the infimum and supremum are over

singleton sets. Consequently, the condition $h_{\mathfrak{J}\mathfrak{B}}/\sigma_{\mathfrak{B}} > h_{\mathfrak{J}\mathfrak{E}}/\sigma_{\mathfrak{E}}$ ensures that there is some $\mathcal{R}$ satisfying (4.14).

Fix γ_1', γ_3' as claimed to exist in Theorem 4, and also fix γ_1, γ_2 with $0 < \gamma_2 < \gamma_1'$ and $0 < \gamma_1 < \gamma_3'$.

Note that in the AWGN channel, $s_1, \ldots, s_K$ correspond to a shift of the output distribution of the channel, and therefore, the variational distance that appears in (4.15) is independent of $s_1, \ldots, s_K$. For sufficiently large M, we can therefore fix a codebook $\mathcal{C}$ from the $(P, M, \mathcal{R})$-ensemble such that for all $s_1, \ldots, s_K$, neither one of the error events described in (4.15) and (4.16) occurs.

Let the jamming strategy be induced by $\mathcal{C}_{\mathfrak{E},\mathfrak{c}}$ and let $d : \mathcal{Z}^M \to [-1, 1]$ be an estimator for $\mathfrak{E}$. We can now bound the MSE of d:

$$\begin{aligned}
&\mathbb{E}_{\hat{R}_{W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)^M}}\left(\left(d(Z^M) - f(s_1,\ldots,s_K)\right)^2\right) \\
&= \int_0^4 \hat{R}_{W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)^M}\left(\left(d(Z^M) - f(s_1,\ldots,s_K)\right)^2 > a\right) da \\
&\overset{(4.15)}{\geq} \int_0^4 \left(R^M_{P,W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)}\left(\left(d(Z^M) - f(s_1,\ldots,s_K)\right)^2 > a\right) - \exp(-M\gamma_1')\right) da \\
&= \mathbb{E}_{R^M_{P,W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)}}\left(\left(d(Z^M) - f(s_1,\ldots,s_K)\right)^2\right) - 4\exp(-M\gamma_1').
\end{aligned}$$

Taking the lower bound for the MSE under $R^M_{P,W_{\mathfrak{E}}(s_1,\ldots,s_K,\cdot,\cdot)}$ from Lemma 12 and noting $\gamma_2 < \gamma_1'$, we arrive at the expression in (4.26) for sufficiently large M.

For the reconstruction strategy at $\mathfrak{B}$, we first let $\mathfrak{B}$ reconstruct the jamming signal as is possible by Theorem 4 and then post-process the received signal as is possible with knowledge of the jamming signal by Lemma 12. Using the error bound in Lemma 12 and observing that the maximum instantaneous square error is 4 since we are constrained to an interval of length 2 and that $\gamma_1 < \gamma_3'$, for sufficiently large M we arrive at (4.25). □

4.6 Proofs

In this section, we prove the Lemmas used in the proof of Lemma 12 and our main results on secure OTA computation and compound channel coding, Theorems 4 and 5, as well as the corollaries that allow for the incorporation of an average cost constraint.

4.6.1 Statistical Preliminaries for the Proof of Lemma 12

In this subsection, we prove the two lemmas used for the proof of Lemma 12.

Proof of Lemma 13. It is known [Jay03, eq. (6.92)] that the MSE is minimized by the

mean of the posterior probability distribution. We can therefore calculate the minimum MSE estimator given the observations $v_1, \ldots, v_M$ as follows, where we use r with random variables in the index to denote (conditional) densities.

$$\begin{aligned}
\hat{\mathcal{U}} &= \int_a^b g r_{\mathcal{U}|\mathcal{V}_1,\ldots,\mathcal{V}_M}(g|v_1,\ldots,v_M)dg \\
&\overset{(a)}{=} \int_a^b g \frac{r_{\mathcal{V}_1,\ldots,\mathcal{V}_M|\mathcal{U}}(v,\ldots,v_M|g) r_{\mathcal{U}}(g)}{r_{\mathcal{V}_1,\ldots,\mathcal{V}_M}(v_1,\ldots,v_M)} dg \\
&= \frac{\int_a^b g r_{\mathcal{V}_1,\ldots,\mathcal{V}_M|\mathcal{U}}(v_1,\ldots,v_M|g) r_{\mathcal{U}}(g) dg}{\int_a^b r_{\mathcal{V}_1,\ldots,\mathcal{V}_M|\mathcal{U}}(v_1,\ldots,v_M|g) r_{\mathcal{U}}(g) dg} \\
&\overset{(b)}{=} \frac{\int_a^b g \exp\left(-\frac{1}{2\sigma^2}\sum_{m=1}^M (v_m - g)^2\right) dg}{\int_a^b \exp\left(-\frac{1}{2\sigma^2}\sum_{m=1}^M (v_m - g)^2\right) dg} \\
&= \frac{\int_a^b g \exp\left(-\frac{1}{2\sigma^2/M}\left(\frac{1}{M}\sum_{m=1}^M v_m^2 - 2g\bar{v} + g^2\right)\right) dg}{\int_a^b \exp\left(-\frac{1}{2\sigma^2/M}\left(\frac{1}{M}\sum_{m=1}^M v_m^2 - 2g\bar{v} + g^2\right)\right) dg} \\
&\overset{(c)}{=} \frac{\int_a^b g \exp\left(-\frac{1}{2\sigma^2/M}\left(\bar{v} - g\right)^2\right) dg}{\int_a^b \exp\left(-\frac{1}{2\sigma^2/M}\left(\bar{v} - g\right)^2\right) dg}
\end{aligned}$$

For (a), we have applied Bayes' rule. (b) is by observing that $r_{\mathcal{U}}(g) = 1/(b-a)$ is independent of g in $[a,b]$ and $r_{\mathcal{V}_1,\ldots,\mathcal{V}_M|\mathcal{U}}$ is the normal density. (c) is by multiplying

$$\exp\left(-\frac{1}{2\sigma^2/M}\left(\bar{v}^2 - \frac{1}{M}\sum_{m=1}^M v_m^2\right)\right)$$

on both sides of the fraction to complete the binomials.

The term we have calculated for $\hat{\mathcal{U}}$ is the mean of a normal distribution centered at $\bar{v}$ with variance σ^2/M truncated in $[a,b]$. This is a distribution with a known mean [JKB94, eq. 13.134], and hence we arrive at (4.6). □

Proof of Lemma 14. Based on the representation (4.6), we calculate the MSE as follows. We use the substitution rule, substituting $v' := \frac{\bar{v}-a}{\sigma/\sqrt{M}}$ in (a) and $g' := \frac{g-a}{\sigma/\sqrt{M}}$ in (b).

$$\begin{aligned}
\mathbb{E}\left(\left(\mathcal{U} - \hat{\mathcal{U}}\right)^2\right) = \int_a^b \int_{-\infty}^{\infty} \Bigg(\bar{v} + \frac{\sigma}{\sqrt{M}} \cdot \frac{\varphi_{\mathcal{N}}\left(\frac{a-\bar{v}}{\sigma/\sqrt{M}}\right) - \varphi_{\mathcal{N}}\left(\frac{b-\bar{v}}{\sigma/\sqrt{M}}\right)}{\Phi_{\mathcal{N}}\left(\frac{b-\bar{v}}{\sigma/\sqrt{M}}\right) - \Phi_{\mathcal{N}}\left(\frac{a-\bar{v}}{\sigma/\sqrt{M}}\right)} - g \Bigg)^2 \\
\cdot \frac{1}{b-a} \cdot \frac{1}{\sigma/\sqrt{M}} \varphi_{\mathcal{N}}\left(\frac{g-\bar{v}}{\sigma/\sqrt{M}}\right) d\bar{v} dg
\end{aligned}$$

$$\stackrel{(a)}{=}\int_a^b \int_{-\infty}^{\infty} \left(\frac{\sigma}{\sqrt{M}} \left(v' + \frac{\varphi_{\mathcal{N}}(-v') - \varphi_{\mathcal{N}}\left(\frac{b-a}{\sigma/\sqrt{M}} - v'\right)}{\Phi_{\mathcal{N}}\left(\frac{b-a}{\sigma/\sqrt{M}} - v'\right) - \Phi_{\mathcal{N}}(-v')} \right) + a - g \right)^2 \cdot \frac{1}{b-a} \cdot \varphi_{\mathcal{N}}\left(\frac{g-a}{\sigma/\sqrt{M}} - v'\right) dv' dg$$

$$\stackrel{(b)}{=}\int_0^{\frac{b-a}{\sigma/\sqrt{M}}} \int_{-\infty}^{\infty} \left(v' + \frac{\varphi_{\mathcal{N}}(-v') - \varphi_{\mathcal{N}}\left(\frac{b-a}{\sigma/\sqrt{M}} - v'\right)}{\Phi_{\mathcal{N}}\left(\frac{b-a}{\sigma/\sqrt{M}} - v'\right) - \Phi_{\mathcal{N}}(-v')} - g' \right)^2 \cdot \left(\frac{\sigma}{\sqrt{M}}\right)^3 \cdot \frac{1}{b-a} \cdot \varphi_{\mathcal{N}}(g' - v') dv' dg'$$

$$= \frac{\sigma^2}{M} \Psi\left(\frac{b-a}{\sigma/\sqrt{M}}\right),$$

concluding the proof of the lemma. □

4.6.2 Proof of Theorem 4

In order to prove Theorem 4, we decompose the system depicted in Fig. 4.2 into smaller (and more easily analyzed) subsystems by considering only a subset of the depicted terminals at a time.

1. *Considering the terminals* $\mathfrak{A}_1, \ldots, \mathfrak{A}_K, \mathfrak{B}$. This is the system summarized in Section 2.2.1. The rationale is that the results specialize to the setting in Section 4.3 as well as, e.g., to the fast-fading setting treated in Chapter 2. This part of the system consists of transmitters $(\mathfrak{A}_k)_{k=1}^K$ each of which holds a value $s_k \in \mathcal{S}_k$ and a receiver $\mathfrak{B}$ which has the objective of estimating $f(s_1, \ldots, s_K)$. To this end, each transmitter $\mathfrak{A}_k$ passes s_k through a pre-processor E_k independently M times yielding a sequence T_k^M of channel inputs. These are transmitted through M independent uses of the channel, generating a sequence Y^M of channel outputs. The receiver passes this sequence through a post-processor D^M which generates an approximation $\tilde{f}$ of $f(s_1, \ldots, s_K)$. As mentioned, the design of the pre- and post-processors depends heavily on the channel model and a particular class of functions f. The idea is that the pre-processors, the channel and the post-processor work together to mimic the function f, and any approach following this idea will be highly dependent on the particular structure of the channel and f. In Theorem 4, it is assumed that such a system is already in place and an augmentation is proposed which makes it more secure. A property of the system described in Section 2.2.1 necessary for our purposes and heavily exploited in this work is that the pre-processing is i.i.d., i.e.,

each pre-processor E_k is a stochastic kernel mapping from $\mathcal{S}_k$ to $\mathcal{T}_k$ and an M-fold product E_k^M of it is used to generate the channel input sequence.

2. *Considering the terminals* $\mathfrak{A}_1, \ldots, \mathfrak{A}_K, \mathfrak{J}, \mathfrak{E}$. In this setting, we assume that the transmitters $\mathfrak{A}_1, \ldots, \mathfrak{A}_K$ run a scheme of the kind described under item 1. Instead of the legitimate receiver, there is now an eavesdropper $\mathfrak{E}$. The objective is then to limit the usefulness of the eavesdropper's received signal Z^M. To this end, we add a friendly jammer $\mathfrak{J}$ to the system which transmits, according to a certain strategy, a word X^M. In this work, any jamming strategy we consider is induced by a codebook $\mathcal{C}$ of words of length M through the rule that the jammer chooses an element of the codebook uniformly at random and transmits it. We use existing results on *channel resolvability* to derive a bound on the usefulness of the signal Z^M received at $\mathfrak{E}$.

3. *Considering the terminals* $\mathfrak{A}_1, \ldots, \mathfrak{A}_K, \mathfrak{J}, \mathfrak{B}$. This is the setting from item 1 with an additional transmitter $\mathfrak{J}$. Here we assume that $\mathfrak{J}$ uses a jamming strategy induced by a codebook $\mathcal{C}$ as described under item 2 and use Theorem 5 on compound channel coding to argue that for suitable choices of $\mathcal{C}$, $\mathfrak{B}$ is able to fully reconstruct the jamming signal X^M. This enables $\mathfrak{B}$ to perform a cancellation of the jamming signal before it applies the post-processor D^M it would use in the setting of item 1. How this cancellation works depends on the particularities of the channel considered, but if, e.g., the jamming signal is simply added to the channel output as in the AWGN example in Section 4.3, it is possible to cancel it entirely by subtracting it from the received signal. So in this case the post-processor would consist of a reconstruction of the jamming signal, the subtraction of this signal from the received one and a post-processing step identical to that from item 1.

4. *Combining settings of item 2 and 3.* The goal here is to argue the existence of a codebook $\mathcal{C}$ which achieves both of the objectives described under item 2 and item 3. It will turn out that this can be achieved by a standard random codebook construction.

The main result of this work, Theorem 4, formulates conditions under which there are codebooks in the $(P, M, \mathcal{R})$-ensemble of which the $(\mathfrak{c}, \mathfrak{C})$-cost constrained versions simultaneously achieve the goals set forth under 2) and 3).

Proof of Theorem 4. An application of Corollary 7 yields (4.16), and (4.15) follows from Corollary 8. □

4.6.3 Proof of Theorem 5

We first pick parameters η, ε, β_1 and β_2 in sequence according to the following scheme, where (4.17) and the previous choices ensure that these intervals are all nonempty.

$$\eta \in \left(0, \frac{\inf_{s\in\mathcal{S}} \mathbf{I}_{P,W_s} - \mathcal{R}}{3}\right) \tag{4.27}$$

$$\varepsilon \in \left(2\eta, \inf_{s\in\mathcal{S}} \mathbf{I}_{P,W_s} - \mathcal{R} - \eta\right) \tag{4.28}$$

$$\beta_1 \in (\eta, \varepsilon - \eta) \tag{4.29}$$

$$\beta_2 \in (0, \varepsilon - \eta - \beta_1) \tag{4.30}$$

Fix a sequence $(\hat{W}_{\eta,j})_{j=1}^{J(\eta)}$ which $(\eta, J(\eta))$-approximates $(W_s)_{s\in\mathcal{S}}$.

We use a joint typicality decoder, i.e. if there is a unique $\mathfrak{m}$ such that

$$\exists j \in \{1, \dots, J(\eta)\} : \ \mathbf{i}_{P,\hat{W}_{\eta,j}}(\mathcal{C}(\mathfrak{m}); Y^M) \geq M(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon),$$

the decoder declares that message $\mathfrak{m}$ has been sent; otherwise it declares an error (or that message 1 has been sent).

We denote the transmitted message with $\mathfrak{M}$, the message declared by the decoder with $\hat{\mathfrak{M}}$ and define error events

$$\mathcal{E} := \{\mathfrak{M} \neq \hat{\mathfrak{M}}\} \tag{4.31}$$

$$\mathcal{E}_1 := \left\{\forall j \in \{1, \dots, J(\eta)\} \ \mathbf{i}_{P,\hat{W}_{\eta,j}}(\mathcal{C}(\mathfrak{M}); Y^M) < M(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon)\right\} \tag{4.32}$$

$$\mathcal{E}_2 := \left\{\exists \mathfrak{m} \neq \mathfrak{M} \ \exists j \in \{1, \dots, J(\eta)\} \ \mathbf{i}_{P,\hat{W}_{\eta,j}}(\mathcal{C}(\mathfrak{m}); Y^M) \geq M(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon)\right\}. \tag{4.33}$$

We note that $\mathcal{E} \subseteq \mathcal{E}_1 \cup \mathcal{E}_2$ and consequently

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2). \tag{4.34}$$

So we can bound these two errors separately and then combine them.

We start with bounding the expectation of the first summand, using the definition (4.32) and $\mathcal{C}$, as well as an addition of zero. Pick j such that $\hat{W}_{\eta,j}$ satisfies (4.9) – (4.12) with

respect to the realization W_s of the compound channel. Then we have

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}(\mathbb{P}(\mathcal{E}_1)) &\leq \mathbb{E}_{\mathcal{C}}\left(\mathbb{P}\left(\mathbf{i}_{P,\hat{W}_{\eta,j}}(\mathcal{C}(\mathfrak{M});Y^M) < M(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon)\right)\right)\\ &= Q_{P,W_s}^M\left(\mathbf{i}_{P,\hat{W}_{\eta,j}}(X^M;Y^M) < M(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon)\right)\\ &= Q_{P,W_s}^M\left(\sum_{m=1}^M \log\left(\frac{d\hat{W}_{\eta,j}(X_m,\cdot)}{dR_{P,\hat{W}_{\eta,j}}}(Y_m)\right) < M(\mathbf{I}_{P,\hat{W}_{\eta,j}} + \mathbf{I}_{P,W_s} - \mathbf{I}_{P,W_s} - \varepsilon)\right)\end{aligned} \tag{4.35}$$

The Radon-Nikodym derivative can be split as

$$\frac{d\hat{W}_{\eta,j}(X_m,\cdot)}{dR_{P,\hat{W}_{\eta,j}}} = \frac{d\hat{W}_{\eta,j}(X_m,\cdot)}{dW_s(X_m,\cdot)} \cdot \frac{dR_{P,W_s}}{dR_{P,\hat{W}_{\eta,j}}} \cdot \frac{dW_s(X_m,\cdot)}{dR_{P,W_s}}. \tag{4.36}$$

This is possible because $\hat{W}_{\eta,j}(x,\cdot) \ll W_s(x,\cdot)$ by (4.11), $R_{P,W_s} \ll R_{P,\hat{W}_{\eta,j}}$ by (4.9) and the joint convexity of Kullback-Leibler divergence in its arguments, and $W_s(x,\cdot) \ll R_{P,W_s}$ for P-almost all x by the properties of the marginalization.

We next bound tail probabilities corresponding to the three factors in (4.36) separately, starting with the first. To this end, we introduce a number $\alpha_1 > 1$ and argue, using Markov's inequality and the definition of Rényi divergence, that

$$\begin{aligned}&Q_{P,W_s}^M\left(\sum_{m=1}^M \log\frac{dW_s(X_m,\cdot)}{d\hat{W}_{\eta,j}(X_m,\cdot)}(Y_m) \geq M\beta_1\right)\\ &= Q_{P,W_s}^M\left(\exp\left((\alpha_1-1)\sum_{m=1}^M \log\frac{dW_s(X_m,\cdot)}{d\hat{W}_{\eta,j}(X_m,\cdot)}(Y_m)\right) \geq \exp((\alpha_1-1)M\beta_1)\right)\\ &\leq \mathbb{E}_{Q_{P,W_s}^M}\left(\left(\prod_{m=1}^M\left(\frac{dW_s(X_m,\cdot)}{d\hat{W}_{\eta,j}(X_m,\cdot)}(Y_m)\right)^{\alpha_1-1}\right)\right)\exp(-(\alpha_1-1)M\beta_1)\\ &= \exp\left(\sum_{m=1}^M \log\left(\mathbb{E}_{Q_{P,W_s}^M}\left(\left(\frac{dW_s(X_m,\cdot)}{d\hat{W}_{\eta,j}(X_m,\cdot)}\right)^{\alpha_1-1}\right)\right)\right)\exp(-(\alpha_1-1)M\beta_1)\\ &= \exp\Big(-(\alpha_1-1)M\left(\beta_1 - \mathbb{E}_P\mathbf{D}_{\alpha_1}\left(W_s(X,\cdot)||\hat{W}_{\eta,j}(X,\cdot)\right)\right)\Big).\end{aligned} \tag{4.37}$$

For the second factor, we argue in an analogous way, but using $\alpha_2 > 0$.

$$
\begin{aligned}
&R^M_{P,W_s}\left(\sum_{m=1}^{M}\log\frac{dR_{P,\hat{W}_{\eta,j}}}{dR_{P,W_s}}(Y_m)\geq M\beta_2\right)\\
&=R^M_{P,W_s}\left(\exp\left(\alpha_2\sum_{m=1}^{M}\log\frac{dR_{P,\hat{W}_{\eta,j}}}{dR_{P,W_s}}(Y_m)\right)\geq\exp(\alpha_2 M\beta_2)\right)\\
&\leq\mathbb{E}_{R^M_{P,W_s}}\left(\prod_{m=1}^{M}\left(\frac{dR_{P,\hat{W}_{\eta,j}}}{dR_{P,W_s}}(Y_m)\right)^{\alpha_2}\right)\exp(-\alpha_2 M\beta_2)\\
&=\exp\left((\alpha_2-1)M\mathbf{D}_{\alpha_2}\left(R_{P,\hat{W}_{\eta,j}}||R_{P,W_s}\right)-\alpha_2 M\beta_2\right).
\end{aligned}
\tag{4.38}
$$

Finally, for the third factor, we use $\alpha_3<1$.

$$
\begin{aligned}
&Q^M_{P,W_s}\Big(\mathbf{i}_{P,W_s}(X^M;Y^M)<M(\mathbf{I}_{P,W_s}-\varepsilon+\beta_1+\beta_2+\eta)\Big)\\
&=Q^M_{P,W_s}\big(\exp\left((\alpha_3-1)\mathbf{i}_{P,W_s}(X^M;Y^M)\right)>\exp\left((\alpha_3-1)M(\mathbf{I}_{P,W_s}-\varepsilon+\beta_1+\beta_2+\eta)\right)\big)\\
&\leq\mathbb{E}_{Q^M_{P,W_s}}\left(\prod_{m=1}^{M}\left(\frac{dW_s(X_m,\cdot)}{dR_{P,W_s}}(Y_m)\right)^{\alpha_3-1}\right)\exp\left(-(\alpha_3-1)M(\mathbf{I}_{P,W_s}-\varepsilon+\beta_1+\beta_2+\eta)\right)\\
&=\exp\Big(-(1-\alpha_3)M\big(\mathbf{D}_{\alpha_3}\left(Q_{P,W_s}||PR_{P,W_s}\right)+\varepsilon-\mathbf{I}_{P,W_s}-\beta_1-\beta_2-\eta\big)\Big),
\end{aligned}
\tag{4.39}
$$

Clearly, by (4.36), the union bound and (4.12), (4.35) is upper bounded by the sum of (4.37), (4.38) and (4.39). Next, we argue that these expressions all vanish exponentially with $M\to\infty$, using the continuity of Rényi divergence in the order which is shown in [vEH14, Theorem 7].

From (4.10), the theorem of monotone convergence and (4.9), we can conclude that

$$
\lim_{\alpha_1\searrow 1}\mathbb{E}_P\mathbf{D}_{\alpha_1}\left(W_s(X_m,\cdot)||\hat{W}_{\eta,j}(X_m,\cdot)\right)=\mathbb{E}_P\mathbf{D}_1\left(W_s(X_m,\cdot)||\hat{W}_{\eta,j}(X_m,\cdot)\right)\leq\eta,
$$

so, (4.29) allows us to fix α_1 at a value greater than 1 such that

$$
\beta_1-\mathbb{E}_P\mathbf{D}_{\alpha_1}\left(W_s(X_m,\cdot)||\hat{W}_{\eta,j}(X_m,\cdot)\right)>0
$$

and hence, (4.37) vanishes exponentially.

(4.38) is true for all $\alpha_2<1$. Since the inequalities are not strict, we can take the limit $\alpha_2\nearrow 1$ and argue that the statement is also valid for $\alpha_2=1$.

$\mathbf{D}_{\alpha_3}\left(Q_{P,W_s}||PR_{P,W_s}\right)$ converges to $\mathbf{I}_{P,W_s}$ from below for $\alpha_3\nearrow 1$ and so (4.30) allows us

to fix α_3 at a value less than 1 such that

$$\mathbf{D}_{\alpha_3}\left(Q_{P,W_s}||PR_{P,W_s}\right)+\varepsilon-\mathbf{I}_{P,W_s}-\beta_1-\beta_2-\eta>0$$

and therefore, (4.39) also vanishes exponentially.

For the second summand in (4.34), we use the definition (4.33) to argue that $\mathbb{E}_{\mathcal{C}}(\mathbb{P}(\mathcal{E}_2))$ is upper bounded by

$$\exp(M\mathcal{R})\sum_{j=1}^{J(\eta)}P^M R_{P,W_s}^M\left(\mathbf{i}_{P,\hat{W}_{\eta,j}}(X^M;Y^M)\geq M\left(\mathbf{I}_{P,\hat{W}_{\eta,j}}-\varepsilon\right)\right). \tag{4.40}$$

We define the indicator function

$$\mathrm{ind}(x^M,y^M):=\begin{cases}1, & \mathbf{i}_{P,\hat{W}_{\eta,j}}(x^M;y^M)\geq M\left(\mathbf{I}_{P,\hat{W}_{\eta,j}}-\varepsilon\right)\\ 0, & \text{otherwise.}\end{cases}$$

Using the definition of information density for a change of measure and multiplying one, we rewrite the probability that appears in (4.40) as

$$\begin{aligned}
&P^M R_{P,W_s}^M\left(\mathbf{i}_{P,\hat{W}_{\eta,j}}(X^M;Y^M)\geq M\left(\mathbf{I}_{P,\hat{W}_{\eta,j}}-\varepsilon\right)\right)\\
&=\int\limits_{\mathcal{X}^M\times\mathcal{Y}^M}\mathrm{ind}(x^M,y^M)\cdot P^M R_{P,W_s}^M(dx^M,dy^M)\\
&=\int\limits_{\mathcal{X}^M\times\mathcal{Y}^M}\exp\left(-\mathbf{i}_{P,W_s}(x^M;y^M)\right)\mathrm{ind}(x^M,y^M)Q_{P,W_s}^M(dx^M,dy^M)\\
&=\int\limits_{\mathcal{X}^M\times\mathcal{Y}^M}\exp\left(-\mathbf{i}_{P,W_s}(x^M;y^M)+\mathbf{i}_{P,\hat{W}_{\eta,j}}(x^M;y^M)-\mathbf{i}_{P,\hat{W}_{\eta,j}}(x^M;y^M)\right)\\
&\qquad\qquad\cdot\mathrm{ind}(x^M,y^M)Q_{P,W_s}^M(dx^M,dy^M)
\end{aligned}$$

Because of the presence of the indicator, we can uniformly bound

$$\mathbf{i}_{P,\hat{W}_{\eta,j}}(x^M;y^M)\geq M\left(\mathbf{I}_{P,\hat{W}_{\eta,j}}-\varepsilon\right)$$

and the indicator itself can be upper bounded by 1. This yields

$$\begin{aligned}
&P^M R^M_{P,W_s}\left(\mathbf{i}_{P,\hat{W}_{\eta,j}}(X^M;Y^M) \geq M\left(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon\right)\right) \\
\leq\ &\exp\left(-M\left(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon\right)\right) \\
&\int\limits_{\mathcal{X}^M\times\mathcal{Y}^M} \exp\Big(-\mathbf{i}_{P,W_s}(x^M;y^M) + \mathbf{i}_{P,\hat{W}_{\eta,j}}(x^M;y^M)\Big) Q^M_{P,W_s}(dx^M,dy^M)
\end{aligned}$$

We expand the definition of information density and apply Fubini's Theorem to rewrite the integral as

$$\int\limits_{\mathcal{Y}^M}\left(\int\limits_{\mathcal{X}^M}\frac{d\hat{W}_{\eta,j}{}^M(x^M,\cdot)}{dR^M_{P,\hat{W}_{\eta,j}}}(y^M)P^M(dx^M)\right)R^M_{P,W_s}(dy^M)$$

and observe that it equals 1.

Combining with (4.40) and applying (4.13), we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}(\mathbb{P}(\mathcal{E}_2)) &\leq \exp(M\mathcal{R})\sum_{j=1}^{J(\eta)}\exp\left(-M\left(\mathbf{I}_{P,\hat{W}_{\eta,j}} - \varepsilon\right)\right) \\
&\leq \exp(M\mathcal{R})\sum_{j=1}^{J(\eta)}\exp\left(-M\left(\inf_{s\in\mathcal{S}}\mathbf{I}_{P,W_s} - \varepsilon - \eta\right)\right) \\
&= \exp\left(-M\left(\inf_{s\in\mathcal{S}}\mathbf{I}_{P,W_s} - \varepsilon - \mathcal{R} - \eta - \frac{\log J(\eta)}{M}\right)\right). \qquad (4.41)
\end{aligned}$$

We observe that by (4.28), $\inf_{s\in\mathcal{S}}\mathbf{I}_{P,W_s} - \varepsilon - \mathcal{R} - \eta > 0$.

Finally, we pick

$$\begin{aligned}
\gamma \in \Bigg(0, \min\Bigg(&(\alpha_1-1)\cdot\left(\beta_1 - \mathbb{E}_P\mathbf{D}_{\alpha_1}\left(W_s(X_m,\cdot)||\hat{W}_{\eta,j}(X_m,\cdot)\right)\right), \\
&\beta_2, \\
&(1-\alpha_3)\big(\mathbf{D}_{\alpha_3}\left(Q_{P,W_s}||PR_{P,W_s}\right) + \varepsilon - \mathbf{I}_{P,W_s} - \beta_1 - \beta_2 - \eta\big), \\
&\inf_{s\in\mathcal{S}}\mathbf{I}_{P,W_s} - \varepsilon - \mathcal{R} - \eta\Bigg)\Bigg).
\end{aligned}$$

Since the exponent in (4.41) is then negative for sufficiently large M, we can combine it with (4.37), (4.38) and (4.39) to obtain (4.18). □

4.6.4 Proof of Theorem 6

In order to prove the theorem, given a codebook $\mathcal{C}$, we write the variational distance as

$$\begin{aligned}\|\hat{R}_{W,\mathcal{C}} - R_{P,W}^M\|_{\mathrm{TV}} &= \sup_{\substack{A\subseteq\mathcal{X}^M \\ \text{measurable}}} \left(\hat{R}_{W,\mathcal{C}}(A) - R_{P,W}^M(A)\right) \\ &= \sup_{\substack{A\subseteq\mathcal{X}^M \\ \text{measurable}}} \int_A \left(\frac{d\hat{R}_{W,\mathcal{C}}}{dR_{P,W}^M}(y^M) - 1\right) R_{P,W}^M(dy^M) \\ &= \mathbb{E}_{R_{P,W}^M}\left[\frac{d\hat{R}_{W,\mathcal{C}}}{dR_{P,W}^M}(y^M) - 1\right]^+. \end{aligned} \tag{4.42}$$

Note that throughout the proofs, we only consider codebooks $\mathcal{C}$ for which $\hat{R}_{W,\mathcal{C}}$ is absolutely continuous with respect to $R_{P,W}^M$. We can do this because the existence of a finite mutual information implies that $W(x,\cdot)$ is absolutely continuous with respect to $R_{P,W}$ for almost every x, and so the probability of drawing a codebook for which $\hat{R}_{W,\mathcal{C}}$ is not absolutely continuous with respect to $R_{P,W}^M$ is 0. Similarly, we assume the existence of the other Radon-Nikodym derivatives that appear.

We define the typical set

$$\mathfrak{T}_\varepsilon := \left\{(x^M, y^M) : \frac{1}{M}\mathbf{i}_{P,W}(x^M; y^M) \leq \mathbf{I}_{P,W} + \varepsilon\right\} \tag{4.43}$$

and split $\hat{R}_{W,\mathcal{C}}$ into two measures

$$\hat{R}_{1,W,\mathcal{C}}(A) := \exp(-M\mathcal{R}) \sum_{\mathfrak{m}=1}^{\exp(M\mathcal{R})} W^M\left(\mathcal{C}(\mathfrak{m}), A \cap \{y^M : (\mathcal{C}(\mathfrak{m}), y^M) \in \mathfrak{T}_\varepsilon\}\right) \tag{4.44}$$

$$\hat{R}_{2,W,\mathcal{C}}(A) := \exp(-M\mathcal{R}) \sum_{\mathfrak{m}=1}^{\exp(M\mathcal{R})} W^M\left(\mathcal{C}(\mathfrak{m}), A \cap \{y^M : (\mathcal{C}(\mathfrak{m}), y^M) \notin \mathfrak{T}_\varepsilon\}\right). \tag{4.45}$$

We observe $\hat{R}_{W,\mathcal{C}} = \hat{R}_{1,W,\mathcal{C}} + \hat{R}_{2,W,\mathcal{C}}$, which allows us to split (4.42) into a typical and an atypical part

$$\begin{aligned}\|\hat{R}_{W,\mathcal{C}} - R_{P,W}^M\|_{\mathrm{TV}} &= \mathbb{E}_{R_{P,W}^M}\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M) + \frac{d\hat{R}_{2,W,\mathcal{C}}}{dR_{P,W}^M}(y^M) - 1\right]^+ \\ &\leq \mathbb{E}_{R_{P,W}^M}\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M) - 1\right]^+ + \hat{R}_{2,W,\mathcal{C}}(\mathcal{Y}^M). \end{aligned} \tag{4.46}$$

We next state and prove two lemmas that we will use as tools to bound the typical and

atypical parts of this term separately.

Lemma 16 (Bound for atypical terms)**.** *Suppose* $Q_{P,W}^M(\mathcal{X}^M \times \mathcal{Y}^M \setminus \mathfrak{T}_\varepsilon) \leq a$ *and* $\delta \in [0,1]$*. Then*

$$\mathbb{P}_{\mathcal{C}}\big(\hat{R}_{2,W,\mathcal{C}}(\mathcal{Y}^M) > a(1+\delta)\big) \leq \exp\left(-\frac{1}{3}\delta^2 a \exp(M\mathcal{R})\right).$$

Proof. Observe $\mathbb{E}_{\mathcal{C}}(\hat{R}_{2,W,\mathcal{C}}(\mathcal{Y}^M)) = Q_{P,W}^M(\mathcal{X}^M \times \mathcal{Y}^M \setminus \mathfrak{T}_\varepsilon) \leq a$ and bound

$$\begin{aligned}
&\mathbb{P}_{\mathcal{C}}\left(\hat{R}_{2,W,\mathcal{C}}(\mathcal{Y}^M) > a(1+\delta)\right) \\
&= \mathbb{P}_{\mathcal{C}}\left(\exp(M\mathcal{R})\hat{R}_{2,W,\mathcal{C}}(\mathcal{Y}^M) > a\exp(M\mathcal{R})(1+\delta)\right) \\
&= \mathbb{P}_{\mathcal{C}}\left(\sum_{\mathfrak{m}=1}^{\exp(M\mathcal{R})} W^M\left(\mathcal{C}(\mathfrak{m}), \{y^M : (\mathcal{C}(\mathfrak{m}), y^M) \notin \mathfrak{T}_\varepsilon\}\right) > a\exp(M\mathcal{R})(1+\delta)\right) \\
&\leq \exp\left(-\frac{1}{3}\delta^2 a\exp(M\mathcal{R})\right).
\end{aligned}$$

The inequality follows from the Chernoff-Hoeffding bound [DP09, Ex. 1.1] by noting that we sum probabilities (i.e. values in $[0,1]$) on the left side, that these probabilities are independently distributed under $\mathbb{P}_{\mathcal{C}}$ and that by the hypothesis of the lemma the expectation of the term on the left is bounded by $a\exp(M\mathcal{R})$. □

Lemma 17 (Bound for typical terms)**.** *Let* $\delta, \lambda > 0$ *and define*

$$\mathbf{r} := \exp(M(\mathcal{R} - I(X;Y) - \varepsilon)). \tag{4.47}$$

Suppose $\mathbf{r}/(6\lambda) \geq 1$*. Then*

$$\begin{aligned}
&\mathbb{P}_{\mathcal{C}}\left(\mathbb{E}_{R_{P,W}^M}\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M) - 1\right]^+ > \delta\right) \\
&\qquad\qquad \leq \left(1 + \sqrt{\frac{3\pi}{2}}\exp\left(\frac{3\lambda^2}{4\mathbf{r}}\right)\frac{\lambda}{\sqrt{\mathbf{r}}} + \exp(-\lambda)\right)\exp(-\delta\lambda),
\end{aligned} \tag{4.48}$$

where $\pi \approx 3.14159$ *is the area of the unit circle.*

Before we prove this lemma, we make an observation that we need in the proof.

Lemma 18. *Let* Ξ *be a measurable function mapping codebooks and elements of* $\mathcal{Y}^M$ *to the nonnegative reals and let* $\lambda, \delta > 0$*. Then*

$$\mathbb{P}_{\mathcal{C}}\left(\mathbb{E}_{Y^M}\Xi(\mathcal{C}, Y^M) > \delta\right) \leq \mathbb{E}_{Y^M}\mathbb{E}_{\mathcal{C}}\Big(\exp(\lambda\Xi(\mathcal{C}, Y^M))\Big)\exp(-\delta\lambda). \tag{4.49}$$

Proof. An application of the Chernoff bound yields

$$\mathbb{P}_{\mathcal{C}}\left(\mathbb{E}_{Y^M}\Xi(\mathcal{C},Y^M)>\delta\right)\leq\mathbb{E}_{\mathcal{C}}\left(\exp\left(\lambda\mathbb{E}_{Y^M}\Xi(\mathcal{C},Y^M)\right)\right)\exp(-\delta\lambda).$$

We can then prove (4.49) by successive applications of Jensen's inequality and Fubini's theorem. □

Proof of Lemma 17. We begin by examining parts of the term in (4.48) for fixed, but arbitrary $\mathcal{C}$ and y^M and rewrite

$$\mathbf{r}\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M)=\sum_{\mathfrak{m}=1}^{\exp(M\mathcal{R})}\exp\left(M(-I(X;Y)-\varepsilon)\right)\frac{dW^M(\mathcal{C}(\mathfrak{m}),\cdot)}{dR_{P,W}^M}(y^M)\mathbf{1}_{(\mathcal{C}(\mathfrak{m}),y^M)\in\mathfrak{T}_\varepsilon}.$$

Now, we observe that the indicator function bounds the relative density to be at most $\exp(M(I(X;Y)+\varepsilon))$ and thus every term in the sum to range within $[0,1]$ and that furthermore

$$\mathbb{E}_{\mathcal{C}}\left(\mathbf{r}\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M)\right)\leq\exp\left(M(-I(X;Y)-\varepsilon)\right)\sum_{\mathfrak{m}=1}^{\exp(M\mathcal{R})}\mathbb{E}_{\mathcal{C}}\left(\frac{dW^M(\mathcal{C}(\mathfrak{m}),\cdot)}{dR_{P,W}^M}(y^M)\right)=\mathbf{r}.$$

We then use these observations to yield, for any $\xi>0$,

$$\begin{aligned}\mathbb{P}_{\mathcal{C}}\left(\exp\left(\lambda\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M)-1\right]^+\right)>\exp(\lambda\xi)\right)&=\mathbb{P}_{\mathcal{C}}\left(\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M)>1+\xi\right)&(4.50)\\&=\mathbb{P}_{\mathcal{C}}\left(\mathbf{r}\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M)>(1+\xi)\,\mathbf{r}\right)\\&\leq\exp\left(-\frac{\xi^2}{2\left(1+\frac{\xi}{3}\right)}\mathbf{r}\right),&(4.51)\end{aligned}$$

where (4.50) holds because the two measured events are equal and (4.51) follows by the Chernoff-Hoeffding bound [McD98, Theorem 2.3b]. (4.51) can be upper bounded by

$$\exp\left(-\frac{\xi^2}{3}\mathbf{r}\right)\tag{4.52}$$

for $\xi\leq 1$ (in particular) and by

$$\exp\left(-\frac{\xi}{3}\mathbf{r}\right)\tag{4.53}$$

for $\xi \geq 1$ (in particular). We will in the following use the substitutions

$$a := \exp(\lambda\xi) \tag{4.54}$$

$$b := \frac{\log(a)}{\lambda}\sqrt{\frac{2\mathbf{r}}{3}} - \sqrt{\frac{3}{2\mathbf{r}}}\lambda. \tag{4.55}$$

Since we will be using (4.55) for integration by substitution, we note that it implies

$$\frac{d}{db}a = \exp\left(b\lambda\sqrt{\frac{3}{2\mathbf{r}}} + \lambda^2\frac{3}{2\mathbf{r}}\right)\lambda\sqrt{\frac{3}{2\mathbf{r}}}. \tag{4.56}$$

We have, e.g. by [Bil95, Eq. 21.9],

$$\mathbb{E}_{\mathcal{C}}\left(\exp\left(\lambda\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR^M_{P,W}}(y^M) - 1\right]^+\right)\right) = \int_0^\infty \mathbb{P}_{\mathcal{C}}\left(\exp\left(\lambda\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR^M_{P,W}}(y^M) - 1\right]^+\right) > a\right) da$$

and upper bound this integral by splitting the integration domain into three parts: The integration over $[0, 1]$ can be upper bounded by 1 (since the integrand is a probability). The integration over $[1, \exp(\lambda)]$ can be upper bounded as

$$\begin{aligned}
&\int_1^{\exp(\lambda)} \mathbb{P}_{\mathcal{C}}\left(\exp\left(\lambda\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR^M_{P,W}}(y^M) - 1\right]^+\right) > a\right) da \\
&\leq \int_1^\infty \exp\left(-\frac{(\log a)^2}{3\lambda^2}\mathbf{r}\right) da && (4.57)\\
&= \int_0^\infty \exp\left(-\frac{b^2\lambda^2\frac{3}{2\mathbf{r}} + 2b\lambda^3\left(\frac{3}{2\mathbf{r}}\right)^{\frac{3}{2}} + \lambda^4\left(\frac{3}{2\mathbf{r}}\right)^2}{3\lambda^2}\mathbf{r} + b\lambda\sqrt{\frac{3}{2\mathbf{r}}} + \lambda^2\frac{3}{2\mathbf{r}}\right)\lambda\sqrt{\frac{3}{2\mathbf{r}}} db && (4.58)\\
&= \int_0^\infty \exp\left(-\frac{b^2}{2}\right) db \cdot \exp\left(\frac{3\lambda^2}{4\mathbf{r}}\right)\lambda\sqrt{\frac{3}{2\mathbf{r}}}. && (4.59)
\end{aligned}$$

(4.57) follows by substituting (4.52) as well as (4.54) and enlarging the integration domain to $[1, \infty)$, which can be done because the integrand is nonnegative. (4.58) follows by the rule for integration by substitution using (4.55).

The integration over $[\exp(\lambda), \infty)$ can be upper bounded as

$$\int_{\exp(\lambda)}^{\infty} \mathbb{P}_{\mathcal{C}} \left(\exp\left(\lambda \left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^{M}}(y^M) - 1 \right]^+ \right) > a \right) da \leq \int_{\exp(\lambda)}^{\infty} \exp\left(-\frac{\log a}{3\lambda} \mathbf{r} \right) da \quad (4.60)$$

$$= \int_{\exp(\lambda)}^{\infty} a^{-\mathbf{r}/(3\lambda)} da$$

$$= \frac{\exp(\lambda(1 - \mathbf{r}/(3\lambda)))}{\mathbf{r}/(3\lambda) - 1}$$

$$\leq \exp(-\lambda), \quad (4.61)$$

where (4.60) is by (4.53) and (4.61) is true because $\mathbf{r}/(6\lambda) \geq 1$. We now apply Lemma 18 with $\Xi(\mathcal{C}, y^M) := \left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^{M}}(y^M) - 1 \right]^+$. In the resulting bound, we substitute the bound of 1 for integration domain $[0, 1]$ as well as (4.59) and (4.61), substitute back (4.47) and note that $\exp(-b^2/2)$ is the well-known unnormalized standard normal density, and get (4.48). □

Proof of Theorem 6. In order to bound the atypical term in the sum (4.46), note first that for any $\alpha > 1$,

$$\begin{aligned}
&Q_{P,W}^{M}(\mathcal{X}^M \times \mathcal{Y}^M \setminus \mathfrak{T}_\varepsilon) \\
&= Q_{P,W}^{M}\left(\left\{ (x^M, y^M) : i(x^M, y^M)/M > I(X;Y) + \varepsilon \right\} \right) \\
&= Q_{P,W}^{M}\left(\left\{ (x^M, y^M) : \exp\left((\alpha - 1) i(x^M, y^M) \right) > \exp\left((\alpha - 1) M \left(I(X;Y) + \varepsilon \right) \right) \right\} \right) \\
&\leq \int_{\mathcal{X}^M \times \mathcal{Y}^M} \exp\left((\alpha - 1) i(x^M, y^M) \right) Q_{P,W}(d(x^M, y^M)) \cdot \exp\left(-(\alpha - 1) M \left(I(X;Y) + \varepsilon \right) \right) \quad (4.62) \\
&= \exp \log \left(\int_{\mathcal{X}^M \times \mathcal{Y}^M} \left(\frac{dW^M(\mathcal{C}(\mathfrak{m}), \cdot)}{dR_{P,W}^{M}}(y^M) \right)^{\alpha - 1} \cdot Q_{P,W}(d(x^M, y^M)) \right) \\
&\qquad \cdot \exp\left(-M(\alpha - 1) \left(I(X;Y) + \varepsilon \right) \right) \\
&= \exp\left(-M(\alpha - 1) \left(I(X;Y) + \varepsilon - \mathbf{D}_\alpha \left(Q_{P,W} || P R_{P,W} \right) \right) \right) \quad (4.63) \\
&\leq \exp(-M\beta_1), \quad (4.64)
\end{aligned}$$

where (4.62) follows by applying Markov's inequality and (4.64) as long as

$$\beta_1 \leq (\alpha - 1) \left(I(X;Y) + \varepsilon - \mathbf{D}_\alpha \left(Q_{P,W} || P R_{P,W} \right) \right). \quad (4.65)$$

Note that since the moment-generating function $\mathbb{E}_{Q_{P,W}} \exp(a \cdot i(X,Y))$ exists and is finite for some $a > 0$, there is some $\alpha' > 1$ such that $\mathbf{D}_{\alpha'} \left(Q_{P,W} || P R_{P,W} \right)$ is finite,

and thus $\mathbf{D}_\alpha\left(Q_{P,W}||PR_{P,W}\right)$ is finite and continuous in α for $\alpha \leq \alpha'$ [vEH14]. Since $\mathbf{D}_\alpha\left(Q_{P,W}||PR_{P,W}\right) \to I(X;Y)$ for $\alpha \to 1$, we can choose $\alpha > 1$, but sufficiently close to 1 such that the bound on β_1 is positive.

We can now apply Lemma 16 with $a := \exp(-M\beta_1)$ and $\delta := 1$ and get

$$\mathbb{P}_{\mathcal{C}}\big(\hat{R}_{2,W,\mathcal{C}}(\mathcal{Y}^M) > 2\exp(-M\beta_1)\big) \leq \exp\left(-\frac{1}{3}\exp(M(\mathcal{R}-\beta_1))\right). \tag{4.66}$$

To bound the typical term in (4.46), we apply Lemma 17 with $\lambda := \exp(M\beta_2)$ and $\delta := \exp(-M\beta_1)$, which yields

$$\begin{aligned}
&\mathbb{P}_{\mathcal{C}}\left(\mathbb{E}_{R_{P,W}^M}\left[\frac{d\hat{R}_{1,W,\mathcal{C}}}{dR_{P,W}^M}(y^M) - 1\right]^+ > \exp(-M\beta_1)\right) \\
&\leq \left(1 + \sqrt{\frac{3\pi}{2}}\exp\left(\frac{3}{4}\exp(-M(\mathcal{R}-I(X;Y)-\varepsilon-2\beta_2)) - \frac{1}{2}M(\mathcal{R}-I(X;Y)-\varepsilon-2\beta_2)\right)\right. \\
&\qquad \left. + \exp(-\exp(M\beta_2))\right)\exp\left(-\exp(M(\beta_2-\beta_1))\right)
\end{aligned} \tag{4.67}$$

as long as M is sufficiently large such that $\exp(M(\mathcal{R}-I(X;Y)-\varepsilon))/6 \geq 1$.

We are now ready to put everything together: Considering (4.46), (4.66) and (4.67), an application of the union bound yields the sum of (4.66) and (4.67) as an upper bound for $\mathbb{P}_{\mathcal{C}}\left(\|\hat{R}_{W,\mathcal{C}} - R_{P,W}^M\|_{\mathrm{TV}} > 3\exp(-M\beta_1)\right)$.

We choose $\varepsilon < \mathcal{R} - I(X;Y)$, then $\beta_1 < (\mathcal{R}-I(X;Y)-\varepsilon)/2$ small enough to satisfy (4.65), then β_2 such that $\beta_1 < \beta_2 < (\mathcal{R}-I(X;Y)-\varepsilon)/2$, and finally we choose $\gamma_1 < \beta_1$ and $\gamma_2 < \min(\mathcal{R}-\beta_1, \beta_2-\beta_1)$. With these choices, we get (4.20) for all sufficiently large M, thereby concluding the proof. □

4.6.5 Cost Constraint in Compound Channel Coding and Resolvability

In this section, we prove Corollaries 7 and 8 which essentially state that the conclusions of Theorems 5 and 6 are also true for the case of cost-constrained codebooks. The approach used is similar to the one in [EGK11, Section 3.3], but we include the adapted derivations in full here for the sake of self-containedness. We begin with a series of preliminary lemmas and conclude the section with the proofs of the corollaries.

Lemma 19. *Let $(\mathcal{U}_i)_{i\geq 1}$ be a sequence of i.i.d. random variables such that the moment generating function $\varphi(\lambda) := \mathbb{E}\exp(\lambda\mathcal{U}_1)$ exists on an interval containing 0 in its interior.*

Let $\mathfrak{C} > \mathbb{E}\mathcal{U}_1$. Then there exists $\gamma > 0$ such that

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathcal{U}_i > n\mathfrak{C}\right) \leq \exp(-n\gamma).$$

Proof. We can without loss of generality assume that $\mathfrak{C} = 0$ and $\mathbb{E}(\mathcal{U}_1) < 0$, because otherwise we could consider the random variables $(\mathcal{U}_i - \mathfrak{C})_{i\geq 1}$ instead.

Clearly, $\varphi(0) = 1$ and $\varphi'(0) = \mathbb{E}(\mathcal{U}_1) < 0$, so we can find some $\lambda > 0$ sufficiently small such that $\varphi(\lambda) < 1$. With this choice of λ, we can apply Markov's inequality and get

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} \mathcal{U}_i > 0\right) &= \mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^{n} \mathcal{U}_i\right) > 1\right) \\ &\leq \mathbb{E}\left(\exp\left(\lambda \sum_{i=1}^{n} \mathcal{U}_i\right)\right) \\ &= \varphi(\lambda)^n\end{aligned}$$

so the lemma follows by choosing $\gamma := -\log \varphi(\lambda)$. □

Lemma 20. *Let $\mathfrak{N}$ be a Bernoulli random variable with $\exp(M\mathcal{R})$ trials and success probability $p \leq \exp(-M\beta_1)$ where $\beta_1 < \mathcal{R}/2$. Then there are $\gamma_1, \gamma_2 > 0$ such that for sufficiently large M,*

$$\mathbb{P}(\mathfrak{N} > \exp(M(\mathcal{R} - \gamma_1))) \leq \exp(-\exp(M\gamma_2)). \tag{4.68}$$

Proof. We choose γ_1, γ_2 and β_2 such that $0 < \gamma_1 < \beta_1 < \beta_2 < \mathcal{R}/2$ and $\gamma_2 < \mathcal{R} - 2\beta_2$. Then

$$\begin{aligned}&\mathbb{P}\big(\mathfrak{N} > \exp(M(\mathcal{R} - \gamma_1))\big) \\ &= \mathbb{P}\big(\mathfrak{N} > p\exp(M\mathcal{R}) + (\exp(-M\gamma_1) - p)\exp(M\mathcal{R})\big) \\ &\leq \mathbb{P}\big(\mathfrak{N} > \mathbb{E}\mathfrak{N} + (\exp(-M\gamma_1) - \exp(-M\beta_1))\exp(M\mathcal{R})\big) \\ &\leq \mathbb{P}\big(\mathfrak{N} > \mathbb{E}\mathfrak{N} + \exp(-M\beta_2)\exp(M\mathcal{R})\big) &(4.69)\\ &\leq \exp\left(-2\frac{\big(\exp(-M\beta_2)\big)^2\big(\exp(M\mathcal{R})\big)^2}{\exp(M\mathcal{R})}\right) &(4.70)\\ &= \exp\left(-2\exp(M(\mathcal{R} - 2\beta_2))\right) &(4.71)\\ &\leq \exp(-\exp(M\gamma_2)), &(4.72)\end{aligned}$$

where (4.70) follows by the Chernoff-Hoeffding bound as stated for instance in [DP09,

Theorem 1.1, eq. (1.6)]. □

Lemma 21. *Let P be a probability distribution on $\mathcal{X}$. Assume moreover that $\mathfrak{c}(X)$ has a moment generating function defined in an interval with 0 in its interior and that $\mathfrak{C} > \mathbb{E}_P\mathfrak{c}(X)$. Denote the number of bad codewords in $\mathcal{C}$ with*

$$\mathfrak{N} := \sum_{\mathfrak{m}=1}^{\exp(M\mathcal{R})} \mathbf{1}_{\sum_{m=1}^{M}\mathfrak{c}(\mathcal{C}(\mathfrak{m})(m))>M\mathfrak{C}}.$$

Then there are $\gamma_1, \gamma_2 > 0$ such that

$$\mathbb{P}_{\mathcal{C}}\left(\mathfrak{N} > \exp(M(\mathcal{R}-\gamma_1))\right) \leq \exp(-\exp(M\gamma_2)). \tag{4.73}$$

Proof. Since the codeword components are i.i.d., we can apply Lemma 19 and obtain an arbitrarily small $\beta_1 > 0$ such that for all $\mathfrak{m}$,

$$p := \mathbb{P}_{\mathcal{C}}\left(\sum_{m=1}^{M}\mathfrak{c}(\mathcal{C}(\mathfrak{m})(m)) > M\mathfrak{C}\right) \leq \exp(-M\beta_1).$$

So since the codewords are independent, $\mathfrak{N}$ is a Bernoulli variable with $\exp(M\mathcal{R})$ trials and success probability p, and an application of Lemma 20 proves the conclusion. □

Proof of Corollary 7. Assume throughout the proof that M is sufficiently large. By Lemma 21, we have $\hat{\gamma_1}, \hat{\gamma_2} \in (0, \infty)$ with

$$\mathbb{P}_{\mathcal{C}}(\hat{\mathcal{E}}) \leq \exp(-\exp(M\hat{\gamma_2})), \tag{4.74}$$

where

$$\hat{\mathcal{E}} := \{\mathbb{P}_{\mathfrak{M}}(\mathcal{C}(\mathfrak{M}) \neq \mathcal{C}_{\mathfrak{c},\mathfrak{C}}(\mathfrak{M})) > \exp(-M\hat{\gamma_1})\}.$$

We denote the error of $\mathcal{C}$ with $\delta_{\mathcal{C}}$ and the error of $\mathcal{C}_{\mathfrak{c},\mathfrak{C}}$ with $\delta_{\mathcal{C}_{\mathfrak{c},\mathfrak{C}}}$. By Theorem 5 and Markov's inequality, we have, for some $\hat{\gamma} \in (0, \infty)$ given by the theorem and with choices $\tilde{\gamma_1} \in (0, \min(\hat{\gamma}, \hat{\gamma_1}))$, $\tilde{\gamma_2} \in (0, \hat{\gamma} - \tilde{\gamma_1})$,

$$\begin{aligned}\mathbb{P}_{\mathcal{C}}(\delta_{\mathcal{C}} \geq \exp(-M\tilde{\gamma_1})) &\leq \mathbb{E}_{\mathcal{C}}\delta_{\mathcal{C}}\exp(M\tilde{\gamma_1}) \\ &\leq \exp(-M(\hat{\gamma}-\tilde{\gamma_1})) \\ &\leq \exp(-M\tilde{\gamma_2}).\end{aligned} \tag{4.75}$$

Conditioned on the complement of $\hat{\mathcal{E}}$, we have

$$\begin{aligned}
\delta_{\mathcal{C}_{\mathfrak{c},\mathfrak{e}}} &\overset{(a)}{=} \sup_{s\in\mathcal{S}} \mathbb{E}_{\mathfrak{M}}\Big(\mathbb{P}_s\big(\mathfrak{M} \neq D^M(Y^M)|X^M = \mathcal{C}_{\mathfrak{c},\mathfrak{e}}(\mathfrak{M})\big)\Big) \\
&= \sup_{s\in\mathcal{S}} \sum_{\mathfrak{m}=1}^{\exp(M\mathcal{R})} \exp(-M\mathcal{R})\mathbb{P}_s\big(\mathfrak{m} \neq D^M(Y^M)|X^M = \mathcal{C}_{\mathfrak{c},\mathfrak{e}}(\mathfrak{m})\big) \\
&\overset{(b)}{\leq} \sup_{s\in\mathcal{S}} \sum_{\substack{\mathfrak{m}=1 \\ \mathcal{C}_{\mathfrak{c},\mathfrak{e}}(\mathfrak{m})=\mathcal{C}(\mathfrak{m})}}^{\exp(M\mathcal{R})} \exp(-M\mathcal{R})\mathbb{P}_s\big(\mathfrak{m} \neq D^M(Y^M)|X^M = \mathcal{C}_{\mathfrak{c},\mathfrak{e}}(\mathfrak{m})\big) + \sum_{\substack{\mathfrak{m}=1 \\ \mathcal{C}_{\mathfrak{c},\mathfrak{e}}(\mathfrak{m})\neq\mathcal{C}(\mathfrak{m})}}^{\exp(M\mathcal{R})} \exp(-M\mathcal{R}) \\
&\overset{(a)}{\leq} \delta_{\mathcal{C}} + \exp(-M\hat{\gamma}_1), \qquad (4.76)
\end{aligned}$$

where the steps marked with (a) are by the definition of compound coding error, and (b) is by upper bounding some of the probabilities in the sum with 1. We can now choose $\gamma_1 \in (0, \tilde{\gamma}_1)$ and obtain

$$\begin{aligned}
\mathbb{P}_{\mathcal{C}}\big(\delta_{\mathcal{C}_{\mathfrak{c},\mathfrak{e}}} \geq \exp(-M\gamma_1)\big) &\overset{(a)}{\leq} \mathbb{P}_{\mathcal{C}}\big(\delta_{\mathcal{C}_{\mathfrak{c},\mathfrak{e}}} \geq \exp(-M\gamma_1)|\neg\hat{\mathcal{E}}\big) + \mathbb{P}_{\mathcal{C}}(\hat{\mathcal{E}}) \\
&\overset{(4.76)}{\leq} \mathbb{P}_{\mathcal{C}}\big(\delta_{\mathcal{C}} + \exp(-M\hat{\gamma}_1) \geq \exp(-M\gamma_1)|\neg\hat{\mathcal{E}}\big) + \mathbb{P}_{\mathcal{C}}(\hat{\mathcal{E}}) \\
&\overset{(a)}{\leq} \frac{\mathbb{P}_{\mathcal{C}}\big(\delta_{\mathcal{C}} \geq \exp(-M\gamma_1) - \exp(-M\hat{\gamma}_1)\big)}{1 - \mathbb{P}_{\mathcal{C}}(\hat{\mathcal{E}})} + \mathbb{P}_{\mathcal{C}}(\hat{\mathcal{E}}) \\
&\overset{(b)}{\leq} \frac{\mathbb{P}_{\mathcal{C}}\big(\delta_{\mathcal{C}} \geq \exp(-M\tilde{\gamma}_1)\big)}{1 - \mathbb{P}_{\mathcal{C}}(\hat{\mathcal{E}})} + \mathbb{P}_{\mathcal{C}}(\hat{\mathcal{E}}) \\
&\overset{(4.74),(4.75)}{\leq} \frac{\exp(-M\tilde{\gamma}_2)}{1 - \exp(-\exp(M\hat{\gamma}_2))} + \exp(-\exp(M\hat{\gamma}_2)) \\
&\overset{(c)}{\leq} \exp(-M\gamma_2),
\end{aligned}$$

where the steps marked with (a) are by the law of total probability, step (b) is by the choices of $\gamma_1, \tilde{\gamma}_1$, and step (c) is valid for any choice of $\gamma_2 \in (0, \tilde{\gamma}_2)$. □

Proof of Corollary 8. By Lemma 21, we pick $\hat{\gamma}_1, \hat{\gamma}_2$ satisfying (4.73) and by Theorem 6, we pick $\tilde{\gamma}_1, \tilde{\gamma}_2$ satisfying (4.20).

We use the observation that $\mathfrak{N} \leq \exp((\mathcal{R} - \hat{\gamma}_1)M)$ implies

$$\|\hat{R}_{\mathcal{C}_{\mathfrak{c},\mathfrak{e}}} - \hat{R}_{\mathcal{C}}\|_{\mathrm{TV}} \leq \frac{\mathfrak{N}}{\exp(M\mathcal{R})} \leq \exp(-\hat{\gamma}_1 M) \qquad (4.77)$$

and observe that, as long as $\gamma_1 < \hat{\gamma}_1, \tilde{\gamma}_1$ and $\gamma_2 < \hat{\gamma}_2, \tilde{\gamma}_2$ and M is sufficiently large,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{C}_{\mathfrak{c},\mathfrak{E}}}\left(\|\hat{R}_{\mathcal{C}_{\mathfrak{c},\mathfrak{E}}} - Q_{P,W}^{M}\|_{\mathrm{TV}} > \exp(-\gamma_1 M)\right) \\
&\overset{(a)}{\leq} \mathbb{P}_{\mathcal{C}}\left(\|\hat{R}_{\mathcal{C}_{\mathfrak{c},\mathfrak{E}}} - \hat{R}_{\mathcal{C}}\|_{\mathrm{TV}} + \|\hat{R}_{\mathcal{C}} - Q_{P,W}^{M}\|_{\mathrm{TV}} > \exp(-\gamma_1 M)\right) \\
&\overset{(b)}{\leq} \mathbb{P}_{\mathcal{C}}\left(\|\hat{R}_{\mathcal{C}_{\mathfrak{c},\mathfrak{E}}} - \hat{R}_{\mathcal{C}}\|_{\mathrm{TV}} > \exp(-\hat{\gamma}_1 M)\right) + \mathbb{P}_{\mathcal{C}}\left(\|\hat{R}_{\mathcal{C}} - Q_{P,W}^{M}\|_{\mathrm{TV}} > \exp(-\tilde{\gamma}_1 M)\right) \\
&\overset{(c)}{<} \exp(-\exp(\hat{\gamma}_2 M)) + \exp(-\exp(\tilde{\gamma}_2 M)) \\
&\overset{(d)}{\leq} \exp(-\exp(\gamma_2 M)),
\end{aligned}$$

where (a) is by the triangle inequality, (b) is by the union bound and the choice of γ_1, (c) is due to (4.73), (4.77) and (4.20), and (d) is by the choice of γ_2. □

Publication List

[1] Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Resolvability on continuous alphabets. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2037–2041. IEEE, 2018.

[2] Igor Bjelaković, Matthias Frey, and Sławomir Stańczak. Distributed approximation of functions over fast fading channels with applications to distributed learning and the max-consensus problem. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1146–1153, Sep. 2019.

[3] Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Over-the-air computation in correlated channels. In *2020 IEEE Information Theory Workshop (ITW)*. IEEE, 2021.

[4] Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Over-the-air computation in correlated channels. *IEEE Transactions on Signal Processing*, 69:5739–5755, 2021.

[5] Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Towards secure over-the-air computation. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 700–705. IEEE, 2021.

[6] Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Towards secure over-the-air computation. *Submitted to IEEE Transactions on Information Theory*, 2022. Preprint available at arXiv:2001.03174.

[7] Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Over-the-air computation for distributed machine learning and consensus in large wireless networks. In Gitta Kutyniok, Holger Rauhut, and Robert J. Kunsch, editors, *Compressed Sensing in Information Processing*. Springer, 2021. To appear.

[8] Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. MAC resolvability: First and second order results. In *2017 IEEE Conference on Communications and Network Security (CNS)*, pages 560–564. IEEE, 2017.

[9] Navneet Agrawal, Matthias Frey, and Sławomir Stańczak. A scalable max-consensus protocol for noisy ultra-dense networks. In *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5. IEEE, 2019.

[10] Zoran Utkovski, Patrick Agostini, Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Learning radio maps for physical-layer security in the radio access. In *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5. IEEE, 2019.

[11] Gutierrez-Estevez, Miguel A., Zoran Utkovski, Patrick Agostini, Daniel Schäufele, Matthias Frey, Igor Bjelaković, and Sławomir Stańczak. Quality-of-service prediction for physical-layer security via secrecy maps. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2867–2871. IEEE, 2020.

[12] Zoran Utkovski, Matthias Frey, Patrick Agostini, Igor Bjelaković, and Sławomir Stańczak. Semantic security based secrecy maps for vehicular communications. In *25th International ITG Workshop on Smart Antennas (WSA)*, pages 307–311. ITG, 2021.

[13] Matthias Frey, Igor Bjelaković, Janis Nötzel, and Sławomir Stańczak. Semantic security with infinite dimensional quantum eavesdropping channel. In *Submitted to 2022 IEEE Information Theory Workshop (ITW)*. IEEE, 2022.

Bibliography

[ADG19] M. M. Amiri, T. M. Duman, and D. Gündüz. Collaborative machine learning at the wireless edge with blind transmitters. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019.

[AG19] M. M. Amiri and D. Gündüz. Computation scheduling for distributed machine learning with straggling workers. *IEEE Transactions on Signal Processing*, 67(24):6270–6284, 2019.

[AG20a] M. M. Amiri and D. Gündüz. Federated learning over wireless fading channels. *IEEE Transactions on Wireless Communications*, 19(5):3546–3557, 2020.

[AG20b] M. M. Amiri and D. Gündüz. Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air. *IEEE Transactions on Signal Processing*, 68:2155–2169, 2020.

[Ahl67] R. Ahlswede. Certain results in coding theory for compound channels. In *Proceedings of the Colloquium on Information Theory, Debrecen, Hungary*, pages 35–60, 1967.

[AOGE20] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin. Hierarchical federated learning across heterogeneous cellular networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8866–8870. IEEE, 2020.

[Arn57] V. Arnold. On functions of three variables. *Doklady Akademii Nauk SSSR*, 114:679–681, 1957.

[ASK19] J.-H. Ahn, O. Simeone, and J. Kang. Wireless federated distillation for distributed edge learning with heterogeneous data. In *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2019.

[BBT59] D. Blackwell, L. Breiman, and A. J. Thomasian. The capacity of a class of channels. *The Annals of Mathematical Statistics*, 30(4):1229–1241, 1959.

[BCD89] R. Bhatia, M.-D. Choi, and C. Davis. Comparing a matrix to its off-diagonal part. In H. Dym, S. Goldberg, M. A. Kaashoek, and P. Lancaster, editors, *The Gohberg Anniversary Collection: Volume I: The Calgary Conference and Matrix Theory Papers*, pages 151–164. Birkhäuser, Basel, Switzerland, 1989.

[BCH08] L. Brunet, H.-L. Choi, and J. How. Consensus-based auction approaches for decentralized task assignment. In *AIAA guidance, navigation and control conference and exhibit*, 2008.

[Bha97] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, New York, United States of America, 1997.

[Bil95] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, New York, United States of America, 3rd edition, 1995.

[BK00] V. Buldygin and Y. Kozachenko. *Metric Characterization of Random Variables and Random Processes*. Cross Cultural Communication. American Mathematical Society, Providence, Rhode Island, United States of America, 2000.

[BL13] M. R. Bloch and J. N. Laneman. Strong secrecy from channel resolvability. *Transactions on Information Theory*, 59(12):8077–8098, 2013.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities. A Nonasymptoitic Theory of Independence*. Oxford University Press, Oxford, United Kingdom, 2013.

[BRB93] K. L. Blackard, T. S. Rappaport, and C. W. Bostian. Measurements and models of radio frequency impulsive noise for indoor wireless communications. *IEEE Journal on Selected Areas in Communications*, 11(7):991–1001, 1993.

[BS92] J. A. Benediktsson and P. H. Swain. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(4):688–704, 1992.

[BTV12] M. Bellare, S. Tessaro, and A. Vardy. Semantic security for the wiretap channel. In *Advances in Cryptology – CRYPTO 2012*, pages 294–311. Springer, Berlin and Heidelberg, Germany, 2012.

[Buc76] R. C. Buck. Approximate complexity and functional representation. Technical Report 1656, Wisconsin University Madison Mathematics Research Center, 1976.

[Buc82] R. C. Buck. Nomographic functions are nowhere dense. *Proceedings of the American Mathematical Society*, 85(2):195–199, 1982.

[CH12] A. Christmann and R. Hable. Consistency of support vector machines using additive kernels for additive models. *Computational Statistics & Data Analysis*, 56(4):854–873, 2012.

[CK11] I. Csiszár and J. Körner. *Information theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, Cambridge, United Kingdom, 2011.

[Csi96] I. Csiszár. Almost independence and secrecy capacity. *Problems of Information Transmission*, 32(1):40–47, 1996.

[CT20] W.-T. Chang and R. Tandon. Communication efficient federated learning over multiple access channels. *arXiv preprint arXiv:2001.08737*, 2020.

[Cuf16] P. Cuff. Soft covering with high probability. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2963–2967. IEEE, 2016.

[Dev05] I. Devetak. The private classical capacity and quantum capacity of a quantum channel. *Transactions on Information Theory*, 51(1):44–55, 2005.

[Dob59] R. L. Dobrushin. Optimum information transmission through a channel with unknown parameters. *Radio Engineering and Electronic Physics*, 4(12):1–8, 1959.

[DP09] D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, Cambridge, United Kingdom, 2009.

[DSD20] J. Dong, Y. Shi, and Z. Ding. Blind over-the-air computation and data fusion via provable wirtinger flow. *IEEE Transactions on Signal Processing*, 68:1136–1151, 2020.

[EG59] E. Eisenberg and D. Gale. Consensus of subjective probabilities: The pari-mutuel method. *The Annals of Mathematical Statistics*, 30(1):165–168, 1959.

[EGK11] A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, Cambridge, United Kingdom, 2011.

[FLNS10] H. Ferreira, L. Lampe, J. Newbury, and T. G. Swart, editors. *Power Line Communications. Theory and Applications for Narrowband and Broadband*

Communications over Power Lines. Wiley, Chichester, United Kingdom, 2010.

[Fre85] S. French. Group consensus probability distributions: a critical survey. In J. M. Bemado, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*. Elsevier and Valencia University Press, Amsterdam, The Netherlands and Valencia, Spain, 1985.

[GBS13] M. Goldenbaum, H. Boche, and S. Stańczak. Harnessing interference for analog function computation in wireless sensor networks. *IEEE Transactions on Signal Processing*, 61(20):4893–4906, 2013.

[GBS14] M. Goldenbaum, H. Boche, and S. Stańczak. Nomographic functions: Efficient computation in clustered gaussian sensor networks. *IEEE Transactions on Wireless Communications*, 14(4):2093–2105, 2014.

[GdKS+19] D. Gündüz, P. de Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. van der Schaar. Machine learning in the air. *IEEE Journal on Selected Areas in Communications*, 37(10):2184–2199, 2019.

[Gil11] M. Gil. On Rényi divergence measures for continuous alphabet sources. Master's thesis, Queen's University Kingston, Ontario, Canada, 2011.

[GJRM+16] M. Goldenbaum, P. Jung, M. Raceala-Motoc, J. Schreck, S. Stańczak, and C. Zhou. Harnessing channel collisions for efficient massive access in 5G networks: A step forward to practical implementation. In *2016 9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, pages 335–339. IEEE, 2016.

[GS13] M. Goldenbaum and S. Stańczak. Robust analog function computation via wireless multiple-access channels. *IEEE Transactions on Communications*, 61(9):3863–3877, 2013.

[GS14] M. Goldenbaum and S. Stańczak. On the channel estimation effort for analog computation over wireless multiple-access channels. *IEEE Wireless Communications Letters*, 3(3):261–264, 2014.

[GV03] M. Gastpar and M. Vetterli. Source-channel communication in sensor networks. In F. Zhao and L. Guibas, editors, *Information Processing in Sensor Networks*, pages 162–177, Berlin and Heidelberg, Germany, 2003. Springer.

[Hay06] M. Hayashi. General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel. *Transactions on Information Theory*, 52(4):1562–1575, 2006.

[Hil00] D. Hilbert. Mathematische Probleme. Vortrag, gehalten auf dem internationalen Mathematiker-Kongreß zu Paris 1900. *Nachrichten von der Königl. Gesellschaft der Wissenschaften zu Göttingen*, pages 253–297, 1900.

[Hil27] D. Hilbert. Über die Gleichung neunten Grades. *Mathematische Annalen*, 97(1):243–250, 1927.

[HK14] J. Hou and G. Kramer. Effective secrecy: Reliability, confusion and stealth. In *2014 IEEE International Symposium on Information Theory*, pages 601–605. IEEE, 2014.

[HM16] M. Hayashi and R. Matsumoto. Secure multiplex coding with dependent and non-uniform multiple messages. *IEEE Transactions on Information Theory*, 62(5):2355–2409, 2016.

[HV93] T. S. Han and S. Verdú. Approximation theory of output statistics. *Transactions on Information Theory*, 39(3):752–772, 1993.

[ICJ12] F. Iutzeler, P. Ciblat, and J. Jakubowicz. Analysis of max-consensus algorithms in wireless channels. *IEEE Transactions on Signal Processing*, 60(11):6103–6107, 2012.

[Jay03] E. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, United Kingdom, 2003.

[JKB94] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1 of *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York, New York, United States of America, 2nd edition, 1994.

[Kes61] H. Kesten. Some remarks on the capacity of compound channels in the semicontinuous case. *Information and Control*, 4(2-3):169–184, 1961.

[KM79] J. Korner and K. Marton. How to encode the modulo-two sum of binary sources (corresp.). *IEEE Transactions on Information Theory*, 25(2):219–221, 1979.

[KMRR16] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[Kol57] A. N. Kolmogorov. On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk SSSR*, 114:953–956, 1957.

[LZLV20] W. Liu, X. Zang, Y. Li, and B. Vucetic. Over-the-air computation systems: Optimization, analysis and scaling laws. *IEEE Transactions on Wireless Communications*, 19(8):5488–5502, 2020.

[MASR21] F. Molinari, N. Agrawal, S. Stańczak, and J. Raisch. Max-consensus over fading wireless channels. *IEEE Transactions on Control of Network Systems*, 8(2):791–802, 2021.

[Mau94] U. M. Maurer. The strong secret key rate of discrete random triples. In R. E. Blahut, D. J. Costello, U. Maurer, and T. Mittelholzer, editors, *Communications and Cryptography: Two Sides of One Tapestry*, pages 271–285. Springer, Boston, Massachusetts, United States of America, 1994.

[McD98] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, Berlin and Heidelberg, Germany, 1998.

[MDR19] F. Molinari, A. M. Dethof, and J. Raisch. Traffic automation in urban road networks using consensus-based auction algorithms for road intersections. In *2019 18th European Control Conference (ECC)*, pages 3008–3015. IEEE, 2019.

[Mid99] D. Middleton. Non-gaussian noise models in signal processing for telecommunications: new methods an results for class A and class B noise models. *IEEE Transactions on Information Theory*, 45(4):1129–1149, 1999.

[MMR+17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017*, pages 1273–1282, 2017.

[MNT07] G. Mergen, V. Naware, and L. Tong. Asymptotic detection performance of type-based multiple access over multiaccess fading channels. *IEEE Transactions on Signal Processing*, 55(3):1081–1092, 2007.

[MR17] B. McMahan and D. Ramage. Federated learning: Collaborative machine learning without centralized training data, 2017.

Google AI Blog. Available at `https://ai.googleblog.com/2017/04/federated-learning-collaborative.html`, retrieved 02 March 2021.

[MRT12] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, United States of America and London, United Kingdom, 2012.

[MS93] D. Middleton and A. D. Spaulding. Elements of weak signal detection in non-gaussian noise environments. In V. Poor and J. B. Thomas, editors, *Advances in Statistical Signal Processing*, volume 2, pages 137–215. JAI Press, Greenwich, Connecticut, United States of America, 1993.

[MSR18] F. Molinari, S. Stańczak, and J. Raisch. Exploiting the superposition property of wireless communication for average consensus problems in multi-agent systems. In *2018 European Control Conference (ECC)*, pages 1766–1772. IEEE, 2018.

[MT06] G. Mergen and L. Tong. Type based estimation over multiaccess channels. *IEEE Transactions on Signal Processing*, 54(2):613–626, 2006.

[NCNC16] B. Nazer, V. R. Cadambe, V. Ntranos, and G. Caire. Expanding the compute-and-forward framework: Unequal powers, signal levels, and multiple linear combinations. *IEEE Transactions on Information Theory*, 62(9):4879–4909, 2016.

[NG05] R. Negi and S. Goel. Secret communication using artificial noise. In *IEEE 62nd Vehicular Technology Conference (VTC)*, volume 3, pages 1906–1910. IEEE, 2005.

[NG07] B. Nazer and M. Gastpar. Computation over multiple-access channels. *IEEE Transactions on Information Theory*, 53(10):3498–3516, 2007.

[NG11] B. Nazer and M. Gastpar. Compute-and-forward: Harnessing interference through structured codes. *IEEE Transactions on Information Theory*, 57(10):6463–6486, 2011.

[OS06] R. Olfati-Saber. Flocking for multi-agent dynamic systems: Algorithms and theory. *IEEE Transactions on Automatic Control*, 51(3):401–420, 2006.

[OSFM07] R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[OUG19] E. Ozfatura, S. Ulukus, and D. Gündüz. Distributed gradient descent with coded partial gradient computations. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3492–3496. IEEE, 2019.

[OZE+11] O. Ordentlich, J. Zhan, U. Erez, M. Gastpar, and B. Nazer. Practical code design for compute-and-forward. In *2011 IEEE International Symposium on Information Theory (ISIT)*, pages 1876–1880. IEEE, 2011.

[P+11] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RGS16] K. Ralinovski, M. Goldenbaum, and S. Stańczak. Energy-efficient classification for anomaly detection: The wireless channel as a helper. In *2016 IEEE International Conference on Communications (ICC)*, 2016.

[RV68] W. L. Root and P. P. Varaiya. Capacity of classes of gaussian channels. *SIAM Journal on Applied Mathematics*, 16(6):1350–1393, 1968.

[SC08] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, New York, United States of America, 2008.

[SC20] T. Sery and K. Cohen. On analog gradient descent learning over multiple access fading channels. *IEEE Transactions on Signal Processing*, 68:2897–2911, 2020.

[Sha49] C. E. Shannon. Communication theory of secrecy systems. *Bell Labs Technical Journal*, 28(4):656–715, 1949.

[STL20] M. Seif, R. Tandon, and M. Li. Wireless federated learning with local differential privacy. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2604–2609, 2020.

[SY12] I. Stanojev and A. Yener. Improving secrecy rate via spectrum leasing for friendly jamming. *IEEE Transactions on Wireless Communications*, 12(1):134–145, 2012.

[SZG20] Y. Sun, S. Zhou, and D. Gündüz. Energy-aware analog aggregation for federated learning with redundant data. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020.

[VBBM10] J. P. Vilela, M. Bloch, J. Barros, and S. W. McLaughlin. Friendly jamming for wireless secrecy. In *2010 IEEE International Conference on Communications (ICC)*. IEEE, 2010.

[VBBM11] J. P. Vilela, M. Bloch, J. Barros, and S. W. McLaughlin. Wireless secrecy regions with friendly jamming. *IEEE Transactions on Information Forensics and Security*, 6(2):256–266, 2011.

[vEH14] T. van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[Ver18] R. Vershynin. *High Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, United Kingdom, 2018.

[Wai19] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, United Kingdom, 2019.

[Win68] R. L. Winkler. The consensus of subjective probability distributions. *Management Science*, 15(2):B–61, 1968.

[Wol59] J. Wolfowitz. Simultaneous channels. *Archive for Rational Mechanics and Analysis*, 4(1):371–386, 1959.

[Wyn75a] A. Wyner. The common information of two dependent random variables. *Transactions on Information Theory*, 21(2):163–179, 1975.

[Wyn75b] A. D. Wyner. The wire-tap channel. *Bell Labs Technical Journal*, 54(8):1355–1387, 1975.

[YC16] X. Yin and X. Cheng. *Propagation Channel Characterization, Parameter Estimation and Modelling for Wireless Communications*. Wiley and IEEE Press, Singapore, 2016.

[YJSD20] K. Yang, T. Jiang, Y. Shi, and Z. Ding. Federated learning via over-the-air computation. *IEEE Transactions on Wireless Communications*, 19(3):2022–2035, 2020.

[YLCT19] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[Yos65] K. Yoshihara. Coding theorems for the compound semi-continuous memoryless channels. *Kodai Mathematical Seminar Reports*, 17(1):30–43, 1965.

[ZCL+19] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8):1738–1762, 2019.

[ZDHL20] Q. Zeng, Y. Du, K. Huang, and K. K. Leung. Energy-efficient radio resource allocation for federated edge learning. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020.

[ZLD+20] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang. Toward an intelligent edge: Wireless communication meets machine learning. *IEEE Communications Magazine*, 58(1):19–25, 2020.

[ZLL+18] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[ZNGE09] J. Zhan, B. Nazer, M. Gastpar, and U. Erez. MIMO compute-and-forward. In *2009 IEEE International Symposium on Information Theory (ISIT)*, pages 2848–2852. IEEE, 2009.

[ZWH20] G. Zhu, Y. Wang, and K. Huang. Broadband analog aggregation for low-latency federated edge learning. *IEEE Transactions on Wireless Communications*, 19(1):491–506, 2020.

[ZZ20] A. R. Zhang and Y. Zhou. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314, 2020.

Notations and Symbols

$\bar{\cdot}$ complex conjugate. 37

$\ll$ absolute continuity relation of measures. 79

$\|\cdot\|_{\mathbf{F}}$ Frobenius norm on matrices. 27

$\|\cdot\|_{\mathbf{op}}$ operator norm on matrices. 27

$\|\cdot\|_{\mathrm{TV}}$ total variation norm of signed measures. 72

$\mathbf{1}$. indicator function. 56

$\mathcal{A}$ fading-generating matrix. 22

$\mathfrak{A}_1, \ldots, \mathfrak{A}_K$ transmitters. 74

$\mathcal{B}$ noise-generating matrix. 22

$\mathfrak{B}$ receiver. 67

$(\mathfrak{c}, \mathfrak{C})$ additive input cost constraint for a channel. 79

$\mathcal{C}$ codebook. 79

$\mathcal{C}_{\mathfrak{c},\mathfrak{C}}$ cost-constrained codebook. 80

$\mathbb{C}$ complex numbers. 21

D^M post-processing/decoding operation of the receiver for a scheme with block length M. 24, 78

$\mathbf{D}_\alpha(\mu||\nu)$ Rényi divergence of order α between measures μ and ν. 78

$\mathbf{D}_1(\mu||\nu)$ Kullback-Leibler divergence between measures μ and ν. 78

E_k^M pre-processing/encoding operation of transmitter k for a scheme with block length M. 24, 78

$\mathfrak{E}$ eavesdropper. 71

$\exp(\cdot)$ exponential function with basis e. 20

$e \approx 2.71828$ Euler's number. 5, 20

$f : \mathcal{S}_1 \times \ldots \times \mathcal{S}_K \to \mathbb{R}$ objective function to be approximated. 2, 24

$f_k : \mathcal{S}_k \to \mathbb{R}$ inner function of nomographic representation. 2, 3, 25

$F : A \to \mathbb{R}$ outer function of nomographic representation. 2, 3, 25

$\tilde{f}$ estimator at the receiver for $f(s_1, \ldots, s_K)$.. 25

$\mathcal{F}_{K,\mathbf{lin}}$ class of generalized linear functions. 25

$\mathcal{F}_{\mathbf{mon}}$ class of functions to be approximated with DFA schemes. 25

$\mathcal{G}$ sub-Gaussian vector with independent entries. 22

g eavesdropper's objective. 71

$H_k(m)$ fading coefficient for transmitter k at channel use m. 21

$\mathcal{H}$ vector of all fading coefficients. 21

$\mathfrak{H}$ reproducing kernel Hilbert space. 53

$h_{\mathfrak{AB}}, h_{\mathfrak{AE}}, h_{\mathfrak{JB}}, h_{\mathfrak{JE}}$ fading coefficients of the channels between transmitter / jammer and legitimate receiver/eavesdropper. 74

$\mathbf{i}_{P,W}(x^M; y^M)$ information density of input and output tuples of the channel W under input distribution P. 78

$\mathbf{I}_{P,W}$ mutual information of the input and output of the channel W under input distribution P. 78

$\mathbf{id}_n$ identity matrix of dimension $n \times n$. 30

$.^{\mathbf{Im}}$ imaginary part of a complex number. 21

$\mathfrak{J}$ jammer. 71

K number of transmitters. 6, 21

$\mathfrak{L}$ loss function. 11

$\log(\cdot)$ natural logarithm with basis e. 4

M codebook block length / number of times the channel is used in a particular communication scheme. 7, 21

$M(f, \varepsilon, \delta)$ communication cost for approximating f. 25

$N(m)$ additive noise at channel use m. 21

N vector of all noise realizations. 22

$\mathcal{N}(\mu, \Sigma)$ multivariate normal distribution with mean μ and covariance matrix Σ. 47

$(P, M, \mathcal{R})$-**ensemble** random codebook ensemble with input distribution P, block length M and rate $\mathcal{R}$. 79

$\mathfrak{P}$ power constraint. 21

$\mathfrak{P}_{\mathfrak{A}}$ transmitter power constraint. 74

$\mathfrak{P}_{\mathfrak{J}}$ jammer power constraint. 74

P channel input distribution. 77

$Q_{P,W}$ joint input-output distribution of channel W under input distribution P. 77

$\mathcal{R}$ rate of a codebook. 78

$R_{P,W}$ output distribution of channel W under input distribution P. 77

$\hat{R}_{W,\mathcal{C}}$ marginal distribution of the output of the channel W induced by the codebook $\mathcal{C}$. 82

$\hat{R}_{s_1,\ldots,s_K}$ ideal eavesdropper output distribution induced by DFA scheme and jamming strategy. 71

$\mathbb{R}$ real numbers. 2

$\cdot^{\mathbf{Re}}$ real part of a complex number. 21

Sym_{+}^{n} symmetric, positive semidefinite $n \times n$ matrices with real entries. 84

Sym_{++}^{n} symmetric, positive definite $n \times n$ matrices with real entries. 84

$\cdot^{T}$ transpose of a matrix or vector. 22

$T_k(m), T_k$ channel input symbol. 21, 89

$\mathrm{tr}(\cdot)$ trace of matrices. 39

$U_k(1), \ldots, U_k(M)$ randomness used in pre-processing. 24

W channel. 6, 77

$(W_s)_{s\in\mathcal{S}}$ compound channel. 78

$(\hat{W}_{\eta,j})_{j=1}^{J}$ sequence of channels that approximate the compound channel $(W_s)_{s\in\mathcal{S}}$. 79

$W_{\mathfrak{B}}$ legitimate user's effective channel. 80

$W_{\mathfrak{E}}$ eavesdropper's effective channel. 80

$\mathfrak{X}$ feature alphabet. 11

$\mathcal{X}$ channel input alphabet. 78

$\mathfrak{Y}$ label alphabet. 11

$\mathcal{Y}$ channel output alphabet. 78

$\mathcal{Z}$ eavesdropper's channel output alphabet. 71

$\bar{\Delta}(f)$ total spread of the inner part of f. 26

$\Delta(f)$ maximum spread of the inner part of f. 26

$\Delta(f\|\mathfrak{P})$ relative spread of the inner part of f with respect to power constraint $\mathfrak{P}$. 26

$\theta\,(X)$ sub-exponential norm of the random variable X. 43

ϑ jamming decoder. 71

κ reproducing kernel. 53

$\pi \approx 3.14159$ area of the unit circle. 97

$\sigma^2_{\mathrm{eff},\mathfrak{B}}$ legitimate receiver's effective noise-to-signal ratio. 75

$\sigma^2_{\mathrm{eff},\mathfrak{E}}$ eavesdropper's effective noise-to-signal ratio. 75

$\tau\,(X)$ sub-Gaussian norm of the random variable X. 20, 43

Φ increment majorant. 25

$\Phi_{\mathcal{N}}$ cumulative distribution function of the standard normal distribution. 75

$\varphi_{\mathcal{N}}$ probability density function of the standard normal distribution. 75

Abbreviations

AWGN Additive White Gaussian Noise.

CSI channel state information.

DFA Distributed Function Approximation.

FL Federated Learning.

HFL Horizontal Federated Learning.

i.i.d. independent and identically distributed.

IQ inphase-quadrature.

MAC multiple-access channel.

ML Machine Learning.

MSE mean square error.

OTA Over-the-Air.

SVM Support Vector Machine.

TDMA Time Division Multiple Access.

VFL Vertical Federated Learning.

Index